UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA, TECNOLOGIA E SOCIEDADE

DANILO FORMENTON

IDENTIFICAÇÃO DE CRITÉRIOS DE SELEÇÃO DE CONTEÚDOS PARA O ARQUIVAMENTO DA *WEB*

São Carlos – SP
2023

DANILO FORMENTON

IDENTIFICAÇÃO DE CRITÉRIOS DE SELEÇÃO DE CONTEÚDOS PARA O ARQUIVAMENTO DA *WEB*

Tese apresentada ao Programa de Pós-Graduação em Ciência, Tecnologia e Sociedade, do Centro de Educação e Ciências Humanas, da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Doutor em Ciência, Tecnologia e Sociedade.

Linha de pesquisa: Gestão Tecnológica e Sociedade Sustentável

Orientadora: Profa. Dra. Luciana de Souza Gracioso

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

São Carlos – SP
2023

Formenton, Danilo

Identificação de critérios de seleção de conteúdos para o arquivamento da Web / Danilo Formenton -- 2023.
369f.

Tese de Doutorado - Universidade Federal de São Carlos, campus São Carlos, São Carlos
Orientador (a): Luciana de Souza Gracioso
Banca Examinadora: Luciana de Souza Gracioso, Miguel Angel Márdero Arellano, Sonia Araújo de Assis Boeres, Ariadne Chloe Mary Furnival, Luzia Sigoli Fernandes Costa
Bibliografia

1. Arquivamento da Web. 2. Preservação digital. 3. Critérios de seleção. I. Formenton, Danilo. II. Título.

Ficha catalográfica desenvolvida pela Secretaria Geral de Informática (SIn)

DADOS FORNECIDOS PELO AUTOR

Bibliotecário responsável: Ronildo Santos Prado - CRB/8 7325

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Educação e Ciências Humanas
Programa de Pós-Graduação em Ciência, Tecnologia e Sociedade

Folha de Aprovação

Defesa de Tese de Doutorado do candidato Danilo Formenton, realizada em 29/06/2023.

Comissão Julgadora:

Profa. Dra. Luciana de Souza Gracioso (UFSCar)

Prof. Dr. Miguel Angel Márdero Arellano (IBICIT)

Profa. Dra. Sonia Araújo de Assis Boeres (IBICIT)

Profa. Dra. Ariadne Chloe Mary Furnival (UFSCar)

Profa. Dra. Luzia Sigoli Fernandes Costa (UFSCar)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência, Tecnologia e Sociedade.

DEDICATÓRIA

Dedico este trabalho aos meus pais, a minha noiva e ao meu futuro.

AGRADECIMENTOS

Primeiro, agradeço a Deus pelas bênçãos em minha vida, por iluminar as minhas decisões e por me dar força e sabedoria ao longo de todo o trajeto percorrido para a concretização de mais este sonho.

À minha família, sobretudo aos meus queridos pais, Noé e Rosemary, exemplos de coragem, amor e união, por acreditarem e torcerem por mim durante a minha formação acadêmica e por me dar condições de fazer o doutorado.

À minha amada noiva Cione, fonte de amor, energia, carinho e compreensão, por toda a sua paciência comigo nos períodos de ausência e pela sua cumplicidade, companheirismo, apoio constante e segurança incondicionais. Parceira de todas as horas, porto seguro e de tranquilidade. Te amo muito!

À minha orientadora Profa. Dra. Luciana de Souza Gracioso, pela amizade e parceria de estudos desde os tempos de graduação, por sua confiança, paciência e dedicação depositada neste trabalho, pelas suas contribuições e sabedoria, pelo aprendizado obtido, por seu constante incentivo, e pelos generosos convites para ministrar aulas e palestras como participar de projetos e eventos no decurso do doutorado.

Aos membros titulares da Comissão Examinadora de Qualificação e de Defesa, em especial, a Profa. Dra. Ariadne Chloe Mary Furnival da Universidade Federal de São Carlos (UFSCar), a Profa. Dra. Luzia Sigoli Fernandes Costa da UFSCar, o Prof. Dr. Miguel Ángel Márdero Arellano do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) e a Profa. Dra. Sonia Araújo de Assis Boeres, além dos membros suplentes da Comissão Examinadora de Defesa Prof. Dr. Felipe Augusto Arakaki da Universidade de Brasília (UnB), a Profa. Dra. Márcia Regina da Silva da Universidade de São Paulo (USP) e o Prof. Dr. Roniberto Morato do Amaral da UFSCar, pela nossa amizade, pelo interesse, prontidão e disposição em aceitarem compor a banca deste trabalho e pelas valiosas recomendações para o seu desenvolvimento.

À CAPES pelo financiamento desta pesquisa.

Aos professores, colegas de turma e funcionários do Programa de Pós-Graduação em Ciência, Tecnologia e Sociedade (PPGCTS) da UFSCar, pela convivência harmônica e pelos valores, conhecimentos e ensinamentos compartilhados que levarei para a toda a vida.

Finalmente, a todos que direta ou indiretamente colaboraram para a realização desta pesquisa.

“Nunca deixe ninguém te dizer que não pode fazer algo. Se você tem um sonho, tem que protegê-lo. As pessoas que não podem fazer algo por elas mesmas, lhe dirão que você não o pode fazer. Se você quer uma coisa, vá buscá-la. Ponto final.”

À Procura da Felicidade (2006)

RESUMO

Com o significativo aumento na produção, difusão e consumo de conteúdos digitais no ambiente da *Web*, ações de recuperação, preservação e comunicação de dados e informações, de valor científico, cultural e histórico produzidos em *websites*, mídias sociais e em outros conteúdos baseados na *Web*, estão sendo criadas e estudadas defronte à efemeridade da *Internet* que impõe rápidas alterações ou, mesmo, perdas permanentes aos recursos informacionais. A preservação digital e arquivamento de conteúdos publicados na *Web* é uma temática recente no Brasil, especificamente discutida na Ciência da Informação e, de modo geral, não contemplada no campo da Ciência, Tecnologia e Sociedade (CTS) sendo necessário o trânsito interdisciplinar com a Ciência da Informação. Nesse contexto, objetiva-se verificar como critérios de seleção de conteúdos da *Web* no âmbito da preservação digital e do arquivamento da *Web* têm sido discutidos pela Ciência da Informação e áreas afins (Arquivologia, Biblioteconomia e Museologia), apontando como estes critérios poderiam atender às demandas de estruturação de arquivos da *Web* institucionais de maneira mais adequada para a preservação a longo prazo de informações digitais. Como metodologia, destaca-se a realização de um estudo exploratório, pautado em pesquisa bibliográfica e documental com revisão de literatura narrativa referente ao arquivamento da *Web* no escopo da preservação digital. Aplicou-se a análise de conteúdo dos critérios e políticas de seleção de conteúdos da *Web* identificados na literatura para fins de preservação digital e arquivamento de longo prazo. Como resultado apresenta-se um quadro referencial teórico, técnico e sistematizado de critérios de seleção aplicáveis à preservação digital e arquivamento da *Web* com base na literatura científica levantada e em políticas de várias iniciativas de arquivos da *Web* no mundo, como critérios para *sites* "antigos" e atuais, critérios geográficos e linguísticos, critérios relativos ao assunto, critérios pautados nos usos do arquivo da *Web* e em seus usuários ou especialistas, ou critérios relativos ao formato e, simultaneamente, que podem ser relevantes para a modelagem das coleções de arquivos da *Web* institucionais. Constatou-se que a imperfeição do processo de arquivamento da *Web* e a incompletude dos seus produtos, isto é, dos arquivos da *Web*, em termos de integralidade dos *sites* arquivados, decorre não apenas de limites técnicos, temporais, orçamentários ou legais, mas também de limites atrelados a tomada de decisões de seleção. As decisões de seleção adotadas nas políticas de arquivos da *Web* trazem algum grau de subjetividade e deliberação, o que pode ser evitado com a justificação explícita dos critérios em uma política de seleção, tornando o arquivamento e o seu produto final (coleções) coerentes. Além disto, dentro do desenvolvimento de uma política de seleção de arquivamento da *Web*, foi possível verificar que nos escopos dos métodos de seleção adotados na criação de coleções *Web*, no escopo, cobertura do arquivamento e o alvo (extensão) da coleção em si, na definição do tipo e dos meios de acesso do arquivo da *Web*, e na definição dos usos esperados do arquivo da *Web* e dos usuários pretendidos, delineiam-se diferentes critérios para a seleção de conteúdos acordados na política.

Palavras-chave: Arquivamento da *Web*. Preservação digital. Gestão de conteúdos na *Web*. Critérios de seleção. Desenvolvimento de coleções.

ABSTRACT

With the significant increase in production, diffusion and consumption of digital content in the Web environment, actions for the recovery, preservation and communication of data and information, of scientific, cultural and historical value produced in websites, social media and other Web-based content, are being created and studied in face of the ephemerality of the Internet that imposes rapid changes or even permanent losses to informational resources. Digital preservation and archiving of Web-based content is a recent theme in Brazil, specifically discussed in Information Science and, in general, not contemplated in the field of Science, Technology and Society (STS), being necessary the interdisciplinary transit with Information Science. In this context, we aim to verify how criteria for selecting Web content in the context of digital preservation and Web archiving have been discussed by Information Science and related areas (Archivology, Librarianship and Museology), pointing out how these criteria could meet the demands for structuring institutional Web archives in a more adequate way for the long-term preservation of digital information. As methodology, an exploratory study was carried out, based on bibliographic and documentary research with a narrative literature review concerning Web archiving in the scope of digital preservation. Content analysis was applied to Web content selection criteria and policies identified in the literature for digital preservation and long-term archiving purposes. As a result, a theoretical, technical and systematized framework of selection criteria applicable to Web archiving and digital preservation is presented, based on the surveyed scientific literature and on policies from various Web archiving initiatives around the world, such as criteria for "old" and "current" sites, geographic and linguistic criteria, subject-related criteria, criteria based on Web archiving uses and its users or experts, or format-related criteria, and, simultaneously, that may be relevant for modeling institutional Web archive collections. It was found that the imperfection of the Web archiving process and the incompleteness of its products, that is, the Web archives, in terms of the completeness of the archived sites, stems not only from technical, temporal, budgetary or legal limits, but also from limits tied to selection decisions. The selection decisions adopted in Web archiving policies bring some degree of subjectivity and deliberation, which can be avoided by explicitly justifying the criteria in a selection policy, making the archiving and its final product (collections) coherent. Furthermore, within the development of a Web archiving selection policy, it was possible to verify that in the scopes of the selection methods adopted in the creation of Web collections, in the scope, coverage of the archiving and the target (extent) of the collection itself, in the definition of the type and means of access of the Web archive, and in the definition of the expected uses of the Web archive and the intended users, different criteria for the selection of contents agreed upon in the policy are delineated.

Keywords: Web archiving. Digital preservation. Web content management. Selection criteria. Collection management.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SIGLAS

AACR2	*Anglo-American Cataloguing Rules*
AAT	*Art & Architecture Thesaurus*
ABC	*Australian Broadcasting Corporation*
ABJC	Associação Brasileira de Jornalismo Científico
ACT-IAC	*American Council for Technology-Industry Advisory Council*
ADS	*Archaeology Data Service*
AGWA	*Australian Government Web Archive*
AHRC	*Arts and Humanities Research Council*
ANSI	*American National Standards Institute*
API	*Application Programming Interface*
ARL	*Association of Research Libraries*
ASIS&T	*Association for Information Science and Technology*
*Audio*MD	*Audio Technical Metadata Extension Schema*
AWA	*Australian Web Archive*
BDTD	Biblioteca Digital Brasileira de Teses e Dissertações
BIBFRAME	*Bibliographic Framework*
BnF	*Bibliothèque Nationale de France*
Brexit	*British exit*
BVS	Biblioteca Virtual em Saúde
C&T	Ciência e Tecnologia
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
Cariniana	Rede Brasileira de Serviços de Preservação Digital
CCO	*Cataloging Cultural Objects*
CCSDS	*Consultative Committee for Space Data Systems*
CDG	*Content Development Working Group*
CDL	*California Digital Library*
CDWA	*Categories for the Description of Works of Art*
CERTH	*Centre for Research and Technology-Hellas*
CGEE	Centro de Gestão e Estudos Estratégicos
CIDOC	*International Committee for Documentation*
CLIR	*Council on Library and Information Resources*
CLOCKSS	*Controlled* LOCKSS
CONARQ	Conselho Nacional de Arquivos
CORD-19	COVID-19 *Open Research Dataset*
COSMOS	*Collaborative Online Social Media Observatory Software*
CRL	*Center for Research Libraries*
CRM	*Conceptual Reference Model*
CSS	*Cascading Style Sheets*
CST	*Council for Science and Technology*
CTS	Ciência, Tecnologia e Sociedade
CTS	*Ciência, Tecnología y Sociedad*
CUL	*Columbia University Libraries*

DACS	*Describing Archives: a Content Standard*
DADVSI	*droits d'auteur et les droits voisins dans la société de l'information*
DArcMail	*Digital Archives of Email*
DAVID	*Digitale archivering in/voor Vlaamse instellingen en diensten*
DC	*Dublin Core*
DCMES	*Dublin Core Metadata Element Set*
DCMI	*Dublin Core Metadata Initiative*
DLF	*Digital Library Federation*
DNS	*Domain Name System*
DOI	*Digital Object Identifier*
DPC	*Digital Preservation Coalition*
DSpace	*Durable Space*
DTD	*Document Type Definition*
EAC-CPF	*Encoded Archival Context – Corporate Bodies, Persons, and Families*
EAD	*Encoded Archival Description*
ECHO DEPository	*Exploring Collaborations to Harness Objects in a Digital Environment for Preservation*
eDAVID	*expertisecentrum* DAVID *vzw*
EJC	*Electronic Journal Center*
EML	*Internet Message Format*
ePADD	*Email: Process, Appraise, Discover, Deliver*
FABICO	Faculdade de Biblioteconomia e Comunicação
FCCN	Fundação para a Computação Científica Nacional
FCT	Fundação para a Ciência e a Tecnologia
Fedora	*Flexible Extensible Digital Object Repository Architecture*
FelixArchief	*Antwerp City Archives*
FINEP	Financiadora de Estudos e Projetos
Fiocruz	Fundação Oswaldo Cruz
FOAF	*Friend of a Friend*
FOIA	*Freedom of Information Act*
FTP	*File Transfer Protocol*
HTML	*HyperText Markup Language*
HTTP	*Hypertext Transfer Protocol*
IBICT	Instituto Brasileiro de Informação em Ciência e Tecnologia
ICA	*International Council on Archives*
ICOM	*International Council of Museums*
IDF	*International* DOI *Foundatione*
IEC	*International Electrotechnical Commission*
IEEE	*Institute of Electrical and Electronics Engineers*
IES	Instituições de Ensino Superior
IETF	*Internet Engineering Task Force*
IFLA	*International Federation of Library Associations and Institutions*
IIPC	*International Internet Preservation Consortium*
IMLS	*Institute of Museum and Library Services*

INA	*Institut national de l'audiovisuel*
IRIS	*Institutional Repository for Information Sharing*
ISA	*Interoperability Standards Advisor*
ISAAR (CPF)	*International Standard Archival Authority Record for Corporate Bodies, Persons and Families*
ISAD(G)	*General International Standard Archival Description*
ISBD	*International Standard Bibliographic Description*
ISO	*International Organization for Standardization*
ISSN	*International Standard Serial Number*
ISTA	*Information Science & Technology Abstracts*
JISC	*Joint Information Systems Committee*
JSON	*JavaScript Object Notation*
JSON-LD	*JavaScript Object Notation for Linked Data*
JSTOR	*Journal Storage*
K.U. *Leuven*	*Katholieke Universiteit Leuven*
KB	*Koninklijke Bibliotheek*
LCGFT	*Library of Congress Genre/Form Terms*
LCNAF	*Library of Congress Name Authority File*
LCSH	*Library of Congress Subject Headings*
LGPD	Lei Geral de Proteção de Dados Pessoais
LIDO	*Lightweight Information Describing Objects*
LIPA	*Legal Information Preservation Alliance*
LISA	*Library & Information Science Abstracts*
LISTA	*Library, Information Science & Technology Abstracts with Full Text*
LOCKSS	*Lots of Copies Keep Stuff Safe*
LOM	*Learning Object Metadata*
LSTA	*Library Services and Technology Act*
MADS	*Metadata Authority Description Schema*
MAG	*Microsoft Academic Graph*
MARC	*Machine Readable Cataloging*
MCTIC	Ministério da Ciência, Tecnologia, Inovações e Comunicações
MeSH	*Medical Subject Headings*
METS	*Metadata Encoding and Transmission Standard*
MINERVA	*Mapping the Internet Electronic Resources Virtual Archive*
MIT	*Massachusetts Institute of Technology*
MIX	NISO *Metadata for Images in* XML *Schema*
MOA2	*Making of America* II
MODS	*Metadata Object Description Schema*
MPEG	*Moving Picture Experts Group*
MPEG-7	*Multimedia Content Description Interface*
MSG	*Microsoft Outlook Item*
MUSE	*Memories USing Email*
NCBI	*National Center for Biotechnology Information*
NCSA	*National Center for Supercomputing Applications*

NDIIPP	*National Digital Information Infrastructure and Preservation Program*
NDSA	*National Digital Stewardship Alliance*
NHPRC	*National Historical Publications and Records Commission*
NHS	*National Health System*
NIC.br	Núcleo de Informação e Coordenação do Ponto BR
NISO	*National Information Standards Organization*
NLM	*National Library of Medicine*
NLP	*Natural Language Processing*
NUAWEB	Núcleo de Pesquisa em Arquivamento da *Web* e Preservação Digital
NUK	*Narodne in Univerzitetne Knjižnice*
NYARC	*New York Art Resources Consortium*
OAC	*Online Archive of California*
OAI	*Open Archives Initiative*
OAI-PMH	*Open Archives Initiative Protocol for Metadata Harvesting*
OAIS	*Open Archival Information System*
OASIS	*Online Archiving & Searching Internet Sources*
OCLC	*Online Computer Library Center*
OCUL	*Ontario Council of University Libraries*
*Ohio*LINK	*Ohio Library and Information Network*
OJS	*Open Journal Systems*
ONC	*Office of the National Coordinator for Health Information Technology*
OSZK	*Országos Széchényi Könyvtár*
PADICAT	*Patrimoni Digital de Catalunya*
PAHO	*Pan American Health Organization*
PAI	Pacote de Arquivamento de Informação
PANDORA	*Preserving and Accessing Networked DOcumentary Resources of Australia*
Archive	
PDF	*Portable Document Format*
PDI	Pacote de Disseminação de Informação
PKP	*Public Knowledge Projec*
PMC	*PubMed Central*
PPGCTS	Programa de Pós-Graduação em Ciência, Tecnologia e Sociedade
PREMIS	*PREservation Metadata: Implementation Strategies*
PSI	Pacote de Submissão de Informação
PST	*Personal Storage Table*
PURL	*Persistent Uniform Resource Locator*
PUS	*Public Understanding of Science*
RDA	*Resource Description and Access*
RDBCI	Revista Digital de Biblioteconomia e Ciência da Informação
RDF	*Resource Description Framework*
RDFa	RDF *in Attributes*
REDES	*Revista de Estudios Sociales de Ciencia*
RFC	*Request for Comments*
RLG	*Research Libraries Group*

RODA	Repositório de Objetos Digitais Autênticos
RSC	RDA *Steering Committee*
RSS	*Really Simple Syndication*
RTS	Revista Tecnologia e Sociedade
SAA	*Society of American Archivists*
SAAI	Sistema Aberto de Arquivamento de Informação
SARA	*Search and Retrieval Application*
SBPC	Sociedade Brasileira para o Progresso da Ciência
SciELO	*Scientific Electronic Library Online*
SFM	*Social Feed Manager*
SGML	*Standard Generalized Markup Language*
SHU	*Seton Hall University*
SIA	*Smithsonian Institution Archives*
SNPC	*Servicio Nacional del Patrimonio Cultural*
SODATO	*Social Data Analytics Tool*
SPAR	*Système de Préservation et d'Archivage Réparti*
SPHSU	*Social and Public Health Sciences Unit*
SSS	*Social Studies of Science*
SUCHO	*Saving Ukraine Cultural Heritage Online*
SUL	*Stanford University Libraries*
SWAP	*Stanford Web Archive Portal*
TDIC	Tecnologias Digitais de Informação e Comunicação
TDR	*Trusted Digital Repository Checklist*
TEI	*Text Encoding Initiative*
*Text*MD	*Technical Metadata for Text*
TGM	*Thesaurus for Graphic Materials*
TGN	*Thesaurus of Geographic Names*
TOMES	*Transforming Online Mail with Embedded Semantics*
TRAC	*Trustworthy Repositories Audit & Certification: Criteria and Checklist*
TS-DACS	*Technical Subcommittee for* DACS
TS-EAS	*Technical Subcommittee for Encoded Archival Standards*
UC3	*University of California Curation Center*
UFG	Universidade Federal de Goiás
UFMG	Universidade Federal de Minas Gerais
UFPB	Universidade Federal da Paraíba
UFRGS	Universidade Federal do Rio Grande do Sul
UFSC	Universidade Federal de Santa Catarina
UFSCar	Universidade Federal de São Carlos
UFSM	Universidade Federal de Santa Maria
UKGWA	UK *Government Web Archive*
UKWA	UK *Web Archive*
ULAN	*Union List of Artist Names*
UML	*University of Manchester Library*
UnB	Universidade de Brasília

UNESP	Universidade Estadual Paulista
UNICAMP	Universidade Estadual de Campinas
URI	*Uniform Resource Identifier*
URL	*Uniform Resource Locator*
URN	*Uniform Resource Name*
US	*United States*
USP	Universidade de São Paulo
UTSA	*University of Texas at San Antonio*
*Video*MD	*Video Technical Metadata Extension Schema*
VRA	*Visual Resources Association*
VRA *Core*	*Visual Resources Association Core*
W3C	*World Wide Web Consortium*
WAHR	*Web Archives for Historical Research*
WAM	*Web Archiving Metadata*
WARC	*Web ARChive*
WARP	*Web Archiving Project*
WAS	*Web Archive Singapore*
WAS	*Web Archiving Service*
WAX	*Web Archive Collection Service*
WHO	*World Health Organization*
XHTML	*Extensible Hypertext Markup Language*
XML	*Extensible Markup Language*
XSD	XML *Schema Definition*

SUMÁRIO

1 INTRODUÇÃO

1.1 Apresentação do tema de investigação, questão de pesquisa, hipótese e justificativas

As sociedades contemporâneas são grandes produtoras e consumidoras de informação, constituindo-se, esta, em um bem imprescindível para o seu desenvolvimento. O uso intensivo de Tecnologias Digitais de Informação e Comunicação (TDIC) associado a *World Wide Web* criada por Tim Berners-Lee em 1989 e a *Internet*, propiciou uma explosão informacional mundial com a rápida produção, difusão e consumo de conteúdos sobretudo na *Web*, seja nas esferas públicas ou privadas. As páginas *Web* (com áudios, vídeos, textos e documentos digitais ou eletrônicos integrados) e também as mídias sociais provém uma imediata publicação e acesso a materiais informativos por *hiperlinks*, multimídias e linguagens de marcação. No entanto, a dinamicidade da *Internet* impondo alterações ou perdas rápidas de informações somado ao fato de grande parte dos eventos e conteúdos serem produzidos, publicados e comunicados apenas no ambiente *Web* obrigaram as instituições de memória (arquivos, bibliotecas e museus) e as universidades, a desenvolverem então iniciativas de preservação digital e arquivamento de conteúdos na *Web*.

Até a década de 1970 as tomadas de decisões relativas à ciência e tecnologia (C&T) refletiam fundamentalmente aos anseios e as vontades dos próprios cientistas ou especialistas, todavia, após as sequelas deixadas pelas guerras e pela devastação do meio ambiente, o apoio incondicional e o aspecto puramente romântico, positivo e otimista da qual se tinha acerca dos avanços em C&T acabou se alterando. Parte das sociedades passaram a dispor então de uma visão mais crítica sobre o assunto de modo a inseri-lo em movimentos sociais e debates políticos e, sobretudo, a questionar os progressos e as aplicações em C&T que não estavam satisfazendo as demandas e o desenvolvimento do bem estar social de toda a população. Neste cenário, iniciou-se no âmbito acadêmico o denominado movimento ou enfoque CTS.

O CTS, campo de estudos e de trabalho acadêmico interdisciplinar surtido nas linhas de investigação advindas das áreas de Filosofia e de Sociologia da Ciência, hoje consolidado institucionalmente, objetiva "[...] ressaltar a importância social da ciência e da tecnologia, de forma a enfatizar a necessidade de avaliações críticas e análises reflexivas sobre a relação científico-tecnológica e a sociedade." (PINHEIRO; SILVEIRA; BAZZO, 2007, p. 74). Como um dos princípios primordiais dos estudos CTS, está a defesa do acesso das sociedades às informações quanto aos avanços científico-tecnológicos e a contextualização das atividades de C&T como processos sociais resultantes de fatores culturais, políticos e econômicos e, segundo

Auler e Bazzo (2001), a reivindicação por decisões mais democráticas e menos tecnocráticas, com a atuação de um maior número de atores sociais nas tomadas de decisões referentes à C&T.

Ao retratar o campo de estudos das inter-relações entre CTS, López Cerezo (2002, p. 6) citando Barnes[1] (1987) e Latour[2] (1992) pontua que:

> O ponto-chave é a apresentação da ciência-tecnologia não como um processo ou atividade autônoma, que segue uma lógica interna de desenvolvimento em seu funcionamento ótimo, mas como um processo ou produto inerentemente social, em que os elementos não técnicos (por exemplo, valores morais, convicções religiosas, interesses profissionais, pressões econômicas, etc.) desempenham um papel decisivo em sua gênese e consolidação. A complexidade dos problemas abordados e sua flexibilidade interpretativa, a partir de distintos marcos teóricos, fazem necessária a presença desses elementos não técnicos, na forma de valores ou de interesses contextuais.

Existe uma variedade de programas de colaboração multidisciplinar que constituem os estudos CTS que, de acordo com López Cerezo (2002, p. 9), focam a dimensão social da C&T, pois compartilham em comum "(a) a rejeição da imagem da ciência como uma atividade pura; (b) a crítica da concepção da tecnologia como ciência aplicada e neutra; e (c) a condenação da tecnocracia". Para Santos e Ichikawa (2002, p. 240) as tradições teóricas em CTS são duas:

> - A tradição européia, que nasceu com os "Programas Fortes" de sociologia do conhecimento científico, e que centra seu estudo na análise dos antecedentes ou os condicionantes da ciência; e
> - A tradição norte-americana, que centra seus estudos nas conseqüências sociais e ambientais do conhecimento científico.

A tradição europeia, originada nos "programas fortes" concluídos na década de 70 por autores como David Bloor, Barry Barnes ou Steven Shapin e marcada por uma interpretação da obra de Thomas Kuhn, é definida mais pela tradição de pesquisa acadêmica do que educativa ou divulgativa; já a tradição americana, o qual teve seu início ligado ao movimento pragmático norte-americano e a obra de ativistas ambientais e sociais como Rachel Carson e E. F. Schumacher[3], é muito mais ativista, firmando-se institucionalmente mediante o ensino e a reflexão política (GONZÁLEZ GARCÍA; LÓPEZ CEREZO; LUJÁN[4], 1996, citado por CEREZO, 2002). Os estudos e programas CTS, de acordo com Santos e Ichikawa (2002), estão

[1] BARNES, Barry. **Sobre ciencia**. Barcelona: Labor, 1987.
[2] LATOUR, Bruno. **Ciencia en acción**. Barcelona: Labor, 1992.
[3] Apesar desta afirmação dos autores, E. F. Schumacher nasceu na Alemanha e se radicou no Reino Unido.
[4] GONZÁLEZ GARCÍA, M.; LÓPEZ CEREZO, J. A.; LUJÁN, J. L. **Ciencia, tecnología y sociedad**: una introducción al estudio social de la ciencia y la tecnología. Madrid: Tecnos, 1996.

sendo desenvolvidos em três direções, a saber: no campo da investigação, das políticas públicas técnicos-científicas e da educação, introduzindo as discussões do enfoque CTS.

Nesta perspectiva, compreendemos que a preservação digital e o arquivamento da *Web*, investigada pela Ciência da Informação, encontra funções e relações recíprocas nos objetivos pretendidos pelos estudos CTS, oportunizando aos indivíduos de hoje e as futuras gerações a formação de uma memória digital e a preservação do acesso aos conteúdos de C&T produzidos, publicados e difundidos na *Web*. Todavia, prover e assegurar o acesso às informações não é o suficiente pois, ao contextualizar o enfoque CTS para o ensino médio, Pinheiro, Silveira e Bazzo (2007) destacam que é preciso que a população tenha também meios que lhe permita avaliar e participar das decisões que venham a interferir no âmbito do qual esteja presente.

Considerando aderir ao movimento CTS no contexto educacional brasileiro frente a nossa inexperiência democrática constatada na trajetória histórica do país, Auler e Bazzo (2001, p. 12) salientam que "[...] além de conhecimentos/informações, necessários para uma participação mais qualificada da sociedade, necessitamos, também, iniciar a construção de uma cultura de participação". Isto posto, dois caminhos são essenciais para uma maior participação qualificada da população: a inserção do enfoque CTS na educação em C&T oferecida nas escolas e a educação não-formal proveniente dos museus e centros de ciência.

Estas ponderações se justificam, pois dados obtidos na enquete nacional de percepção pública da C&T de 2010 pelo Ministério da Ciência, Tecnologia, Inovações e Comunicações (MCTIC) e pelo Museu da Vida da Fundação Oswaldo Cruz (Fiocruz) refutam a hipótese de que um maior nível de instrução ou informação desencadearia a atitudes em geral mais positivas acerca do papel da C&T na sociedade, pois: I) Em torno de 60% dos brasileiros declararam ter um grande interesse por assuntos de C&T (e também uma postura positiva e otimista), porém, possuem uma carência de conhecimento destes assuntos e acessam pouca informação científica; e II) Com o aumento da informação, os indivíduos passam a valorizar a potência ligada ao conhecimento científico e tecnológico, entretanto, enfocando também os riscos e perigos (CASTELFRANCHI *et al*., 2013).

Por sua vez, dados de 2019, do estudo de percepção pública de C&T do Centro de Gestão e Estudos Estratégicos (CGEE) e do MCTIC, permitem conhecer a visão, o interesse (ou atribuição de importância) e o grau de informação dos brasileiros sobre C&T, onde: I) Cerca de 73% e 62% dos brasileiros declararam ter uma visão otimista sobre os efeitos da C&T para a sociedade e um grande interesse (em particular, entre pessoas de alta renda e escolaridade) por temas de cunho técnico científico, respectivamente; II) Em oposição, temos um escasso acesso e apropriação do saber científico pelos brasileiros, incluindo uma baixa

visitação/participação de atividades em espaços de C&T (maiormente, os museus de ciências e entre pessoas de baixa renda familiar) e demandas por controle e participação social diante dos aspectos éticos e da compreensão aos riscos socioambientais e de saúde resultantes dos avanços tecnocientíficos; e III) O consumo de informação de C&T na *Internet* se manteve baixo ainda que cerca de 70% dos brasileiros declararam acessar a *Internet* todos os dias e 19% declararam não ter acesso a mesma, aliás, a maioria dos brasileiros disseram "nunca" ou "raramente" procurar informação sobre o tema em qualquer mídia, sendo que na *Internet*, o acesso a informações tecnocientíficas por parte dos brasileiros é dominado por *websites* de busca (21%) e as plataformas do *Facebook* (13%) e do *Youtube* (11%) (CENTRO DE GESTÃO E ESTUDOS ESTRATÉGICOS, 2019).

Em adição, Medeiros e Costa (2017) trazem ainda um estudo recente do uso de redes sociais na C&T, com a análise da divulgação científica (e os seus impactos na sociedade) no ambiente digital das redes, por meio das teorias da Sociologia da ciência. Tendo em vista que o impacto social da C&T nas redes é determinado por todos os usuários-atores participantes das redes sociais, tais como, cientistas, universidades, jornalistas, instituições, leigos etc., os autores demonstraram que existe uma escassez na adoção destas ferramentas de comunicação (em específico, o *Facebook*) para divulgação científica entre as principais revistas de ciência no Brasil de maior acesso na *Scientific Electronic Library Online* (SciELO)[5], onde a atividade e/ou a interação das páginas das revistas no *Facebook* com demais usuários desta rede social são inexpressivas.

Assim, a preservação digital e o arquivamento da *Web* (com a definição de critérios para a seleção dos conteúdos) torna-se um grande desafio e oportunidade para o movimento CTS. O acesso aos conteúdos da *Web* sobre C&T junto com a criação, a avaliação e a melhoria de políticas técnico-científicas abertas à atuação da população e de modelos de programas práticos de popularização, comunicação e educação em C&T, contribuirá numa aproximação da ciência com o público não especialista e na formação de indivíduos menos passivos e alienados intelectualmente que exerçam, numa democracia, os seus factuais direitos e deveres como cidadãos nas sociedades.

À vista disto, o público geral e especialista poderá obter então um posicionamento mais autônomo, pertinente e crítico-avaliativo, assim como uma participação mais ativa, efetiva e de maior qualidade nas opiniões ou tomadas de decisões na área de C&T, proporcionando, além do mais, a compreensão, o questionamento e a reflexão crítica a respeito dos princípios, dos

[5] Disponível em: https://scielo.org/. Acesso em: 4 jun. 2023.

valores, dos interesses, das causas, das consequências sociais e ambientais e, também, das questões culturais, políticas, econômicas e/ou éticas intrínsecas nas interações entre CTS.

A Ciência da Informação, campo de pesquisa científica e de prática profissional interdisciplinar definida pelo imperativo tecnológico (SARACEVIC, 1996), propõem-se a investigar "[...] as propriedades e o comportamento da informação, as forças que governam o fluxo e o uso da informação e as técnicas, tanto manuais quanto mecânicas, de processamento de informação para otimizar o armazenamento, a recuperação e a disseminação.", conforme Borko (1968, p. 5, tradução nossa). Dessa maneira, o enfoque CTS constitui um caminho teórico próspero, procurando a promoção do acesso aos conteúdos que motivem uma visão crítica-avaliativa das interações entre CTS.

Na literatura científica do campo da Ciência da Informação existem algumas definições de preservação digital. Em Duranti (2010, p. 157, tradução nossa) a preservação digital consiste no "[...] conjunto de princípios, políticas, regras e estratégias destinadas a prolongar a existência do objeto digital, mantendo-o em condições adequadas para uso [...]" e "[...] protegendo a identidade e integridade do objeto, ou seja, sua autenticidade." Estes objetos digitais, nascidos digitais ou digitalizados, são todos os tipos de conteúdos em meio digital, como textos, imagens, vídeos, áudios, jogos, *sites*, mídias sociais, *e-mails* etc., dos quais a preservação digital pode agir, constituindo "[...] itens na forma digital que requerem um computador para dar suporte à sua existência e apresentação visual." (PINHEIRO; FERREZ, 2014, p. 163) e que, para Baucom (2019, p. 5, tradução nossa), são compostos por "[...] cadeias de uns e zeros, que requerem componentes específicos de *software* e *hardware* para permanecerem acessíveis aos usuários."

Por sua vez, na literatura específica do campo CTS, as temáticas de preservação digital e de arquivamento da *Web* não estão suficientemente contempladas, pois em um breve levantamento realizado em maio de 2023 com os termos de busca "preservação digital", "*digital preservation"*, "*preservación digital*", "arquivamento da *Web*", "*Web archiving*", "*archivado Web*", "preservação da *Web*", "*preserving the Web*" e "*preservación de la Web*" nos principais periódicos nacionais e internacionais relacionados com a área (por exemplo, *Social Studies of Science* – SSS[6], *Revista Iberoamericana de Ciência, Tecnología y Sociedad* – CTS[7], *Revista de Estudios Sociales de Ciencia* – REDES[8], *Public Understanding of Science* – PUS[9] e Revista Tecnologia e Sociedade – RTS[10]) não houve ocorrência de quaisquer resultados. Diante disso,

[6] Disponível em: https://journals.sagepub.com/home/sss. Acesso em: 4 jun. 2023.
[7] Disponível em: http://www.revistacts.net/. Acesso em: 4 jun. 2023.
[8] Disponível em: http://iec.unq.edu.ar/index.php/en/publications/redes-journal. Acesso em: 4 jun. 2023.
[9] Disponível em: https://journals.sagepub.com/home/pus. Acesso em: 4 jun. 2023.
[10] Disponível em: https://periodicos.utfpr.edu.br/rts. Acesso em: 4 jun. 2023.

torna-se necessário realizar este trânsito interdisciplinar com o campo da Ciência da Informação e as suas áreas afins.

Para solucionar os desafios da preservação digital, um conjunto de estratégias têm sido propostas pela comunidade científica, que podem ser reunidas em duas categorias genéricas, a saber: as estratégias estruturais, que tratam dos investimentos iniciais advindos das instituições com o intuito de construir um ambiente adequado para implementação da preservação digital, tal como a adoção de padrões de metadados para preservação; e as estratégias operacionais, que compreendem as medidas reais de preservação digital a serem desenvolvidas pelas instituições com o propósito de preservar por longo prazo objetos digitais, como a migração, a emulação, a conservação de tecnologia, a transferência para suportes analógicos e a determinação do meio de armazenamento (BULLOCK, 1999; FORMENTON, 2015; FORMENTON; GRACIOSO; CASTRO, 2015; MÁRDERO ARELLANO, 2008; THOMAZ, 2004; THOMAZ; SOARES, 2004).

Uma destas estratégias estruturais trata-se do uso de manuais e guias criados em várias partes do mundo por iniciativas, programas e projetos (FORMENTON, 2015; FORMENTON; GRACIOSO; CASTRO, 2015), que propiciam diretrizes à preservação digital e o arquivamento de conteúdos, como as informações publicadas na *Web*. Dentre estas iniciativas existentes estão aquelas dirigidas a prestar recomendações para o desenvolvimento de políticas de preservação digital, possibilitando amparar o gerenciamento da preservação e do acesso utilizável em longo prazo de conteúdos digitais produzidos, selecionados, coletados e armazenados em instituições, com garantias de autenticidade, integridade e não perdas permanentes de dados. Uma política desta natureza deve ser dinâmica e readequada regularmente às demandas atuais e futuras e ao contexto da instituição, considerando aspectos administrativos, econômicos, legais, culturais e técnicos que são fundamentais para a consecução dos procedimentos e objetivos institucionais.

Algumas publicações nacionais do campo da Ciência da Informação têm trazido relatos de experiências e metodologias para a elaboração, implantação e manutenção de políticas de preservação digital em Instituições de Ensino Superior (IES), os quais refletem o contexto legal, político, econômico e social no país, conforme podemos constatar em Almeida, Cendón e Souza (2012), Boeres e Faria (2012), Ferreira, Gadelha e Gamba (2012), Grácio (c2012), Grácio, Fadel e Valentim (2013), Santos, Passos e Sae (2012), Schäfer e Constante (2013) e Silva Junior e Mota (2012).

Grácio (c2012) e Grácio, Fadel e Valentim (2013) propõem um modelo de política e administração da preservação a longo prazo da informação digital em IES baseado em quinze elementos fundamentais, inter-relacionados e embasados nas TIC e na cultura informacional (e

organizacional) da instituição, que são distribuídos em três aspectos: organizacional, legal e técnico. Somando-se as considerações de determinados estudos analisados neste trabalho, os aspectos podem ser interpretados conjuntamente e descritos da seguinte forma:

a) Aspectos organizacionais – inclui elementos de gestão para fixação e estabilização institucional sobre a política e o desenvolvimento das medidas de preservação, como:

- Determinação da missão, da visão e dos objetivos institucionais com a inserção da necessidade da implantação de uma política de preservação digital. Os tipos de informação digital a serem preservadas devem ser definidas nesta categoria, refletindo o requisito para a preservação digital "Fixar os limites do objeto a ser preservado" apontado por Bullock (1999), Formenton, Gracioso e Castro (2015) e Thomaz (2004), o qual representa a definição clara de quais elementos do objeto digital serão realmente mantidos mediante uma política de seleção.

b) Aspectos legais - inclui questões legais concernentes ao âmbito institucional e a legislação vigente a nível nacional e internacional, como:

- Adesão às leis nacionais sobre os aspectos que envolvam a preservação digital, a fim de assegurar a legalidade e a padronização dos processos e os direitos de propriedade intelectual aos autores/produtores dos objetos e sua autenticidade, além de suporte profissional jurídico à IES para o conhecimento da legislação internacional vigente e de suas normas. Como exemplo, ressaltamos a Lei nº 12.527[11], de 2011, que regulamenta o direito de acesso dos cidadãos a informações produzidas ou sob a guarda de entidades públicas, como as IES; a Lei nº 9.610[12], de 1998, que defini os direitos autorais no país; e a Lei nº 13.853[13], de 2019, ou Lei Geral de Proteção de Dados Pessoais (LGPD), para dispor sobre a proteção de dados pessoais. Os direitos autorais trazem impasses legais e éticos na coleta, cópia e uso de conteúdos para fins de preservação e acesso produzidos em ambientes digitais, como a *Web* (ROCKEMBACH, 2018; SANTOS, 2020).

c) Aspectos técnicos - inclui elementos técnicos sobre os fluxos, os processos e as medidas de preservação digital, como:

- Criação e revisão contínua de critérios e diretrizes para seleção e descarte de objetos digitais pelas instituições, pautando-se na missão e nos objetivos institucionais, na legislação vigente, nos atos normativos internos, nas demandas da comunidade e, ainda,

[11] Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 4 jun. 2023.
[12] Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9610.htm. Acesso em: 4 jun. 2023.
[13] Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2019/lei/l13853.htm. Acesso em: 4 jun. 2023.

> em condições de definição de prioridades e de custo-benefício de realização da preservação. É primordial estabelecer uma criteriosa política de seleção para se determinar o que será preservado para um específico fim e período, considerando questões éticas, sociais e legais, pois a salvaguarda permanente de tudo e para todos torna-se inviável e nulo em razão, por exemplo, dos custos e da complexidade (BOERES; MÁRDERO ARELLANO, 2005; ROCKEMBACH; PAVÃO, 2018).

Nesse domínio, a pesquisa sobre preservação digital e arquivamento de conteúdos da *Web* está sendo um tema de reflexão e discussão da Ciência da Informação. Trata-se de um problema complexo, atual e insólito nas publicações nacionais desta ciência, necessitando de análises interdisciplinares e soluções multidisciplinares.

A partir deste cenário, a questão de pesquisa a que esta investigação se propôs a responder é: quais critérios de seleção de conteúdos da *Web* poderiam ser considerados pelas instituições que estão desenvolvendo os seus arquivos da *Web*, para que estas pudessem contemplar a preservação digital e o arquivamento da *Web*?

Parte-se da hipótese de que há uma certa carência, no âmbito dos estudos nacionais de preservação digital e arquivamento da *Web* na área da Ciência da Informação de modo específico, e no campo CTS de maneira geral, que analisem profundamente e sistematizem os critérios de seleção aplicáveis à preservação digital e o arquivamento de conteúdos publicados na *Web*. Para mais, supõem-se que não existirá nenhuma ou, talvez, poucas iniciativas nacionais práticas e efetivas de arquivamento da *Web*, haja vista o aspecto atual e singular deste tema nas publicações científicas da Ciência da Informação no Brasil. À vista disso, iremos prever que no contexto internacional existirá vários estudos sobre a temática e, assim, será mais propício de se analisar critérios de seleção de conteúdos da *Web* que têm sido articulados, testados e implantados em importantes iniciativas ao redor do mundo. Frente à carência de pesquisas sobre os dois assuntos no campo CTS brasileiro e ibero-americano, constatada pelo levantamento junto as principais revistas da área, iremos também concentrar a investigação diretamente ao campo da Ciência da Informação e suas áreas afins.

Como pressuposto cada iniciativa terá uma política própria que reflita as diretrizes, a missão e os objetivos das instituições o qual se inserem, porém a adoção de padrões comuns e correlações poderão ocorrer entre os critérios de seleção já que estes irão se fundamentar na análise do assunto dos materiais e das necessidades da comunidade interna e/ou da sociedade. Além disto, aspectos como disposição de equipe qualificada e de recursos tecnológicos para preservação, custo-benefício no armazenamento a longo prazo e restrições dos direitos autorais no uso de materiais poderão influenciar e fazer parte dos critérios de seleção de conteúdos *Web*.

A presente tese está inserida na linha de investigação “Gestão Tecnológica e Sociedade Sustentável” do PPGCTS da UFSCar, que foca “[...] compreender as oportunidades e desafios tecnológicos presentes e futuros, enfrentados por organizações empresariais e públicas [...]” tanto “[...] para formulação de estratégias para desenvolvimento sustentável, social, econômico e ambiental [...]” como “[...] para elaboração de políticas públicas em ciência, tecnologia e inovação.” (UNIVERSIDADE FEDERAL DE SÃO CARLOS, [2023], não paginado). Sendo assim, o diagnóstico, a percepção e a definição de critérios de seleção de conteúdos da *Web* no âmbito da preservação digital e do arquivamento da *Web*, sob o enfoque CTS e da Ciência da Informação, se tornará relevante para: a) entender as oportunidades e os desafios impostos hoje e no futuro sobre o tema, como o tamanho, a efemeridade e a dinamicidade da *Web*, a gestão dos avanços tecnológicos e sua obsolescência que faz recursos *Web* inacessíveis, e as restrições de direitos autorais e outras proteções; e b) contribuir na elaboração de estratégias sustentáveis para políticas institucionais públicas em benefício da salvaguarda e do acesso democrático por longo prazo da sociedade sobre as informações técnico-científicas e culturais na *Web* brasileira.

O arquivamento da *Web* pode ser entendido como um processo que inclui a seleção e coleta, o armazenamento e a recuperação da informação da *World Wide Web*. Assim, visando impedir a perda permanente dos *websites* devido à dinamicidade da *Internet* e desenvolver a preservação digital dos conteúdos, várias iniciativas de preservação digital e arquivamento da *Web* vêm se manifestando no mundo, possuindo diferentes formações e fins com abordagens a nível global, nacional, regional e local. Para auxiliar na proteção e no desenvolvimento da *Web*, consórcios internacionais também foram fundados com o intuito de definir padrões e diretrizes para a expansão e o arquivamento da *Web* ao longo do tempo (ROCKEMBACH, 2018).

Em relação à seleção do que será preservado, Boeres e Márdero Arellano (2005) indicam que um critério básico advém de serviços como o “Serviço de disseminação da informação” das bibliotecas universitárias que se comunica com os usuários para obtenção de suas demandas de informação. Neste aspecto, há duas classificações para a coleta e os tipos de arquivamento dos conteúdos *Web*: a seleção extensiva, onde a coleta engloba os domínios nos seus primeiros níveis apresentando um panorama abrangente da *Web* e, a seleção intensiva, cuja coleta enfoca poucos *websites* a fim de arquivar a maior quantia de níveis, assegurando toda a hierarquia e deslocação entre os *hiperlinks* do arquivo *Web* (ROCKEMBACH; PAVÃO, 2018).

1.2 Objetivos

1.2.1 Objetivo geral

Neste sentido, esta tese teve o objetivo geral de verificar como critérios de seleção de conteúdos da *Web* no âmbito da preservação digital e do arquivamento da *Web* têm sido discutidos pela Ciência da Informação e suas áreas afins (Arquivologia, Biblioteconomia e Museologia), apontando como estes critérios poderiam atender às demandas de estruturação de arquivos da *Web* institucionais de maneira mais adequada para a preservação a longo prazo de informações digitais. O resultado deste mapeamento e análise vislumbrou contribuir para possíveis delimitações do conjunto de diretrizes e políticas a serem adotadas por instituições de patrimônio cultural (bibliotecas, arquivos e museus) e pelas universidades públicas nacionais interessadas e/ou envolvidas com a seleção, coleta, armazenamento, recuperação, preservação e acesso permanente de informações produzidas, publicadas e difundidas no ambiente da *Web*.

1.2.2 Objetivos específicos

Para tal fim foram determinados os seguintes objetivos específicos:

a. Identificar as relações entre a preservação digital e a comunicação científica relatadas nos estudos CTS e na Ciência da Informação, expondo as iniciativas de preservação digital em longo prazo de conteúdos publicados, produzidos e/ou difundidos na *Web*;
b. Identificar na literatura internacional e nacional as pesquisas em preservação digital na Ciência da Informação e áreas afins, oferecendo uma visão ampla e reflexiva dos seus principais problemas, princípios, estratégias e produção científica;
c. Identificar, sistematizar e analisar os padrões e esquemas de metadados articulados pela Ciência da Informação no âmbito da preservação digital de conteúdos da *Web*, sinalizando os metadados aplicáveis à preservação digital no arquivamento da *Web*;
d. Identificar na literatura internacional e nacional as pesquisas em arquivamento da *Web* na Ciência da Informação e áreas afins, oferecendo uma visão ampla e reflexiva dos seus desafios, métodos de coleta e políticas de seleção de conteúdos da *Web*; e
e. Sinalizar, a partir das estratégias, políticas de seleção e experiências identificadas, os critérios de seleção que são potencialmente aplicáveis na estruturação de diretrizes e políticas institucionais de preservação digital e arquivamento de conteúdos da *Web*.

1.3 Metodologia

1.3.1 Procedimentos metodológicos

Para atingir os objetivos propostos optou-se metodologicamente pelo desenvolvimento de uma investigação exploratória, pautada em pesquisa bibliográfica e documental com revisão de literatura narrativa concernente ao arquivamento da *Web* no escopo da preservação digital.

Selltiz *et al*.[14] (1967 citado por GIL, 2010, p. 27), entende que os estudos exploratórios buscam "[...] proporcionar maior familiaridade como o problema, com vistas a torná-lo mais explícito ou a construir hipóteses. Seu planejamento tende a ser bastante flexível, pois interessa considerar os mais variados aspectos relativos ao fato ou fenômeno estudado". Por sua vez, as investigações bibliográficas são efetuadas através de materiais já publicados, como livros, teses, dissertações, artigos etc., utilizando-se de informações ou de categorias teóricas documentadas e manipuladas por outros pesquisadores; e as pesquisas documentais abrangem como fonte uma ampla gama de documentos criados para propósitos variados, tais como material institucional, jurídico, de divulgação, iconográfico, itens pessoais etc. (GIL, 2010; SEVERINO, 2016).

À medida que revisão de literatura é o processo de busca, análise e descrição de um corpo do conhecimento a procura de resposta a uma pergunta específica, a revisão narrativa é um tipo de revisão adequada para a fundamentação teórica de artigos, dissertações, teses etc. que não usa critérios explícitos e sistemáticos para a busca e a análise crítica da literatura (isto é, livros, artigos de periódicos ou de jornais, registros históricos, relatórios governamentais, dentre outros), não havendo o esgotamento das fontes de informação e a adoção de estratégias de busca refinadas e exaustivas, onde a seleção dos estudos e a interpretação dos dados podem estar submetidas à subjetividade dos autores (UNIVERSIDADE DE SÃO PAULO, [2022]).

Ainda no que diz respeito à revisão de literatura narrativa, Cordeiro *et al*. (2007, p. 429-430) destacam que:

> A revisão da literatura narrativa ou tradicional [...] apresenta uma temática mais aberta; dificilmente parte de uma questão específica bem definida, não exigindo um protocolo rígido para sua confecção; a busca das fontes não é pré-determinada e específica, sendo frequentemente menos abrangente. A seleção dos artigos é arbitrária, provendo o autor de informações sujeitas a viés de seleção, com grande interferência da percepção subjetiva.

[14] SELLTIZ, Claire. *et al*. **Métodos de pesquisa nas relações sociais**. São Paulo, SP: Herder, 1967.

1.3.2 Forma de análise dos resultados

Como forma de análise de resultados dos materiais, aplicou-se a análise de conteúdo dos critérios e das políticas de seleção de conteúdos da *Web* identificados na literatura para fins de preservação digital e arquivamento de longo prazo. Enquanto método de coleta, organização e análise de dados na pesquisa qualitativa que é aplicável às investigações nas Ciências Sociais Aplicadas em que se precisa identificar a presença ou ausência de certos aspectos investigados (CAVALCANTE; CALIXTO; PINHEIRO, 2014; SILVA; VALENTIM, 2019), a análise de conteúdo trata de "[...] um conjunto de técnicas de análise das comunicações visando obter, por procedimentos, sistemáticos e objectivos de descrição do conteúdo das mensagens, indicadores (quantitativos ou não) [...]" os quais possibilitam "[...] a inferência de conhecimentos relativos às condições de produção/recepção (variáveis inferidas) destas mensagens.", conforme Bardin (2009, p. 44). E, para o mesmo autor (2016, p. 48), integram o domínio da análise de conteúdo todas as iniciativas que proporcionam a "[...] explicitação e sistematização do conteúdo das mensagens e da expressão deste conteúdo, com o contributo de índices passíveis ou não de quantificação, a partir de um conjunto de técnicas, que, embora parciais, são complementares."

Segundo Oliveira (2008, p. 570) a análise de conteúdo concede:

> [...] o acesso a diversos conteúdos, explícitos ou não, presentes em um texto, sejam eles expressos na axiologia subjacente ao texto analisado; implicação do contexto político nos discursos; exploração da moralidade de dada época; análise das representações sociais sobre determinado objeto; inconsciente coletivo em determinado tema; repertório semântico ou sintático de determinado grupo social ou profissional; análise da comunicação cotidiana, seja ela verbal ou escrita, entre outros.

A análise de conteúdo permite o exame dos fenômenos sociais associados a um objeto como as suas interações, e a escolha deste método deve-se "[...] pela necessidade de ultrapassar as incertezas conseqüentes das hipóteses e pressupostos, pela necessidade de enriquecimento da leitura por meio da compreensão das significações [...]" ou, ainda, "[...] pela necessidade de desvelar as relações que se estabelecem além das falas propriamente ditas." (CAVALCANTE; CALIXTO; PINHEIRO, 2014, p. 14). Para Bardin (1977, 2009) a análise de conteúdo organiza-se em torno de três fases: I) a pré-análise, o qual se destina a sistematizar as ideias iniciais para o desenvolvimento das fases seguintes e que inclui a escolha dos documentos para análise, a formulação das hipóteses e dos objetivos, e a elaboração de indicadores que fundamentem a interpretação final; II) a exploração do material, ou seja, a codificação do material que integra a coletânea de análise, em função de regras antecipadamente formuladas; III) o tratamento dos

resultados, que condensa e evidencia as informações dadas pela análise, podendo estabelecer figuras, modelos etc.; e a inferência e a interpretação, onde após os resultados serem submetidos a testes de validação, se propõe inferências e interpretações a propósito dos objetivos previstos.

1.3.3 Resultados esperados

Os resultados provenientes do trabalho irão apoiar as discussões CTS quanto à análise e as correlações possíveis entre os padrões de critérios de seleção de conteúdos identificados em iniciativas de políticas de preservação digital e arquivamento da *Web*. Se refletirá como estes critérios podem vir a interferir no acesso democrático aos conhecimentos produzidos e na atuação crítica-avaliativa da sociedade sobre C&T, uma vez que o arquivamento de conteúdos *Web* implica se confrontar com assuntos complexos, como as restrições de direitos autorais, a confidencialidade e a privacidade de dados, o proveito e a perda permanente de informações por razões variadas. Os critérios serão aplicáveis na estruturação de políticas de preservação digital e arquivamento da *Web* em instituições de memória e universidades públicas do país, visando atender às necessidades de reconhecimento de critérios sistematizados neste contexto.

Como resultados esperados do trabalho, estará uma contribuição importante no avanço das pesquisas nacionais de preservação digital e arquivamento da *Web* sob a perspectiva CTS, permitindo amenizar ou resolver os problemas da sociedade contemporânea com novos focos de atuação; uma significativa contribuição dos fundamentos CTS e da Ciência da Informação no mapeamento de critérios para a seleção de conteúdos da *Web*; a aplicação destes critérios de seleção nos conteúdos coletados pelas universidades em suas políticas de preservação digital e arquivamento da *Web*; uma melhor padronização e eficácia na seleção dos conteúdos da *Web*, permitindo atender às demandas de averiguação de critérios neste domínio; além do diagnóstico das iniciativas internacionais e nacionais com mapeamento e análise dos critérios de seleção praticados e das implicações éticas, políticas, econômicas, sociais e legais na seleção, coleta, armazenamento, preservação e acesso aos conteúdos sobre C&T produzidos no ambiente *Web*.

1.4 Estrutura do trabalho

Considerando que os procedimentos metodológicos usados serão apresentados em cada capítulo precisamente, esta tese está organizada da seguinte maneira:

O presente primeiro capítulo faz uma introdução com relação ao tema de investigação, elucida a questão de pesquisa, a hipótese, os objetivos, a metodologia e as justificativas.

No segundo capítulo discute-se as relações entre a preservação digital e a comunicação científica, partindo da discussão da história, evolução e definição do conceito de comunicação científica e da exploração das aplicações da preservação digital neste tema através da estratégia de metadados para preservação digital e das ferramentas, iniciativas e estratégias de preservação digital de conteúdos baseados na *Web* (especialmente, a preservação digital de *sites* arquivados ou arquivamento da *Web*, de conteúdos em mídias sociais, de periódicos científicos eletrônicos e de *e-mails* arquivados).

O terceiro capítulo aborda as dificuldades, os requisitos e as estratégias de preservação digital (em particular, adoção de padrões abertos; documentos de políticas e estratégias institucionais; orçamentos e custos da preservação digital; seleção para preservação digital e conformidade legal; treinamento e desenvolvimento de pessoal; metadados para preservação digital; investimento e montagem de infraestrutura tecnológica; formação de redes de colaboração; definição do meio de armazenamento; migração; transferência para suportes analógicos; emulação; conservação de tecnologia; arqueologia digital; e arquivamento da *Web*), além da produção científica recente desse tema de pesquisa.

No quarto capítulo debate-se os metadados no contexto da preservação digital a partir da abordagem do arquivamento da *Web*, incluindo a definição, a categorização e as funções dos metadados; o conceito de metadados de preservação e as informações documentadas por metadados que apoiam a gestão da preservação digital de longo prazo e o arquivamento da *Web*; e os principais padrões de metadados usados na descrição e na preservação digital de conteúdos da *Web* arquivados, isto é, o *Dublin Core* (DC), o *Metadata Object Description Schema* (MODS), o *Encoded Archival Description* (EAD), o *Visual Resources Association Core* (VRA *Core*), o *PREservation Metadata: Implementation Strategies* (PREMIS) e o *Metadata Encoding and Transmission Standard* (METS).

O quinto capítulo trata do processo de arquivamento da *Web* no âmbito da preservação digital, abrangendo os conceitos de arquivamento da *Web* e de arquivo da *Web*; as motivações para se arquivar *sites* a partir de uma vasta gama de casos de uso para os arquivos da *Web* e o arquivamento da *Web*; e o processo de seleção, destacando a definição de uma política de seleção para conteúdo da *Web*, a seleção e as suas fases de preparação, descoberta e filtragem, a documentação do processo de seleção e dos próprios critérios de seleção através de metadados, a manutenção e o contexto da política de seleção, os principais métodos de seleção (em especial, abordagem não seletiva; temática baseada em assunto, criador, gênero e domínio; seletiva baseada em um critério pré-estabelecido ou evento; e de depósito), a definição dos

critérios de seleção e a determinação dos limites do recurso da *Web* selecionado (incluindo os limites técnicos para a captura), através de várias iniciativas de arquivos da *Web* no mundo.

E, no sexto capítulo, as considerações finais da tese, expondo as reflexões dos resultados obtidos e analisados, as contribuições trazidas por eles e sugestões para futuras investigações.

2 USOS DA PRESERVAÇÃO DIGITAL NA COMUNICAÇÃO CIENTÍFICA: reflexões a partir dos estudos CTS e da Ciência da Informação

Atualmente, a *Internet* é o meio que reúne o conjunto mais multifacetado de materiais-fonte que registram os fenômenos sociais, culturais e políticos modernos; além do mais ela se faz crucial para o desenvolvimento geral dos meios de comunicação, incluindo as mídias de massa e uma variedade de dispositivos digitais (BRÜGGER; FINNEMANN, 2013). Como um dos serviços baseados nesta rede, a *World Wide Web*, ou *Web*, do mesmo modo tornou-se meio vital de facilitar a comunicação global sendo especialmente importante para a comunicação da ciência, publicação, comercio eletrônico e muito mais, segundo Day (2003a). Aliás, a *Web*, desde da década de 90, evoluiu para talvez o canal de comunicação mais dominante do mundo, o qual detém um caráter democrático em que todos podem publicar quaisquer tipos de informações usando inúmeras formas de mídia ou serviços, tais como opiniões públicas, notícias, *blogs* e *wikis*, sendo que uma porção destas informações é única e historicamente tão valiosa no futuro quanto os manuscritos antigos são hoje (BROWN, c2006; COSTA, GOMES; SILVA, 2017).

A "geração *Google*"[15], ou a que cresceu num mundo regido pela *Internet* e que depende dos meios tecnológicos, não só busca informações na *Web*, mas também a utiliza para realizar várias tarefas, como fazer compras, se relacionar e assim por diante; igualmente, os provedores de serviços adotaram a *Web* excluindo métodos tradicionais como foi o caso, por exemplo, do crescimento dos periódicos eletrônicos, conforme Brown (c2006). Com um volume infinito de conteúdo criado, a *Web* teve um grande impacto no nosso espaço de informação, em que muito do nosso discurso cultural ocorre neste ambiente e a sua preservação é uma pré-condição vital para pesquisas nos diferentes campos científicos, como História, Sociologia e outras disciplinas (ALNOAMANY; WEIGLE; NELSON, 2016; NELSON, 2012; SHIOZAKI; EISENSCHITZ, 2009). Para mais, a *Web*, com a sua evolução de um meio de publicação para de comunicação, apresenta uma vasta coleção de fontes primárias do nosso passado, e também não existe apenas como um meio de acesso à literatura científica, mas ainda como campo de investigação, mais ou menos formal, em acordo com Stirling, Chevallier e Illien (2012) e Vlassenroot *et al*. (2019).

[15] A atual geração, ou seja, os que nasceram depois de 2010, é classificada como "geração Alpha", e a geração anterior a esta, conhecida como "Geração Z" ou aqueles que nasceram entre 1995 e 2010, passou a substituir recentemente o uso do buscador *Google* pelo aplicativo de mídia *TikTok* como ferramenta de busca e descoberta de informações na *Internet*. Disponível em: https://www.terra.com.br/byte/internet/para-a-geracao-z-o-tiktok-e-a-nova-ferramenta-de-busca,0c557458700435dd2b2a073dfd605fe2vawk55s5.html. Acesso em: 6 jun. 2023.

No entanto, o surgimento da rede mundial de computadores foi da mesma forma seguido de preocupações com a preservação de seu conteúdo pelo fato da transitoriedade, efemeridade e dinâmica serem umas de suas propriedades definidoras, como Brown (c2006) e Costa, Gomes e Silva (2017). O que estava na *Web* ontem ou um ano atrás não está mais lá, e o conteúdo da *Web online* é continuamente atualizado, alterado, substituído ou perdido em um curto período de tempo e, muitas vezes, sem deixar vestígios, representando assim uma ameaça a persistência de acesso aos dados produzidos que pode criar uma lacuna de saber e informação sobre o tempo presente (BRÜGGER, c2018; COSTA, GOMES; SILVA, 2017; GOMES; FREITAS; SILVA, 2006; POST, 2017). Na literatura acadêmica, segundo *Digital Preservation Coalition* (c2015), com muita frequência os endereços da *Web* (URLs) não nos levam ao recurso referenciado que desejamos devido a, por exemplo, falhas de servidores; ademais, o tempo de vida dos *sites* é inadequado para verificação científica e para referência de longa duração (MASANES, 2005).

Logo, "[...] a facilidade com que os conteúdos podem ser disponibilizados via *Web*, aliada à fragilidade desses conteúdos num mundo em constante mudança tecnológica, engendra um ambiente informacional que pode ser positivamente hostil à sustentabilidade a longo prazo." (BROWN, c2006, p. 3, tradução nossa). Para a manutenção da persistência do acesso aos recursos ao longo do tempo, a preservação digital integra "[...] estratégias de preservação que lidam com a obsolescência tecnológica dos objetos digitais[16] de forma a assegurar, no futuro, o acesso aos mesmos." (PINHEIRO; FERREZ, 2014, p. 176). Neste processo, a preservação dos recursos *Web* não é muito diferente da preservação de outros recursos digitais e, como *Digital Preservation Coalition* (c2015), qualquer objeto digital pode ser considerado, isto é, nato digital ou digitalizado, corporativo ou pessoal, inovador ou rotineiro, abrangendo desde textos, filmes, músicas assim como bancos de dados, *e-mails*, mídia social, domínios inteiros da *Web* e outros.

Estes materiais/objetos digitais são um bem vital e fonte cada vez mais útil à indústria, ao comércio e aos governos de sociedades, assim como para os cientistas, médicos e demais profissionais (DIGITAL PRESERVATION COALITION, c2015) que pesquisam e analisam conteúdos digitais para objetivos diversos, como prever crises climáticas e descobrir a cura para doenças. Por essa razão, vários projetos internacionais, geralmente em bibliotecas e arquivos nacionais, universidades, órgãos de governos, consórcios de organizações e instituições do setor privado (incluindo as editoras e sociedades científicas e os financiadores), foram empreendidos

[16] Objetos digitais são "[...] itens na forma digital que requerem um computador para dar suporte à sua existência e apresentação visual." (PINHEIRO; FERREZ, 2014, p. 163), ou de acordo com Conselho Nacional de Arquivos (2020, p. 37) diz respeito a "unidade de informação em formato digital composta de uma ou mais cadeia de *bits* e de metadados que a identificam e descrevem suas propriedades."

para preservar e arquivar a longo prazo materiais digitais. Apesar disto, consoante Masanès (c2006b), a ideia de se preservar o que é produzido ou publicado na *Web* vêm sendo questionada e ainda não é aceita por todos. O autor indica os principais argumentos contra o arquivamento:

I) A qualidade do conteúdo encontrado na *Web* que não atende aos padrões de preservação, em que seria preciso uma seleção manual dos conteúdos, mas isso é incompatível considerando a amplitude de dados na *Web*, tal como o crescimento das publicações científicas;

II) A concepção de que a *Web* se autopreserva, onde os recursos que merecem ser preservados serão mantidos nos servidores, e outros desaparecerão à vontade do criador original, não sendo assim preciso tal processo; e

III) A consideração de que o arquivamento da *Web* não é possível.

Em vista disso, este capítulo tem por objetivo aprofundar as discussões até então pouco explicitadas pela comunidade científica sobre as relações entre a preservação digital, enquanto conjunto de processos, requisitos e estratégias para a garantia do acesso contínuo e utilizável de recursos digitais e eletrônicos, e a comunicação científica, como prática de disseminação dos saberes em ciência para os cientistas e o público leigo. À medida que os conteúdos vêm sendo produzidos, difundidos e compartilhados maiormente na *Web* ou em outros ambientes digitais, julgamos que a explanação dos usos da preservação digital na comunicação científica possibilita a sustentação do papel das instituições de patrimônio cultural (bibliotecas, arquivos e museus) como gestoras, curadoras e custodiadoras das publicações de valor histórico e dos registros do saber humano no século XXI. Tais materiais ao serem arquivados para que estejam disponíveis de modo consistente em prol das investigações no futuro contribuirá não só na preservação da capacidade dos pesquisadores de apoiar e outros avaliarem as suas descobertas, mas também na guarda de uma memória digital e na promoção e avanço do ensino acadêmico e da ciência.

Faz-se então uma revisão documental e da literatura específica nacional e internacional sobre estes dois objetos de estudo expondo descritivamente experiências da preservação digital de *sites* arquivados (ou arquivamento da *Web*), de conteúdos em mídias sociais, de periódicos científicos eletrônicos e de *e-mails* arquivados, e a importância dos metadados na preservação digital e na comunicação científica. Analisa dados do levantamento bibliográfico assistemático de materiais publicados nos últimos vinte anos como artigos de periódicos, monografias e livros na área da Ciência da Informação e do campo CTS buscados na SciELO e no *Google Scholar*, ou indexados nas bases *Scopus*, *ScienceDirect*, *Emerald Insight* e *Web of Science* disponíveis via Portal de Periódicos CAPES[17], além dos conteúdos de *sites*, relatórios e guias de iniciativas,

[17] Disponível em: https://www-periodicos-capes-gov-br.ezl.periodicos.capes.gov.br/index.php. Acesso em: 6 jun. 2023.

que se relacionam aos assuntos ‘preservação digital’, ‘comunicação científica’, ‘arquivamento da *Web*’, ‘periódico científico eletrônico’, ‘mídia social’ e ‘*e-mail*’, representando um trabalho de reflexões (GIL, 2010, 2012; LUNA, 1997; SEVERINO, 2016; SILVA; MENEZES, 2005).

Diferente dos seus elementos constitutivos – isto é, História, Filosofia e Sociologia da Ciência e da Tecnologia – (BAUCHSPIES; CROISSANT; RESTIVO, c2006), o campo CTS é, como indicam os editores Felt *et al*. (c2017, p. 1, tradução nossa) da edição mais recente do “*Handbook of Science and Technology Studies*” “[...] um campo interdisciplinar que investiga as instituições, as práticas, os significados e os resultados da ciência e da tecnologia e seus múltiplos envolvimentos com os mundos que as pessoas habitam, suas vidas e seus valores.” Com uma evidente tendência a refutar muitos dos elementos da percepção comum da ciência, o campo CTS procura estudar “[...] como o conhecimento científico e os artefatos tecnológicos são construídos.”, sendo estes dois produtos humanos “[...] marcados pelas circunstâncias de sua produção.”, de acordo com Sismondo (c2010, p. 11, tradução nossa). E segundo os editores Hackett *et al*. (c2008, p. 1, tradução nossa), da edição anterior do livro supracitado, o campo CTS “[...] está criando uma compreensão integrativa das origens, dinâmicas e consequências da ciência e da tecnologia.”, não constituindo um “[...] esforço estritamente acadêmico [...]”.

Assim, o capítulo procura apresentar os resultados e as análises dos conteúdos coletados partindo da discussão da história, evolução e definição do conceito de comunicação científica e da exploração das aplicações da preservação digital neste tema pelas estratégias identificadas e supracitadas para conteúdos baseados na *Web* apoiando-se na essencialidade da difusão e uso de informações nas atividades científicas, no processo cumulativo e probatório da produção da ciência e no destaque dos documentos digitais para transmissão de informações com a *Internet*.

2.1 Comunicação científica e a preservação digital: caminhos possíveis

O campo CTS no Brasil também se caracteriza por estudos sob a ótica da comunicação científica os quais, segundo Hayashi, Hayashi e Furnival (2008), focam a divulgação

científica[18], a compreensão e participação pública da ciência e o jornalismo científico[19]; igualmente, o tema é uma subárea da Ciência da Informação no país, como descrito por Pinheiro (2012). No cenário internacional, Davies (2022, p. 307, tradução nossa) da mesma maneira discute as relações entre a temática e o campo CTS e pontua que este último, enquanto enfoca como o conhecimento científico é criado, interage e é moldado por vários contextos, transita e é refeito e remodelado, nos "[...] promove uma abordagem exploratória e descritiva para a comunicação científica que prioriza a compreensão das práticas e significados dos envolvidos em seus próprios termos.", permitindo realizar perguntas abertas e não estruturadas sobre o que estamos fazendo e por quê.

Por sua vez, dentro da literatura das áreas de Biblioteconomia e Ciência da Informação, a comunicação científica ainda pode ser considerada uma expressão sinônima do conceito de comunicação acadêmica (FLEMING-MAY, 2023). Como definição do termo, Borgman (1989) interpreta que a comunicação acadêmica (*scholarly communication*) consiste no estudo de como os acadêmicos em qualquer campo utilizam e disseminam informações por canais formais e informais de comunicação, incluindo o crescimento das informações acadêmicas, as relações entre métodos de comunicação ou entre áreas e disciplinas de pesquisa (por exemplo, Ciências Humanas e Sociais, Tecnologia etc.), e as necessidades e os usos de informações de usuários.

Sendo assim, tão crucial quanto a pesquisa em si, a comunicação viabiliza a existência e a legitimidade da ciência pela a sua avaliação por pares, sendo que as atividades mais remotas com impacto na comunicação científica pertencem ao povo grego da Atenas Antiga (séculos

[18] Julgando que a divulgação científica tem em vista à comunicação para o público diverso e além da comunidade científica, no qual vulgarização científica e popularização da ciência são seus sinônimos, Valério e Pinheiro (2008, p. 160) citando Bueno (1985) apontam que a divulgação científica se refere "[...] a comunicação de informações científicas para o público não especializado, fazendo uso da recodificação da linguagem e tornando os termos acessíveis ao entendimento comum." Assim, a divulgação científica – que se faz pela imprensa, TV, palestras etc. – cumpre a função de democratizar o acesso ao saber científico e de definir meios para a alfabetização científica, contribuindo na inclusão dos cidadãos (que não possuem uma formação técnico-científica, não admitem o caráter coletivo da produção da ciência, imaginam que C&T não se realizam num contínuo e que avançam aos saltos por *insights* de mentes singulares etc.) nos debates de temas especializados (Covid-19, mudanças climáticas etc.) que impactam as suas vidas como permitindo que eles captem as descobertas e o progresso científico (BUENO, 2010).

[19] Jornalismo científico se trata de um "processo social que se articula, a partir da relação [...] entre organizações formais (editoras e emissoras) e coletividades (público, receptores), através de canais de difusão (jornal, revista, rádio, televisão e cinema) que asseguram a transmissão de informações [...]" científicas e tecnológicas à vista de interesses e expectativas (BUENO, 1985, p. 22, citado por CARIBÉ, 2011). Isto posto, no jornalismo científico o emissor é o jornalista e o receptor é o público leigo, integrando um tipo de divulgação científica e, em geral, uma espécie de difusão científica – isto é, todo e qualquer processo e/ou recurso usado na propagação de informações tecnocientíficas –, conforme Bueno (1985, 2010) e Caribé (2011). Ainda, segundo Massarani e Moreira (2004), Moreira e Massarani (2002) e Mueller e Caribé (2010), dentre os fatos marcantes para o jornalismo científico e a profissionalização na área estão, por exemplo, a abertura, em 1837, das sessões e atas da *Académie des Sciences* na França para os jornalistas; a criação, no século XX, da *Science Server* nos Estados Unidos, uma agência de notícias científicas feita por e para os jornalistas científicos; e a fundação, em 1977, da Associação Brasileira de Jornalismo Científico (ABJC) no Brasil, o qual possuía entre seus propósitos a democratização do saber científico.

IV e V a. C.) em que pessoas se reuniam para debater questões filosóficas (MEADOWS, 1999; ZIMAN, 1979). Porém, através de Mueller e Caribé (2010), podemos entender que na Europa do século XV temos as primeiras iniciativas de divulgação científica para leigos, que ocorreram junto ao avanço da ciência e da imprensa, onde o documento escrito passou a exercer um papel vital na transmissão dos saberes.

Baseando-se em Caribé (2011), Freitas (2006), Hayashi e Ferreira Jr (2006), Meadows (1999), Moreira e Massarani (2002), Mueller e Caribé (2010), Valente, Cazelli e Alves (2005) e Ziman (1981), é possível expor alguns fatos notáveis da história e evolução da comunicação da ciência moderna à sociedade, que se deram na Europa entre os séculos XVI e XX, com breve alusão a iniciativas no Brasil como ainda nos Estados Unidos em séculos mais recentes, a saber:

- Século XVI – surgimento das primeiras academias de ciência pela Europa. Por exemplo, a *Accademia Secretorum Naturae* ou *Accademia dei Segreti* e a *Accademia del Cimento*, ambas na Itália, nesta ordem, em 1560 e 1657; a *Royal Society for the Improvement of Natural Knowledge*, na Inglaterra, em 1622; e a *Académie des Sciences*, na França, em 1666; onde, por exemplo, o projeto científico baconiano[20] teve o seu impulso na Inglaterra e o modelo cartesiano da ciência foi adotado na França, e os cientistas-membros destas sociedades reuniam-se em colégios invisíveis (*invisible college*)[21] e comunicavam seus resultados por correspondências particulares (ou cartas) que, somado aos registros impressos (tidos como anais/atas) do que era tratado nas reuniões, geraram os primeiros periódicos científicos. Como uma colônia portuguesa, no Brasil dos séculos XVI, XVII e XVIII as atividades científicas ou a difusão de ideias modernas eram quase que nulas.

[20] A estruturação da ciência como conhecemos hoje – a investigação científica – remete ao século XVII com a ideia exposta na obra "Nova Atlântida" do filosofo propulsor do pensamento científico atual Francis Bacon (BACON, 1979), com sua primeira publicação em 1627, o qual é um clássico da língua inglesa que apresenta uma perspectiva profética e empirista iniciando um novo entendimento do mundo e da realidade, onde: a ciência trata-se de uma obra coletiva, exigindo de vários pesquisadores que colham material a fim que seja analisado por especialistas; a priori, a ciência não é realizada por afirmações teóricas e, sim, pelo contato com os fenômenos reais mediante a pesquisa empírica; e a ciência possui propósito basicamente prático como, por exemplo, tratar doenças, estender a longevidade e fabricar equipamentos de diversos tipos para voar, navegar etc. (HAYASHI; FERREIRA JR, 2006).

[21] Colégio invisível consistia mais numa comunidade intelectual do que em um grupo de instituições ou construções materiais, sendo que as relações entre os seus membros não se fundavam em normas, deveres legais ou transações financeiras e, sim, na troca de informações e saberes, segundo Caribé (2011) e Mueller e Caribé (2010) respaldados em Ziman (1981). Sendo os grupos de elite que se formam no "cume" da comunidade científica e em redor de uma frente de pesquisa, num colégio invisível os membros também são "um grupo de poder" por serem passíveis de controlar a gerência de fundos de pesquisa, de laboratórios, as novas ideias científicas etc., do qual cada cientista permanece informado do trabalho dos outros (antes deste ser publicado) via conferências, seminários etc. fechados, acrescidos por um intercâmbio informal de documento escrito (HAYASHI; FERREIRA JR, 2006; PRICE, 1986).

- Séculos XVII e XVIII – advento dos primeiros periódicos científicos em 1665, quando apareceram juntos o *Journal des Sçavans* da *Académie des Sciences* em Paris, enquanto pioneiro do periódico moderno de humanidades; e *Philosophical Transactions* da *Royal Society of London*, na Inglaterra, como precursor do moderno periódico científico. Para mais, houve a publicação de livros julgados instrumentos predecessores da divulgação científica, como a obra *Dialoghi sopra i due massimi sistemi del mondo, tolemaico e copernicano*, de Galileu Galilei, em 1632; bem como tivemos as primeiras conferências científicas públicas não universitárias de modo a divulgar o conhecimento científico à sociedade como, por exemplo, as conferências científicas populares patrocinadas pela *Royal Institution of Great Britain*, uma entidade criada em 1799, na cidade de Londres.
- Século XIX – consolidação das disciplinas científicas e da especialização; otimismo em relação aos benefícios do avanço científico-técnico, exposto pelas grandes Exposições Universais, iniciadas em Londres no ano de 1851; além da criação das associações para o progresso da ciência (por exemplo, a *British Association*, de 1831) e da consolidação de periódicos científicos nacionais no estrito da ciência (por exemplo, as revistas *Nature* e *Science* criadas, nessa ordem, na Inglaterra em 1869 e nos Estados Unidos em 1880). Com a chegada da Corte portuguesa em 1808, no Brasil houve os primeiros indícios de instituição da ciência com a criação da Academia Real Militar (1810), da primeira editora Imprensa Régia (1810) e do Museu Nacional (1818); interesse pelas aplicações práticas de ciência, expresso nas Exposições Nacionais da Indústria feitas em 1861 e 1866; além da criação dos primeiros jornais e periódicos com a publicação de notícias e/ou artigos ligados à ciência (por exemplo, O Patriota de 1813 e o *Miscelanea scientifica* de 1835).
- Século XX – com a vinda dos novos meios de comunicação, como o cinema, o rádio e a televisão, estes passam a ser explorados na divulgação científica mormente nos países desenvolvidos; valorização da educação e divulgação da ciência, expressa nos primeiros museus/centros interativos de ciências e tecnologia (o *Deutsches Museum* em Munique, criado em 1903 e o *Exploratorium*, em São Francisco, de 1969, por exemplo). A partir dos anos 20, há o aumento no Brasil das ações de divulgação científica junto à formação de sua comunidade científica, mostrado pela criação da Sociedade Brasileira de Ciências (1916), da Sociedade Brasileira para o Progresso da Ciência – SBPC (1948), e a primeira rádio do país, a Rádio Sociedade do Rio de Janeiro (1923); além da criação das primeiras faculdades de ciências e institutos de pesquisa; de seções de ciência em jornais diários

e revistas gerais; de livros, filmes e programas de TV voltados para ciência; e de centros e museus de ciência para educação científica, que ajudaram na popularização da ciência.

Como definição conceitual de comunicação científica, Bueno (2010, p. 2) interpreta que o tema se remete "[...] à transferência de informações científicas, tecnológicas ou associadas a inovações e que se destinam aos especialistas em determinadas áreas do conhecimento.", sendo que o seu público-alvo "[...] não ignora o fato de que a produção da ciência está respaldada num processo cumulativo, que se refina ao longo do tempo, pela ação daqueles que a protagonizam (pesquisadores/cientistas)." como admite "[...] que ela precisa ser validada pela demonstração rigorosa e/ou pela comprovação empírica." Também, consoante Valerio e Pinheiro (2008) com base em Garvey e Griffith (1979), a comunicação científica é um processo que reúne atividades ligadas à produção, disseminação e uso da informação, iniciado com a pesquisa e findo com as descobertas inclusas ao saber científico, onde na fase da pesquisa dá-se a geração da informação e a disseminação se faz pela transmissão da informação via diferentes canais de comunicação.

De forma objetiva, podemos dizer que este processo intenciona, principalmente, "[...] à disseminação de informações especializadas entre os pares, com o intuito de tornar conhecidos, na comunidade científica, os avanços obtidos [...]" "[...] em áreas específicas ou à elaboração de novas teorias ou refinamento das existentes." (BUENO, 2010, p. 5). Para tal, a comunicação científica se realiza por canais formais, como os periódicos científicos, livros, relatórios e outros meios escritos, ou informais, como os eventos científicos, *e-mails*, conversas e mais meios orais e escritos. No caso dos primeiros, eles têm como prós a probabilidade de atingir público mais amplo, a armazenagem e a recuperação mais seguras e a maior rigidez e controle por avaliação prévia das informações; já no caso dos segundos, dentre as suas principais vantagens incluem a probabilidade de maior atualização e rapidez das informações transferidas, em conformidade com Garvey e Griffith (1979), Meadows (1999), Targino (2000) e Valerio e Pinheiro (2008).

Contudo, é válido frisar que os processos, relações e natureza da comunicação científica são descritos na literatura usando vários termos e conceitos relacionados – como, disseminação, divulgação, difusão e popularização científica; jornalismo científico; educação e alfabetização científica e percepção pública da ciência –, inferindo-se que comunicação científica é um termo genérico que "[...] engloba todas as demais formas de comunicação que variam de acordo com o tipo de linguagem utilizada ou com o tipo de entidade do processo de comunicação ao qual se encontra relacionado." e se estende "[...] tanto a comunicação interna dirigida à comunidade científica quanto a externa, destinada ao público leigo." (CARIBÉ, 2015, p. 101). Isto reflete a proximidade entre a comunicação e a divulgação da ciência e tecnologia que teve o seu impulso

com as transformações suscitadas pelo advento de novas tecnologias de comunicação no século XIX-XX e, sobretudo, da *Internet* no final do século XX que se difundiu na contemporaneidade.

Entre os efeitos do surgimento da *Internet* como principal meio da comunicação pública de ciência e tecnologia e da comunicação científica, Trench (2008) observa os distúrbios muito significativos causados pela publicação eletrônica no importante campo das revistas científicas. O autor sugere que estes e outros desenvolvimentos se tornaram mais inteiramente permeáveis do que antes as fronteiras entre comunicação científica profissional e pública, possibilitando o acesso público a espaços anteriormente privados. De fato, a *Internet* trata-se da rede que conecta os computadores ao redor do mundo a qual pode ser usada para uma série de serviços, incluindo transferência de arquivos, *e-mail*, mensagens instantâneas e a *World Wide Web*. Este último diz respeito a um meio de comunicação integrado por páginas e demais conteúdos ligados entre si (por exemplo, imagens e vídeos) e que tem se efetivado como o meio central para publicações das sociedades modernas (GOMES, 2010; SOCIETY OF AMERICAN ARCHIVISTS, c2022h).

Para mais, em acordo com Valerio e Pinheiro (2008), a *Internet* promoveu novos fluxos de informação à ciência por permitir que bilhões de novos usuários da informação naveguem pela rede em qualquer hora, expandindo demais o público alcançável ao acesso da comunicação e da informação. Também o avanço na ciência e tecnologia propiciou a comunicação eletrônica que, por seu lado, seguiu para as conexões em redes, suscitando desde alterações em práticas estabelecidas como variadas espécies de relacionamento interpessoal que incitam a espiral de produção do saber. Nas redes eletrônicas, o conhecimento incluído pela literatura científica é disponibilizado em periódicos científicos eletrônicos – a expressão máxima do sistema formal de comunicação da ciência –, os quais trouxeram novas formas de comunicação e de informação científicas, ampliando a audiência e atingindo demais públicos, criando uma alta convergência com públicos não especialistas/acadêmicos (VALERIO; PINHEIRO, 2008; TARGINO, 2000).

Como caracterização dos canais eletrônicos de comunicação, Targino (2000) indica que a comunicação eletrônica abriga atributos dos sistemas formal e informal, pois em geral atinge um público grande, possibilita acesso a informações recentes, não possui rigidez e controle por avaliação prévia das informações pela comunidade científica e, ainda, apresenta complexidade de armazenagem e de recuperação das informações eletrônicas, onde estas não têm a brevidade das conversas, exposições orais etc. já que é provável a sua impressão, garantindo a conservação e o uso seguido. Para a autora a disposição de recursos em redes eletrônicas de informação, em máxime a *Internet*, reflete no ciclo da informação eletrônica e, por consequência, nos domínios formal e informal da comunicação científica, trazendo a necessidade de discussão de questões

no que concerne, a título de exemplo, à natureza, fidedignidade, consistência, armazenamento e preservação das informações distribuídas e à sua validade ou não como referencial confiável.

O desenvolvimento das novas TIC, acima de tudo, a rede *Internet*, favoreceu o acesso à comunicação científica com a formação de ambientes que reúnem um volume expressivo de periódicos de acesso aberto, como é o caso da SciELO que é uma fonte útil para pesquisadores (BUENO, 2010). Aliás, a tecnologia expandiu as formas e os canais de comunicação científica como o jeito de processar os dados por bibliotecas e editoras; a consulta, percepção e apreensão do saber pelo público por intermédio da velocidade, amplidão e conveniência do acesso remoto na disposição e procura das informações, conforme relatam Ferreira, Martins e Rockembach (2018), Meadows (1999) e Mueller (2007). Estas e outras alterações nas características e padrões da comunicação científica pela comunicação por TIC apontam uma proximidade entre a comunicação e a divulgação científica, mormente na relação com o seu público que, hoje, é visto com novas dimensões propiciado pelo alcance das TIC (VALERIO; PINHEIRO, 2008).

Sob o enfoque histórico, comunicação e divulgação científica têm dialogado de maneira retribuidora e, particularmente, o empenho de interação com o público leigo se fez por figuras renomadas da comunidade científica (BUENO, 2010). No século XX, após a Primeira Guerra Mundial, surgiu em todo mundo um novo tipo de divulgação científica quando cientistas, como Albert Einstein e Marie Curie, adquiriram uma imagem notável diante do público, sendo que a defesa da ciência pura passou a caracterizar a divulgação da ciência na época; por sua vez, na América Latina foram os próprios cientistas que se comprometeram com a divulgação científica desde o século XIX, os quais as atividades visavam ampliar a sua presença dentro da sociedade, conforme Caribé (2011) e Massarani e Moreira (2004). Porém, mesmo com o rádio, TV, cinema e imprensa mais apurada que fizeram do século XX a era da informação, nenhuma tecnologia trouxe o impacto da *Internet* na divulgação da ciência cujo todas as formas de comunicação se reúnem e a informação científica faz-se mais acessível ao cidadão (MULLER; CARIBÉ, 2010).

Hoje, vários estudos na Ciência da Informação e de outras áreas abordam a comunicação e divulgação científica de conteúdos na *Web* sob diferentes perspectivas. Podemos indicar como exemplos dessas pesquisas, as que tratam dos formatos de *podcasts* de divulgação científica e os estilos nessa produção de conteúdo (DANTAS; DECCACHE-MAIA, 2022; FIGUEIRA; BEVILAQUA, 2022); ou da divulgação científica de conteúdos sobre inteligência artificial pelo *Instagram* (AZEVEDO *et al.*, 2022); ou do uso do *Facebook* e *Twitter* no acesso à informação científica por parte de acadêmicos (GALLOTTI; BORGES, 2019); ou do uso do *Facebook* por instituições de ciência como forma de diálogo com a população e de ação contra os boatos sobre doenças epidêmicas (BARRETO *et al.*, 2020); ou da opinião de jornalistas e de comunicadores

sobre a adaptação do jornalismo científico ao ambiente digital, incluindo diante do potencial do *TikTok* como plataforma informativa (MARTIN NEIRA; TRILLO DOMÍNGUEZ; OLVERA LOBO, 2023); ou também das transformações ocorridas na comunicação científica no campo da saúde durante a pandemia decorrente da Covid-19 (CARVALHO; LIMA; MACÊDO, 2022).

Diante disto tudo, é certo que as atividades da comunicação científica geram um volume grande de informação; por esse motivo, e da relevância de eternizar o saber às gerações futuras, é preciso guardar o que se deixou de herança, em especial quando se trata na era da informação e da *Web*, segundo Ferreira, Martins e Rockembach (2018) apoiados em Meadows (1999). Em reconhecimento do desafio global no século XXI de garantir a preservação e o acesso contínuo à uma memória do que foi produzido e difundido em meios digitais, incluindo dados, pesquisas e publicações científicas; ou até mesmo de que a informação se constitui num bem essencial ao avanço econômico, político, cultural e de conhecimento das sociedades modernas, organizações – em especial, instituições memorialísticas e universidades – e profissionais de múltiplas áreas do conhecimento no mundo têm estudado e implementado métodos e tecnologias para a criação, gestão e preservação de diferentes tipos de recursos/objetos digitais utilizáveis em longo prazo.

A preservação digital é um tema de pesquisa atual na Ciência da Informação, como indicado. Trata-se de um "conjunto de ações gerenciais e técnicas exigidas para superar as mudanças tecnológicas e a fragilidade dos suportes, garantindo o acesso e a interpretação de documentos digitais pelo tempo que for necessário." (CONSELHO NACIONAL DE ARQUIVOS, 2020, p. 39), que para Santos (2016, p. 451) ao se destinar a "[...] manutenção da ciência, cultura e do conhecimento humano, repete em parte o debate da preservação da informação registrada em suportes mais estáveis que o eletrônico". Assim, entre as ações/estratégias de preservação digital descritas na literatura específica, destacamos as seguintes para os objetivos do trabalho: o uso de metadados (*metadata*) para preservação digital; a preservação digital do conteúdo de *sites*, ou arquivamento da *Web* (*Web archiving*); a preservação digital em mídias sociais (*preserving social media*), de periódicos científicos eletrônicos (*electronic jornal*, *e-journals*) e, ainda, de mensagens de correio eletrônico (*e-mail*).

Em seguida, discutiremos as estratégias operacionais e estruturais de preservação digital aludidas, destacando as suas capacidades e limitações sob o escopo da comunicação científica.

2.1.1 Metadados para preservação digital

Como uma estratégia estrutural de preservação digital – isto é, os esforços iniciais das instituições, a fim de preparar um ambiente ideal para este processo (MÁRDERO ARELLANO,

2008, THOMAZ; SOARES, 2004) – e um aspecto técnico a ser julgado na elaboração de uma política de preservação digital (GRÁCIO, c2012, 2020), a utilização de padrões e esquemas de metadados é uma das questões para a sustentação da preservação de informações digitais, como conteúdos em *sites* e mídias sociais, publicações científicas na *Web* e *e-mail*, de forma a apoiar a gerencia do seu arquivamento e acesso utilizável em longo prazo. Estas estruturas formais de descrição chamadas de padrões de metadados (*metadata standards*) definirão a identificação e a persistência, a coerência e a interpretação, o acesso e a representação, as funcionalidades, a autenticidade, o contexto e a proveniência do recurso/objeto digital no ambiente informacional, além da sua interoperabilidade[22], de acordo com Formenton (2015) e Formenton *et al*. (2017).

Enquanto informações que descrevem as características (ou propriedades) significativas de um recurso digital (DIGITAL PRESERVATION COALITION, c2015), todas as classes de metadados – descritivos, estruturais, administrativos e linguagens de marcação (RILEY, c2017) – são essenciais para obtenção da preservação digital requerendo o uso combinado de esquemas visto a complexidade dos processos envolvidos e os vários tipos de conteúdo a serem descritos, conforme Formenton *et al*. (2017) e Silva e Silva (2020). Assim, entre os padrões de metadados aplicáveis à preservação digital, destacamos: o DC[23], para descoberta de recursos na *Web* e que é amplamente usado no arquivamento da *Web* (VENLET *et al*., c2018); o MODS[24], derivado do MARC 21[25] e que é utilizado por bibliotecas digitais; o PREMIS[26], um padrão de fato para metadados de preservação que aplica o *Open Archival Information System* (OAIS)[27]; o METS[28], para codificação de metadados sobre objetos complexos, como páginas *Web* com vídeos etc.; e o formato *Extensible Markup Language* (XML)[29], para interoperabilidade de dados na *Web* e que propicia o uso conjunto de diferentes esquemas.

---

[22] Interoperabilidade (*interoperability*) é "[...] a troca efetiva de conteúdo entre sistemas [...]" o qual "[...] depende de metadados que descrevem esse conteúdo para que os sistemas envolvidos possam efetivamente traçar o perfil do material recebido e combiná-lo com suas estruturas internas." (RILEY, c2017, p. 6-7, tradução nossa).
[23] Disponível em: https://dublincore.org/. Acesso em: 7 jun. 2023.
[24] Disponível em: http://www.loc.gov/standards/mods/. Acesso em: 7 jun. 2023.
[25] Disponível em: https://www.loc.gov/marc/. Acesso em: 7 jun. 2023.
[26] Disponível em: https://www.loc.gov/standards/premis/. Acesso em: 7 jun. 2023.
[27] O modelo de referência OAIS é "[...] uma estrutura conceitual que descreve o ambiente, componentes funcionais e objetos de informação associados a um sistema responsável pela preservação a longo prazo." que tem o objetivo principal de "[...] fornecer um conjunto comum de conceitos e definições que podem auxiliar na discussão entre setores e grupos profissionais e facilitar a especificação de arquivos e sistemas de preservação digital." (DIGITAL PRESERVATION COALITION, c2015, não paginado, tradução nossa). Aprovado pela primeira vez como norma ISO 14721 em 2002 e com uma segunda edição publicada em 2012, o OAIS está traduzido e adaptado pela norma nacional em vigor NBRISO14721 – Sistemas espaciais de transferência de dados e de informação – SAAI – Modelo de referência. Disponível em: https://www.normas.com.br/visualizar/abnt-nbr-nm/13167/abnt-nbriso14721-sistemas-espaciais-de-transferencia-de-dados-e-de-informacao-sistema-aberto-de-arquivamento-de-informacao-saai-modelo-de-referencia. Acesso em: 7 jun. 2023.
[28] Disponível em: http://www.loc.gov/standards/mets/. Acesso em: 7 jun. 2023.
[29] Disponível em: https://www.w3.org/XML/. Acesso em: 7 jun. 2023.

Quanto aos metadados de preservação (*preservation metadata*), estes são uma forma de metadados administrativos "[...] que contém informações necessárias para arquivar e preservar um recurso." (NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004, p. 1, tradução nossa) como, por exemplo, consoante *Digital Preservation Coalition* ([2018a]) e Riley (c2017), o *hardware* e o *software* exigidos para abrir (ou acessar e renderizar – exibir, executar etc. –) e usar um arquivo digital, e também soma de verificação (*checksum*)[30] e uma licença *Creative Commons* ou outros direitos de propriedade intelectual associados ao conteúdo, onde os usos primários são interoperabilidade, gerenciamento de objetos digitais e preservação. Esses metadados foram delineados conceitualmente no modelo de informação OAIS e implementados na prática através do dicionário de dados do padrão PREMIS e outras iniciativas de modelos de esquemas de metadados de preservação, tal como da *National Library of New Zealand* (2003).

Definidos assim, as funções dos metadados de preservação abrangem todo o ciclo de vida (criação, seleção, descrição etc.) dos recursos informacionais e se traduzem nos requisitos da preservação digital com base no OAIS que incluem, por exemplo, o registro de informações que fixem e validem a autenticidade[31], a origem e o contexto tecnológico de criação do objeto digital e as suas relações com outros objetos, além dos motivos de sua seleção, informações de conteúdo e de sua recuperação, identificação única e localização persistente. Sobre as razões da importância dos metadados e padrões de metadados na preservação digital, *Digital Preservation Coalition* ([2018a]) e *National Library of New Zealand* (2003) indicam que padrões e esquemas asseguram que os metadados possam ser interoperáveis e, por sua vez, os metadados permitem aos profissionais tomar decisões sobre como ou por que preservar um objeto digital, ou também que futuros usuários serão capazes de abrir, renderizar e entender corretamente o seu conteúdo.

[30] Soma de verificação (*checksum*) alude "[...] uma assinatura numérica única retirada de um arquivo." (DIGITAL PRESERVATION COALITION, c2015, não paginado, tradução nossa), ou segundo *Society of American Archivists* (c2022a) refere-se a um valor alfanumérico exclusivo que constitui o fluxo de *bits* de um arquivo individual de computador ou conjunto de arquivos, onde na preservação digital é usado durante a transferência/armazenamento de arquivos para apurar se os arquivos foram alterados indevidamente; por exemplo, um arquivista pode comparar *checksums* gerados antes e depois da transferência do arquivo para saber se ele manteve o mesmo valor durante este processo.

[31] Autenticidade é a "[...] capacidade de garantir que o objeto digital seja autêntico, ou seja, que reflita o conteúdo original de sua criação [...]" (GRÁCIO, c2012, p. 62), ou para Márdero Arellano (2008, p. 350), trata-se da "[...] comprovação de autoria do documento por meio de mecanismos de verificação como o *layout*, tipologia de fontes, vocabulários controlados da época e assinatura digital." Em outras palavras, a autenticidade (*authenticity*) remete que o material digital é o que ele pretende ser, sendo que para registro eletrônico, se refere à confiabilidade deste como um registro; e no caso de materiais "nascidos digitais" e digitalizados, diz respeito ao fato de que o que está sendo citado é o mesmo que era quando foi criado pela primeira vez, a menos que os metadados que o acompanham indiquem quaisquer mudanças realizadas ao longo do tempo (DIGITAL PRESERVATION COALITION, c2015).

Para isto, os metadados devem ser regidos por normas e boas práticas que proporcionam uma melhor qualidade e consistência de descrição (GILLILAND, c2016; GRÁCIO, c2012). A *National Information Standards Organization* (NISO)[32], através do *Framework of Guidance for Building Good Digital Collections*[33] que está na terceira edição de 2007, articula princípios para metadados de qualidade, como o uso de padrões de conteúdo[34] e de vocabulários controlados[35] e a detenção de verificabilidade, de acordo com *National Information Standards Organization* (c2004). A qualidade na construção de metadados é igualmente um requisito vital para dar visibilidade, acessibilidade e utilidade aos periódicos eletrônicos – parte importante no fluxo de comunicação científica –, visto o modelo tecnológico para revistas científicas e acadêmicas de acesso aberto na *Web* (como o protocolo *Open Archives Initiative Protocol for Metadata Harvesting* – OAI-PMH[36] e o *software Open Journal Systems* – OJS[37]) que se concentra na interoperabilidade de metadados. Para a avaliação da qualidade dos metadados em periódicos científicos eletrônicos temos os critérios de qualidade do Modelo de Qualidade de Dados da *International Organization for Standardization* (ISO)[38]/*International Electrotechnical Commission* (IEC) 25012[39] (BENTANCOURTI; ROCHA, 2012).

---

[32] A NISO, fundada em 1939, é uma associação sem fins lucrativos, credenciada pelo *American National Standards Institute* (ANSI), que identifica, cria, mantém e publica padrões técnicos para gerir informações (BACA, c2016; NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004), tal como o ANSI/NISO Z39.85-2012 *The DC Metadata Element Set*. Disponível em: https://www.niso.org/welcome-to-niso. Acesso em: 7 jun. 2023.

[33] Disponível em: https://www.niso.org/sites/default/files/2017-08/framework3.pdf. Acesso em: 7 jun. 2023.

[34] Padrão de conteúdo (*content standard*) é "[...] um conjunto de regras formais que especificam o conteúdo, a ordem e a sintaxe das informações para promover a consistência." (SOCIETY OF AMERICAN ARCHIVISTS, c2022b, não paginado, tradução nossa). Através de Gilliland (c2016), a *Resource Description and Access* (RDA), a *Anglo-American Cataloguing Rules* (AACR2), a *Describing Archives: a Content Standard* (DACS), o *Cataloging Cultural Objects* (CCO) e a *International Standard Bibliographic Description* (ISBD) são alguns exemplos de padrões de conteúdo de dados (regras e códigos de catalogação).

[35] Vocabulário controlado (*controlled vocabulary*) é "[...] uma lista enumerada de termos pré-selecionados da linguagem natural e usados principalmente para ajudar na descoberta de sistemas de recuperação de informações" (SOCIETY OF AMERICAN ARCHIVISTS, c2022c, não paginado, tradução nossa). Com base em Gilliland (c2016), o *Library of Congress Subject Headings* (LCSH), o Getty *Thesaurus of Geographic Names* (TGN), o *Library of Congress Name Authority File* (LCNAF) e o *Iconclass* são alguns exemplos de padrões de valor de dados (vocabulários controlados, tesauros, listas controladas).

[36] Publicado pela OAI que se constitui numa "[...] organização que desenvolveu padrões de interoperabilidade para facilitar a disseminação eficiente de conteúdo *online*, especialmente *EPrints*.", o protocolo OAI-PMH "[...] é usado por provedores de dados, que expõem metadados sobre informações mantidas em um repositório, e por provedores de serviços, que usam esses metadados para criar serviços de valor agregado." (SOCIETY OF AMERICAN ARCHIVISTS, c2022j, não paginado, tradução nossa). Disponível em: http://www.openarchives.org/pmh/. Acesso em: 7 jun. 2023.

[37] Elaborado e lançado pela *Public Knowledge Projec* (PKP) em 2001 para melhorar o acesso à pesquisa, o OJS é um *software* livre e plataforma para gerir e publicar periódicos acadêmicos de acesso aberto, com mais de vinte e cinco mil periódicos utilizando-o em todo o mundo. Disponível em: https://pkp.sfu.ca/ojs/. Acesso em: 7 jun. 2023.

[38] Fundada em 1947, a ISO é uma rede global voluntária de institutos nacionais de normalização, onde os órgãos de normalização agem com organizações internacionais, governos, empresas e representantes dos consumidores para definir padrões em comum e promover seu uso com o objetivo de facilitar o comércio e atender às demandas amplas da sociedade (BACA, c2016). Disponível em: https://www.iso.org/about-us.html. Acesso em: 2 abr. 2021.

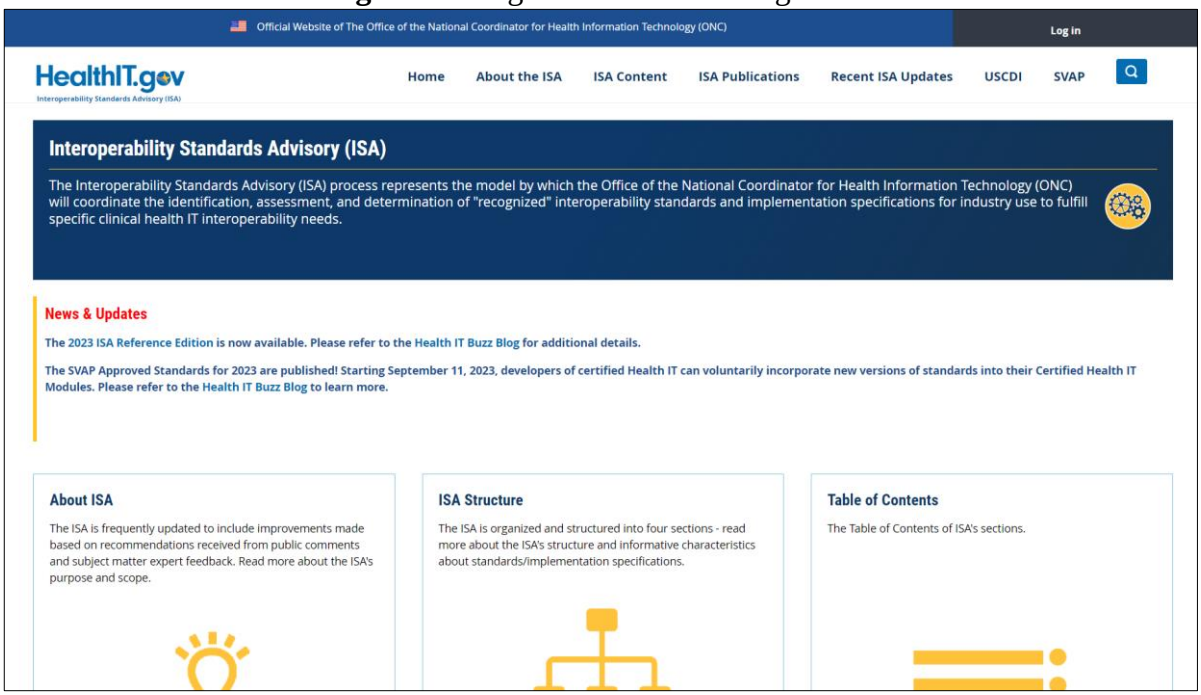
[39] Disponível em: https://www.iso.org/standard/35736.html. Acesso em: 7 jun. 2023.

Desta forma, a conferência de qualidade permite ter metadados consistentes e credíveis, assegurando maior eficiência e estruturação na recuperação de informações e uma melhora no rastreamento de citações e referências e na autenticidade e confiabilidade de artigos publicados em acesso aberto nos vários ambientes digitais (como OJS, SciELO etc.), além de favorecer a preservação digital dos periódicos e a interoperabilidade entre sistemas (GULKA; SILVEIRA, 2019). Sobre exemplos de uso de metadados na comunicação científica, em *Allen Institute for AI* ([2021a], [2021b]), Castro (2020), InformaSUS-UFSCar (c2021), *Internet Archive* ([2021a], 2021b), *Office of the National Coordinator for Health Information Technology* (2021a, [2021b]), *Pan American Health Organization* ([2021]) e Tainacan ([2021]), notamos iniciativas internacionais e nacionais que lidam com a adoção de padrões de metadados para o tratamento de dados e informações em saúde (sobretudo, acerca do Covid-19), que garantem a persistência, a coerência e a interoperabilidade dos dados entre os órgãos de saúde e/ou governos e facilitam o acesso e a recuperação de informações confiáveis por especialistas e o público leigo, a saber:

- *Interoperability Standards Advisor* (ISA)[40] – criado para ser um catálogo ou material informativo, o ISA é o modelo pelo qual o *Office of the National Coordinator for Health Information Technology* (ONC)[41] faz a identificação, avaliação e consciência pública dos padrões e especificações de implementação para atender necessidades específicas de interoperabilidade de informações em saúde clínica eletrônicas, como sobre o Covid-19[42], nos Estados Unidos.

[40] Disponível em: https://www.healthit.gov/isa/. Acesso em: 7 jun. 2023.
[41] Disponível em: https://www.healthit.gov/topic/about-onc. Acesso em: 7 jun. 2023.
[42] Disponível em: https://www.healthit.gov/isa/covid-19. Acesso em: 7 jun. 2023.

Figura 1 – Página inicial do catálogo ISA.



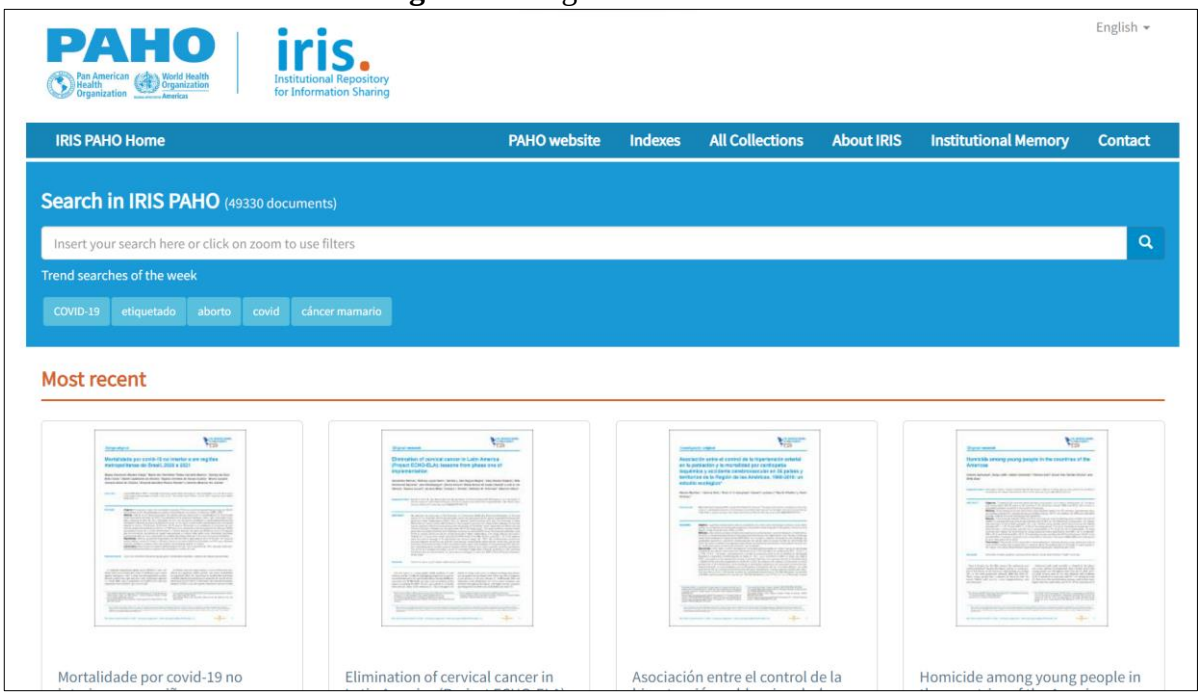
Fonte: *Office of the National Coordinator for Health Information Technology* ([2023?]).

O ISA é ordenado em quatro seções: Vocabulário, conjuntos de códigos, padrões de terminologia e especificações de implementação (isto é, dados da "semântica" do recurso); Padrões de conteúdo/estrutura e especificações (ou dados da "sintaxe" do recurso); Padrões e especificações de implementação para serviços (quer dizer, elementos de infraestrutura implantados e usados para atender as necessidades de interoperabilidade); e Padrões administrativos e especificações de implementação (ou seja, pagamento, operações e outras necessidades de interoperabilidade "não clínicas").

- *Institutional Repository for Information Sharing* (IRIS)[43] – criado pela *Pan American Health Organization* (PAHO)[44], o IRIS visa facilitar o livre acesso em formato digital à produção técnico-científica da PAHO e da *World Health Organization* (WHO)[45], sendo fundamental para assegurar a disponibilização e a preservação dos materiais através da garantia de disposição permanente dos arquivos, que é feito na plataforma *DSpace*[46] e está associado a uma das suas funcionalidades, assim como da realização de *backup* de forma periódica e externamente a este *software*.

[43] Disponível em: https://iris.paho.org/. Acesso em: 7 jun. 2023.
[44] Disponível em: https://www.paho.org/en. Acesso em: 7 jun. 2023.
[45] Disponível em: https://www.who.int/. Acesso em: 7 jun. 2023.
[46] O DSpace, abreviação de *Durable Space*, foi desenvolvido no *Massachusetts Institute of Technology* (MIT) e constitui "[...] um sistema de gerenciamento de conteúdo especializado que permite que diferentes comunidades usem a *Web* para capturar, distribuir e preservar obras digitais e fornecer acesso a essas obras por meio de metadados." (SOCIETY OF AMERICAN ARCHIVISTS, c2022g, não paginado, tradução nossa). Disponível em: https://duraspace.org/dspace/. Acesso em: 7 jun. 2023.

Figura 2 – Página inicial do IRIS.



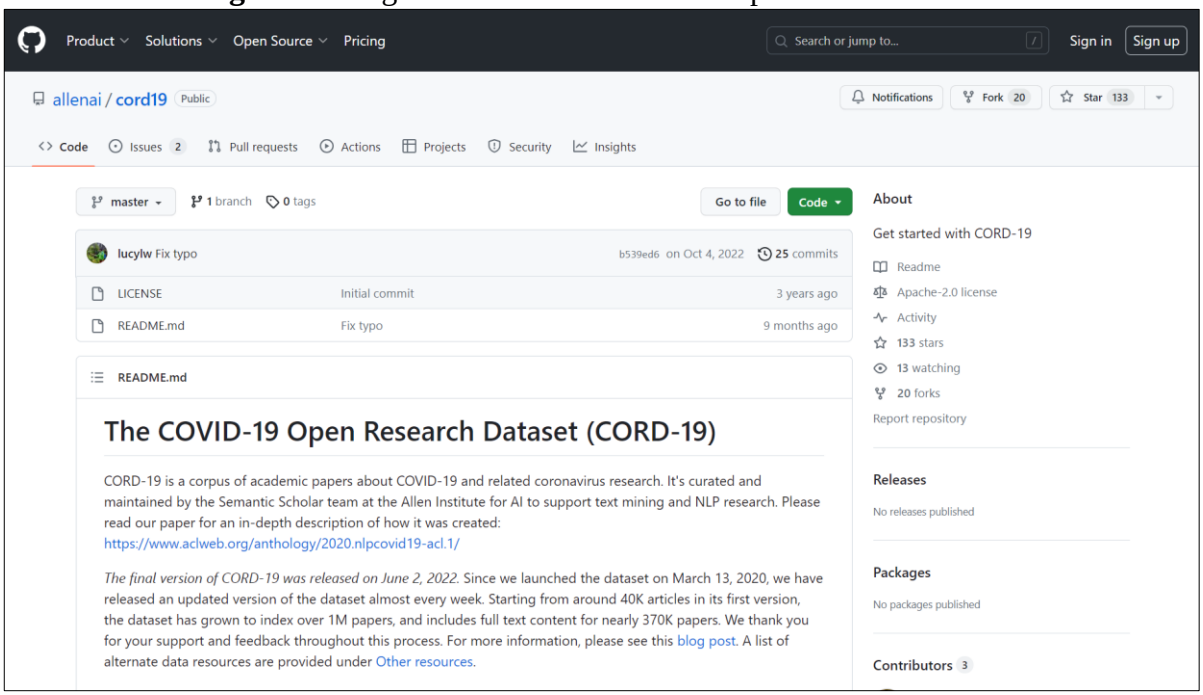
Fonte: *Pan American Health Organization* ([2023?]).

O IRIS também fornece a recuperação de informações em navegadores *online* (*browsers*) e a integração com outros bancos de dados e repositórios como, por exemplo, o *Global Health Observatory*[47] da WHO e a Biblioteca Virtual em Saúde (BVS)[48], mediante o protocolo OAI-PMH. Implementado mundialmente em conjunto com o *DSpace*, o formato DC é utilizado no IRIS para a descrição bibliográfica e a indexação dos documentos de acesso livre e gratuito.

- COVID-19 *Open Research Dataset* (CORD-19)[49] – criado pela *Semantic Scholar*[50] do *Allen Institute for* AI[51] nos Estados Unidos que é uma plataforma gratuita de pesquisa baseada em inteligência artificial o qual auxilia os pesquisadores a descobrir a literatura científica mais útil para o seu trabalho, o Cord-19 objetiva fornecer aos pesquisadores ferramentas e conjuntos de dados gratuitos e abertos para encontrar novos *insights* sobre o Covid-19, ajudando acadêmicos a superar o problema da sobrecarga de informações na ciência.

[47] Disponível em: https://www.who.int/data/gho. Acesso em: 7 jun. 2023.
[48] Disponível em: http://red.bvsalud.org/. Acesso em: 7 jun. 2023.
[49] Disponível em: https://www.semanticscholar.org/cord19. Acesso em: 7 jun. 2023.
[50] Disponível em: https://www.semanticscholar.org/about. Acesso em: 7 jun. 2023.
[51] Disponível em: https://allenai.org/. Acesso em: 7 jun. 2023.

Figura 3 – Página inicial do CORD-19 na plataforma *GitHub*.



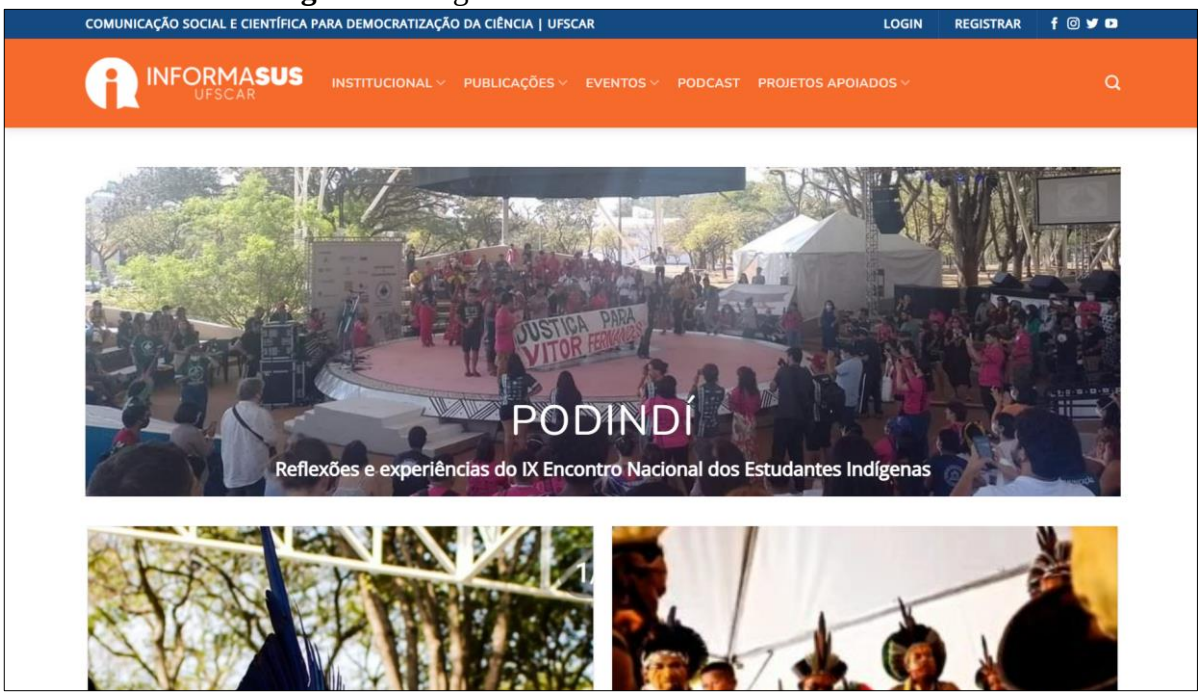
Fonte: *Allen Institute for AI* (c2023).

Com mais de 280 mil artigos científicos para uso pela comunidade global de pesquisa, a modelagem deste repositório digital é formada de vinte e dois elementos de metadados provenientes da *Microsoft Academic Graph* (MAG)[52] e do mapeamento dos dados via códigos fonte cedidos pela WHO. Apesar de não usar um padrão de metadados específico, a estruturação do Cord-19 apresenta o requisito computacional do formato XML e está em acesso aberto, propiciando a interoperabilidade e a integração dos dados.

- UFSCar[53] – constituindo uma instituição federal brasileira de ensino superior no Estado de São Paulo, a UFSCar tem iniciativas de pesquisadores inquietos com a crise sanitária do coronavírus. Uma delas é o Projeto de Extensão "Comunicação Social no Contexto da Covid-19" tido de InformaSUS-UFSCar[54], atrelado ao Departamento de Medicina, que foi criado em 2020 pelo esforço coletivo entre docentes, alunos etc. para promover a pesquisa, organização, checagem e produção de conteúdos científicos para imprensa, *Internet* e mídias sociais, qualificando as informações difundidas ao público e apoiando no controle da pandemia e ao combate de notícias falsas.

[52] Disponível em: https://docs.microsoft.com/en-us/academic-services/graph/reference-data-schema. Acesso em: 22 dez. 2021.
[53] Disponível em: https://www.ufscar.br/. Acesso em: 7 jun. 2023.
[54] Disponível em: https://www.informasus.ufscar.br/. Acesso em: 7 jun. 2023.

Figura 4 – Página inicial do InformaSUS-UFSCar.



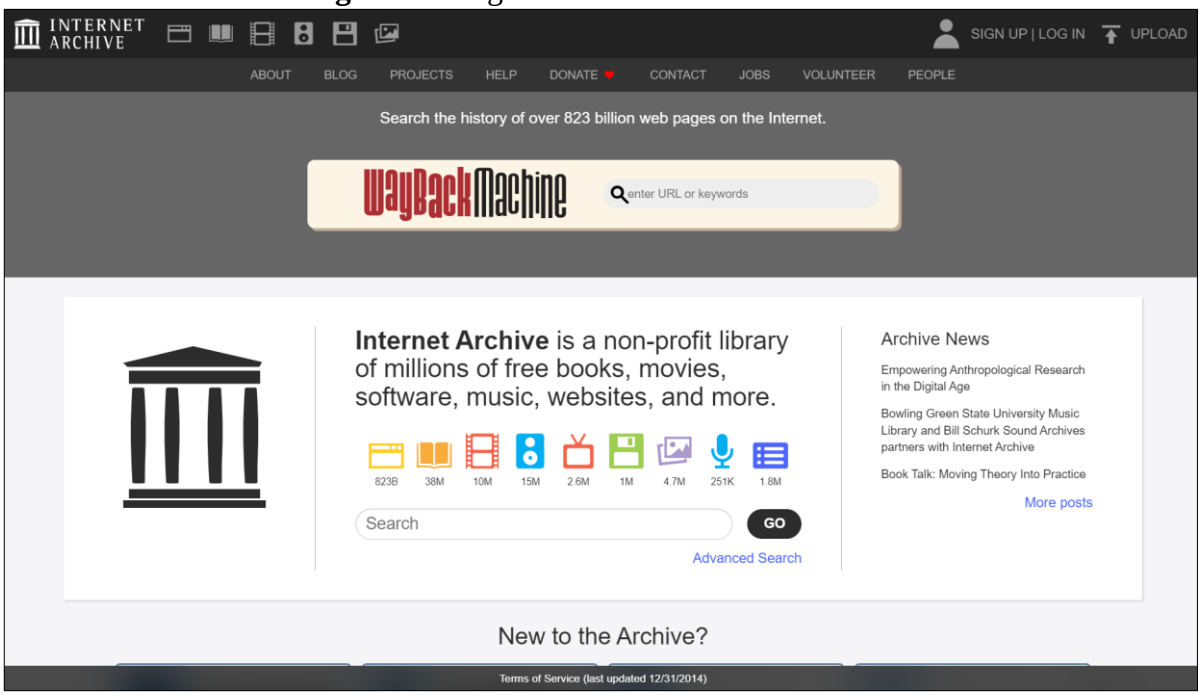
Fonte: InformaSUS-UFSCar (c2023).

Para a pesquisa, recuperação e acesso adequado dos conteúdos, as coleções de publicações no *website* do projeto vêm sendo catalogadas através do Tainacan[55], um *software* livre para criação de repositórios de acervos digitais em *WordPress*[56] que permite dispor as coleções em vários formatos, como o OAI-PMH, e cria-las com um conjunto de metadados a partir do DC.

- *Internet Archive*[57] – sendo uma organização sem fins lucrativos iniciada em 1996 e que fornece acesso gratuito e universal a uma biblioteca digital com *sites*, *ebooks* etc., o *Internet Archive* lançou em 2006 o *Archive-It*[58] como solução de arquivamento da *Web* para que instituições criem, armazenem e deem o acesso a coleções de conteúdos *Web*.

[55] Disponível em: https://tainacan.org/. Acesso em: 7 jun. 2023.
[56] Surgido em 2003, o *software WordPress* é um projeto de código aberto que permite criar *websites*, *blogs* ou aplicativos. Disponível em: https://br.wordpress.org/. Acesso em: 7 jun. 2023.
[57] Disponível em: https://archive.org/about/. Acesso em: 7 jun. 2023.
[58] Disponível em: https://archive-it.org/. Acesso em: 7 jun. 2023.

Figura 5 – Página inicial do *Internet Archive*.



Fonte: *Internet Archive* ([2023?]).

Como exemplos destas coleções *Web*, temos o "*Global Health Events Web archive*" (https://archive-it.org/collections/4887) da NLM que abrange conteúdos de sites e mídias sociais sobre a doença pelo vírus Zika, o surto por Coronavírus etc., e o "*Global Social Responses to Covid-19 Web Archive*" (https://archive-it.org/collections/14022) da *Ivy Plus Libraries Confederation*[59] em que inclui *sites*, *blogs* etc. que registram respostas regionais e sociais ao Covid-19. Além da função central de capturar e preservar conteúdos *Web*, este serviço possibilita que os usuários adicionem, importem e exportem metadados descritivos, e cede tanto campos de metadados do DC como a capacidade de adicionar campos personalizados.

Com efeito, os padrões e esquemas de metadados, sejam eles para descrição, gerência ou preservação de objetos digitais, são recursos tecnológicos chave na interoperabilidade. Visto que a ciência se faz mais orientada à dados, colaborativa e interdisciplinar crescendo a demanda por interoperabilidade entre dados, serviços e ferramentas, Edwards *et al*. (c2011) notam que os metadados (ou informações sobre um conjunto de dados, livros, artigos etc.) podem ser uma fonte de atrito ou de dificuldades entre os cientistas de duas ou mais disciplinas que trabalham juntos em problemas associados – chamada de "fricção científica" (*science friction*) –, inibindo o reuso e o compartilhamento interdisciplinar de dados. Para superar tal fato, os autores alegam que os produtos de metadados (ou seja, conjuntos de descritores, *tags* XML, *links*, catálogos e

[59] Disponível em: https://ivpluslibraries.org/. Acesso em: 7 jun. 2023.

demais registros fixos e estruturados) devem ser completados com processos de metadados – conversas, *e-mails*, anotações e mais meios informais e efêmeros de comunicação –, indicando o papel enorme das práticas *ad hoc* e informações incompletas no trabalho científico cotidiano.

A metáfora do atrito de dados (*data friction*) de Edwards (2010) explica o que acontece nas interfaces entre as "superfícies" de dados, ou melhor, os pontos em que os dados se movem entre pessoas, organizações e máquinas – de uma disciplina para outra, de um laboratório para outro, de um monitor para um computador, ou de um formato de dados (como planilhas *Excel*) para outro (como uma base de dados científica) –, no qual cada interface entre grupos, máquinas etc. é um ponto de resistência onde os dados podem ser alterados, mal interpretados ou perdidos, conforme Edwards *et al*. (c2011). Para os autores o atrito de dados resulta na fricção científica, e os metadados são uma forma de comunicação científica onde os produtos de metadados bem codificados aumentam a precisão com que um conjunto de dados científicos pode ser ajustado para fins aos quais não foi destinado originalmente, ou pode ser reutilizado por pessoas que não participaram da sua criação; e conjuntamente, os processos de metadados atuam como práticas cujo cientistas e o público podem galgar uma comunicação científica desarticulada e imprecisa.

Corroborando com Edwards (2010) e Edwards *et al*. (c2011), Mayernik (2019) entende que os metadados são partes integrantes essenciais para a produção do conhecimento científico no mundo digital, onde se os dados são entidades usadas como prova de fenômenos em apoio a um argumento científico, então os metadados são produtos e processos que permitem que tais entidades sejam responsabilizadas como evidência. Para o autor tanto os produtos formais de metadados (isto é, descrições escritas padronizadas e estruturadas de dados, anotações textuais, documentos formalizados etc., criados e aplicados para gerenciar, descobrir, acessar, utilizar e preservar recursos informacionais, os quais aumentam a precisão das interações de pesquisa, facilitam a interoperabilidade e a legibilidade por máquinas) como os processos informais de metadados (ou *e-mails* pessoais, discussões face a face, documentos *ad hoc* etc., que viabilizam a produção, comunicação e colaboração relacionadas aos dados) nos propiciam que algo exista como dados (ou sirva a um papel de evidência) em qualquer forma ou configuração disponível.

Em sua pesquisa, o autor analisa também a criação de metadados na prática científica cotidiana, focando em como os cientistas descrevem, registram, anotam, organizam e gerem os seus dados, seja para o seu próprio uso ou, até, para o uso de pesquisadores fora de seu projeto. Mayernik (2019) expõe que o estudo e a conceituação de metadados na Sociologia da ciência e Sociologia dos dados pode fornecer uma visão sobre a produção de evidências científicas, isto é, como algo que podemos chamar de "dados" torna-se capaz de servir a um papel de evidência (por exemplo, relatos, descrições e instruções por metadados em conjunto com procedimentos

de trabalho e colaboração científica permitem que medições, amostras, observações de pesquisa etc. tenham valor probatório no contexto de um argumento científico); assim como fornecer um meio para mostrar pelo que/quais os pesquisadores são responsáveis, e com o que eles obtêm essa responsabilidade (como alguns cientistas que são responsáveis pelos dados e metadados, outros não, onde muitas vezes as responsabilidades são designadas informalmente ou de fato).

2.1.2 Arquivamento da *Web*

Sendo uma estratégia operacional de preservação digital – ou seja, as medidas reais de preservação física, lógica ou conceitual dos objetos digitais a serem realizadas pelas instituições (MÁRDERO ARELLANO, 2008, THOMAZ; SOARES, 2004) –, o arquivamento da *Web*, ou a preservação da *Web* (*preserving the Web*), diz respeito ao "[...] processo de coleta de porções da *World Wide Web*, preservando as coleções em formato de arquivo, e em seguida fornecendo os arquivos para o acesso e a utilização." (INTERNATIONAL INTERNET PRESERVATION CONSORTIUM, c2022a, não paginado, tradução nossa). De acordo com *Digital Preservation Coalition* ([2018f], p. 1, tradução nossa) corresponde, por sua vez, na "[...] prática de obter uma cópia de um *website* ou de um conteúdo particular publicado na *Web* para servir de registro.", onde os registros da *Web* podem [...] consistir de um *site* inteiro ou apenas do texto de algumas páginas." bem como "[...] necessitam de atenção urgente já que a *Web* por natureza é efêmera."

A *Internet* nos possibilitou uma era sem precedentes em termos de compartilhamento de saber, criatividade, inovação e conexão, no qual a *Web* é um recurso dinâmico e único de comunicação com alto valor para pesquisas atuais e futuras; no entanto, também criou novos desafios para as organizações cuja missão é documentar e preservar o conhecimento e a cultura contemporâneos, como as instituições patrimoniais que coletam publicações acadêmicas, obras de arte, materiais governamentais e demais itens que costumavam ser impressos e estão agora disponíveis apenas no ambiente da *Web* (INTERNATIONAL INTERNET PRESERVATION CONSORTIUM, c2022a). Ademais, consoante *Library of Congress* ([2022b]), os *websites* – que registram eventos atuais, organizações, reações públicas, informações culturais, acadêmicas e do governo sobre assuntos diversos – são efêmeros e tidos como conteúdos de

risco, pois novos *sites* se formam constantemente, os *Uniform Resource Locator* (URL)[60] e conteúdos mudam e, às vezes, os *sites* desaparecem.

Uma variedade de estudos aborda o problema da efemeridade da informação produzida na *Web*. Como levantado por Costa, Gomes e Silva (2017), em torno de 80% das páginas da *Web* e 11% dos recursos de mídia social, tais como os postados no *Twitter*, estarão perdidos ou não disponíveis na sua forma original depois de um ano publicados, e 13% das referências *Web* em artigos acadêmicos desaparecem após 27 meses. Através de Webster (2015) as estatísticas do arquivo da *Web* do Reino Unido, do inglês UK *Web Archive* (UKWA)[61], demostram que após um ano apenas 10% da *Web* permanece viva e inalterada. Para mais, enquanto as citações na *Web* se tornaram comuns com o aumento da quantidade de literatura *online*, o uso de *links* não persistentes e que decaem com o tempo provoca problemas de acessibilidade nas citações *Web* usadas em artigos científicos (GUL; MAHAJAN; ALI, 2014; KUMAR; KUMAR, 2017; SADAT-MOOSAVI; ISFANDYARI-MOGHADDAM; TAJEDDINI, 2012; SIFE; LWOGA, 2017; TAJEDDINI *et al*., 2011) ou em teses (KUSHKOWSKI, 2005; SIFE; BERNARD, 2013).

Há uma série de razões pelas quais o conteúdo da *Web* é arquivado através de iniciativas surgidas nas últimas décadas em todo o mundo de arquivos da *Web* (*Web archives*), isto é, as organizações que se dedicam "[...] principalmente à coleta e preservação do conteúdo da *Web* [...]" ou "[...] as cópias preservadas do conteúdo *Web* ao vivo coletado para retenção e acesso permanente [...]" (SOCIETY OF AMERICAN ARCHIVISTS, c2022k, não paginado, tradução nossa). Em Costa, Gomes e Silva (2017), *Digital Preservation Coalition* ([2018f]) e Webster (c2020) indicam-se fatores para os quais informações publicadas na *Web* são arquivadas, como para servir de prova para apoiar o cumprimento de políticas organizacionais e exigências legais e normativas (onde *sites* organizacionais são registros oficiais e estão sujeitos a demandas de livre acesso à informação), para servir como memória corporativa ou para explicar as histórias do passado e prever eventos futuros (doenças, desastres naturais etc.) via extração e análise da evolução dos dados, os quais retratam a relevância dos arquivos *Web* como fontes para pesquisa.

Visto que referenciar fontes – como um item vital do discurso acadêmico – e a espera de que as fontes referenciadas sejam verificadas por outros (para a interpretação correta dos

[60] Como um tipo de *Uniform Resource Identifier* (URI), o URL consiste em um endereço de *Internet* que identifica e informa aos usuários como e onde localizar um arquivo específico na *Web*, sendo que ele abrange o nome de um arquivo, o nome do computador hospedeiro (*host*, do qual *hostname* refere-se a um identificador para uma máquina específica na *Internet*, e sua sub-rede e domínio), o caminho do diretório para chegar a esse arquivo e o protocolo exigido para usá-lo (por exemplo, http://www.getty.edu/research/publications/electronic_publications/index.html indica que o HTTP deve ser usado para recuperar o documento "index.html" do *host* "www.getty.edu" no diretório "research/publications/electronic_publications/index.html") (BACA, c2016).

[61] Disponível em: https://www.webarchive.org.uk/en/ukwa/about. Acesso em: 7 jun. 2023.

dados comunicados e o apoio a reprodução dos resultados) passam a ser um problema assíduo no ambiente virtual, Ferreira, Martins e Rockembach (2018), a partir do valor do registro e da armazenamento de conteúdos *Web* para acesso futuro e devida comunicação científica, julgam que o arquivamento da *Web* é uma das ações que garante o potencial informacional e probatório das referências que fornecem subsídios e sustentam as investigações científicas. Para os autores as iniciativas nesta área concedem origem a referências persistentes, que formam uma memória virtual, assegurando o seu potencial informacional; e permitem a validação durante o tempo, garantindo o seu potencial probatório (incluindo a capacidade de acesso para a validação e a preservação do conteúdo da referência como era no ato da coleta dos dados para uma pesquisa).

Porém, o arquivamento da *Web* é ainda uma área pouco explorada internacionalmente, máxime nas universidades e no campo científico e, igualmente, se nota a carência de pesquisas sobre este tema na América Latina, incluindo o Brasil (ROCKEMBACH, 2018; FERREIRA; MARTINS; ROCKEMBACH, 2018). Aliás, ao passo que iniciativas internacionais preservam conteúdos brasileiros de forma dispersa, o Brasil até então não realiza o arquivamento da *Web* de maneira sistemática, em conformidade com Rockembach (2018, 2019) e Rockembach e Pavão (2018). Dos exemplos recentes de iniciativas brasileiras na área, destacamos o estudo de Melo (2020) o qual concluiu que os *websites* do Governo Federal Brasileiro (domínio gov.br) são arquiváveis sem a perda de informações importantes havendo a necessidade de uma política pública para sistematizar o arquivamento de *sites* governamentais; assim como os grupos de pesquisa formados nos últimos anos que investigam a temática, como é o caso do Núcleo de Pesquisa em Arquivamento da *Web* e Preservação Digital (NUAWEB)[62] da Universidade Federal do Rio Grande do Sul (UFRGS), que mantém a plataforma Arquivo da *Web* Brasileira[63] com a ajuda de voluntários.

Ferramentas

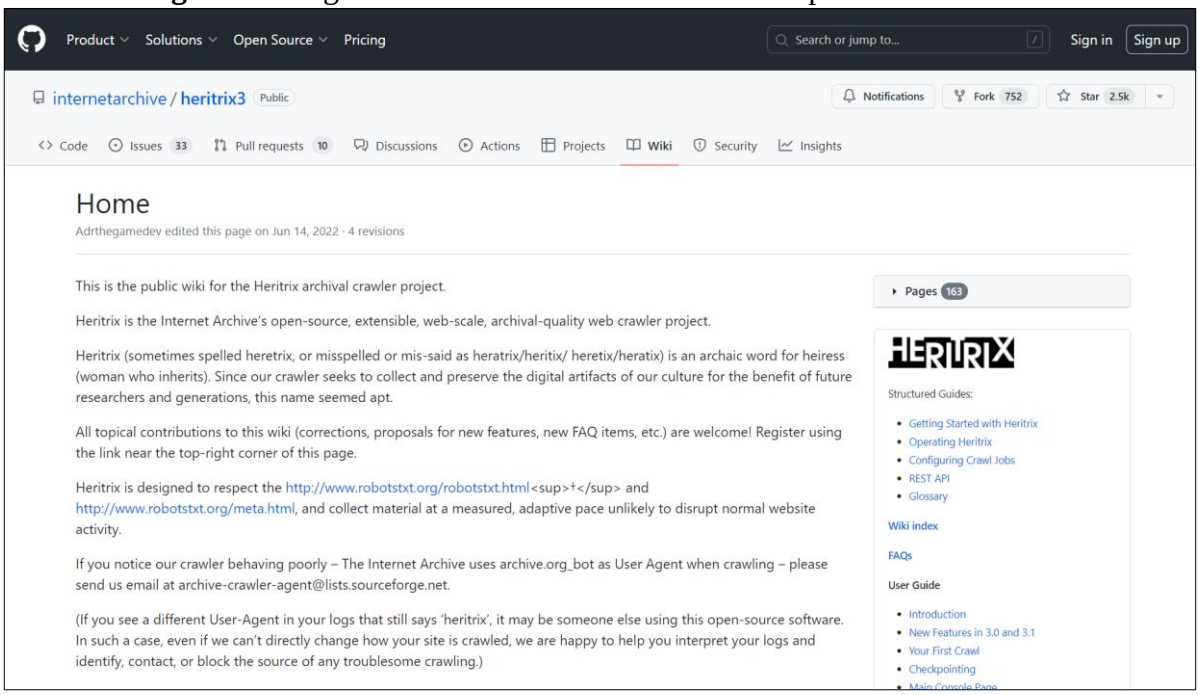
Através de *Digital Preservation Coalition* (c2015), *International Internet Preservation Consortium* (c2022c), *Internet Archive* ([2021a], c2022b), *Memento Project* (2016), *National Library of the Netherlands* e *National Library of New Zealand* (2020), *Rhizome.org* ([2022]),

[62] Disponível em: https://www.ufrgs.br/nuaweb/. Acesso em: 7 jun. 2023.
[63] Disponível em: http://www.arquivo.ong.br/. Acesso em: 7 jun. 2023.

Roche (c2022), Samouelian e Dooley (c2018) e Sørensen e Have (2020), identificamos algumas ferramentas[64] ou soluções para a coleta, arquivamento e preservação de páginas da *Web*, a saber:

- *Heritrix*[65] – rastreador *Web* (*Web crawler*) de código aberto, extensível, escalável e com qualidade de arquivamento do *Internet Archive*.

Figura 6 – Página inicial da ferramenta *Heritrix* na plataforma *GitHub*.



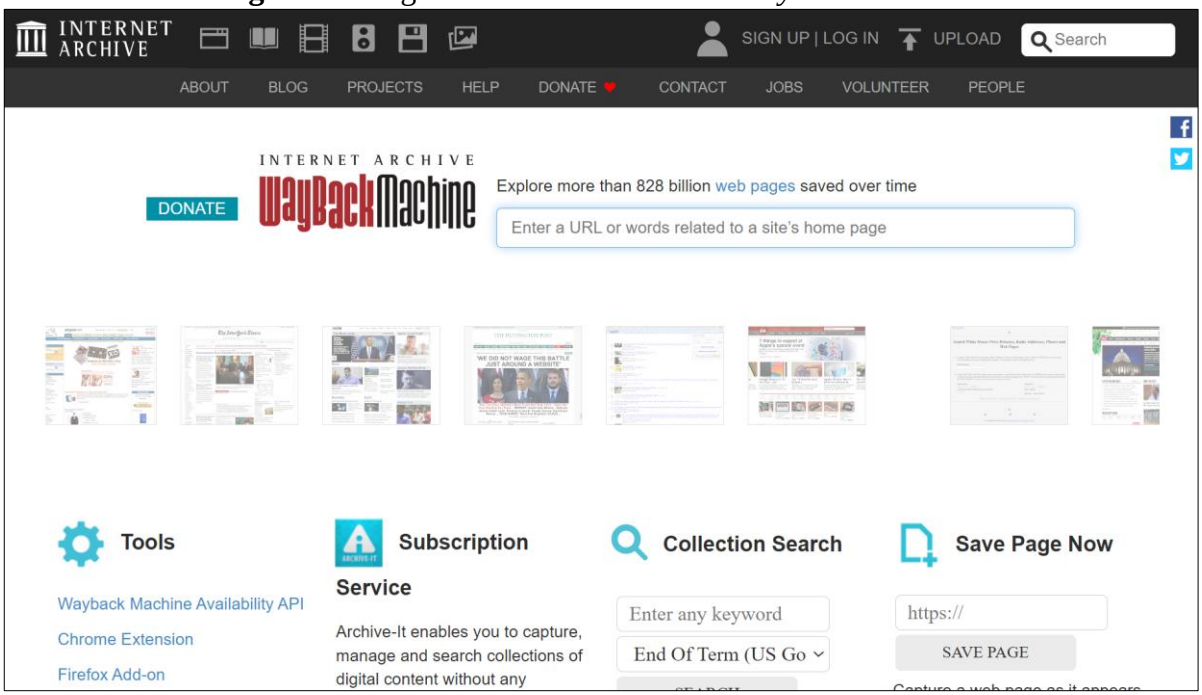
Fonte: *Internet Archive* (c2023).

Trata-se de uma das principais soluções de captura adotadas pela organização (incluindo na *Wayback Machine* e no *Archive-It*) e por muitas outras iniciativas de arquivamento da *Web* no mundo para a coleta de *sites*. Este *software* é projetado para recolher o material a um ritmo medido e adaptável, pouco susceptível de perturbar a atividade normal do *website*, e permite produzir arquivos no formato *Web ARChive* (WARC)[66], um padrão de arquivo para conteúdo da *Web*, porém não admiti a entrada/geração de metadados descritivos adicionais dentro da ferramenta.

- *Wayback Machine*[67] – *software* do *Internet Archive* que reproduz páginas *Web* de *sites* arquivados.

[64] Deve-se levar em conta que grande parte dos *softwares* identificados está em desenvolvimento ativo e os estudos citados ocorreram ao longo de vários anos, assim, pode haver inconsistências na forma como as ferramentas são descritas pelas diferenças entre versões ou, mesmo, pelas constantes atualizações e melhorias de funcionalidades.
[65] Disponível em: https://webarchive.jira.com/wiki/spaces/Heritrix/overview. Acesso em: 7 jun. 2023.
[66] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 28500**: Information and documentation: WARC file format, Geneva: ISO, 2017. Disponível em: https://www.iso.org/standard/68004.html. Acesso em: 7 jun. 2023.
[67] Disponível em: http://web.archive.org/. Acesso em: 7 jun. 2023.

Figura 7 – Página inicial da ferramenta *Wayback Machine*.



Fonte: *Internet Archive* (2014b).

É um conjunto de aplicações que indexam e recuperam o conteúdo da *Web* capturado; renderizam arquivos WARC[68] e ARC[69]; e exibem o conteúdo numa interface de usuário na *Web*. Na interface pública (https://web.archive.org/) pode-se pesquisar por URL ou palavras ligadas a *homepage* de um *site* e no *Archive-It* há o suporte para busca por texto completo, cujo no caso do URL de um *site* tenha sido recuperada o arquivo do *site* pode ser navegado por data de captura. Ela captura metadados descritivos (URL original etc.) e administrativos (data, URL direcionada, nome do arquivo WARC etc.).

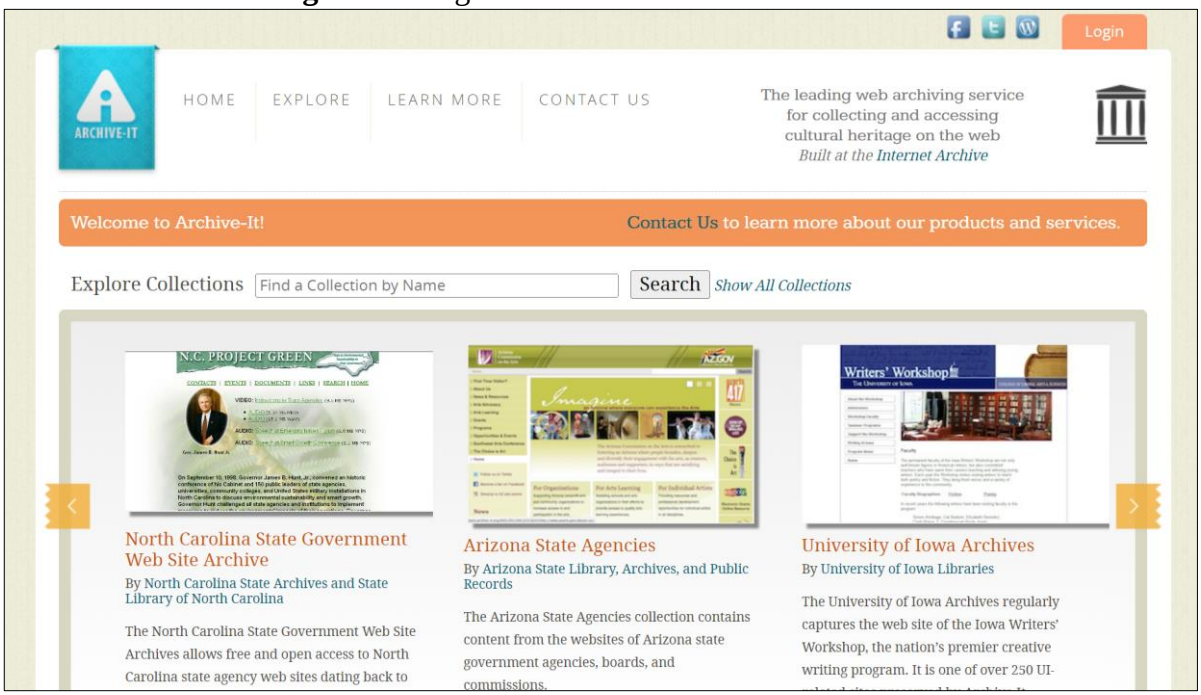
- *Archive-It*[70] – serviço de arquivamento *Web* por assinatura e de várias funcionalidades do *Internet Archive*.

[68] O WARC é um "formato de arquivo que especifica um método para combinar múltiplos recursos digitais em um arquivo agregado junto com as informações relacionadas" (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2013, não paginado, tradução nossa) ou, segundo Pennock (c2013, p. 36, tradução nossa), refere-se a "um formato de contêiner para *sites* arquivados [...]".

[69] ARC é um "formato de contêiner para *sites* criado pelo *Internet Archive*, substituído pelo WARC" (PENNOCK, c2013, p. 35, tradução nossa). Disponível em: https://archive.org/web/researcher/ArcFileFormat.php. Acesso em: 7 jun. 2023.

[70] Disponível em: https://archive-it.org/. Acesso em: 7 jun. 2023.

Figura 8 – Página inicial da ferramenta *Archive-It*.



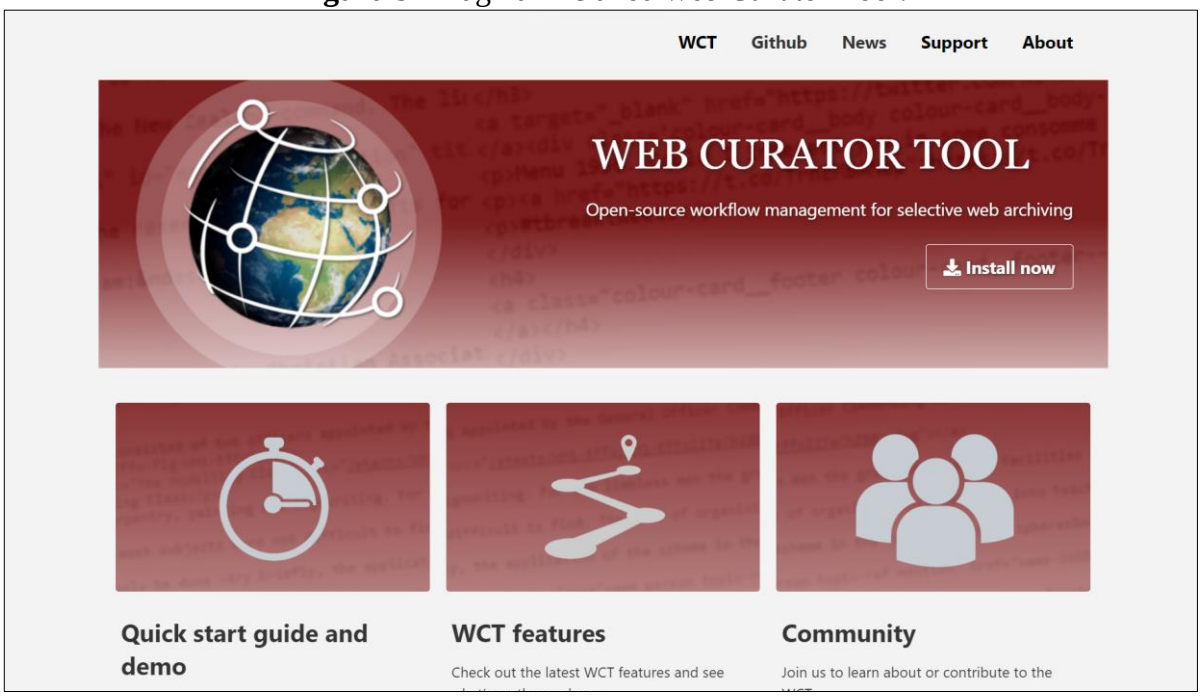
Fonte: *Internet Archive* (2014a).

Além da função de captura de *sites* usando o *Heritrix*; o *Archive-It* é uma ferramenta administrativa que permite coletar, catalogar e gerenciar as coleções arquivadas de instituições acadêmicas e de patrimônio cultural; e preservar os arquivos da *Web* destas instituições armazenando (e tendo a capacidade de baixar) seus arquivos WARC que contêm conteúdo da *Web* capturado nos repositórios do *Internet Archive*. Através do seu *site* público, *sites* arquivados são exibidos com a *Wayback Machine* e a interface de acesso admiti a pesquisa por texto completo e URL. A adição de metadados no nível da coleção é manual e os custos depende das exigências da instituição coletora.

- *Web Curator Tool*[71] – *software* de código aberto de gestão de fluxo de trabalho para que usuários não técnicos gerenciem o processo de arquivamento seletivo da *Web*, criado pela Biblioteca Nacional da Nova Zelândia e a Biblioteca Britânica com a *Oakleigh Consulting* no Reino Unido.

[71] Disponível em: https://webcuratortool.org/. Acesso em: 7 jun. 2023.

Figura 9 – Página inicial da *Web Curator Tool*.



Fonte: *National Library of the Netherlands* e *National Library of New Zealand* (c2021).

A aplicação suporta seleção, coletas, descrição, permissões e outras tarefas incluídas no fluxo de trabalho para apoiar a aquisição e a descrição da *Web*; usa o *Heritrix* para rastrear e pode integra-se com a *Wayback Machine*; e não tenta ser uma ferramenta de acesso, repositório digital, ou um catálogo. Pode captar e gerar WARC, e a maior parte dos metadados descritivos devem ser adicionados pelo usuário.

- *NetarchiveSuite*[72] – *software* de código aberto para gerenciar o arquivamento seletivo e de domínio amplo da *Web*, desenvolvido em 2004 pela Biblioteca Nacional do Reino da Dinamarca. Esta ferramenta consiste de diferentes módulos, incluindo o módulo de coleta (*harvester*) que usa o *Heritrix* e trata da definição, programação e execução de rastreamentos (*crawls*); um módulo de arquivo que serve como sistema de preservação de material coletado; e um módulo de acesso para dar acesso e visualização ao material via uma solução *proxy*. O *NetarchiveSuite* gera arquivos WARC, e dados estruturais e técnicos (a estrutura do *site*, a extensão do conteúdo etc.) são gerados automaticamente.
- *HTTrack*[73] – utilitário de navegador (*browser*) *offline*, de código aberto e de uso para cópia e preservação de *websites*.

[72] Disponível em: https://sbforge.org/display/NAS/NetarchiveSuite. Acesso em: 13 abr. 2022.
[73] Disponível em: https://www.httrack.com/. Acesso em: 7 jun. 2023.

Figura 10 – Página inicial da ferramenta *HTTrack*.

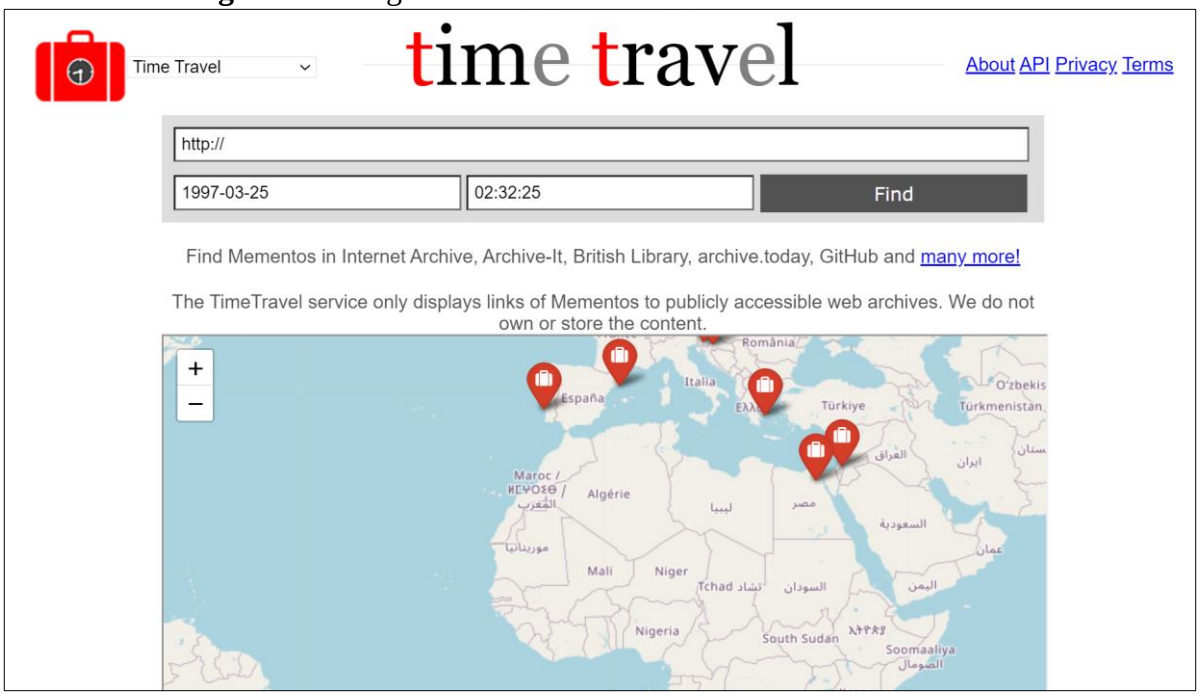
Fonte: Roche (c2022).

Esta ferramenta de captura permite baixar um *website* para um diretório local em seu computador, gerar uma hierarquia de pastas e salvar o conteúdo que espelha a estrutura original do *site*. Sendo configurável e tendo um sistema de ajuda integrado, o *HTTrack* organiza a estrutura de *links* do *site* original. Assim, ao abrir uma página do *site* "espelhado" (*mirrored*), ele permite navegar *link* por *link* como visualizamos o *site online*, mas não gera WARCs e não permite a entrada de metadados descritivos e a extração de metadados técnicos e de preservação do conteúdo espelhado.

- *Memento Time Travel*[74] – ferramenta de código aberto para acessar e visualizar versões de páginas *Web* que existiam em algum momento no passado.

[74] Disponível em: https://timetravel.mementoweb.org/. Acesso em: 7 jun. 2023.

Figura 11 – Página inicial da ferramenta *Memento Time Travel*.



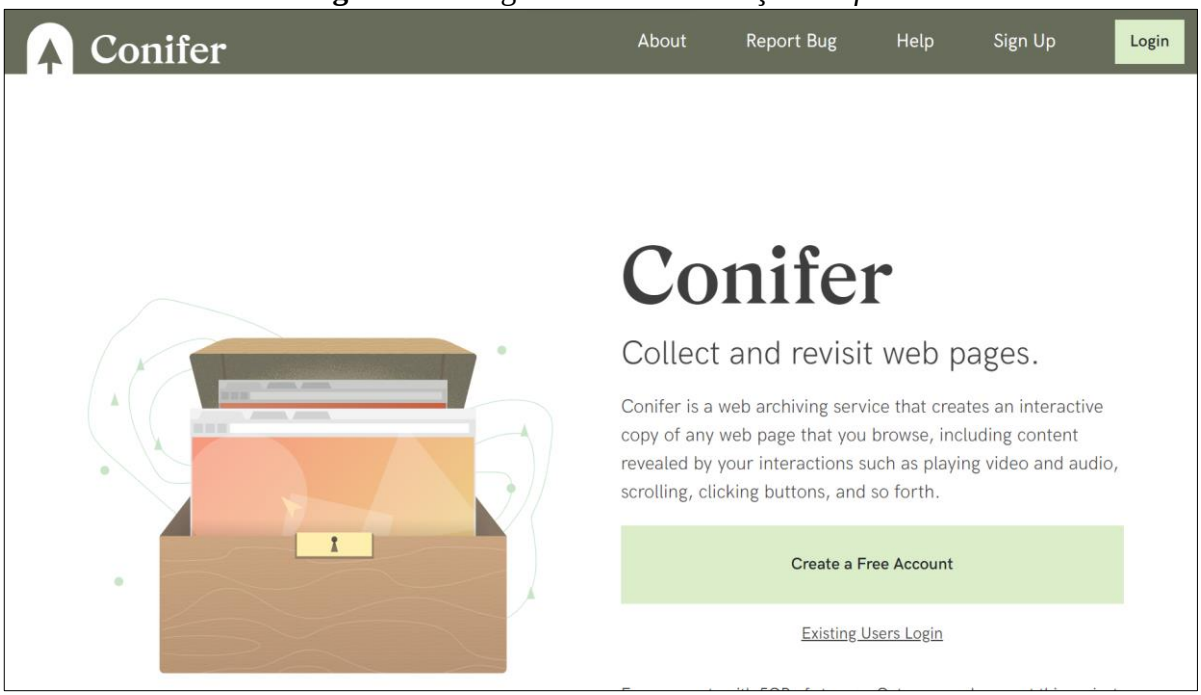
Fonte: *Memento Project* ([2023]).

As versões anteriores de páginas *Web* são chamadas de *Mementos* e, ao passo que inserimos o endereço da *Web* de uma página (que existe ou que já desapareceu) e uma hora no passado, os *Mementos* dessa página que datam da época de sua escolha são localizados pelo *Time Travel Find* em arquivos da *Web*. Vários componentes que formam o *Memento* da página (o HTML, imagens etc.) são extraídos destes sistemas compatíveis com o protocolo *Memento Request for Comments* (RFC) 7089[75] que se centra na navegação do conteúdo arquivado por meio da URL e data de captura. Falta-lhe a função de colher metadados descritivos para melhorar a descoberta.

- *Conifer*[76] – serviço gratuito que captura a ordem de navegação de uma série de páginas *Web* de qualquer *website*, preservando a experiência do usuário.

[75] Disponível em: https://datatracker.ietf.org/doc/html/rfc7089. Acesso em: 7 jun. 2023.
[76] Disponível em: https://conifer.rhizome.org/. Acesso em: 7 jun. 2023.

Figura 12 – Página inicial do serviço *Conifer*.



Fonte: *Rhizome.org* ([2023?]).

> Este *software* de código aberto (chamado antes de *Webrecorder*) cria gravações de alta fidelidade, interativas e contextuais de mídia social e outros conteúdos dinâmicos, como *JavaScript* complexo e vídeos integrados, usando-se de si mesmo para capturar e reproduzir o *site* (abordagem tida como arquivamento simétrico da *Web*). Os metadados descritivos (a data/hora da captura, URLs que o usuário visitou numa sessão de gravação etc.) são gerados de forma automática durante o arquivamento e incorporados num arquivo WARC para *download*.

Iniciativas em bibliotecas nacionais e universitárias, arquivos nacionais, órgãos de pesquisa e empresas

Com base em Arquivo.pt (2021b), *Bibliothèque Nationale de France* (c2022a, c2022b, c2022c), *California Digital Library* (2021), *Columbia University Libraries* (2021b), *Harvard Library* (c2022), *International Internet Preservation Consortium* (c2021a), *Library of Congress* ([2022b], [2022d]), *National Library of New Zealand* ([2022?a], [2022?b], [2022?c]), Pennock (c2013), Rockembach (2018), Rockembach e Pavão (2018), *Stanford Libraries* ([c2022?b]), *The National Archives* ([2022?a], [2022b], [2022?c], [2021]) e *UK Web Archive* ([2022?]), constatamos diversas iniciativas ao redor do mundo de preservação digital de informações da *Web*, que se apoiam na legislação de cada país ou em permissões dos

proprietários de *sites* e no uso das tecnologias de serviços comerciais de arquivamento da *Web*, com alguns exemplos abrangendo coleções pesquisáveis de *sites* arquivados sobre conteúdos criados no Brasil, como:

- *Bibliothèque Nationale de France* (BnF)[77] – desde 2006 o arquivamento da *Web* faz parte da missão de depósito legal[78] da Biblioteca Nacional da França ou BnF, mas o seu programa de arquivamento iniciou em 2002 e as primeiras coleções contaram com a contribuição do *Internet Archive*.

Figura 13 – Página inicial da BnF sobre o arquivo da *Internet* francesa.

Fonte: *Bibliothèque Nationale de France* (c2023).

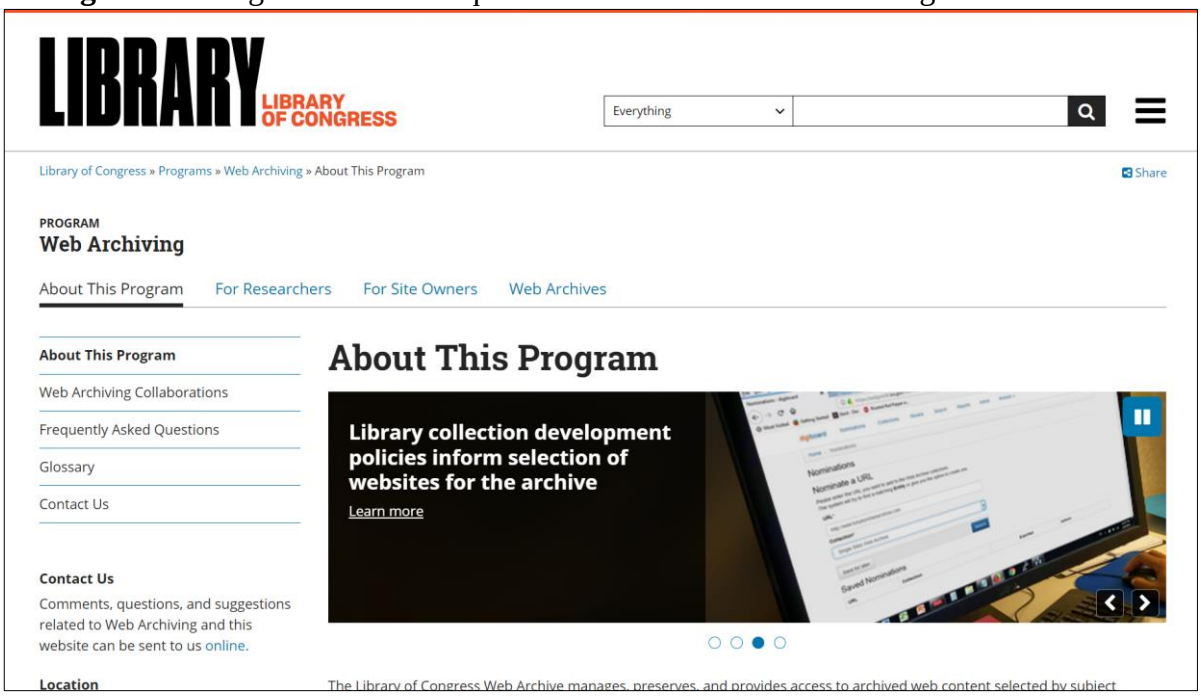
[77] Disponível em: https://netpreserve.org/about-us/members/biblioth%C3%A8que-nationale-de-france-national-library-france/. Acesso em: 7 jun. 2023.

[78] Regido pela Lei do Patrimônio Francês (do francês *code du patrimoine*) e fundado em 1537 por François I, o depósito legal trata-se da obrigação de todo editor, impressor, produtor e importador de depositar cada documento que publica, produz, imprime, distribui ou importa na França, junto da entidade autorizada a receber o depósito em função da natureza do documento a ser coletado, preservado e fornecido consulta (isto é, livros, periódicos, fotografias, músicas, *softwares*, banco de dados, *websites* etc.), de modo a compor uma coleção de referência como elemento vital da memória coletiva e do patrimônio cultural difundido no país. Também definido pela Lei nº 2006-961, de 2006, ou lei relativa aos direitos de autor e direitos conexos na sociedade da informação (do francês *loi relative au droit d'auteur et aux droits voisins dans la société de l'information*), tida como lei DADVSI, o depósito legal da *Web* francesa não pretende ser exaustivo, mas representativo, e é operado diretamente pela BnF, sem que os produtores ou os proprietários de conteúdo tenham que realizar qualquer ação. Para garantir a representatividade dos arquivos coletados, a BnF combina três métodos de coleta, a saber: a coleta "ampla", que consiste na coleta, uma vez por ano, de uma vasta amostra da *Web* francesa; as coletas "direcionadas", mais assíduas, que são baseadas em uma seleção de *sites* escolhidos pela BnF; e a vigilância diária, que permite coletar "notícias efêmeras" para documentar as repercussões de um evento na *Web* e nas mídias sociais, tal como o incêndio da Catedral de Notre-Dame de Paris (BIBLIOTHÈQUE NATIONALE DE FRANCE, c2022b; c2022c; LEROY-TERQUEM, 2019).

Responsável pela coleta, preservação e disposição dos arquivos da *Internet* francesa, a instituição oferece aos usuários/pesquisadores a chance de consultar e acessar em suas instalações páginas *Web* arquivadas que datam de 1996 até hoje. As coletas da *Web* referem-se ao domínio francês (*sites* registrados em *.fr*; sob uma extensão ligada ao território nacional – *.re*, *.bzh* etc.–; ou sob uma extensão genérica – *.com*, *.org* etc. – criados no país ou que seu autor aí resida), não visando ser exaustivas, mas se baseiam no princípio da representatividade. Por exemplo, há a coleta de conteúdo *Web* alusivo à Covid-19 (https://www.bnf.fr/fr/la-bnf-archive-le-web-du-coronavirus)[79] em que documenta a evolução, a propagação e o impacto geral da pandemia na França.

- *Library of Congress Web Archive*[80] – iniciado em 2000, este programa da Biblioteca do Congresso americano destina-se a gerenciar, preservar e fornecer acesso a conteúdo da *Web* arquivado e selecionado para que esteja disponível hoje e no futuro ao Congresso dos Estados Unidos, à pesquisadores interessados e o público em geral.

Figura 14 – Página inicial do arquivo da *Web* da Biblioteca do Congresso americano.



Fonte: *Library of Congress* ([2023a]).

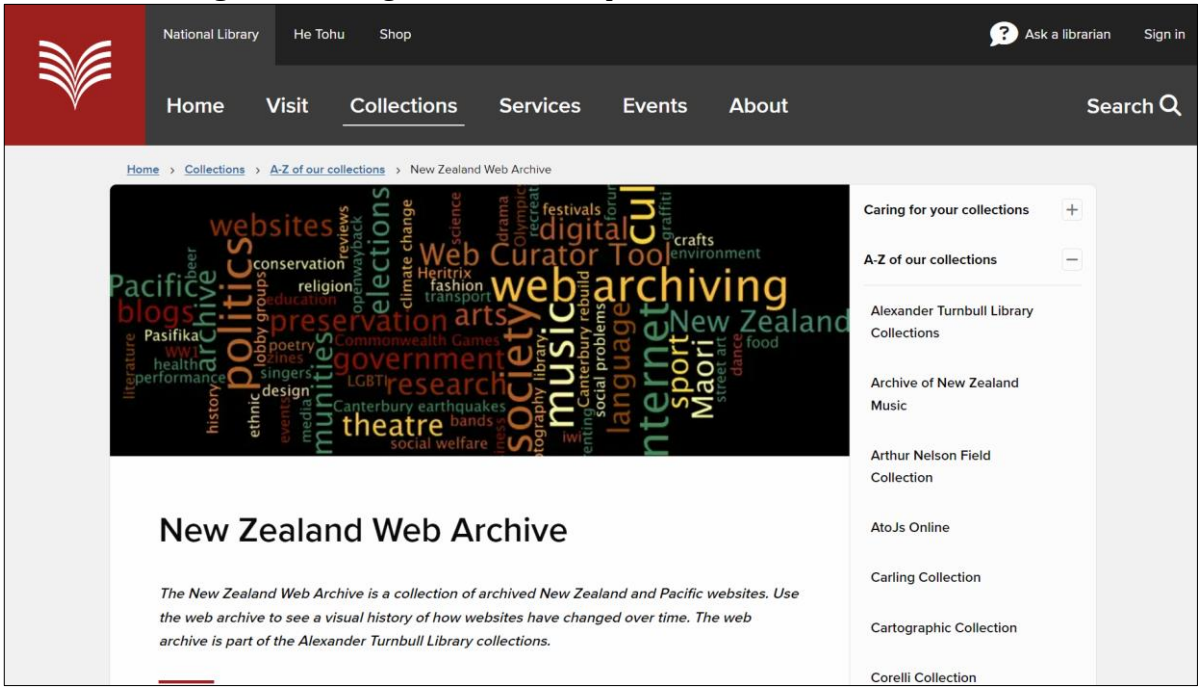
Os arquivos de *websites* coletados e seus conteúdos em vários formatos não só podem ser pesquisados e navegados (https://www.loc.gov/web-archives/) e são dispostos em coleções temáticas e por eventos, e contêm *sites* que documentam uma série de

[79] Disponível em: https://netpreserveblog.wordpress.com/2020/05/27/the-french-coronavirus-covid-19-web-archive-collection/. Acesso em: 7 jun. 2023.
[80] Disponível em: https://www.loc.gov/programs/web-archiving/about-this-program/. Acesso em: 7 jun. 2023.

organizações americanas e internacionais representando diferentes áreas temáticas e assuntos. Como exemplos de coleções com arquivos *Web* criadas, temos dos ataques de 11 de setembro de 2001, das eleições do Afeganistão, e da literatura de cordel e das eleições presidenciais do Brasil.

- *New Zealand Web Archive*[81] – integrando as coleções da Biblioteca Alexander Turnbull na Biblioteca Nacional da Nova Zelândia, este arquivo da *Web* iniciado em 1999 é uma coleção de *sites* arquivados da Nova Zelândia e do Pacífico (*sites* de domínio *.nz*, *.com*, *.org* etc.) para fins de pesquisa e preservação de longo prazo.

Figura 15 – Página inicial do arquivo da *Web* da Nova Zelândia.



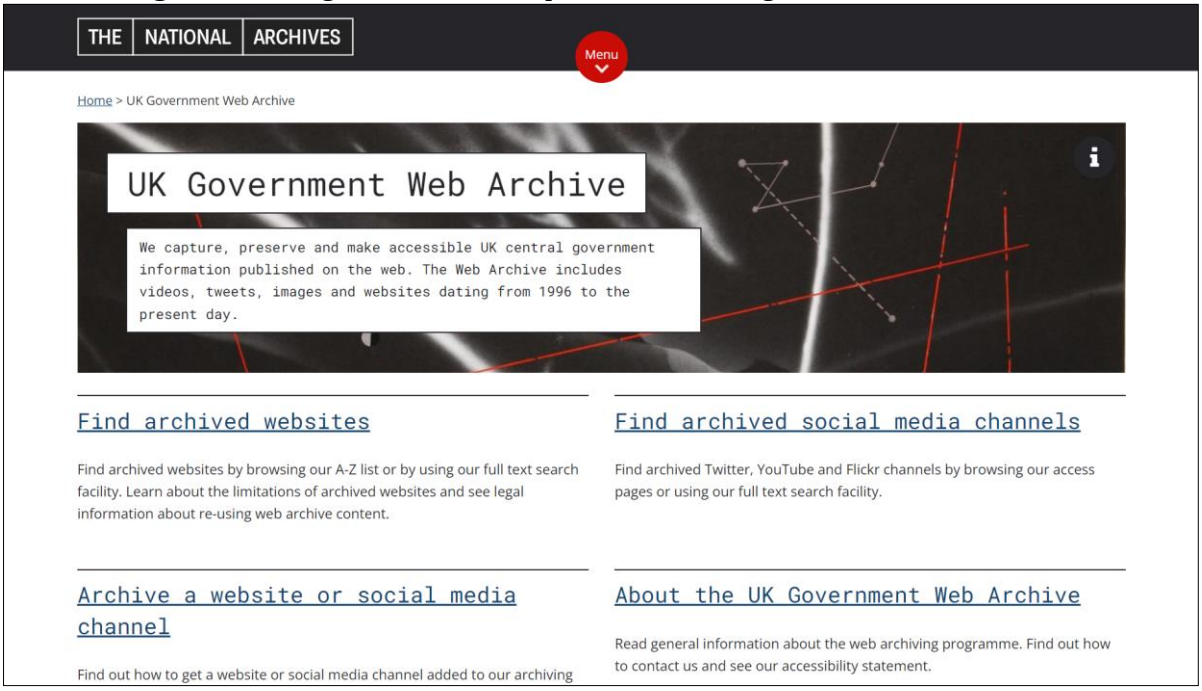
Fonte: *National Library of New Zealand* ([2023?b]).

O arquivo da *Web* da Nova Zelândia tem *sites* selecionados e coletados sobre o país, os neozelandeses e referentes ao Pacífico, que podem ser acessados no catálogo *online* da biblioteca nacional[82]. Com o apoio dos serviços do *Internet Archive*, certas coleções foram criadas neste arquivo da *Web*, tais como *sites* arquivados por e para a cultura e a história dos povos indígenas "*Māori*"; e *sites* de candidatos e partidos, grupos de *lobby*, *blogs* políticos arquivados etc. durante as "*New Zealand General Elections*" (http://t.ly/HH8h) de 1999 em diante.

[81] Disponível em: https://natlib.govt.nz/collections/a-z/new-zealand-web-archive. Acesso em: 7 jun. 2023.
[82] Disponível em: https://natlib-primo.hosted.exlibrisgroup.com/primo-explore/search?vid=NLNZ. Acesso em: 7 jun. 2023.

- *The National Archives* UK[83] – como o arquivo oficial do governo do Reino Unido e da Inglaterra e País de Gales que atua para garantir o futuro dos registros físicos e digitais, o Arquivo Nacional do Reino Unido realiza o arquivamento da *Web* desde 2003 a fim de preservar os *sites* e certas contas de mídia social do governo central (*sites* de domínio *gov.uk, org.uk, .com, .org* etc.) para futuros pesquisadores, historiadores e o público. Os materiais arquivados, o qual incluem páginas *Web* que datam de 1996, são mantidos no UK *Government Web Archive* (UKGWA) e podem ser acessados *online* gratuitamente (https://www.nationalarchives.gov.uk/webarchive/).

Figura 16 – Página inicial do arquivo da *Web* do governo do Reino Unido.



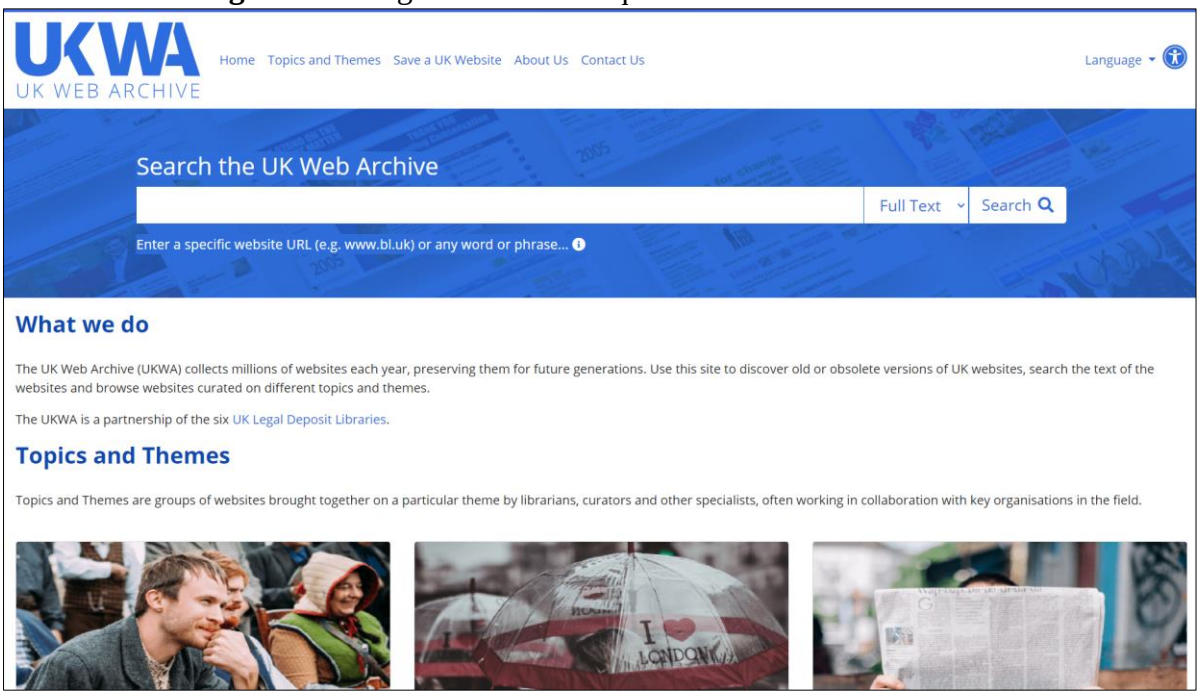
Fonte: *The National Archives* ([2023?a]).

Neste arquivo da *Web* temos *sites* de todos os órgãos do governo central do Reino Unido, suas agências, e órgãos públicos não departamentais como, por exemplo, o *Council for Science and Technology* (CST)[84].

[83] Disponível em: https://www.nationalarchives.gov.uk/about/. Acesso em: 7 jun. 2023.
[84] Disponível em: https://webarchive.nationalarchives.gov.uk/ukgwa/*/http://www.cst.gov.uk/. Acesso em: 7 jun. 2023.

- UKWA – como uma parceria entre as Bibliotecas de Depósito Legal[85] do Reino Unido (isto é, *British Library*, Londres; *National Library of Scotland*, Edimburgo; *Bodleian Library*, *Oxford*; *Cambridge University Library*; *Trinity College*, *Dublin*; e *National Library of Wales*, *Aberystwyth*), o arquivo da *Web* do Reino Unido foi criado em 2004 e objetiva coletar e preservar os *sites* disponíveis publicamente na *Web* aberta do Reino Unido como dar acesso a esses materiais para pesquisadores atuais e futuros.

Figura 17 – Página inicial do arquivo da *Web* do Reino Unido.



Fonte: *UK Web Archive* ([2023]).

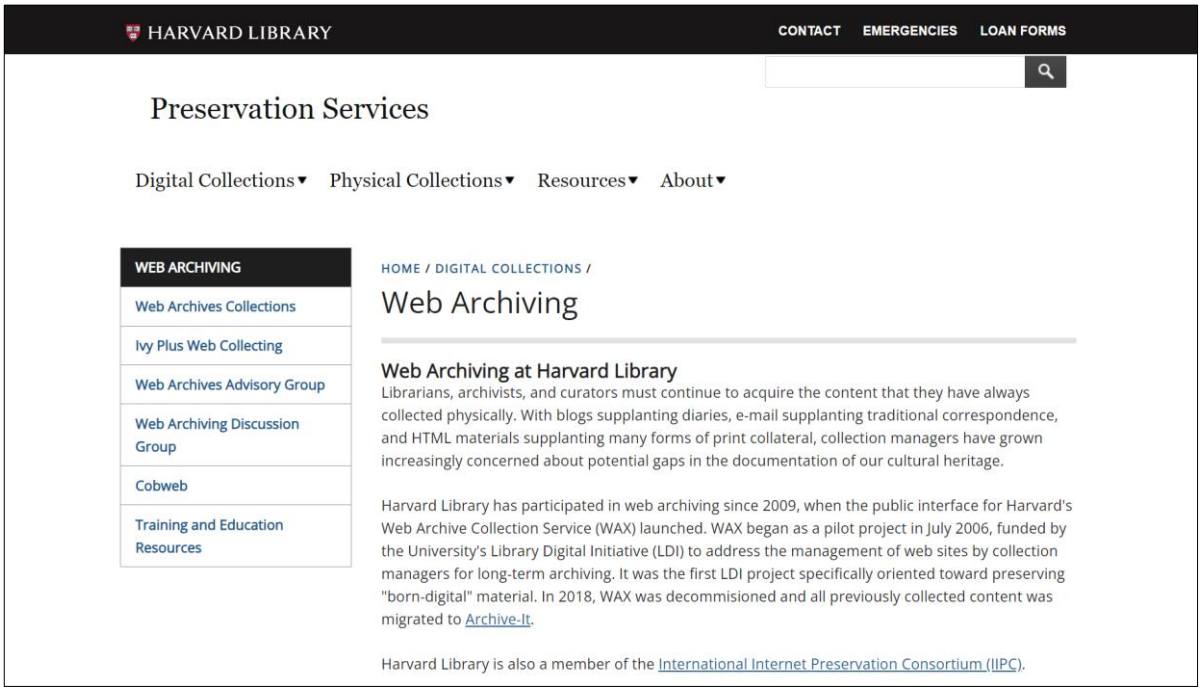
O arquivo possui *sites* arquivados do Reino Unido (*sites* de domínio *.uk*, *.scot*, *.wales*, *.london* etc.) agrupados em tópicos e assuntos (https://www.webarchive.org.uk/en/ukwa/collection), como as comunidades negras,

[85] Existente na legislação inglesa desde 1662, o depósito legal exige que as editoras forneçam uma cópia de cada obra que publicam no Reino Unido à Biblioteca Britânica, e como solicitado por demais bibliotecas de depósito legal, garantindo que a produção publicada do país, em formato impresso e digital, seja coletada sistematicamente, preservada para o uso das gerações futuras, e descoberta e acessada pelos usuários dentro das bibliotecas, tornando-se parte do patrimônio britânico, de acordo com *British Library* ([c2022]). A Lei de Bibliotecas de Depósito Legal (do inglês *Legal Deposit Libraries Act*), de 2003, foi expandida por uma legislação secundária, os Regulamentos de Bibliotecas de Depósito Legal (obras não impressas) (do inglês *Legal Deposit Libraries - Non-Print Works - Regulations*), que defini o depósito de obras publicadas *online*, como *websites*, *blogs*, periódicos eletrônicos etc. Desde que esses regulamentos entraram em vigor em 2013, as bibliotecas de depósito legal estão aptas a coletar e a arquivar qualquer *site* baseado no Reino Unido (por exemplo, *sites* identificados como hospedados num servidor localizado fisicamente no país; *sites* com um endereço postal no país, ou que o seu proprietário confirme residência ou local de negócios no país), com a ressalva de que este material apenas esteja disponível para visualização nas instalações das bibliotecas, a menos que se tenha uma permissão adicional do editor do *site* para tornar o conteúdo mais amplamente disponível. Aliás, para criar coleções de *sites* temáticos abrangentes, ocasionalmente é solicitado permissão do proprietário do *website* para se arquivar *sites* de fora do Reino Unido (UK WEB ARCHIVE, [2022?]).

asiáticas, latinas e LGBTQ+ no Reino Unido; a epidemia do vírus Ebola na África Ocidental; o *Brexit*[86]; o surto do vírus Zika na América do Sul, que iniciou em 2015 no Brasil; a história da Biblioteconomia no Reino Unido, e outros.

- *Harvard Library*[87] – como uma organização central das bibliotecas da Universidade de *Harvard* nos Estados Unidos, a biblioteca de *Harvard* participa do arquivamento da *Web* desde 2009 com o lançamento da interface pública da sua ferramenta de coleta da *Web*, o "*Web Archive Collection Service* (WAX)". Em 2018, o WAX foi desativado e todo o conteúdo anteriormente capturado foi migrado para o serviço *Archive-It*.

Figura 18 – Página inicial sobre o arquivamento da *Web* na biblioteca de *Harvard* dos Estados Unidos.



Fonte: *Harvard Library* (c2022).

Dos exemplos de coleções desenvolvidas, a "*Capturing Women's Voices on the Web*" (https://archive-it.org/collections/8238) é uma coleção da Biblioteca *Schlesinger* com foco em capturar as vozes das mulheres cujos pontos de vista não podem ser encontrados em outros locais, e também documentar o uso de *blogs* por mulheres americanas no início do século XXI que se preocupam com questões de saúde, refletem seu envolvimento com a política etc.

[86] *Brexit*, abreviação para "*British exit*", é uma expressão que se refere a decisão do Reino Unido de deixar a União Europeia a partir de 31 de janeiro de 2020 após a realização de um referendo em junho de 2016 (BREXIT, 2022).
[87] Disponível em: https://library.harvard.edu/about/about-harvard-library. Acesso em: 7 jun. 2023.

- *Columbia University Libraries* (CUL)[88] – através do uso do serviço *Archive-It*, as bibliotecas da Universidade de *Columbia* nos Estados Unidos têm capturado e arquivado *sites* de domínio da instituição, incluindo todos os *sites* com um endereço *Web* "*columbia.edu*" e, além disto, *sites* sem este endereço, como os de grupos e publicações de alunos.

Figura 19 – Página inicial dos arquivos da *Web* nas bibliotecas da Universidade de *Columbia* nos Estados Unidos.

Fonte: *Columbia University Libraries* ([c2023?]).

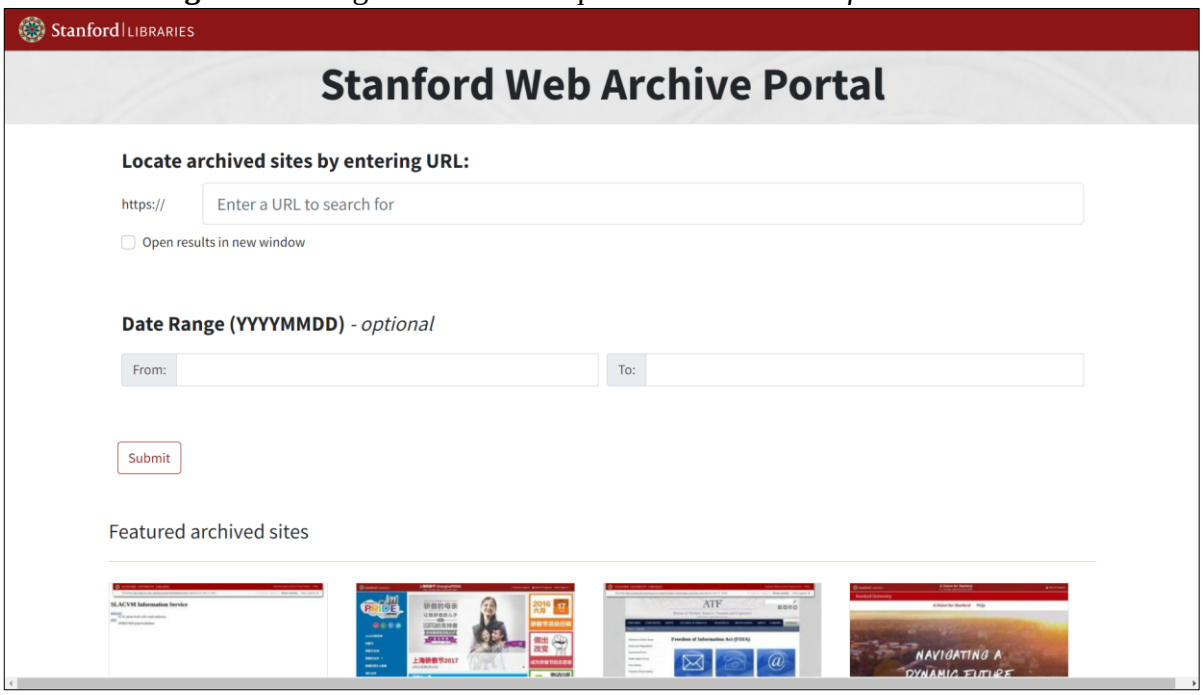
Uma das coleções desenvolvidas no *Archive-It* são os "*University Archives*" (https://archive-it.org/collections/1914), o qual mantém um registro da presença da universidade na *Web* desde 2010. Esta coleção objetiva coletar, descrever, preservar e, se apropriado, pôr à disposição de administradores, pesquisadores e público geral, registros da Universidade de *Columbia* que documentem a sua evolução (ou seja, as contribuições para o ensino e a pesquisa; o desenvolvimento de cursos etc.) e preservam a sua memória institucional.

- *Stanford University Libraries* (SUL)[89] – como o sistema de bibliotecas da Universidade *Stanford* nos Estados Unidos, as bibliotecas de *Stanford* estão incluídas em projetos de arquivamento da *Web* desde 2007. A SUL identifica conteúdo *Web* valioso, armazena

[88] Disponível em: https://archive-it.org/home/Columbia. Acesso em: 7 jun. 2023.
[89] Disponível em: https://netpreserve.org/about-us/members/stanford-university-libraries/. Acesso em: 7 jun. 2023.

os arquivos no *Stanford Digital Repository*[90], fornece descobertas pelo *SearchWorks*[91] e permite a navegação por uma instância local da *Wayback Machine*[92], sendo que uma coleção pesquisável de *sites* arquivados pela universidade está no *Stanford Web Archive Portal* (SWAP)[93].

Figura 20 – Página inicial do arquivo da *Web* de *Stanford* ou SWAP.



Fonte: *Stanford University* ([2023?d]).

Dos exemplos de coleções nas quais a SUL está atuando, a "*Stanford University Website Collection*" (https://archive-it.org/collections/5591) no *Archive-It* é uma coleção de *websites* de domínio da universidade ("*stanford.edu*"), capturados pelos *University Archives*[94], que documenta a sua história administrativa, cultural, social etc.

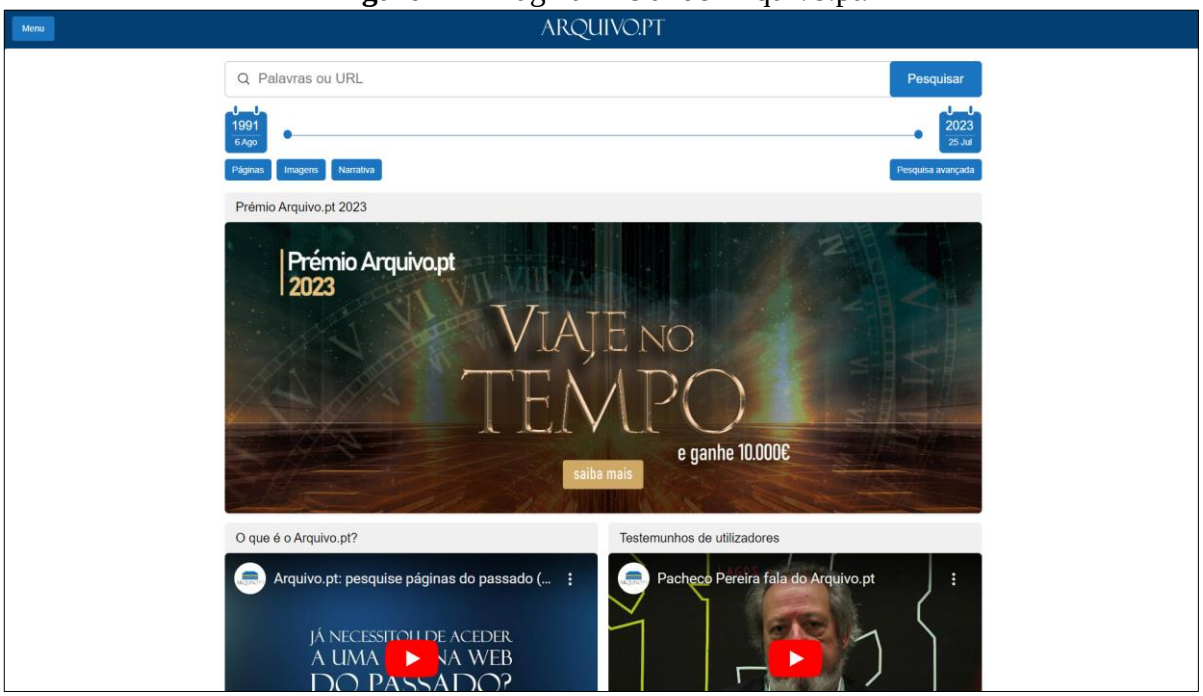
- *California Digital Library* (CDL)[95] – como uma colaboração entre as bibliotecas da Universidade da Califórnia[96] nos Estados Unidos e outros parceiros que consisti numa das maiores bibliotecas de pesquisa digital do mundo, a CDL participa do arquivamento da *Web* desde 2003 com o projeto "*Web-at-Risk*"[97] que resultou na operação do "*Web*

[90] Disponível em: https://library.stanford.edu/research/stanford-digital-repository. Acesso em: 7 jun. 2023.
[91] Disponível em: https://searchworks.stanford.edu/, Acesso em: 7 jun. 2023.
[92] Disponível em: https://web.archive.org/. Acesso em: 7 jun. 2023.
[93] Disponível em: https://swap.stanford.edu/. Acesso em: 7 jun. 2023.
[94] Disponível em: https://library.stanford.edu/spc/university-archives/about-archives. Acesso em: 7 jun. 2023.
[95] Disponível em: https://cdlib.org/about/. Acesso em: 7 jun. 2023.
[96] Disponível em: https://libraries.universityofcalifornia.edu/. Acesso em: 7 jun. 2023.
[97] Disponível em: https://cdlib.org/cdlinfo/2004/10/14/cdl-receives-24-million-library-of-congress-grant/. Acesso em: 7 jun. 2023.

Archiving Service (WAS)"[98] para permitir que curadores da Universidade da Califórnia e demais instituições parceiras construíssem, preservassem e publicassem arquivos da *Web*. Em 2015, a CDL fez a transição de seus clientes do WAS para o serviço *Archive-It*. Dos exemplos de projetos criados, o "*Archive of the California Government Domain, CA.gov*" (https://archive-it.org/collections/5763) é uma coleção que preserva o acesso à *sites* de agências estaduais da Califórnia com comunicados à imprensa, relatórios etc.

- Arquivo.pt.[99] – gerido pela Fundação para a Ciência e a Tecnologia (FCT) do Ministério da Educação e Ciência de Portugal, o Arquivo.pt é uma infraestrutura que, através de coletas exaustivas da *Web*, visa a preservação da informação na *Web* portuguesa – que em pouco tempo não se encontrará disponível e se perderá fatalmente – a fim de que o saber nela contido esteja acessível às futuras gerações para usos educativos, científicos e de investigação.

Figura 21 – Página inicial do Arquivo.pt.



Fonte: Arquivo.pt (2023).

Iniciado em 2008, o Arquivo.pt permite a pesquisa e o acesso gratuito a páginas da *Web* arquivadas desde 1996, sendo que até 2007 o conteúdo foi adquirido via o *Internet Archive*. Apesar de estar fora do seu escopo (isto é, *websites* de domínio .pt), o

[98] Disponível em: http://webarchives.cdlib.org/. Acesso em: 7 jun. 2023.
[99] Disponível em: https://sobre.arquivo.pt/pt/ajuda/o-que-e-o-arquivo-pt/. Acesso em: 12 set. 2021.

Arquivo.pt preservou de forma não exaustiva algumas páginas de *sites* do Brasil (https://arquivo.pt/wayback/20210624211228/https://www.gov.br/pt-br, por exemplo).

- Coca-Cola *Web Archive* – criado para capturar e preservar na sua forma original os *sites* corporativos da Coca-Cola e as páginas em mídias sociais, como *Facebook*, *Twitter* e *blogs*, o arquivo da *Web* da Coca-Cola tem o objetivo de ser o mais abrangente possível incluindo a integridade e a funcionalidade dos *websites* coletados. Os primeiros esforços contaram com a simples captura de arquivos planos, o *Internet Archive*, as soluções da *Microsoft* etc., que não foram suficientes às exigências da empresa. Em 2009, a Coca-Cola passou a colaborar com a *Hanzo Archives* e, hoje, usa o seu serviço comercial de arquivamento da *Web* para uma captura mais completa dos *sites*. O arquivo é acessível só aos funcionários e em alguns equipamentos, sendo que a coleção contém milhões de páginas *Web* e informações acerca de eventos dos quais a companhia foi patrocinadora.

Primeiramente, na análise das experiências levantadas de preservação e arquivamento da *Web*, é fato que os arquivos da *Web* selecionam, coletam em massa e preservam conteúdos da *Web* para o seu acesso e pesquisa em longo prazo, garantindo que uma parte dos *sites*, *blogs* e dados de mídias sociais na *Internet* sejam arquivados em coleções específicas ou abrangentes, formando um todo patrimonial significativo e representativo sobre tópicos e eventos políticos, científicos etc. de interesse para nações, organizações ou indivíduos. Assim, conforme descrito no *UK Web Archive* ([2022?]), páginas da *Web* de *sites* são coletadas em um certo momento no tempo por um *software* e visam refletir da forma mais completa possível o comportamento e a aparência do *site* na *Internet* naquele momento. Também, as iniciativas identificadas originam-se de bibliotecas nacionais e universitárias, arquivos nacionais e empresas com colaborações entre instituições de diferentes países ou de uma mesma nação, como no caso do arquivamento da *Web* no Reino Unido o qual conta com *The National Archives* UK, o UKGWA e o UKWA.

Para mais, em cada arquivo da *Web* temos políticas de seleção e de coleta automatizada por domínios eletrônicos de instituições, governos, territórios etc., além de políticas de acesso e uso aos materiais arquivados que preveem o respeito as permissões de proprietários de direitos autorais dos *sites* a serem preservados e às leis de privacidade de dados. Na Biblioteca Nacional da Nova Zelândia do *New Zealand Web Archive*, as obras digitais fornecidas para depósito legal (incluindo livros, *sites* etc.) e com restrições de acesso à sua publicação (por exemplo, porque estão disponíveis comercialmente) são acessadas apenas *in loco* por pesquisadores em salas de leitura e não podem ser baixadas, impressas, enviadas por *e-mail* ou copiadas por seus usuários (NATIONAL LIBRARY OF NEW ZEALAND, [2022c?]). De outra maneira, no UKWA, para criar coleções de *sites* temáticos abrangentes, ocasionalmente este arquivo da *Web* solicita a

permissão para arquivar *sites* de fora do Reino Unido ao proprietário do *site*; além do mais, esta iniciativa coleta somente páginas da *Web* que estão disponíveis gratuitamente e abertamente na *Internet*, e não coleta conteúdo protegido por senha (a menos que uma permissão especial tenha sido dada) ou páginas em servidores seguros, tal como *e-mail* (UK WEB ARCHIVE, [2022?]).

Em segundo lugar, algumas iniciativas de arquivamento da *Web*, tais como da BnF, do *New Zealand Web Archive* com a *National Library of New Zealand* (*Te Puna Mātauranga o Aotearoa*) *Act* 2003[100], e do UKWA, se baseiam na legislação de depósito legal de seus países para fazer a coleta, preservação para uso das gerações futuras e disponibilização de publicações digitais como parte da memória de uma nação constituindo, segundo *Bibliothèque Nationale de France* (c2022b), numa fonte para complementar as coleções impressas nas bibliotecas. Por sua vez, outras experiências, como na Biblioteca Nacional da Nova Zelândia do *New Zealand Web Archive* (NATIONAL LIBRARY OF NEW ZEALAND, [2022b?]), apoiam-se no dever social das bibliotecas nacionais de preservar a história social e cultural de um país, seja na forma de livros, jornais ou, mesmo, de *sites*, *blogs* etc. Estas iniciativas ainda têm uma forte contribuição dos serviços pagos do *Internet Archive*, o que permite ampliar as capacidades de arquivamento da *Web* nas instituições, mas traz igualmente uma grande dependência em seus serviços que há pouco tempo tem passado por problemas judiciais acerca de violação de direitos autorais[101][102].

Por último, embora não descrito anteriormente, algumas iniciativas sinalizam restrições de acesso e parâmetros de frequência e profundidade de coleta para os seus conteúdos da *Web* selecionados. Na BnF, o acesso as páginas *Web* arquivadas ocorre só pessoalmente em salas de leitura da biblioteca, e a frequência e profundidade de coleta (total/parcial de um *site*) em suas coleções são adaptadas segundo a natureza e o ritmo de atualizações do *site* para manter versões sucessivas representativas de sua evolução, como várias vezes ao dia para mídias sociais e uma vez ao dia para *sites* de imprensa (BIBLIOTHÈQUE NATIONALE DE FRANCE, c2022a, c2022c; ROCKEMBACH, 2018). Já no *Library of Congress Web Archive*, o acesso completo aos conteúdos arquivados ocorre apenas nas instalações da Biblioteca do Congresso americano (porém é possível o acesso *online* fora da biblioteca para navegar por registros descritivos), e a

[100] A obrigação de depósito legal na *National Library of New Zealand (Te Puna Mātauranga o Aotearoa) Act* 2003 (https://legislation.govt.nz/act/public/2003/0019/latest/DLM191962.html) permite que a Biblioteca Nacional da Nova Zelândia colete e preserve para a sua comunidade e as gerações futuras as publicações neozelandesas (isto é, qualquer trabalho físico e digital publicado na Nova Zelândia, incluindo obras publicadas por neozelandeses em *websites* e hospedados em plataformas de auto publicação no exterior) como, por exemplo, livros, *sites*, músicas gravadas, periódicos, mapas, partituras, dentre outros (NATIONAL LIBRARY OF NEW ZEALAND, [2022c?]).
[101] Disponível em: https://blog.archive.org/2020/06/10/temporary-national-emergency-library-to-close-2-weeks-early-returning-to-traditional-controlled-digital-lending/. Acesso em: 7 jun. 2023.
[102] Disponível em: https://www.tecmundo.com.br/voxel/236051-nintendo-derruba-guia-super-mario-64-infracao-copyright.htm. Acesso em: 7 jun. 2023.

maioria dos *sites* em suas coleções são arquivados mais de uma vez a fim de documentar as mudanças destes no tempo, do qual a frequência da coleta varia segundo o *site* e as decisões tomadas quando o *site* é indicado para coleta (LIBRARY OF CONGRESS, [2022b], [2022d]).

Outras iniciativas também indicam tanto as limitações técnicas no arquivamento da *Web* realizado por elas que afetam a visualização completa e correta dos arquivos como os requisitos que permitem que os *sites* sejam arquivados mais facilmente. No *New Zealand Web Archive*, determinadas formas de navegação não funcionam em seus *sites* arquivados, como caixas de pesquisa (*search boxes*) que exigem um mecanismo de busca e essa funcionalidade não pode ser capturada, e vídeos, músicas e demais arquivos inseridos de *sites* externos (*Youtube*, *Vimeo* etc.) que precisam de um reprodutor de mídia (*media player*) externo não são arquivados; além disto, dentre as orientações para tornar os *sites* preserváveis, estão o uso de padrões da *Web*[103], de diretrizes de acessibilidade[104] e de formatos abertos (NATIONAL LIBRARY OF NEW ZEALAND, [2022a?], [2022b?]). De outra forma, no UKWA, mídia de *streaming*[105], *deep Web*[106] e conteúdo de base de dados que requer a entrada do usuário (*user input*) são incapazes de serem capturados por robôs rastreadores da *Web* (*crawlers*); e entre os itens que ajudam a capturar o conteúdo dos *sites* (HTML, textos, arquivos de áudio e vídeo, *links* etc.), está a criação de um mapa do *site* (*sitemap*)[107] para que seu conteúdo seja rastreado (UK WEB ARCHIVE, [2022?]).

2.1.3 Preservação digital de mídias sociais

Como uma estratégia operacional de preservação digital e um aspecto do arquivamento da *Web*, a preservação das mídias sociais envolve "[...] arquivar conjuntos de dados em formato legível por máquina que serão de grande utilidade para os pesquisadores orientados a dados nas ciências [...]", e "o arquivamento de dados de mídia social como conjuntos de dados, ou como

[103] Disponível: https://www.w3.org/standards/. Acesso em: 7 jun. 2023.
[104] Disponível: https://www.w3.org/WAI/standards-guidelines/. Acesso em: 7 jun. 2023.
[105] *Streaming media* é a "[...] multimídia que é entregue e consumida de forma contínua a partir de uma fonte, com pouco ou nenhum armazenamento intermediário em elementos de rede.", e *streaming* concerne "[...] ao método de entrega do conteúdo, e não ao conteúdo em si." (STREAMING MEDIA, 2022, não paginado, tradução nossa).
[106] Alguns estudos têm explorado a viabilidade da coleta da *deep Web*, como é o caso de Masanès (2002).
[107] *Site map* (ou *sitemap*) é uma lista de páginas de um *site* dentro de um nome de domínio (do inglês *domain name*, que é o endereço único que identifica um *site* da *Internet*), sendo de três tipos: mapas de *site* para planejamento de um *site* por seus *designers*; listagens hierárquicas visíveis para humanos, das páginas de um *site*; e listagens estruturadas para rastreadores da *Web*, tais como os mecanismos de pesquisa (BACA, c2016; SITE MAP, 2022).

coleções de *Big Data*[108], permitirá às instituições de pesquisa e coleta arquivar dados de fonte para pesquisa acadêmica e jornalismo e preservá-los para acesso futuro." (THOMSON, c2016, p. 4, tradução nossa). As mídias sociais (isto é, *blogs*, *sites* de redes sociais – *Facebook*, *Twitter* etc. –, e *sites* de compartilhamento de conteúdo – *YouTube* etc. –), como outros conteúdos *Web*, são um canal relevante para que as organizações se comuniquem oficialmente e interajam com o público; estas interações podem criar registros valiosos, onde para arquivar com êxito dados de mídias sociais a fim de reutilização, estes devem ser não só capturados, mas ainda indexados e armazenados em um ambiente pesquisável, segundo *Digital Preservation Coalition* (2018d).

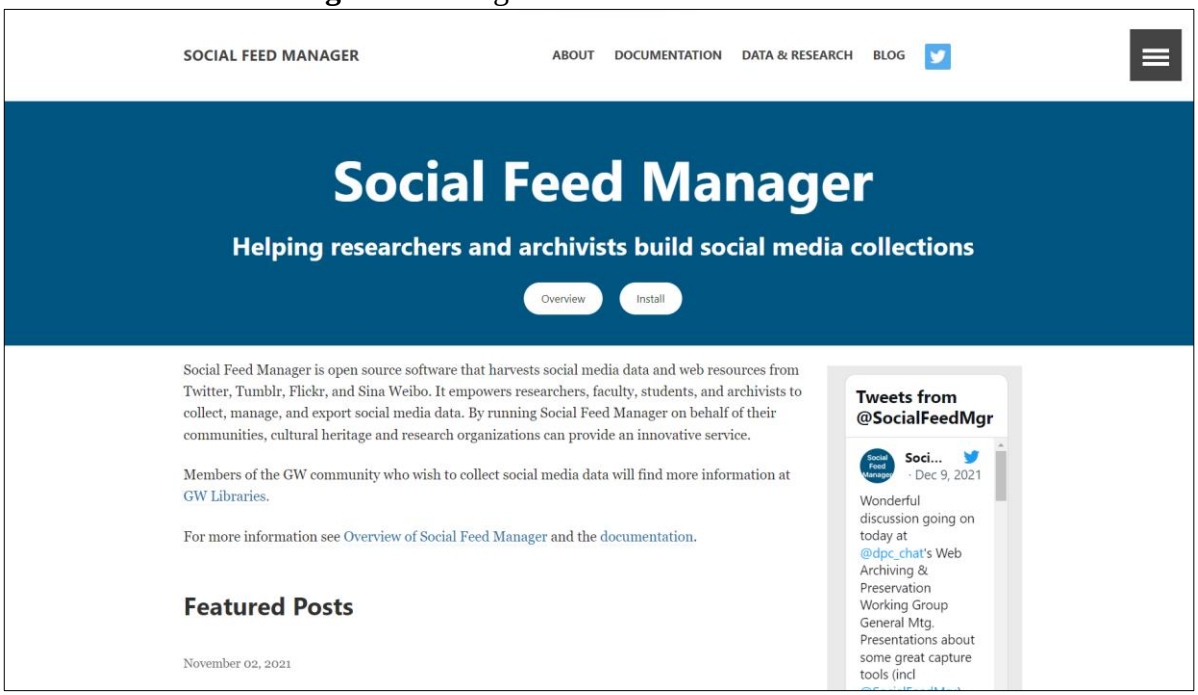
Ferramentas

Em Arcoman (2016), Brooker, Barnett e Cribbin (2016), Burnap *et al.* (2015), *Cardiff University* ([2022]), *Documenting the Now* ([2021]), George Washington *University Libraries* (c2021), Hussain e Vatrapu (2014), *International Internet Preservation Consortium* (c2022c) e Samouelian e Dooley (2018), verificamos determinadas ferramentas[109] ou soluções para a coleta, arquivamento e preservação de conteúdos públicos em plataformas de mídia social, a saber:

- *Social Feed Manager* (SFM)[110] – *software* de código aberto que coleta, gerencia e exportar dados de mídia social e recursos da *Web* do *Twitter*, *Tumblr*, *Flickr* e *Sina Weibo*.

[108] *Big Data* é a área do conhecimento "[...] que trata de maneiras de analisar, extrair sistematicamente informações, ou lidar de outra forma com conjuntos de dados que são muito grandes ou complexos para serem tratados por *softwares* tradicionais de processamento de dados." (BIG DATA, 2022, não paginado, tradução nossa) e que, em concordância com Reis e Sá (2020, p. 247), por intermédio da sua capacidade de relacionar dados, se intenciona a "[...] ser um elemento decisivo para que a empresa obtenha vantagem competitiva ao traduzir o resultado da análise de dados em ajustes na tomada de decisões, criação de novos produtos e adaptação das estratégias em tempo real."
[109] Deve-se levar em conta que grande parte dos *softwares* identificados está em desenvolvimento ativo e os estudos citados ocorreram ao longo de vários anos, assim, pode haver inconsistências na forma como as ferramentas são descritas pelas diferenças entre versões ou, mesmo, pelas constantes atualizações e melhorias de funcionalidades.
[110] Disponível em: https://sfm.readthedocs.io/en/latest/install.html. Acesso em: 7 jun. 2023.

Figura 22 – Página inicial da ferramenta SFM.



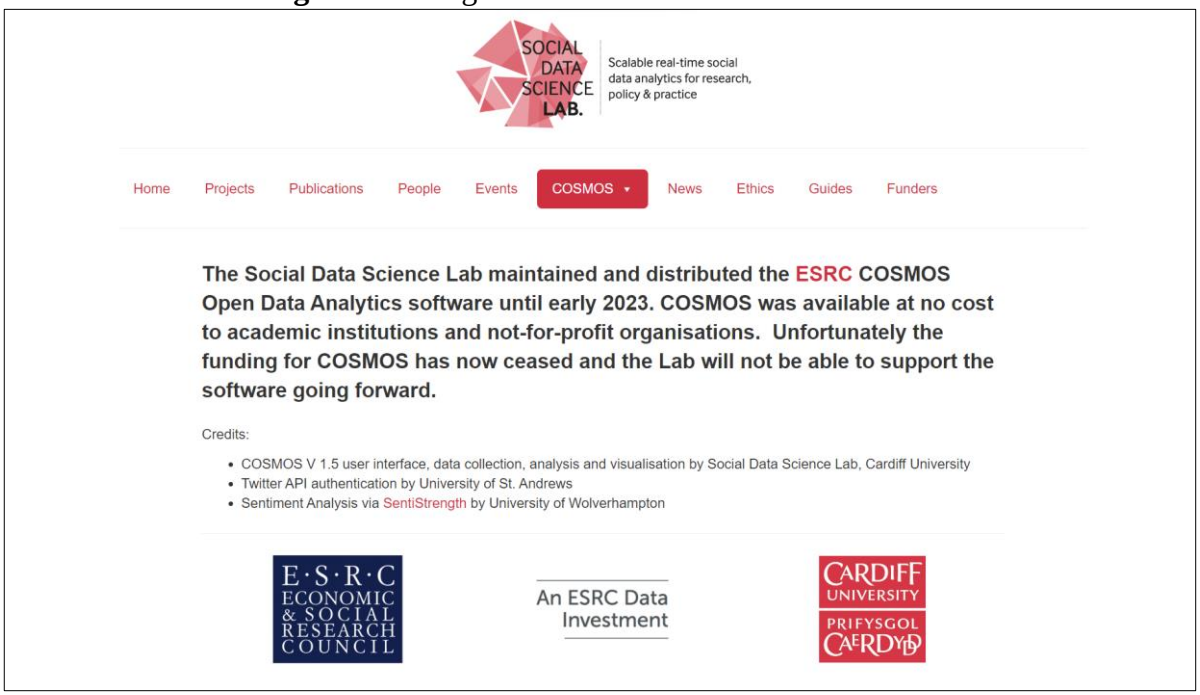
Fonte: George Washington *University Libraries* (c2021).

Se conectando à *Application Programming Interface* (API) públicas de mídia social para coletar dados (e armazená-los em WARC) e usando *Heritrix* para capturar *sites* e mídias incorporadas ou referenciadas nas mídias sociais, a aplicação permite aos usuários criar coleções para pesquisas futuras, incluindo a coleta de dados de mídia social "em risco" em torno de um certo evento ou tema etc. O SFM não é uma ferramenta de preservação ou acesso, assim, não suporta metadados descritivos robustos para uso nesses sistemas, e os metadados sobre postagens e coleta em mídias sociais são gerados automaticamente enquanto que aqueles sobre a seleção de uma coleção devem ser gerados manualmente.

- *Collaborative Online Social Media Observatory Software* (COSMOS)[111] – ferramenta de análise de mídia social que ajuda cientistas sociais a coletar, armazenar, analisar e visualizar grandes conjuntos de dados de mídia social para suas pesquisas.

[111] Disponível em: http://socialdatalab.net/cosmos. Acesso em: 7 jun. 2023.

Figura 23 – Página inicial da ferramenta COSMOS.



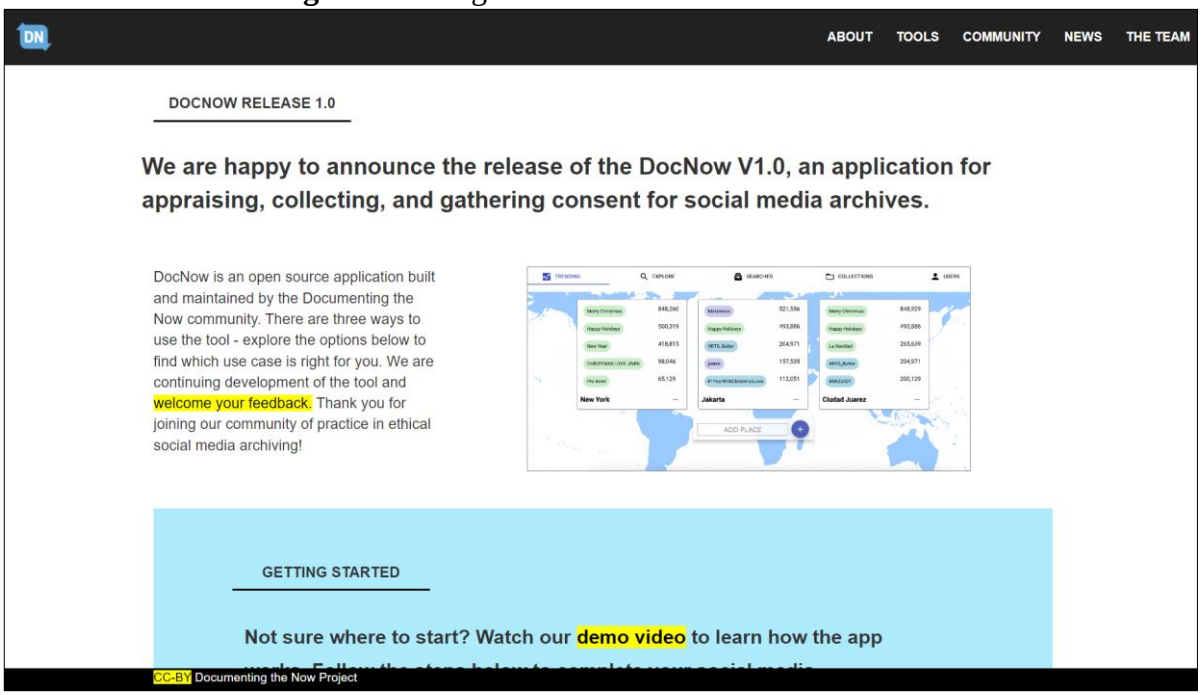
Fonte: *Cardiff University* ([2023?]).

De acesso gratuito para instituições acadêmicas, o COSMOS permite coletar dados do *Twitter* em tempo real via API ou importar grupos de dados coletados antes; e analisar por recursos de filtragem e consulta os dados em nível de *tweet* individual e corpus a partir de certos parâmetros, como detecção de gênero, análise de sentimentos e de frequência de *tweets*, a fim de que pesquisadores entendam mudanças nesses parâmetros (e correlações com humor do público etc.) em torno de um evento ou tópico. Há limitações para a detecção de gênero em que, por exemplo, o *tweeter* (ou pessoa que postou o *tweet*) pode utilizar um nome não-pessoal ou, mesmo, falso que poderia representar falsamente seu gênero.

- *DocNow*[112] – *software* de código aberto desenvolvido pelo projeto *Documenting the Now* nos Estados Unidos para avaliar, coletar e obter autorização para conteúdo do *Twitter*.

[112] Disponível em: https://www.docnow.io/docnow-app/. Acesso em: 7 jun. 2023.

Figura 24 – Página inicial da ferramenta *DocNow*.



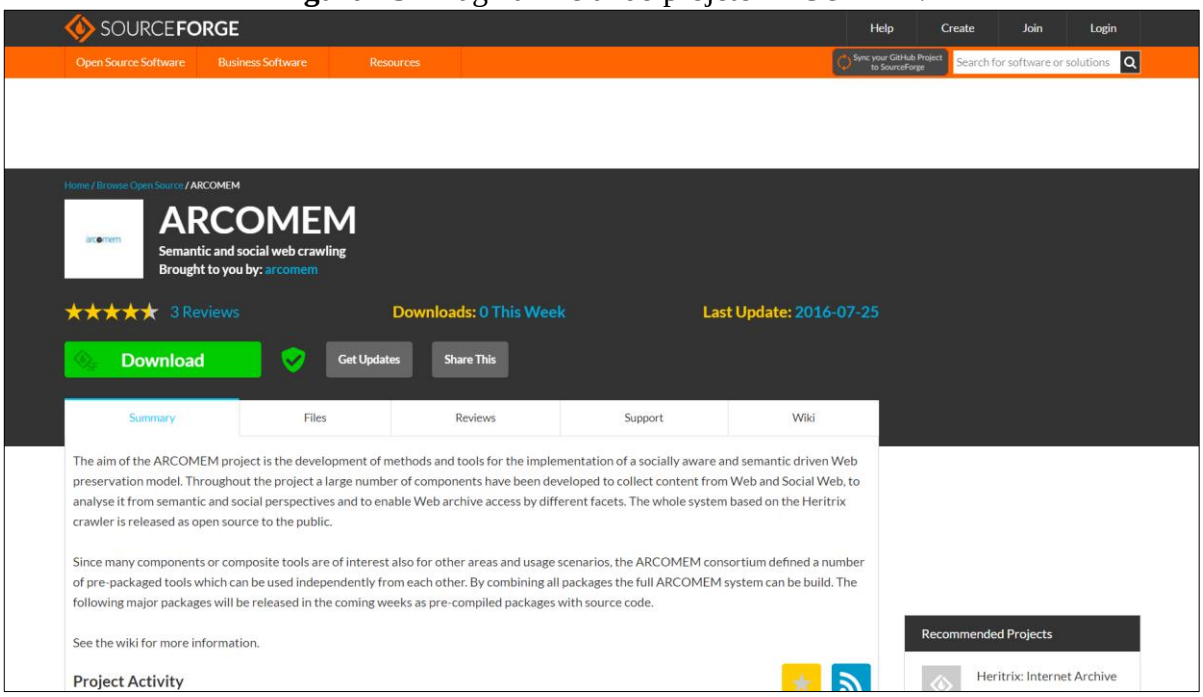
Fonte: *Documenting the Now* ([2023?a]).

O *DocNow* ajuda arquivistas, ativistas e pesquisadores a criar suas próprias coleções de conteúdo do *Twitter* e da *Web*, com atenção especial às práticas éticas de coleta e arquivamento de dados "públicos" de mídia social. Esta ferramenta gratuita permite, por exemplo, testar ao vivo e refinar parâmetros de coleta em *tweets* recentes; explorar conteúdo por usuários, mídia, URLs e *hashtags* relacionadas; obter permissão dos criadores de conteúdos; e compartilhar informações sobre a coleção com o público.

- *Social Data Analytics Tool* (SODATO) – *software* para análise descritiva, preditiva e prescritiva de grandes conjuntos de dados sociais do *Facebook*. O SODATO difere das soluções comerciais existentes visto que tanto implementa procedimentos sistemáticos de coleta de dados, registro e opções de recuperação de erros como pode ser usado como uma ferramenta estratégica para obtenção do registro completo de dados sociais *online* de uma organização na plataforma do *Facebook*. Esta ferramenta permite, por exemplo, buscar e armazenar dados históricos desde o princípio do *Facebook*; fornecer um nível muito alto de transparência na obtenção de dados; determinar a proveniência dos dados; e gerar vários *insights* através das suas técnicas e métodos de análise, tais como análise de sentimentos, de palavras-chave, de desempenho de conteúdo, e de influência social.

- *ARCOMEM*[113] – pacote de ferramentas de código aberto baseado no rastreador *Heritrix* e orientado por semântica para a coleta, rastreamento e arquivamento de conteúdo da *Web* e da *Web* social (*sites* de mídia social, como *Twitter*, *YouTube* ou *Facebook*).

Figura 25 – Página inicial do projeto *ARCOMEM*.



Fonte: Arcoman (2016).

No *ARCOMEM*, há duas aplicações para manipulação de rastreadores (*crawler handling*) e acesso à arquivos *Web*: o *Crawler Cockpit*, que pode gerir a visualização de estatísticas sobre os rastreamentos no arquivamento da *Web*; e a *Search and Retrieval Application* (SARA), que fornece uma interface de usuário intuitiva para a busca e a recuperação de documentos da *Web* arquivados, permitindo que os usuários façam pesquisas por texto completo e consultas semânticas rápidas num arquivo indexado, e os resultados podem ser acessados e refinados em diversas facetas, tais como tópicos, entidades, opiniões etc.

- *Chorus*[114] – solução de coleta de dados e análise visual gratuita projetada para facilitar e possibilitar a pesquisa quanti-qualitativa em Ciências Sociais usando dados do *Twitter*. O pacote *Chorus* é composto por dois programas distintos: o *Tweetcatcher*, que permite aos usuários vasculhar o *Twitter* em busca de dados relevantes, seja por palavras-chave tópicas que aparecem nas conversas do *Twitter* (dados orientados semanticamente) ou

[113] Disponível em: https://sourceforge.net/projects/arcomem/. Acesso em: 7 jun. 2023.
[114] Disponível em: http://chorusanalytics.co.uk/. Acesso em: 7 jun. 2023.

> por identificação de uma rede de usuários do *Twitter* e seguindo suas "vidas no *Twitter*" diárias (dados dirigidos pelo usuário); e o *Tweetvis*, que oferece aos usuários a chance de analisar os dados do *Twitter* ao longo do tempo e visualizar o desenrolar da conversa no *Twitter* de acordo com diferentes métricas como, por exemplo, a frequência de *tweet*, sentimentos, novidade semântica e homogeneidade, palavras colocadas, dentre outras.

Iniciativas em empresas, organizações sem fins lucrativos, instituições de pesquisa, arquivos nacionais, universidades e consórcios

Através de *American Council for Technology* (2011), *Archive Social* (c2021, c2023), Byrne (2016), *Documenting the Now* ([2021]), *International Internet Preservation Consortium* (c2021b), Rezende e Martins (2018, 2019), Rockembach (2018), Rockembach e Pavão (2018), Ruest e Milligan (2016), *The National Archives* (c2014, [2022d]) e Thomson (c2016), notamos várias iniciativas internacionais de preservação digital e arquivamento de conteúdos em mídias sociais, que se pautam na API e na política de uso das próprias plataformas de mídia social, com alguns casos abrangendo coleções sobre eventos e conteúdos produzidos no Brasil, como:

- *Obama White House Social Media Archive*[115] – definido por Barack Obama, o primeiro presidente negro dos Estados Unidos, e realizado pela *ArchiveSocial*[116], o arquivo visa a transparência e a preservação histórica das interações com o público entre 2009 e 2017 por perfis oficiais do presidente e da Casa Branca em mídias sociais (*Twitter*, *Facebook*, *Instagram*, *Pinterest* etc.), garantindo a recuperação e o acesso público a todo o conteúdo original gerado.

[115] Disponível em: https://archivesocial.com/whitehouse/. Acesso em: 7 jun. 2023.
[116] Disponível em: https://archivesocial.com/about-us/. Acesso em: 7 jun. 2023.

Figura 26 – Página inicial do *Obama White House Social Media Archive*.

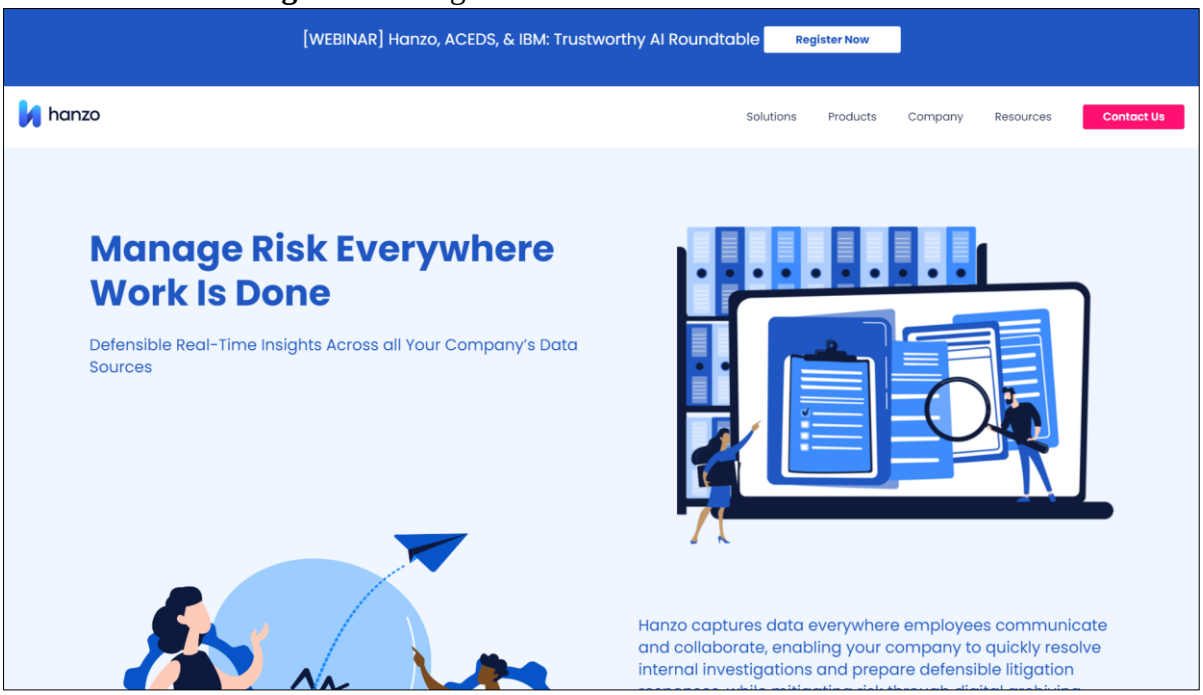
Fonte: *Archive Social* (c2023).

Os registros completos (e os metadados) brutos e nativos em *JavaScript Object Notation* (JSON)[117] ou XML são coletados e autenticados segundo o RFC 3161[118] e, depois, armazenados numa base confiável para preservação digital. Em obediência à política de uso da mídia social, os dados capturados não podem ser armazenados em dispositivos na nuvem ou compartilhados com demais instituições.

- *Hanzo Archives Limited*[119] – como um provedor de serviços com sedes no Reino Unido e nos Estados Unidos e surgido em 2009 de um projeto da Biblioteca Britânica dirigido para capturar a *Web* "para sempre", a *Hanzo Archives* fornece soluções de preservação de *sites* e mídias sociais para fins de patrimônio cultural corporativo, ou também para o gerenciamento de registros e requisitos de conformidade legal e regulatória corporativa que exigem que o conteúdo da *Web* em formato nativo de organizações seja capturado em defesa.

[117] Disponível em: https://www.json.org/json-en.html. Acesso em: 7 jun. 2023.
[118] Disponível em: https://datatracker.ietf.org/doc/html/rfc3161. Acesso em: 7 jun. 2023.
[119] Disponível em: https://www.hanzo.co/about. Acesso em: 7 jun. 2023.

Figura 27 – Página inicial da *Hanzo Archives Limited*.



Fonte: Hanzo (c2022).

Com acesso restrito a clientes e conforme demanda de organizações, a coleta pode abranger *sites* corporativos, incluindo contas de *login* e páginas interativas; redes sociais privadas, como *Chatter* e *Yammer*; sistemas colaborativos, como *SharePoint* e *Wiki*; além de mídias sociais públicas, como *Twitter*, *Facebook*, *Instagram* e *LinkedIn*.

- *Best Practices Study of Social Media Records Policies* – criado pelo *American Council for Technology-Industry Advisory Council* (ACT-IAC)[120] nos Estados Unidos em 2011, este estudo propôs discussões sobre o uso de mídias sociais a fim de auxiliar o governo norte-americano e os cidadãos a se conectarem de forma mais colaborativa, próxima e aberta. Entre as melhores práticas apontadas na pesquisa, podemos citar: a formação de equipe de arquivamento de mídias sociais com arquivistas, gestores *Web*, profissionais de tecnologia da informação e outros atores válidos neste contexto; a utilização de vários recursos informacionais pelas agências governamentais ao elaborarem políticas e boas práticas quanto ao uso e o arquivamento de mídias sociais; e a necessidade de definição de papéis e de responsabilidades relativas ao uso e o armazenamento das mídias sociais.
- GESIS *Leibniz Institute for the Social Sciences*[121] – como a maior instituição autônoma de Ciências Sociais da Alemanha, o GESIS desenvolveu um estudo piloto de coleta e

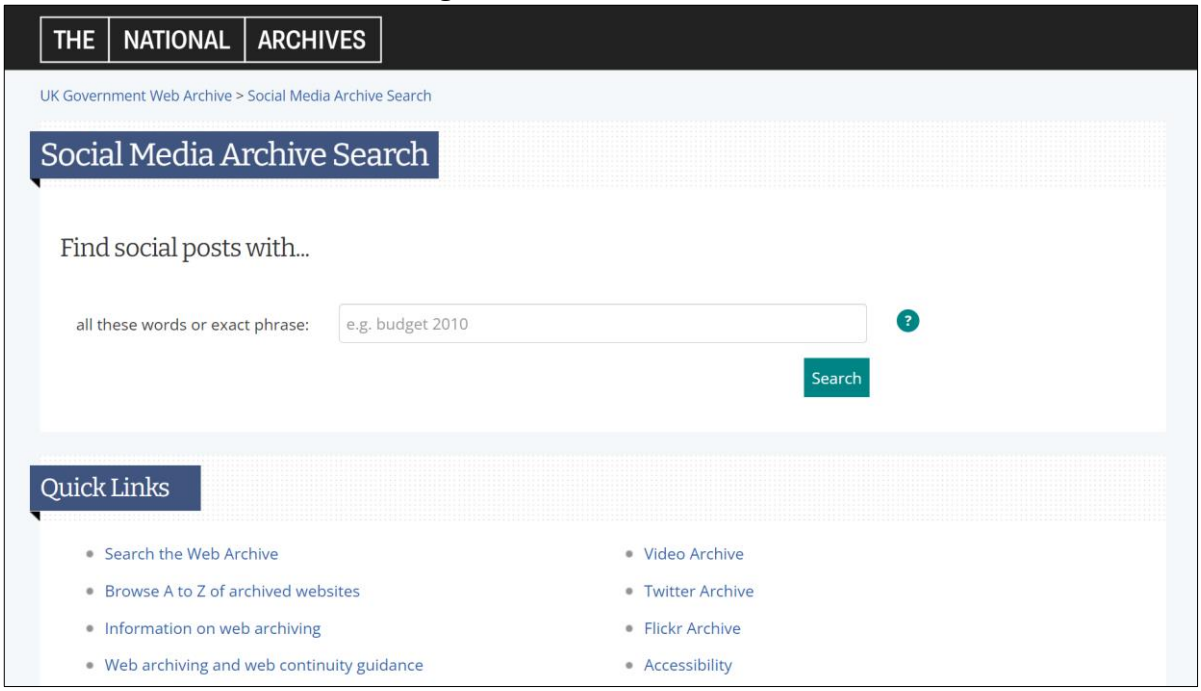
[120] Disponível em: https://www.actiac.org/act-iac-glance. Acesso em: 7 jun. 2023.
[121] Disponível em: https://www.gesis.org/en/institute/. Acesso em: 7 jun. 2023.

arquivamento de dados de mídias sociais durante as eleições parlamentares alemãs de 2013. No caso do *Facebook*, os dados foram coletados e arquivados em formato texto e imagem com o uso do *software* SODATO; já no *Twitter*, *tweets* relevantes foram coletados em formato JSON ou XML, filtrados por *hashtag* e intervalo de datas, e disponibilizados apenas para pesquisadores autorizados em acordo com à política de acesso e uso de ambas as mídias sociais. Entre os desafios enfrentados no projeto, podemos citar a volatilidade e a autenticidade dos dados, sendo assim difícil a reprodução exata das experiências de acesso e utilização dos dados das mídias sociais.

- *Social Repository of Ireland*[122] – criado por um consórcio de instituições de pesquisa e lançado em 2014, este projeto objetiva explorar os desafios do arquivamento de dados de mídias sociais importantes sobre a Irlanda, preservando-os no *Digital Repository of Ireland*[123] a fim de garantir o acesso a longo prazo para jornalistas e pesquisadores. Dados do *Twitter* são coletados com o uso de palavras-chave e de *hashtags*, em especial acerca de figuras públicas irlandesas, localizações geográficas e instituições pertinentes para a Irlanda, atendendo a política de uso da mídia social de coletar no máximo 1% dos dados do *Twitter* por dia etc. Dos exemplos de coleções criadas, citamos os dados de *tweets* com as opiniões dos cidadãos irlandeses em volta do histórico referendo constitucional realizado no país em 2015[124] para reconhecer o casamento entre pessoas do mesmo sexo.
- *'The National Archives' UK Government Social Media Archive*[125] – lançado pelos Arquivos Nacionais do Reino Unido com a *Internet Memory Foundation*, este projeto piloto iniciado em 2011 visou a captura de comunicações do governo inglês por perfis oficiais no *Twitter* e no *Youtube*, assegurando a disposição contínua deste conteúdo.

[122] Disponível em: https://dri.ie/projects. Acesso em: 7 jun. 2023.
[123] Disponível em: https://dri.ie/about-dri. Acesso em: 7 jun. 2023.
[124] Disponível em: https://www.bbc.com/portuguese/institutional/2015/05/150523_resultado_casamento_gay_irlanda_lab. Acesso em: 7 jun. 2023.
[125] Disponível em: https://webarchive.nationalarchives.gov.uk/social/search/. Acesso em: 7 jun. 2023.

Figura 28 – Página inicial dos canais de mídia social arquivados no arquivo da *Web* do governo do Reino Unido.



Fonte: *The National Archives* ([2023?b]).

O estudo criou soluções para o arquivamento automático de dados e metadados originais do *Twitter*, atendendo não só os direitos autorais pela *Copyright, Designs and Patents Act* 1988[126] e a *Public Records Act* 1958[127] que obriga a preservação de todos os registros informacionais do governo; e também às demandas informacionais dos usuários do *UK Government Web Archive*[128], onde as contas governamentais nas duas mídias e os canais oficiais dos Jogos Olímpicos de Londres em 2012 são mantidos como registros públicos.

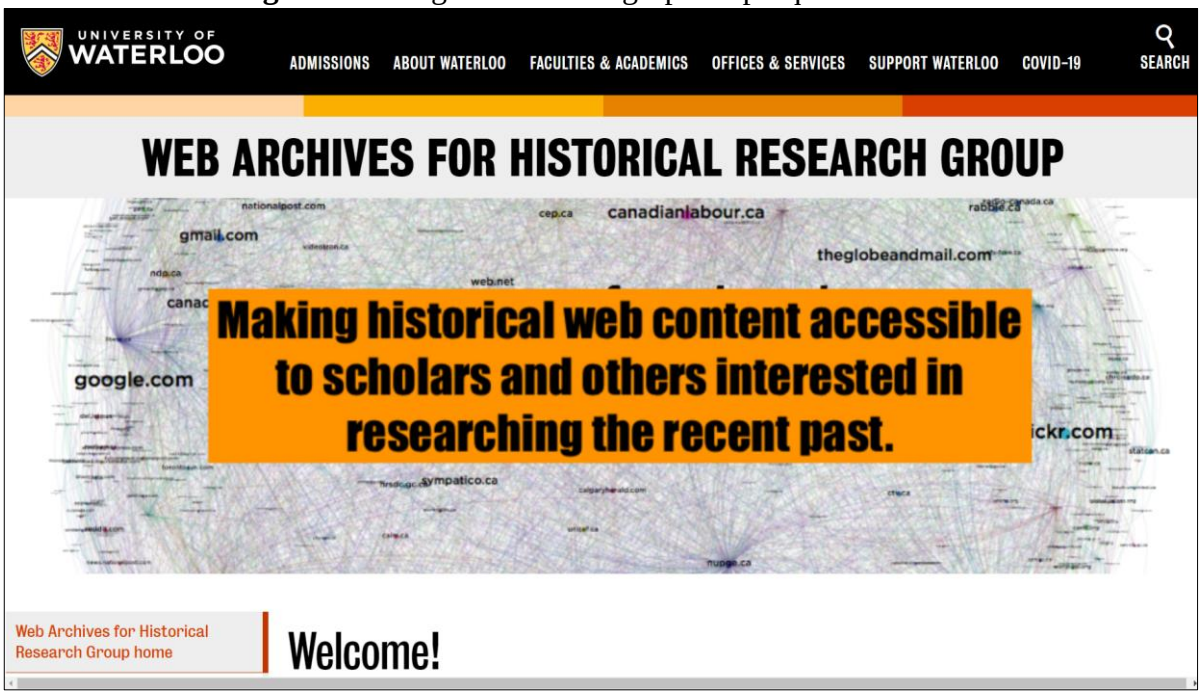
- *Social Media Records in Australia* – os Arquivos Nacionais da Austrália[129] publicaram orientações[130] sobre a aplicação das leis que afetam os registros de mídia social gerados por órgãos governamentais do país, incluindo a *Public Records Act* 2002[131] e a *Archives Act* 1983[132] que definem o registro como um objeto em algum formato. Para os Arquivos Nacionais da Austrália os registros criados como efeitos do uso de mídias sociais estão

[126] Disponível em: http://www.legislation.gov.uk/ukpga/1988/48/contents. Acesso em: 7 jun. 2023.
[127] Disponível em: https://www.legislation.gov.uk/ukpga/Eliz2/6-7/51. Acesso em: 7 jun. 2023.
[128] Disponível em: http://www.nationalarchives.gov.uk/webarchive/. Acesso em: 7 jun. 2023.
[129] Disponível em: https://www.naa.gov.au/about-us. Acesso em: 7 jun. 2023.
[130] Disponível em: https://www.naa.gov.au/information-management/types-information-and-systems/types-information/managing-social-media. Acesso em: 7 jun. 2023.
[131] Disponível em: https://www.legislation.qld.gov.au/view/html/inforce/current/act-2002-011. Acesso em: 7 jun. 2023.
[132] Disponível em: https://www.legislation.gov.au/Details/C2014C00417. Acesso em: 7 jun. 2023.

submetidos as mesmas exigências comerciais e legislativas que os registros criados por outros meios, sendo que estes registros públicos devem ser capturados e armazenados pelas agências que os criaram, com garantias de que sejam mantidos de forma utilizável e acessível pelo tempo que for necessário. A tecnologia de arquivamento de registros de mídia social que vêm sendo empregada na Austrália e está divulgada é da *ArchiveSocial*.

- *Web Archives for Historical Research* (WAHR) *Group*[133] – dirigido na Universidade de *Waterloo* com parcerias com as Universidades de *Western* e de *York* no Canadá, o grupo WAHR visa vincular história e *Big Data* para prover aos historiadores as ferramentas necessárias para encontrar e interpretar fontes digitais de arquivos da *Web*.

Figura 29 – Página inicial do grupo de pesquisa WAHR.



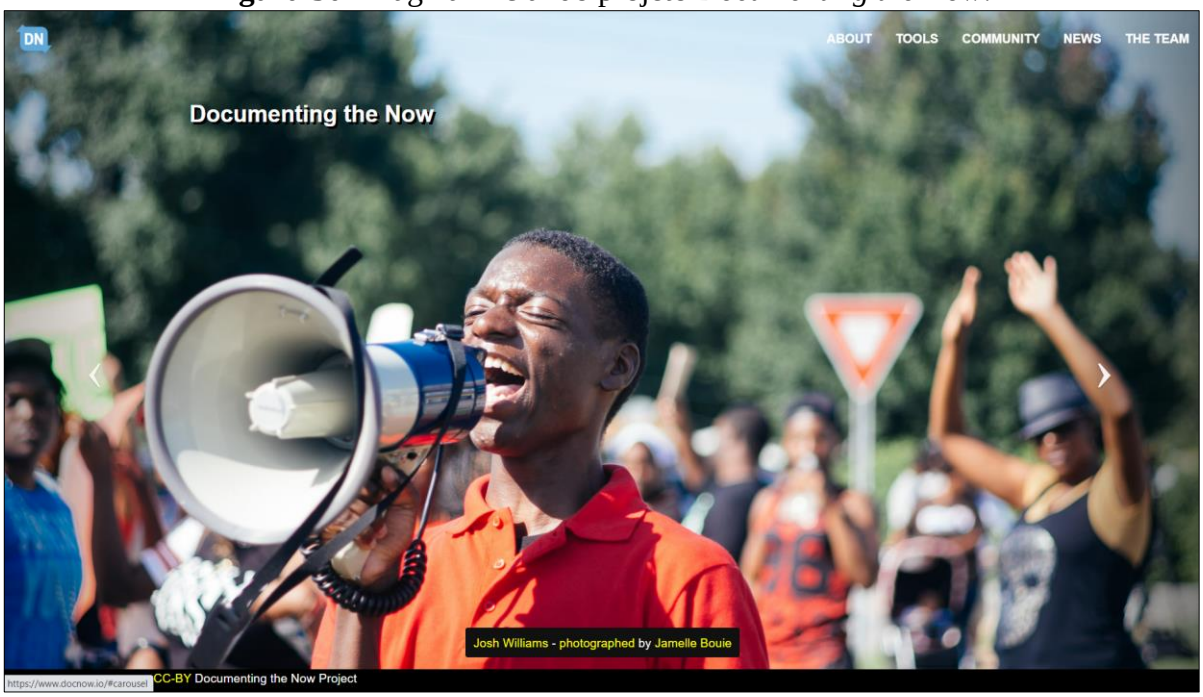
Fonte: *University of Waterloo* ([2023?]).

Este grupo fez um estudo de caso da 42ª eleição federal canadense de 2015, onde aplicou-se a coleta e o arquivamento de *tweets* no *Twitter* como método para documentar informações do evento. Em obediência à política de uso da plataforma de mídia social, no estudo houve o uso da TWARC para o arquivamento dos dados de *tweets* coletados em formato JSON, e também o uso da *hashtag #elxn42* em alusão a eleição que permitiu a participação do público no registro colaborativo de informações e a criação de uma memória do evento.

[133] Disponível em: https://uwaterloo.ca/web-archive-group/. Acesso em: 7 jun. 2023.

- *Documenting the Now*[134] – surgido incitado nas reações no *Twitter* após o tiroteio policial de Michael Brown em *Ferguson*[135] de 2014, tal projeto cooperativo entre universidades norte-americanas destina-se a propiciar que pesquisadores e profissionais da informação coletem e preservem os conteúdos das mídias sociais (especialmente, o *Twitter*) acerca de fatos históricos significativos ao redor do mundo.

Figura 30 – Página inicial do projeto *Documenting the Now*.



Fonte: *Documenting the Now* ([2023?b]).

O *Documenting the Now* prioriza práticas éticas ao coletar e preservar os conteúdos, e ainda usa ferramentas de *software* livre, como o TWARC[136] para baixar dados do *Twitter* em JSON; o *DiffEngine*[137] que possibilita acompanhar as alterações nos artigos de notícias através de seus *feeds Really Simple Syndication* (RSS); e o *Tweet Catalog*[138], um catálogo de conjuntos de dados de identificadores de *tweets* compartilhados publicamente que permite baixá-los em JSON.

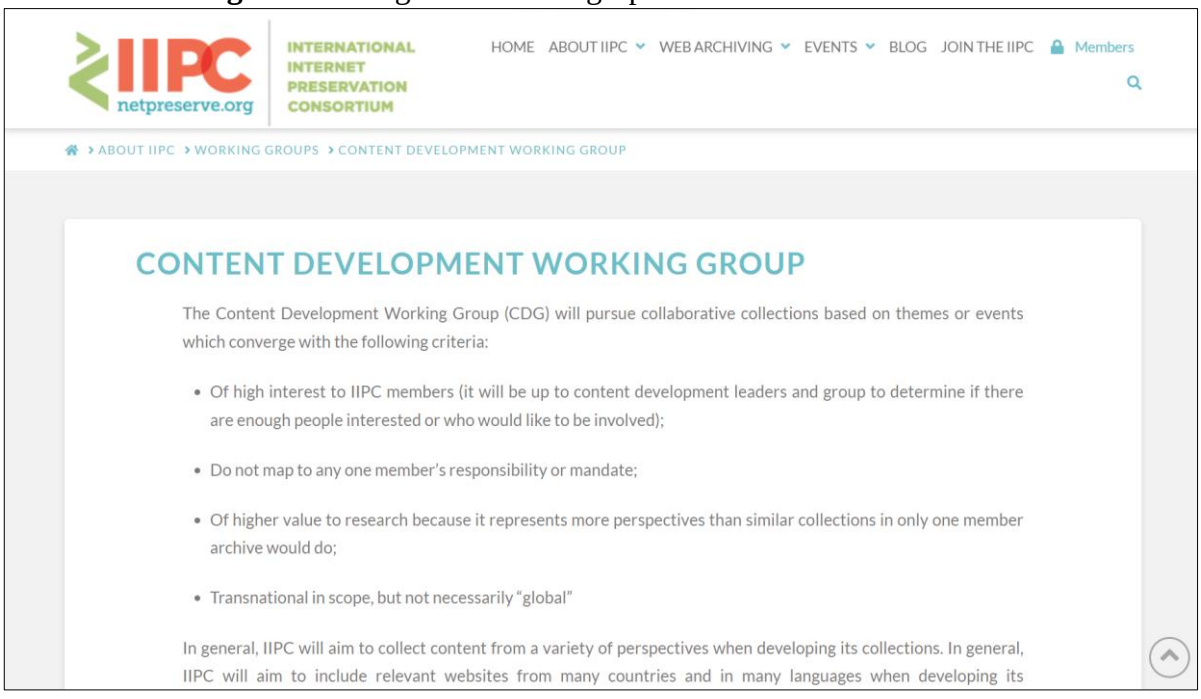
- COSMOS *Platform*[139] – mantida pelo *Social Data Science Lab*[140] da Universidade *Cardiff* no Reino Unido, a plataforma COSMOS propõe facilitar o tratamento e as

[134] Disponível em: https://www.docnow.io/. Acesso em: 7 jun. 2023.
[135] Disponível em: http://g1.globo.com/mundo/noticia/2016/08/vigilia-lembra-dois-anos-da-morte-de-michael-brown-em-ferguson.html. Acesso em: 7 jun. 2023.
[136] Disponível em: https://github.com/DocNow/twarc. Acesso em: 7 jun. 2023.
[137] Disponível em: https://github.com/DocNow/diffengine. Acesso em: 7 jun. 2023.
[138] Disponível em: https://catalog.docnow.io/. Acesso em: 7 jun. 2023.
[139] Disponível em: http://socialdatalab.net/cosmos. Acesso em: 7 jun. 2023.
[140] Disponível em: http://socialdatalab.net/home. Acesso em: 7 jun. 2023.

análises de grandes volumes de dados extraídos de mídias sociais, em especial, o *Twitter*, de forma gratuita e acessível para uso sem fins lucrativos por pesquisadores de Ciências Sociais. Os dados brutos (e os seus metadados), originais em JSON, são extraídos para uma base local da COSMOS, indexados e consultados por uma camada intermediária de *software*. Sob a política de uso da mídia social, a COSMOS está igualmente restrita a coletar por dia só 1% dos dados do *Twitter* etc. Para incrementar as estratégias de armazenamento, coleta e análises destes dados, o *Social Data Science Lab* tem feito parcerias com outros órgãos internacionais, incluindo empresas proprietárias das plataformas de mídia social.

- *Content Development Working Group* (CDG)[141] – ligado ao *International Internet Preservation Consortium* (IIPC)[142], o grupo de trabalho CDG coleta *sites*, artigos, notícias, *blogs* e mídias sociais (*Twitter*, *Facebook* etc.) sobre temas ou eventos relevantes de muitos países e em muitos idiomas, como as Olímpiadas Rio 2016 (https://archive-it.org/collections/7235).

Figura 31 – Página inicial do grupo de trabalho CDG do IIPC.



Fonte: *International Internet Preservation Consortium* (c2023).

[141] Disponível em: https://netpreserve.org/about-us/working-groups/content-development-working-group/. Acesso em: 7 jun. 2023.

[142] Formado em 2003 na BnF, o IIPC dispõe da participação de várias organizações, tais como a Biblioteca Nacional do Chile e a Biblioteca do Congresso dos Estados Unidos, que se dedicam a criação de padrões e ferramentas para o arquivamento da *Web*. Disponível em: http://netpreserve.org/about-us/. Acesso em: 7 jun. 2023.

> Nesta coleção em específico, o grupo criou um formulário público[143] para que as pessoas contribuíssem com a seleção de temas relativos ao evento citado, onde a *hashtag* #RIO2016WA no *Twitter* foi utilizada para acompanhar atualizações desta iniciativa e conectar pessoas propensas a ajudar com o processo de arquivamento. Outra coleção muito importante, criada em 2020 pelo grupo em colaboração com o *Archive-It*, envolve a preservação do conteúdo da *Web* referente à pandemia do novo coronavírus (COVID-19) (https://archive-it.org/collections/13529).

Primeiro, na análise das experiências levantadas de arquivamento e preservação digital de conteúdos públicos em mídias sociais, é verdade que as plataformas de mídia social se fazem cruciais hoje para compartilhar e registrar as opiniões, experiências e atividades de indivíduos, governos, empresas, organizações etc. acerca de eventos e tópicos com valor histórico, cultural e científico significativos no mundo, constituindo-se em fontes de dados úteis para jornalistas, profissionais da informação e pesquisadores de Ciências Sociais e outras ciências. Contudo, em acordo com Brooker, Barnett e Cribbin (2016), *Cardiff University* ([2022]) e Hussain e Vatrapu (2014), coletar, analisar sistematicamente, recuperar, armazenar, preservar, e fornecer o acesso contínuo e utilizável a esses grandes conjuntos de dados sociais, originais em JSON ou XML, como recursos de pesquisa requer habilidades computacionais que se traduzem na adoção de ferramentas gratuitas e comerciais, como *Archive Social*, TWARC, *Hanzo*, *Chorus*, COSMOS e SODATO, além de repositórios digitais e bases de dados confiáveis para preservação digital.

Em segundo lugar, a maioria das iniciativas internacionais analisadas e descritas neste trabalho provém de universidades, grupos de trabalho de consórcios, conselhos, instituições de pesquisa e de patrimônio cultural (em especial, arquivos nacionais), dos quais têm como foco a captura, arquivamento e preservação em longo prazo de dados e metadados de comunicação (as atualizações de *status*, comentários, imagens e vídeos postados etc.) gerados em plataformas de mídia social por contas pessoais ou institucionais, garantindo a recuperação, acesso e análise de *Big Data* de mídia social para usuários atuais e futuros. Porém, este processo está submetido as políticas de coleta, acesso, desenvolvimento e uso das plataformas de mídia social que, de acordo com Ruest e Milligan (2016) e Thomson (c2016), inibem um maior acesso e disposição de dados arquivados de mídia social na *Web*, obrigando as iniciativas a buscarem formas de tornar os dados disponíveis cumprindo as condições das plataformas, como dados brutos sendo disponibilizados apenas para pesquisadores/usuários autorizados ou para acesso e uso no local.

---

[143] Disponível em: https://netpreserveblog.wordpress.com/2016/06/27/2016-rio-games-collection-how-to-get-involved/. Acesso em: 7 jun. 2023.

Por último, apesar da relevância das mídias sociais como registro cultural e histórico de uma nação (REZENDE; MARTINS, 2018) e, também, fonte de pesquisa para que historiadores possam compreender eventos na história e cientistas políticos e sociais entendam e expliquem as formas pelas quais a sociedade contemporânea funciona (CARDIFF UNIVERSITY, [2022]; RUEST; MILLIGAN, 2016), existem muitos outros desafios no arquivamento de conteúdos de mídias sociais. Conforme *American Council for Technology* (2011) e Rezende e Martins (2019) isto pode incluir as dificuldades em se definir esses conteúdos como registros arquivísticos; e o fato da maioria dos conteúdos de mídias sociais serem de domínio público, não estando sob o controle das agências governamentais complicando a sua coleta. Ademais, as iniciativas devem avaliar as questões legais e éticas que surgem na preservação, acesso e uso de dados de mídia social buscando atender tanto as leis de direitos autorais, privacidade e liberdade de informação como a garantia do acesso aos dados para pesquisadores, como o estudo de caso da 42ª eleição federal canadense do grupo WAHR que entende que podemos capturar, arquivar e preservar só *tweets* públicos do *Twitter* segundo a política desta mídia social (RUEST; MILLIGAN, 2016).

2.1.4 Preservação digital de periódicos científicos eletrônicos

Sendo uma estratégia operacional de preservação digital, as questões de preservação e acesso contínuo para periódicos eletrônicos se fazem cada vez mais notáveis às bibliotecas de pesquisa visto que as revistas e os artigos acadêmicos publicados passaram do formato impresso ao eletrônico trazendo grandes transformações nas relações e modelos comerciais de publicação tradicionais (BEAGRIE, c2013). Entre muitas mudanças, houve a de bibliotecas que compram, mantêm e preservam uma revista em papel localmente para o aluguel (licenciamento) de acesso remoto a um periódico eletrônico mantido em plataformas de editores que são muito baseadas internacionalmente em outras jurisdições; em paralelo, temos o crescente movimento de acesso aberto (*Open Acess*)[144] para artigos de periódicos científicos eletrônicos que busca remover as taxas de assinatura para acesso, e as revistas por assinatura, de acesso aberto e híbridos das duas proporcionam um cenário complexo para a preservação e o acesso a longo prazo aos periódicos eletrônicos, em conformidade com Beagrie (c2013) e *Digital Preservation Coalition* (c2015).

[144] Acesso livre, ou Acesso Aberto (*Open Acess*) significa "[...] um modelo de acesso a obras e dados criados que procura eliminar barreiras para os leitores, como taxas de assinatura e mídia física, e barreiras à reutilização, como restrições de direitos autorais e taxas de licenciamento" (SOCIETY OF AMERICAN ARCHIVISTS, c2022i, não paginado, tradução nossa).

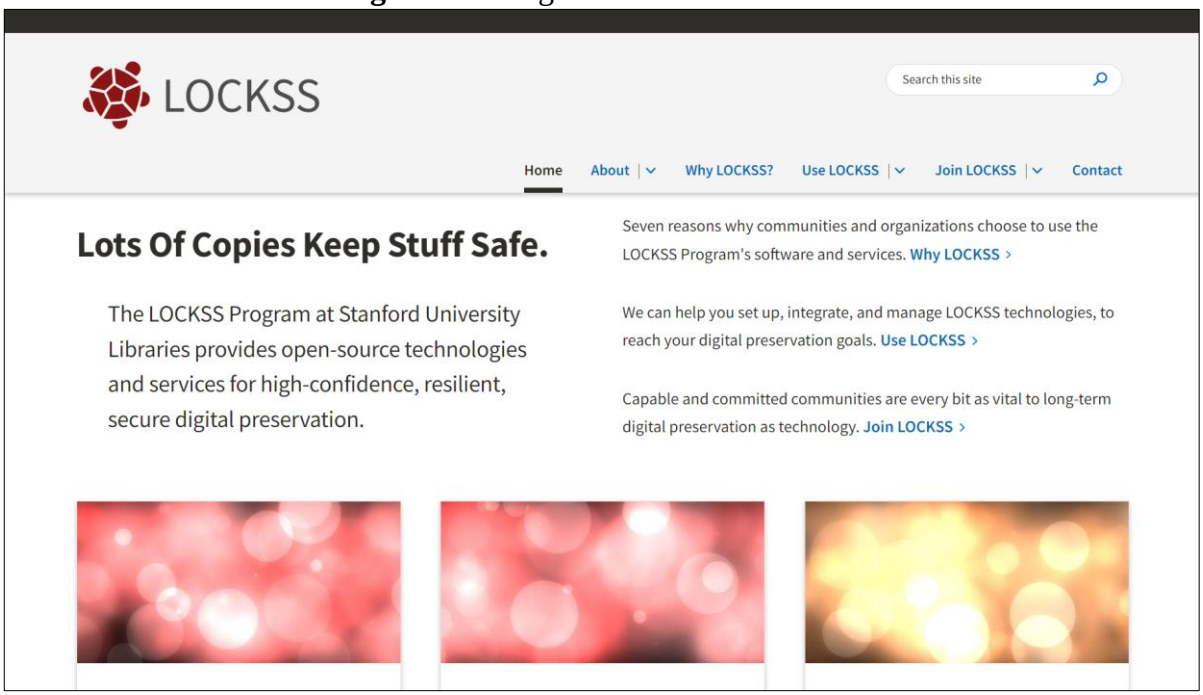
Iniciativas em bibliotecas universitárias, de pesquisa, acadêmicas e nacionais, além de universidades, órgãos de pesquisa e organizações sem fins lucrativos

Em Araujo (2015), *Archaeology Data Service* (2021), Beagrie (c2013), CLOCKSS *Archive* (c2022), Instituto Brasileiro de Informação em Ciência e Tecnologia (2016), ISSN *International Center* ([2022]), *Digital Preservation Coalition* (c2015), *HathiTrust* ([2022]), *International Standard Serial Number Portal* (c2022), Ithaka (c2022a, c2022b), *Koninklijke Bibliotheek* ([c2021]), Moghaddam (2008), *National Center for Biotechnology Information* (2019), *Ontario Council of University Libraries* ([2022]), *Ohio Library and Information Network* (c2022), *Stanford University* ([2021]) e *University of California* (c2022), observamos algumas iniciativas internacionais e nacionais de preservação digital de periódicos científicos eletrônicos (e outros tipos e formatos de publicações acadêmicas na *Web*), que se dirigem às suas comunidades e ao fomento à pesquisa como se pautam em padrões, em ferramentas de auditoria e certificação de repositório digital baseadas, e em licenças e autorizações legais com editoras, autores e instituições detentoras de direitos autorais sobre estes conteúdos abertos ou de assinatura para a sua coleta, preservação a longo prazo e acesso contínuo/perpétuo, como:

- *Lots of Copies Keep Stuff Safe* (LOCKSS, em português Muitas Cópias Mantêm as Coisas Seguras)[145] – lançado em 1999 pela SUL, o Programa LOCKSS oferece acesso perpétuo e pós-cancelamento[146] para publicações eletrônicas (livros, periódicos etc.) de acesso aberto e por assinatura preservadas pelas instituições acadêmicas e de patrimônio cultural.

[145] Disponível em: https://www.lockss.org/about. Acesso em: 7 jun. 2023.
[146] Disponível em: https://www.lockss.org/use-lockss/post-cancellation-and-perpetual-access. Acesso em: 7 jun. 2023.

Figura 32 – Página inicial do LOCKSS.



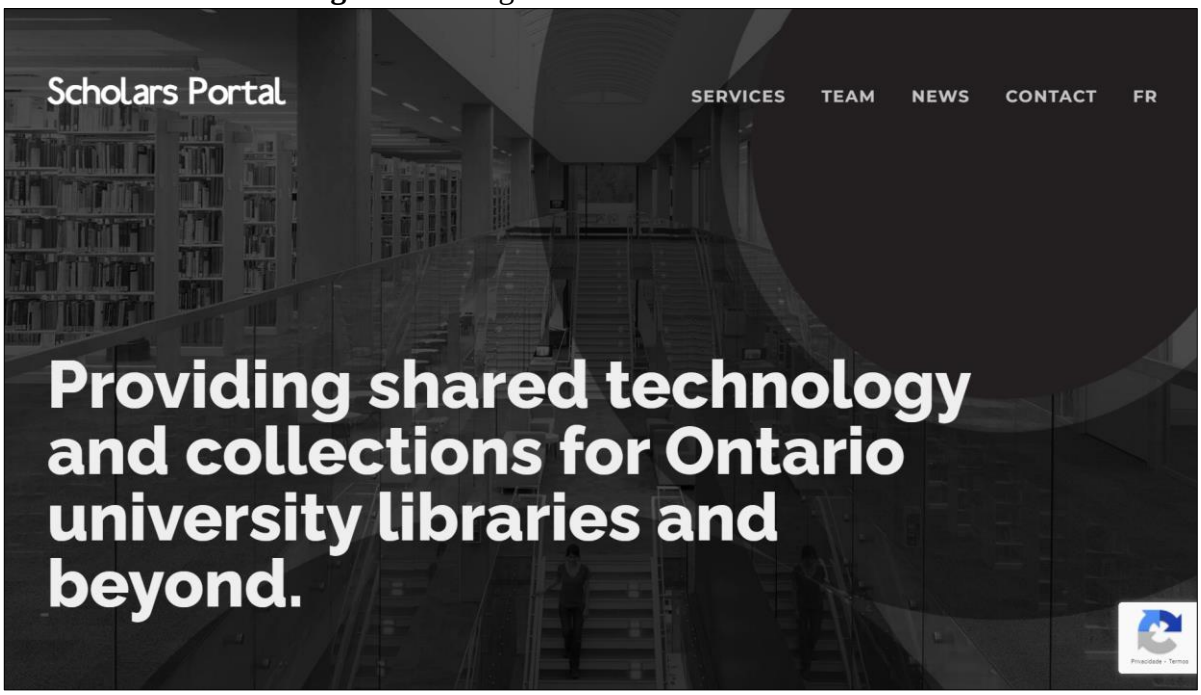
Fonte: *Stanford University* ([2023?b]).

Além de ser um programa provedor de soluções e tecnologia de preservação digital e uma comunidade internacional de instituições e redes que atuam juntas[147], o LOCKSS é um princípio de uso de várias cópias distribuídas e descentralizadas dos dados para uma preservação digital robusta e segura em termos de integridade a longo prazo com possibilidade de reparação se necessário, e um *software* de código aberto que permite coletar e arquivar os conteúdos. Entre as comunidades que usam a rede LOCKSS, estão a Rede Cariniana[148], e o Perma.cc[149] que ajuda acadêmicos, periódicos etc. a criarem registros permanentes e inalteráveis das fontes da *Web* que eles citam.

- *Scholars Portal*[150] – gerido pelo *Ontario Council of University Libraries* (OCUL)[151], esta iniciativa foi fundada em 2002 e provê uma estrutura tecnológica que preserva e fornece o acesso à recursos coletados e partilhados pelas bibliotecas universitárias de *Ontario* no Canadá.

[147] Disponível em: https://www.lockss.org/join-lockss/networks. Acesso em: 7 jun. 2023.
[148] Disponível em: https://cariniana.ibict.br/. Acesso em: 7 jun. 2023.
[149] Disponível em: https://perma.cc/. Acesso em: 7 jun. 2023.
[150] Disponível em: https://journals.scholarsportal.info/about. Acesso em: 7 jun. 2023.
[151] Disponível em: https://ocul.on.ca/. Acesso em: 7 jun. 2023.

Figura 33 – Página inicial do *Scholars Portal*.



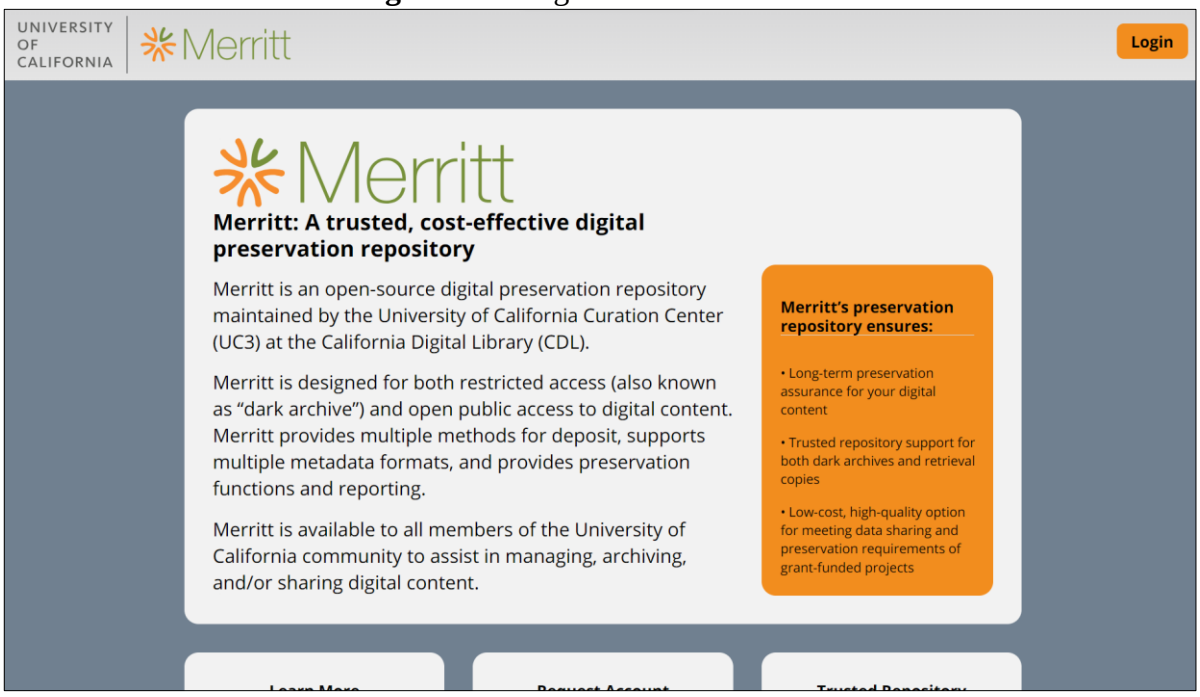
Fonte: *Ontario Council of University Libraries* ([2023?]).

Através dos seus serviços *online*, os alunos, docentes e pesquisadores da região têm acesso a uma ampla coleção de periódicos eletrônicos, livros etc. O *Scholars Portal Journals*[152] é um dos serviços que consisti num sistema com 50 milhões de artigos de 20 mil periódicos em texto completo que cobrem todas as disciplinas acadêmicas e em geral são por assinatura. Em 2013[153], o serviço foi avaliado e certificado externamente pelo *Center for Research Libraries* (CRL)[154] via os critérios da ferramenta T*rustworthy Repositories Audit & Certification: Criteria and Checklist* (TRAC)[155] como sendo um repositório digital confiável para preservação a longo prazo.

- *Merritt*[156] – mantido pelo *University of California Curation Center* (UC3)[157] na CDL, o *Merritt* é um repositório de preservação digital confiável que está disponível a membros da comunidade da Universidade da Califórnia para ajudar na gerência, arquivamento, compartilhamento e acesso de seu conteúdo digital.

[152] Disponível em: https://journals.scholarsportal.info/. Acesso em: 7 jun. 2023.
[153] Disponível em: https://www.crl.edu/reports/scholars-portal-audit-report-2013. Acesso em: 7 jun. 2023.
[154] Disponível em: https://www.crl.edu/about. Acesso em: 7 jun. 2023.
[155] Disponível em: https://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/trac. Acesso em: 7 jun. 2023.
[156] Disponível em: https://merritt.cdlib.org/. Acesso em: 7 jun. 2023.
[157] Disponível em: https://cdlib.org/services/uc3/about/. Acesso em: 7 jun. 2023.

Figura 34 – Página inicial do *Merritt*.



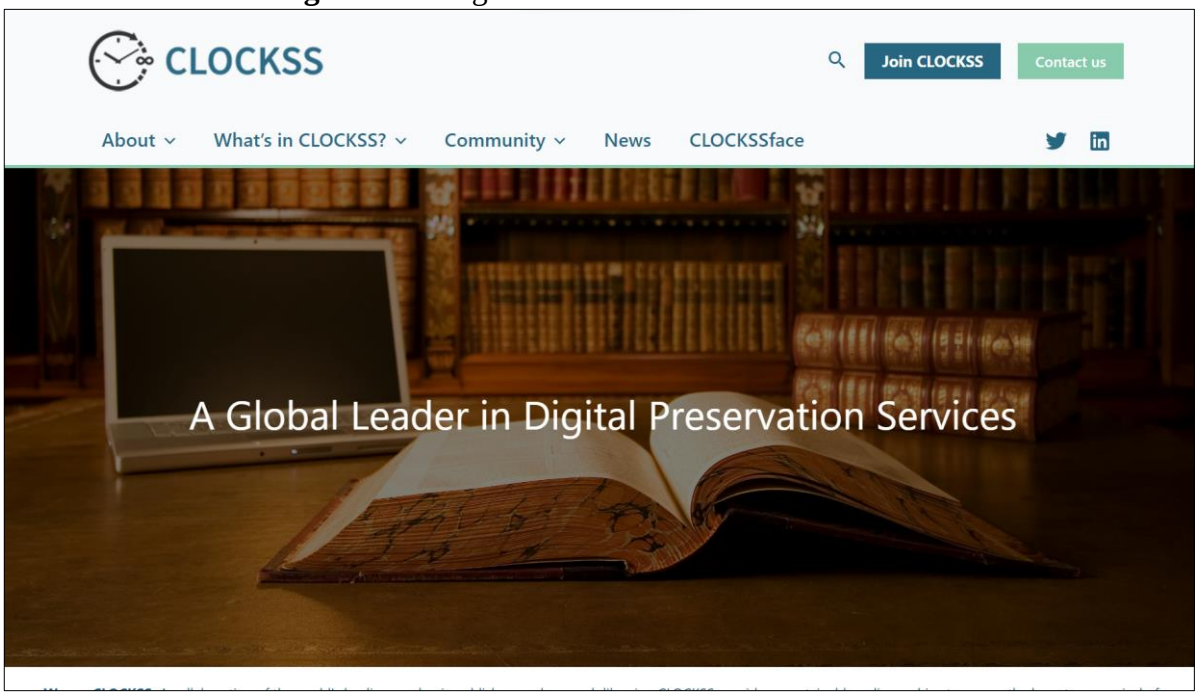
Fonte: *University of California* (c2023).

Certificado pela *CoreTrustSeal*[158], o *Merrit* gere as cópias de preservação de todas as publicações do repositório institucional *e-Scholarship*[159] da CDL de periódicos de acesso aberto. Projetado para acesso restrito (tido como arquivo oculto) e acesso público aberto aos conteúdos por meio de contas de usuário autorizadas, o *Merrit* fornece diversos métodos de depósito como, por exemplo através de *upload* direto do usuário final, além de suportar vários formatos de metadados (incluindo o MARC, o MODS e o METS) e prover funções de preservação e relatórios.

- *Controlled* LOCKSS (CLOCKSS) *Archive*[160] – criado em 2006 e sendo uma colaboração sem fins lucrativos entre as principais bibliotecas de pesquisa e editoras acadêmicas no mundo, o CLOCKSS fornece um arquivo oculto (*dark archive*)[161] baseado na tecnologia LOCKSS e certificado pelo CRL[162] que preserva por longo prazo os periódicos e livros eletrônicos originais assinados por sua comunidade.

[158] Disponível em: https://www.coretrustseal.org/wp-content/uploads/2018/08/UC3-Merritt.pdf. Acesso em: 7 jun. 2023.
[159] Disponível em: https://escholarship.org/aboutEschol. Acesso em: 7 jun. 2023.
[160] Disponível em: https://clockss.org/about/. Acesso em: 7 jun. 2023.
[161] Arquivos escuros ou ocultos (*dark archives*) tratam-se de "[...] um repositório que armazena recursos de arquivo para uso futuro, mas é acessível apenas para seu custodiante [...]" ou "[...] uma coleção de materiais preservados para uso futuro, mas sem acesso atual." (SOCIETY OF AMERICAN ARCHIVISTS, c2022e, não paginado, tradução nossa).
[162] Disponível em: https://clockss.org/2018/11/clockss-announces-its-trac-re-certification/. Acesso em: 7 nov. 2021.

Figura 35 – Página inicial do CLOCKSS *Archive*.



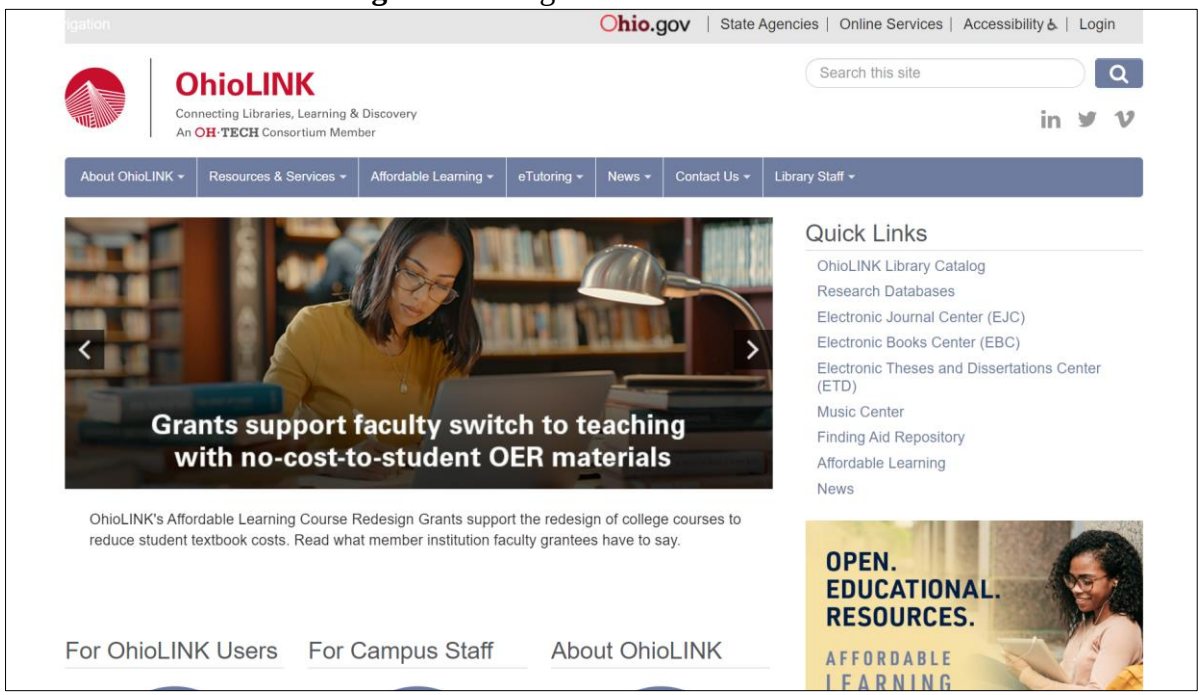
Fonte: CLOCKSS *Archive* (c2023).

O arquivo CLOCKSS cede licenças *Creative Commons Open Access* a todas as publicações de assinatura arquivadas apenas após serem acionadas[163] por situações de indisponibilidade[164] para fins de assegurar que elas sempre permaneçam disponíveis e abertamente acessíveis a todos. Esta organização conta com mais de 46 milhões de artigos de periódico, além de 300 bibliotecas de apoio e 400 editoras participantes, tais como a SciELO e a CAPES do Brasil (ambas até 2016).

- *Ohio Library and Information Network* (*Ohio*LINK)[165] – fundado em 1992 e sendo um consórcio de bibliotecas acadêmicas do Estado de *Ohio* e mais a *State Library of Ohio* nos Estados Unidos, o *Ohio*LINK organiza e presta serviços de hospedagem para acesso e preservação de periódicos eletrônicos.

[163] Disponível em: https://clockss.org/triggered-content/. Acesso em: 7 jun. 2023.
[164] Disponível em: https://clockss.org/faq/. Acesso em: 7 jun. 2023.
[165] Disponível em: https://www.ohiolink.edu/content/about_ohiolink. Acesso em: 7 jun. 2023.

Figura 36 – Página inicial do *Ohio*LINK



Fonte: OHIO *Library and Information Network* (c2023).

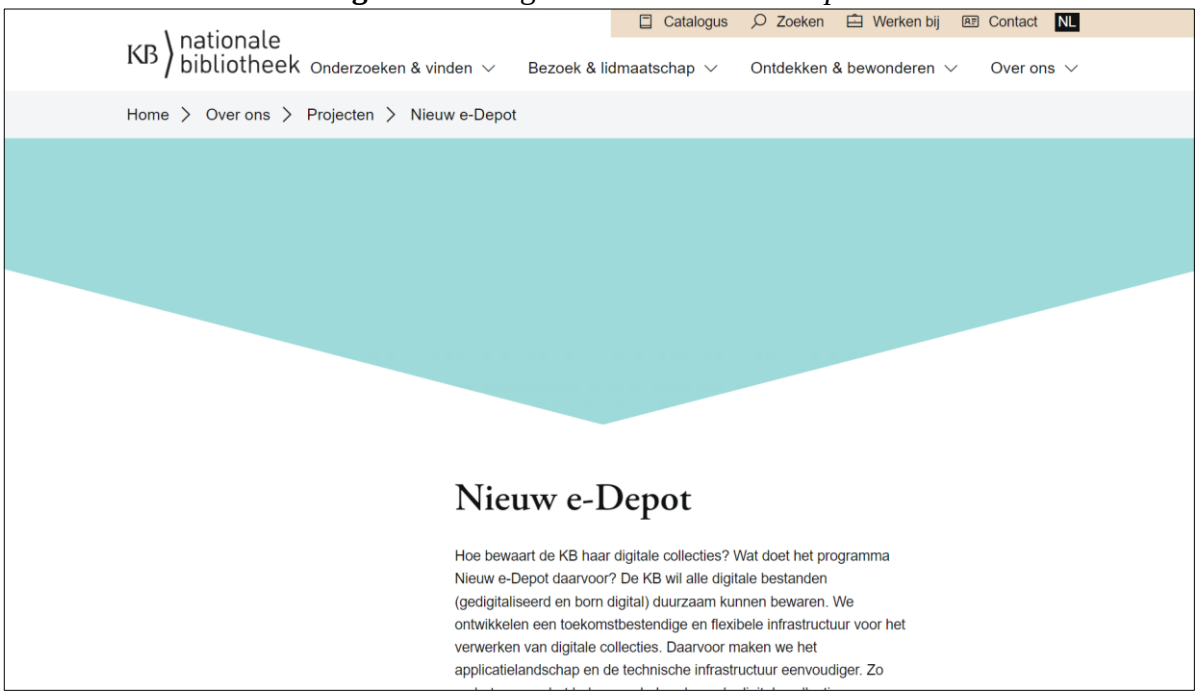
Em 2015[166], o consórcio passou a usar o sistema *Rosetta* da *Ex Libris*[167] para gerenciar, preservar e fornecer acesso a longo prazo às suas coleções digitais[168], como o *OhioLINK Electronic Journal Center* (EJC). Esta coleção é um banco de dados que contém cerca de 33 milhões de artigos em 10 mil periódicos de diversas editoras, do qual os usuários (alunos, docentes, pesquisadores etc.) baixam mais de 1,5 milhão de artigos em texto completo anualmente, e o conteúdo é financiado pela combinação de fundos centrais do *OhioLINK* e contribuições das bibliotecas membros.

- *Koninklijke Bibliotheek* (KB)[169] *e-Depot* – como biblioteca de depósito do patrimônio impresso e digital holandês, a Biblioteca Nacional da Holanda ou KB opera desde 2003 o *e-Depot* que é um arquivo digital para preservação a longo prazo e acesso contínuo e usável de periódicos eletrônicos acadêmicos, publicados dentro e fora de suas fronteiras, baseado em acordos de arquivamento com editoras[170] que permite a KB fornecer o

[166] Disponível em: https://www.ohiolink.edu/press/ohiolink_adopts_ex_libris_rosetta_digital_preservation. Acesso em: 7 jun. 2023.
[167] Disponível em: https://exlibrisgroup.com/products/rosetta-digital-asset-management-and-preservation/. Acesso em: 7 jun. 2023.
[168] Disponível em: https://www.ohiolink.edu/content/ohiolink_resources. Acesso em: 7 jun. 2023.
[169] Disponível em: https://www.kb.nl/en/organisation. Acesso em: 3 nov. 2021.
[170] Disponível em: https://www.kb.nl/organisatie/onderzoek-expertise/e-depot-duurzame-opslag/archiveringsovereenkomsten. Acesso em: 3 nov. 2021.

acesso *in loco* às publicações arquivadas (incluindo quando os títulos já não estão disponíveis na editora ou em outra fonte por circunstâncias específicas[171]).

Figura 37 – Página inicial do KB *e-Depot*.



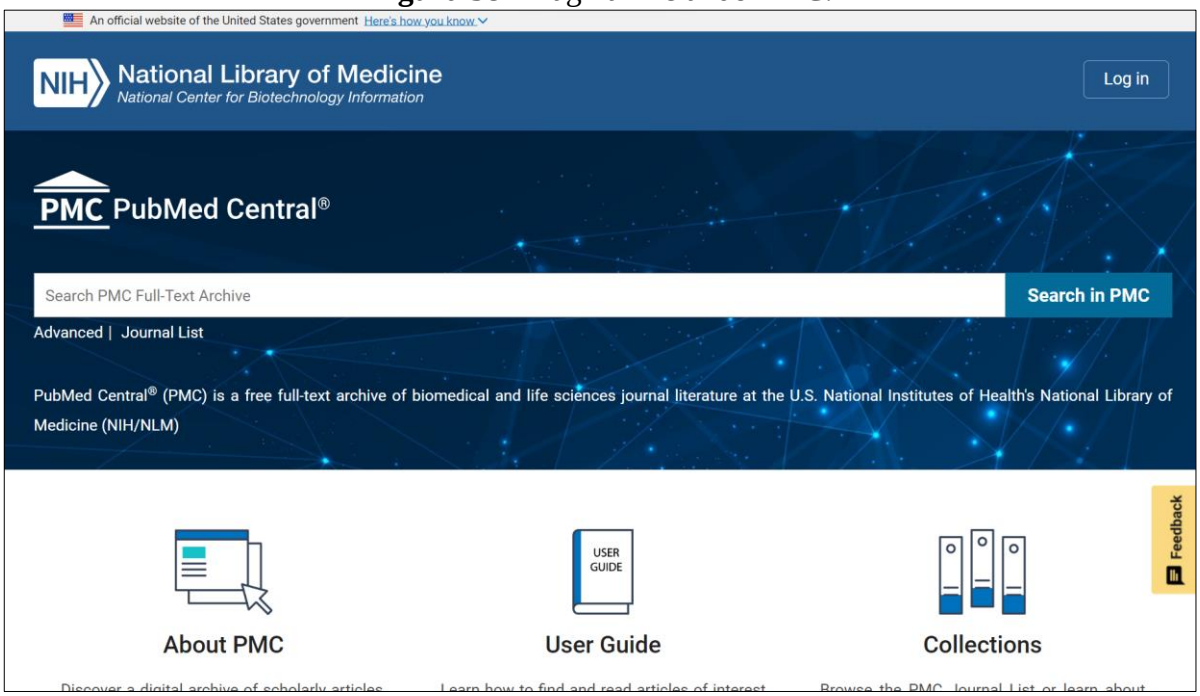
Fonte: *Koninklijke Bibliotheek* ([2023?]).

O sistema adota padrões internacionais, como o OAIS, as diretrizes para repositórios digitais confiáveis da ISO-16363 (ou *Trusted Digital Repository Checklist* – TDR)[172] e os padrões de metadados METS, MODS e PREMIS, sendo que em 2021 o *e-Depot* da KB obteve a certificação independente *CoreTrustSeal*[173] para repositórios digitais confiáveis.

- *PubMed Central* (PMC)[174] – criado em 2000 e elaborado e operado pelo *National Center for Biotechnology Information* (NCBI)[175] da *National Library of Medicine* (NLM)[176] dos Estados Unidos, o PMC é um arquivo eletrônico de artigos de periódicos em texto completo de Biomedicina e Ciências Biológicas, o qual preserva a longo prazo e fornece acesso permanente e gratuito ao seu conteúdo com uso sujeito aos direitos autorais e/ou termos de licença dos autores e editoras.

[171] Disponível em: https://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources/trigger-events. Acesso em: 3 nov. 2021.
[172] Disponível em: https://www.iso.org/standard/56510.html. Acesso em: 7 jun. 2023.
[173] Disponível em: https://www.coretrustseal.org/about/. Acesso em: 7 jun. 2023.
[174] Disponível em: https://www.ncbi.nlm.nih.gov/pmc/. Acesso em: 7 jun. 2023.
[175] Disponível em: https://www.ncbi.nlm.nih.gov/home/about/. Acesso em: 7 jun. 2023.
[176] Disponível em: https://www.nlm.nih.gov/about/index.html. Acesso em: 7 jun. 2023.

Figura 38 – Página inicial do PMC.



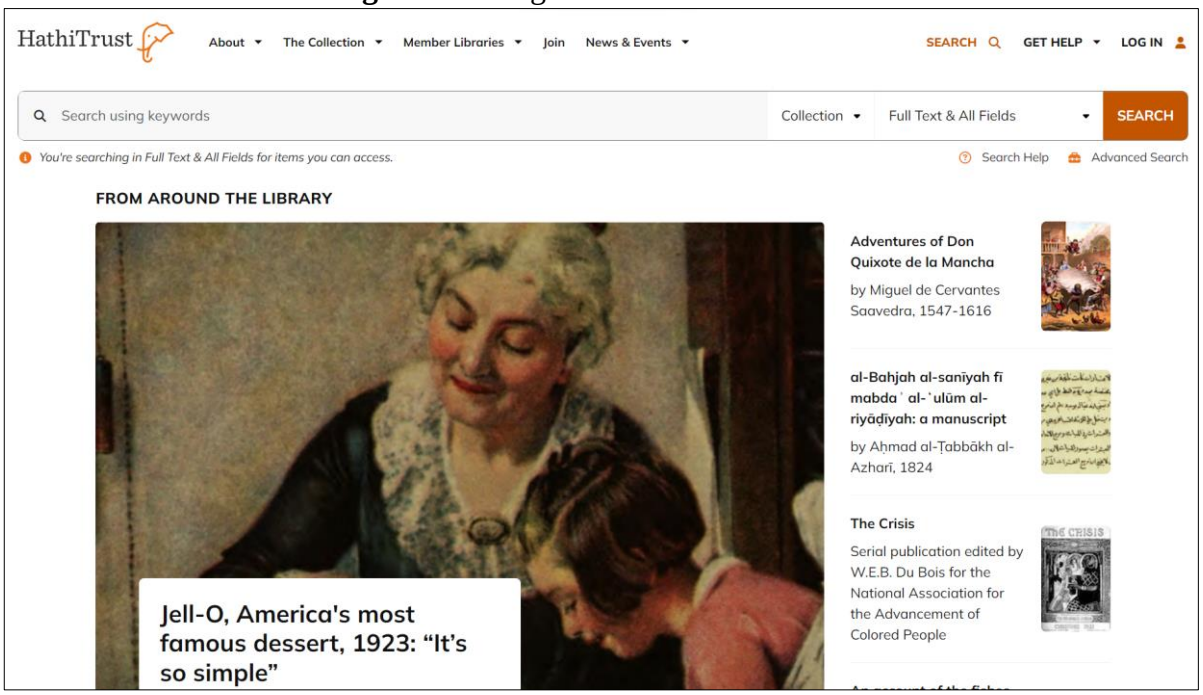
Fonte: *National Center for Biotechnology Information* [2023].

Com mais de 7,5 milhões de artigos arquivados (de 1800 até hoje), os conteúdos em XML são adicionados ao PMC com o depósito de periódicos[177] e manuscritos de autores[178] e por projetos de digitalização[179] via acordos da NLM com editoras, sociedades e agências de financiamento de pesquisas. Em 2020, a NLM passou a usar o PMC para promover o acesso a artigos acerca do coronavírus[180].

- *HathiTrust*[181] – fundada em 2008 e liderada pelas universidades americanas de *Michigan*, de *Indiana* e da Virgínia e o sistema da Universidade da Califórnia, a *HathiTrust* é uma comunidade internacional de bibliotecas acadêmicas e de pesquisa[182] que coleta, arquiva e compartilha coleções digitalizadas de maneira colaborativa.

[177] Disponível em: https://www.ncbi.nlm.nih.gov/pmc/pub/addjournal/. Acesso em: 7 jun. 2023.
[178] Disponível em: https://www.ncbi.nlm.nih.gov/pmc/about/authorms/. Acesso em: 7 jun. 2023.
[179] Disponível em: https://www.ncbi.nlm.nih.gov/pmc/about/scanning/. Acesso em: 7 jun. 2023.
[180] Disponível em: https://www.ncbi.nlm.nih.gov/pmc/about/covid-19/. Acesso em: 7 jun. 2023.
[181] Disponível em: https://www.hathitrust.org/about. Acesso em: 7 jun. 2023.
[182] Disponível em: https://www.hathitrust.org/community. Acesso em: 7 jun. 2023.

Figura 39 – Página inicial da *HathiTrust*.



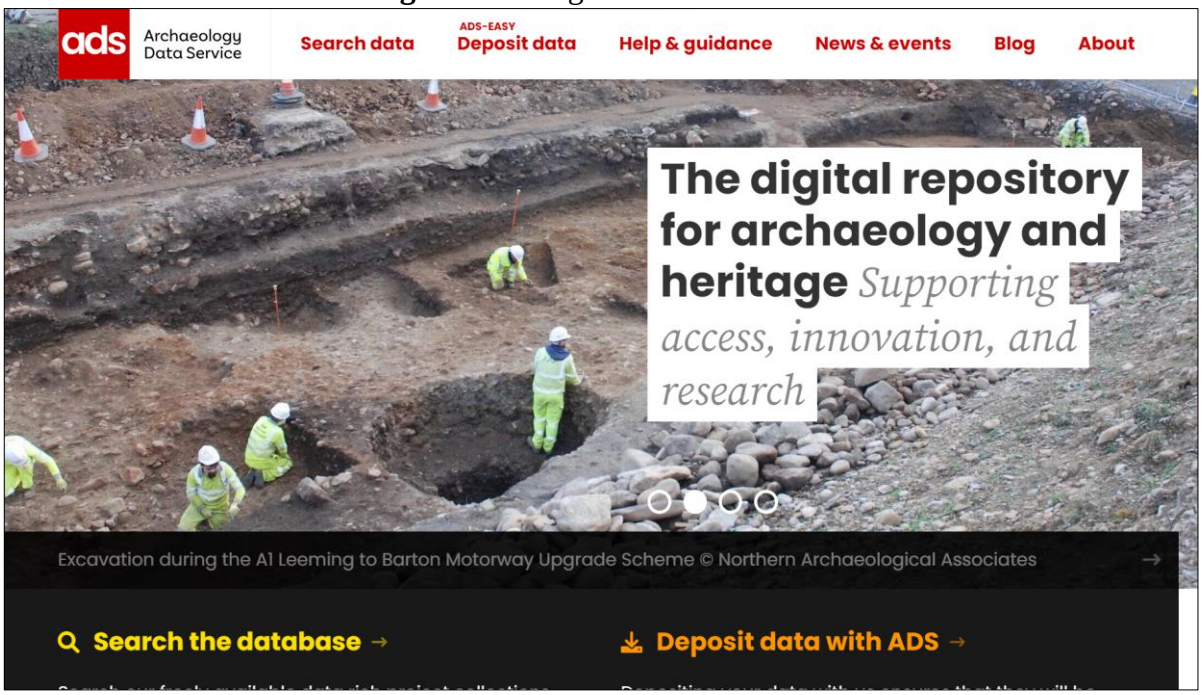
Fonte: *Hathitrust* (c2023).

Um dos seus serviços é a *HathiTrust Digital Library*[183], um repositório certificado pelo CRL[184] que foi criado para fornecer a preservação a longo prazo e, quando possível, o acesso *online* a milhões de materiais digitalizados, sobretudo, livros e periódicos de domínio público ou protegidos por direitos autorais das coleções de instituições parceiras ou via outras fontes, tais como o *Google* e o *Internet Archive*. O acesso aos itens no repositório é definido pela lei norte-americana de direitos autorais[185] e em permissões concedidas por detentores de direitos.

- *Archaeology Data Service* (ADS)[186] – estabelecido em 1996 com financiamento do *Joint Information Systems Committee* (JISC)[187] e do *Arts and Humanities Research Council* (AHRC)[188] e sediada pela Universidade de *York* no Reino Unido, o ADS é o principal repositório digital para dados arqueológicos produzidos por investigações em território britânico.

[183] Disponível em: https://www.hathitrust.org/digital_library. Acesso em: 7 jun. 2023.
[184] Disponível em: https://www.hathitrust.org/hathitrust-certified-as-trustworthy-repository. Acesso em: 3 nov. 2021.
[185] Disponível em: https://www.copyright.gov/title17/. Acesso em: 7 jun. 2023.
[186] Disponível em: https://archaeologydataservice.ac.uk/about.xhtml. Acesso em: 7 jun. 2023.
[187] Disponível em: https://www.jisc.ac.uk/about/who-we-are-and-what-we-do. Acesso em: 7 jun. 2023.
[188] Disponível em: https://ahrc.ukri.org/about/. Acesso em: 7 jun. 2023.

Figura 40 – Página inicial do ADS.



Fonte: *Archaeology Data Service* (c2023).

Com a certificação *CoreTrustSeal*[189], o ADS se centra na preservação a longo prazo de dados digitais em arqueologia que foram depositados aos seus cuidados[190]. Os recursos arquivados e disseminados, incluindo periódicos etc., são de acesso aberto e fornecidos via o seu *site*[191] para a reutilização pelo setor de patrimônio e a comunidade em geral. O ADS é membro da *Digital Preservation Coalition* (DPC)[192] e foi premiado com o DPC *Decennial Award* em 2012 por ser a contribuição mais notável para a preservação digital na última década[193].

- Rede Brasileira de Serviços de Preservação Digital (Cariniana) – gerida pelo IBICT no Brasil, a Cariniana surgiu da exigência de se criar no instituto uma rede de serviços de preservação digital de documentos eletrônicos brasileiros, baseando-se numa estrutura descentralizada usando recursos de computação distribuída para garantir o seu acesso perene por longo prazo.

[189] Disponível em: https://archaeologydataservice.ac.uk/blog/2020/06/we-passed/. Acesso em: 7 jun. 2023.
[190] Disponível em: https://archaeologydataservice.ac.uk/deposit.xhtml. Acesso em: 7 jun. 2023.
[191] Disponível em: https://archaeologydataservice.ac.uk/search.xhtml. Acesso em: 7 jun. 2023.
[192] Disponível em: https://archaeologydataservice.ac.uk/about/accreditation.xhtml. Acesso em: 3 nov. 2021.
[193] Disponível em: https://archaeologydataservice.ac.uk/blog/2012/12/ads-wins-dpc-decennial-award/. Acesso em: 3 nov. 2021.

Figura 41 – Página inicial da Rede Cariniana.

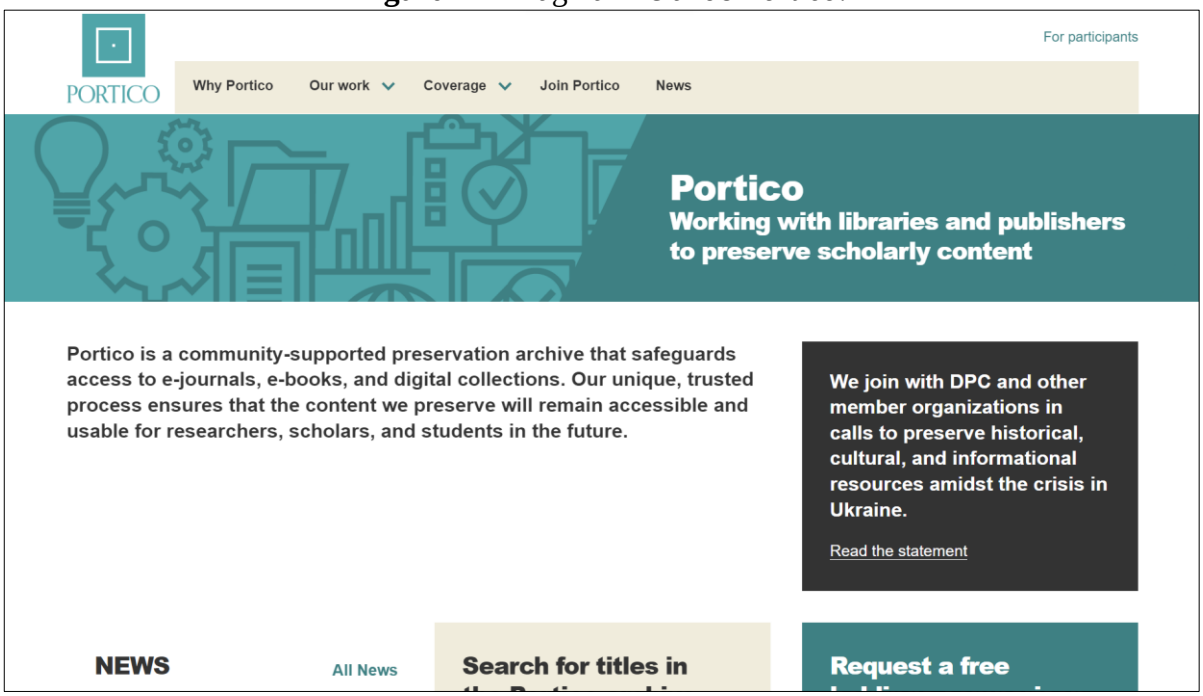
Fonte: Instituto Brasileiro de Informação em Ciência e Tecnologia (c2022).

Em 2012, com o apoio da Financiadora de Estudos e Projetos (FINEP)[194], o IBICT na criação da rede aderiu ao Programa LOCKSS para a preservação das publicações científicas nacionais de acesso aberto (periódicos, teses etc.) hospedadas no OJS e no *DSpace*, dos livros eletrônicos do Portal do Livro Aberto[195] e das monografias da Biblioteca Digital Brasileira de Teses e Dissertações (BDTD)[196] do IBICT. A Cariniana define a participação de universidades brasileiras[197] e outras instituições[198] detentoras desses conteúdos e de sua infraestrutura.

- *Portico*[199] – como um serviço de preservação digital que faz parte da organização norte-americana *ITHAKA*[200] desde 2004, o *Portico* visa preservar periódicos eletrônicos, e-*books* e coleções digitais, garantindo à sua comunidade, o acesso contínuo e utilizável dos seus conteúdos acadêmicos por pesquisadores, alunos, docentes e funcionários no futuro.

[194] Disponível em: http://www.finep.gov.br/a-finep-externo/sobre-a-finep. Acesso em: 7 jun. 2023.
[195] Disponível em: https://livroaberto.ibict.br/Sobre.jsp. Acesso em: 7 jun. 2023.
[196] Disponível em: https://bdtd.ibict.br/vufind/. Acesso em: 7 jun. 2023.
[197] Disponível em: https://cariniana.ibict.br/index.php/parceiros-da-rede/parceiros-integrais. Acesso em: 7 nov. 2021.
[198] Disponível em: https://cariniana.ibict.br/index.php/parceiros-da-rede/parceiros-institucionais. Acesso em: 7 nov. 2021.
[199] Disponível em: https://www.portico.org/. Acesso em: 7 jun. 2023.
[200] Disponível em: https://www.ithaka.org/. Acesso em: 7 jun. 2023.

Figura 42 – Página inicial do *Portico*.



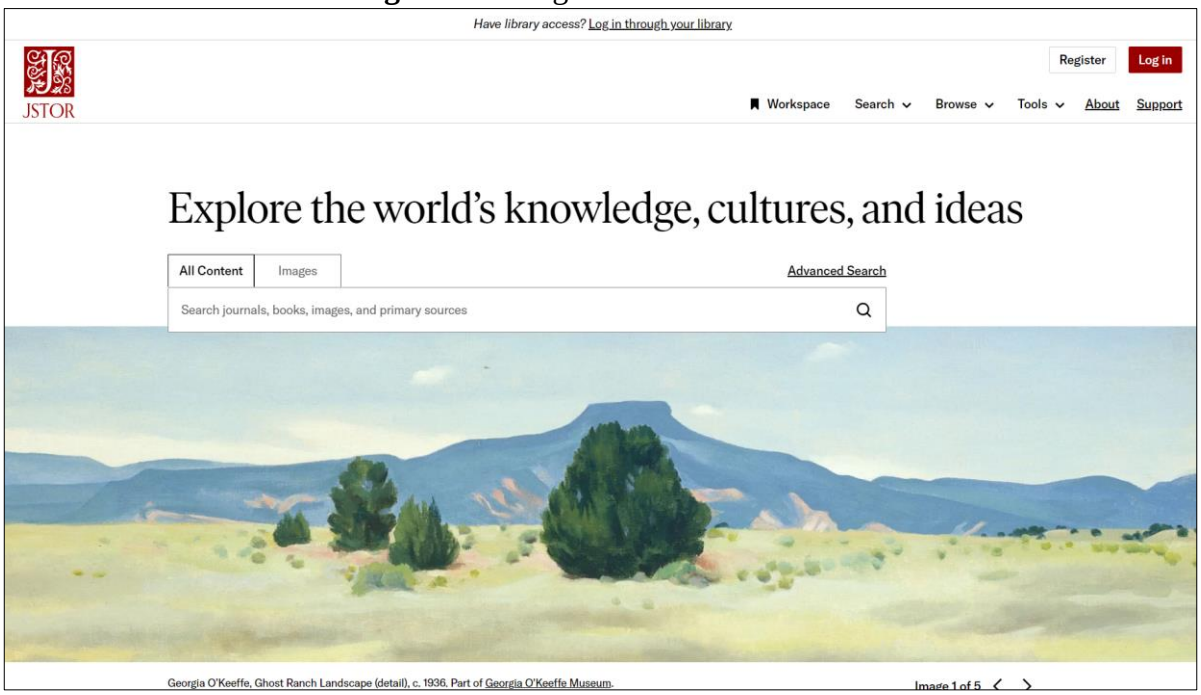
Fonte: *Ithaka* (c2023b).

Com 92 milhões de artigos preservados e tendo a participação de mil bibliotecas e centenas de editoras no mundo (inclusive do Brasil), este arquivo oculto e confiável é certificado pelo CRL[201], adota padrões internacionais, como o OAIS e o METS, possui contratos de licença com editoras para preservar os conteúdos a longo prazo e distribuí-los quando ocorrem eventos que tornem os títulos indisponíveis[202] e ceder o acesso pós-cancelamento/perpétuo a títulos para quais editoras elegeram *Portico* como provedor[203].

- *Journal Storage* (JSTOR)[204] – constituído em 1995 e sendo um serviço que faz parte da *ITHAKA* desde 2009, o JSTOR é uma biblioteca digital que fornece acesso remoto, de baixo custo ou gratuito, a mais de 12 milhões de artigos de periódicos acadêmicos, de 2 milhões de fontes primárias (imagens, relatórios etc.), e de 100 mil *e-books* de editoras renomadas, via uma plataforma única e robusta[205].

[201] Disponível em: https://www.portico.org/news/portico-certified-trustworthy-digital-repository-center-research-libraries/. Acesso em: 7 jun. 2023.
[202] Disponível em: https://www.portico.org/coverage/triggered-content/. Acesso em: 7 jun. 2023.
[203] Disponível em: https://www.portico.org/for-participants/#pca. Acesso em: 7 jun. 2023.
[204] Disponível em: https://about.jstor.org/. Acesso em: 7 jun. 2023.
[205] Disponível em: https://about.jstor.org/platform-features/. Acesso em: 7 jun. 2023.

Figura 43 – Página inicial do JSTOR.



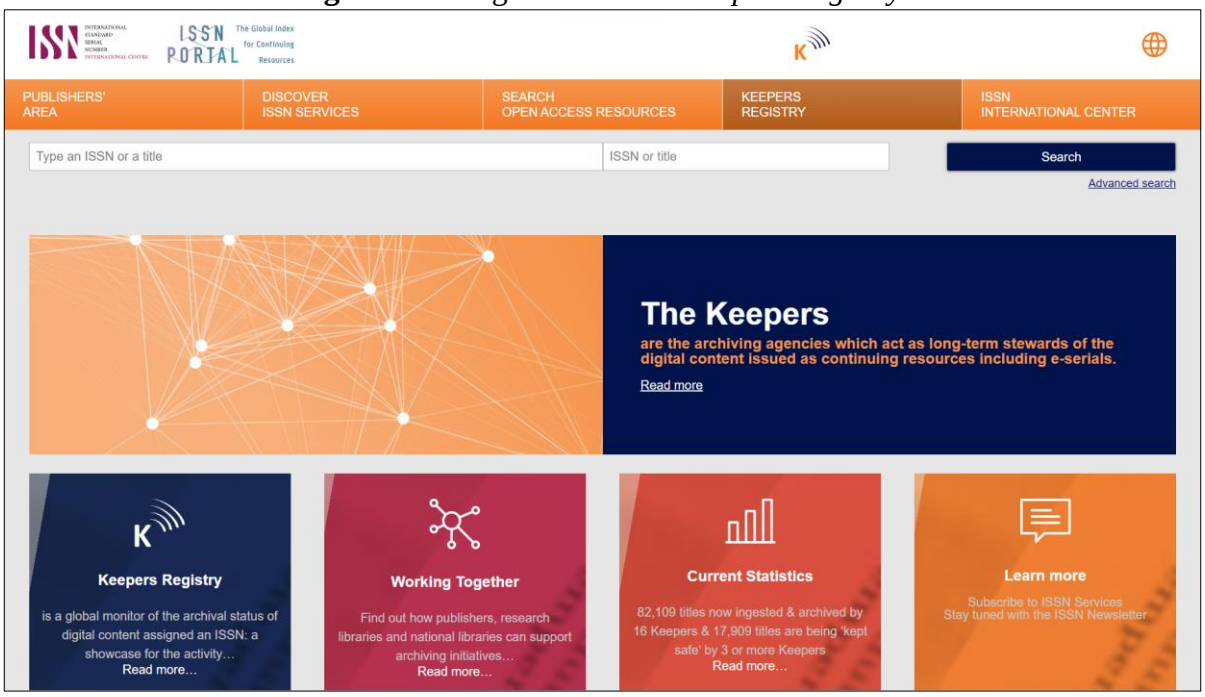
Fonte: *Ithaka* (c2023a).

Com publicações completas de 2.600 periódicos revisados por pares nas áreas das Humanidades, Ciências Sociais e Ciências de vários países e idiomas, os usuários do serviço visualizaram e baixaram 230 milhões de artigos em 2019. As coleções digitais do arquivo JSTOR são preservadas adotando a abordagem e infraestrutura do *Portico*. Devido à crise em curso do COVID-19, o JSTOR aumentou a quantidade de leituras *online* mensais disponíveis para as contas gratuitas[206].

- *Keepers Registry*[207] – custeado pelo JISC de 2016 a 2019 e sendo um serviço *online* do *International Standard Serial Number* (ISSN) *International Centre*, o *Keepers Registry* cede informações sobre a inclusão de periódicos eletrônicos e outros recursos contínuos em iniciativas de preservação a longo prazo (isto é, os *Keepers*), e destaca os periódicos para os quais não existem acordos de arquivamento e, assim, estão "em risco de perda" e precisam ser preservados.

[206] Disponível em: https://about.jstor.org/covid19/. Acesso em: 7 jun. 2023.
[207] Disponível em: https://keepers.issn.org/keepers-registry. Acesso em: 7 jun. 2023.

Figura 44 – Página inicial do *Keepers Registry*.



Fonte: *International Standard Serial Number Portal* (c2023).

> Entre as agências de arquivamento participantes do *Keepers Registry* que fornecem metadados acerca de seus acervos ao ISSN *International Centre* e relatam sobre os materiais arquivados usando o ISSN, estão: a BnF; a KB; a *Library of Congress*; a *British Library*; o *Merritt* na CDL; o ADS, o *Portico*; o *Scholars Portal*; a Rede Cariniana; a *Global* LOCKSS *Network*; o CLOCKSS *Archive*, dentre outras.

Primeiramente, na análise das experiências levantadas de arquivamento de periódicos eletrônicos, nota-se que os serviços se concentram na preservação a longo prazo de periódicos e de outras publicações científicas ou acadêmicas, de acesso aberto e por assinatura, garantindo à sua comunidade (incluindo pesquisadores, professores, alunos e funcionários de instituições patrimoniais e de ensino e pesquisa) o acesso contínuo e usável de conteúdos técnico-científicos produzidos em diferentes áreas de estudo[208]. Para tal, as iniciativas identificadas possuem uma

[208] É importante compreender a distinção terminológica entre acesso contínuo e preservação a longo prazo, pois estas diferenças levam a diversos tipos de serviços e soluções para arquivamento de periódicos eletrônicos; isto posto, o acesso contínuo (*continuing access*) "[...] aplica-se apenas a periódicos de assinatura e garante acesso a longo prazo para seus assinantes [...]" e, em contra partida, a preservação a longo prazo (*long-term preservation*) "[...] aplica-se a conteúdo aberto e subscrito.", conforme *Digital Preservation Coalition* (c2015, não paginado, tradução nossa). Por sua vez, em Jones (2007, não paginado, tradução nossa) o termo acesso perpétuo (*perpetual access*) também "[...] é mais comumente associado a cláusulas de licença de periódico eletrônico destinadas a garantir o acesso contínuo ao material assinado em determinadas circunstâncias, incluindo o pós-cancelamento.", e a preservação a longo prazo diz respeito "[...] aos processos e procedimentos necessários para garantir que o conteúdo permaneça acessível no futuro, independentemente de quaisquer mudanças técnicas ou organizacionais."

forte colaboração entre bibliotecas de pesquisa, universidades e editoras acadêmicas, junto com o financiamento de certas organizações, como o FINEP na Rede Cariniana, e o JISC no *Keepers Registry*. Dentre os exemplos, destacamos o LOCKSS que se baseia no modelo de preservação digital distribuída onde reflete, em certa medida, o requisito para preservação digital de manter a recuperação dos documentos digitais através de uma política de *backup*[209], prezando-se pela replicação do documento (e seus metadados) em local físico separado a fim de assegurar tanto o acesso como a restauração confiável, integra e segura dos dados (INNARELLI, 2009, 2014).

Segundo Souza *et al*. (2012, p. 69) a preservação digital distribuída (*distributed digital preservation*) remete a "[...] uma estratégia focada na distribuição de cópias dos conteúdos em locais geograficamente dispersos, de forma segura e em que seja possível garantir o acesso em longo prazo." tendo em vista as perdas ou danos nos dados por desastres, falhas humanas etc., consistindo num modelo de rede onde "[...] propõe que várias instituições armazenem, ofereçam acesso e criem cópias digitais atualizadas." (MÁRDERO ARELLANO, 2012, p. 84). Skinner e Schultz (c2010) indicam alguns princípios que devem ser seguidos em um modelo deste tipo, como manter as cópias do mesmo conteúdo de modo disperso (de 120 a 200 quilômetros entre os locais de preservação); e replicar o conteúdo por pelo menos três vezes. Além do mais, como notado antes, a Rede Cariniana é o principal exemplo nacional que ilustra o modelo de gestão de preservação distribuída o qual adota o LOCKSS como uma solução em preservação digital distribuída que atende ao modelo OAIS e permite criar e gerenciar redes de preservação digital sem muita complexidade e altos custos (MÁRDERO ARELLANO, 2012; SOUZA *et al*., 2012).

Outra analogia possível no modelo de preservação digital distribuída do LOCKSS está na estratégia de preservação digital de replicação. Consistindo na criação de cópias duplicadas de dados digitais em um ou mais sistemas, a replicação (*replication*) prevê que os dados digitais sobrevivam se forem replicados em vários locais, diferentemente se existissem como uma única cópia num só local tornando-se altamente suscetíveis a falhas de *software*/*hardware*, alteração intencional/acidental e tragédias ambientais; porém, por estarem localizados em muitos lugares,

[209] *Backup* são "[...] cópias adicionais de um recurso digital feitas para proteger contra perda devido à destruição ou corrupção não intencional do conjunto principal de recursos digitais." e "o atributo essencial de uma cópia de segurança é que as informações nela contidas podem ser restauradas caso o acesso à cópia principal seja perdido." (NATIONAL DIGITAL STEWARDSHIP ALLIANCE, 2013, não paginado, tradução nossa).

os dados replicados podem trazer dificuldades de migração[210] e de refrescamento (*refreshing*)[211] ou, mesmo, de versionamento e de controle de acesso (DIGITAL PRESERVATION, 2022, não paginado, tradução nossa). A replicação também é um conjunto de estratégias de preservação digital, incluindo a adesão ao consórcio LOCKSS e a cópia *bit-stream* (ou *bit-stream copying*, conhecida como "*backup up data*" que se refere ao processo de realizar cópias duplicadas exatas de um recurso informacional digital), o qual destinam-se a "[...] modificar a longevidade dos documentos digitais, mantendo sua autenticidade e integridade através do processo de cópia e pelo uso do número de locais de armazenamento." (NAJAR; WANI, 2019, p. 9, tradução nossa).

Em segundo, os serviços identificados se comprometem a manter o acesso contínuo as publicações de assinatura arquivadas após serem acionadas devido a situações que levam a sua indisponibilidade, assegurando o seu acesso por bibliotecas assinantes ou ex-assinantes (e seus usuários alunos, pesquisadores etc.) após o término do contrato de licenciamento com a editora, ou quando o periódico deixa de ser publicado pela editora além de outras situações específicas. Conforme Beagrie (c2013, p. 36, tradução nossa) o acesso contínuo (*continuing access*, às vezes chamado de acesso perpétuo/pós-cancelamento – *post-cancellation/perpetual access* –) alude "[...] ao direito do assinante e de seus usuários de ter acesso permanente e contínuo a materiais eletrônicos que já foram alugados e pagos pelo assinante de uma editora.", sendo que "[...] o assinante/licenciado e a editora/licenciadora, ambos parte da licença, precisam acordar os termos para a concessão de direitos de acesso contínuo ao assinante." Assim, como exemplos:

- No arquivo CLOCKSS, se um conteúdo de periódico científico mantido no arquivo desaparecer (ou estiver prestes a desaparecer) da *Web*, o CLOCKSS o "disparará" para acesso aberto (CLOCKSS ARCHIVE, c2022);
- No *e-Depot* da KB, a KB fornece acesso *in loco* a todas as publicações arquivadas das editoras e, além dos mais, em situações que os títulos não estão mais disponíveis na

[210] De acordo com *Digital Preservation Coalition* (c2015, não paginado, tradução nossa) a migração compreende "[...] um meio de superar a obsolescência tecnológica, transferindo recursos digitais de uma geração de *hardware* ou *software* para a próxima." com o objetivo de "[...] preservar o conteúdo intelectual dos objetos digitais e manter a capacidade dos clientes de recuperá-los, exibi-los e, de outra forma, utilizá-los diante de uma tecnologia em constante mudança."; além disto, esta estratégia de preservação digital "[...] difere do refrescamento das mídias de armazenamento, uma vez que nem sempre é possível produzir uma cópia digital exata ou replicar características e aparência originais e, ao mesmo tempo, manter a compatibilidade do recurso com a nova geração de tecnologia.".

[211] Em Santos e Flores (2017b, p. 35) o refrescamento consiste em "[...] transferir os documentos digitais fixados em um determinado suporte, o qual é considerado obsoleto, para outro suporte considerado atual.", centrando-se "[...] na preservação do objeto físico, ou seja, preserva a forma física do documento digital evitando que o suporte no qual o documento digital está armazenado seja danificado." Para os autores o refrescamento deve atuar de forma tanto complementar apoiando demais estratégias de preservação digital (a migração, emulação etc.) como conjunta ao monitoramento da obsolescência das mídias de armazenamento, dos formatos de arquivo e da degradação física.

editora ou em qualquer outra fonte, o acesso pode ser estendido para um público mais amplo (KONINKLIJKE BIBLIOTHEEK, [c2021]); e

- No *Portico*, os livros e periódicos acionados são acessíveis a todos os seus participantes, enquanto que as coleções digitais acionadas estão disponíveis para todas as bibliotecas que adquiriram as coleções anteriormente (o conteúdo de acesso aberto está disponível para todos), e a editoras podem nomear o serviço como uma fonte de acesso perpétua através da qual seus ex-assinantes (que também devem ser bibliotecas participantes do *Portico*) podem solicitar acesso contínuo ao conteúdo (ITHAKA, c2022b).

Como indicado nos serviços do CLOCKSS, do *e-Depot* da KB e do *Portico*, existem várias situações que levam a indisponibilidade do conteúdo arquivado de periódicos científicos. No CLOCKS, tais situações são tidas de "eventos de gatilho" (*trigger events*)[212] e podem incluir quando, por exemplo, a editora não está mais no mercado ou, até, não está mais no negócio de publicar conteúdo ou fornecer acesso as edições anteriores do conteúdo, e não há interesses de sucessor ou transferência de direitos (CLOCKSS ARCHIVE, c2022). Já no *e-Depot* da KB, os "eventos de gatilho" (ou quando os títulos não estão mais disponíveis na editora ou em qualquer outra fonte) ocorrem após falha catastrófica e contínua da plataforma de distribuição de uma editora, ou quando uma editora interrompe as operações ou, ainda, não oferece mais edições anteriores, e quando uma editora deixa de publicar um título (KONINKLIJKE BIBLIOTHEEK, [c2021]). Também no *Portico*, os "eventos de gatilho" fazem com que os títulos não estejam mais disponíveis pela editora ou por um sucessor e incluem, por exemplo, falha catastrófica e permanente da plataforma de distribuição de uma editora além de 90 dias (ITHAKA, c2022b).

Por fim, os serviços de preservação a longo prazo e de acesso contínuo para periódicos eletrônicos que foram analisados e descritos no trabalho se baseiam, também, em certificações e padrões internacionais para sistemas de repositório digital confiáveis de arquivamento. O *e-Depot* da KB e o ADS, por exemplo, possuem a certificação *CoreTrustSeal* o qual, consoante *CoreTrustSeal* (c2022) e Moore (2020), garante que os resultados de pesquisas depositados em repositórios de dados sejam localizáveis, acessíveis e disponíveis de forma fiável e consistente em formatos adequados, além dos dados serem referenciados de modo eficaz e sustentável ao longo do tempo. Já o *Scholars Portal*, o *CLOCKSS*, a *HathiTrust Digital Library* da *HathiTrust* e o *Portico* têm a certificação CRL por meio da TRAC que, de acordo com *Center for Research*

[212] Evento de gatilho/acionador (*trigger event*) é uma terminologia que, em conformidade com Beagrie (c2013, p. 38, tradução nossa), adota-se "[...] quando condições específicas referentes a um título de periódico eletrônico e sua distribuição contínua aos usuários são atendidas." ou, melhor dizendo, ocorre "se o periódico não estiver mais disponível aos usuários da editora ou de qualquer outra fonte por uma variedade de razões [...]" e "[...] estes podem acionar o acesso dos usuários através de um arquivo onde o periódico eletrônico pode ser preservado digitalmente."

Libraries ([2022?]), contém métricas baseadas no modelo de referência OAIS que auxiliam no julgamento de um repositório nas áreas de gestão de objetos digitais, tecnologias, infraestrutura técnica, segurança etc., assegurando a capacidade deste em preservar o conteúdo digital, como de periódicos eletrônicos, sendo que a ISO 16363 é uma revisão da lista de verificação TRAC.

Estas e outras ferramentas para avaliação de repositórios, como os *Ten Principles*[213] do CRL, oferecem muitos benefícios a um repositório para conteúdo de periódicos eletrônicos e as suas partes interessadas. As certificações cedem aos depositantes a confiança de que os seus dados estão sendo armazenados, geridos, arquivados e compartilhados segundo as boas práticas de preservação digital; dão as agências de fomento garantias de que os resultados de pesquisas financiadas serão preservados e disseminados hoje e futuramente; propiciam aos sistemas um modo de melhorar a qualidade e a transparência dos seus processos, operações e políticas como de aumentar a conscientização e a conformidade com os padrões estabelecidos; e permitem que os dados criados e acessados por pesquisadores permaneçam úteis e significativos no tempo (CENTER FOR RESEARCH LIBRARIES, [2022?]; CORETRUSTSEAL, c2022). Ademais, o *e-Depot* da KB, o *Merrit* e o *Portico* (e o JSTOR que usa a sua estrutura) adotam do mesmo modo o OAIS, e padrões de metadados METS, MODS e PREMIS que, para Formenton (2015), são aplicáveis à preservação de documentos digitais e eletrônicos, como periódicos científicos.

2.1.5 Preservação digital de *e-mail*

Como uma estratégia operacional de preservação digital, a preservação de *e-mail* incide sobre "[...] um tipo de dado complexo que consiste em múltiplas conversas envolvendo muitas pessoas distintas." que pode "[...] incluir anexos (que podem ser de qualquer outro tipo de dados ou formato de arquivo), pode ser grande em tamanho e pode ser um desafio capturá-lo de forma eficaz." (DIGITAL PRESERVATION COALITION; ARTEFACTUAL SYSTEMS, c2021, p. 2, tradução nossa). Constituindo um arquivo digital pessoal, como demais conteúdos digitais criados por indivíduos em seu cotidiano (incluindo *websites*, interações de mídia social etc.), e

[213] Disponível em: https://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re. Acesso em: 7 jun. 2023.

um documento arquivístico[214] (CONSELHO NACIONAL DE ARQUIVOS, 2012; DIGITAL PRESERVATION COALITION, [2018b]), as mensagens de correio eletrônico é um dos meios de comunicação atuais mais usados para interações pessoais ou profissionais que podem prover evidências valiosas de atividades e decisões, segundo *Digital Preservation Coalition* ([2018c]).

Contudo, a preservação de *e-mail* é também um tema carente de pesquisas e iniciativas no Brasil. Um dos poucos exemplos nacionais encontrados na literatura especializada refere-se ao estudo de Luz e Maringeli (2018) o qual traz um relato da estruturação e da definição para a implantação da política de preservação digital da Pinacoteca de São Paulo composta de quatorze partes, sendo uma delas a gestão de correio eletrônico que estabelece a maneira pelo qual os *e-mails* serão tratados por esta instituição como documentos arquivísticos. Assim, de acordo com os autores, nesta política os *e-mails* são considerados documentos arquivísticos quando apoiam as ações e os processos de tomada de decisões, fornecem evidência no caso de litígio, protegem os interesses da instituição e os direitos de seus funcionários e usuários, asseguram as atividades de pesquisa, desenvolvimento e inovação, assim como mantêm a memória corporativa coletiva.

Ferramentas

Respaldando-se em *Coolutils* (c2022), *Council on Library and Information Resources* (2018), *Encryptomatic* LLC (c2022), Hangal, Lam e Heer (2011), *North Carolina Department of Natural and Cultural Resources* ([2018]), Schneider *et al*. (2017, 2019), Simpson (2016), *Smithsonian Institution Archives* ([2022?]), *Stanford Libraries* ([c2022c]), *Stanford University* ([2012?], c2022b) e *Stimulus Software* (c2020), apontamos algumas ferramentas de *software*[215] notáveis para a seleção e coleta, administração, preservação, acesso e pesquisa de coleções de *e-mail* com valor histórico e cultural, que auxiliam instituições patrimoniais e universidades na

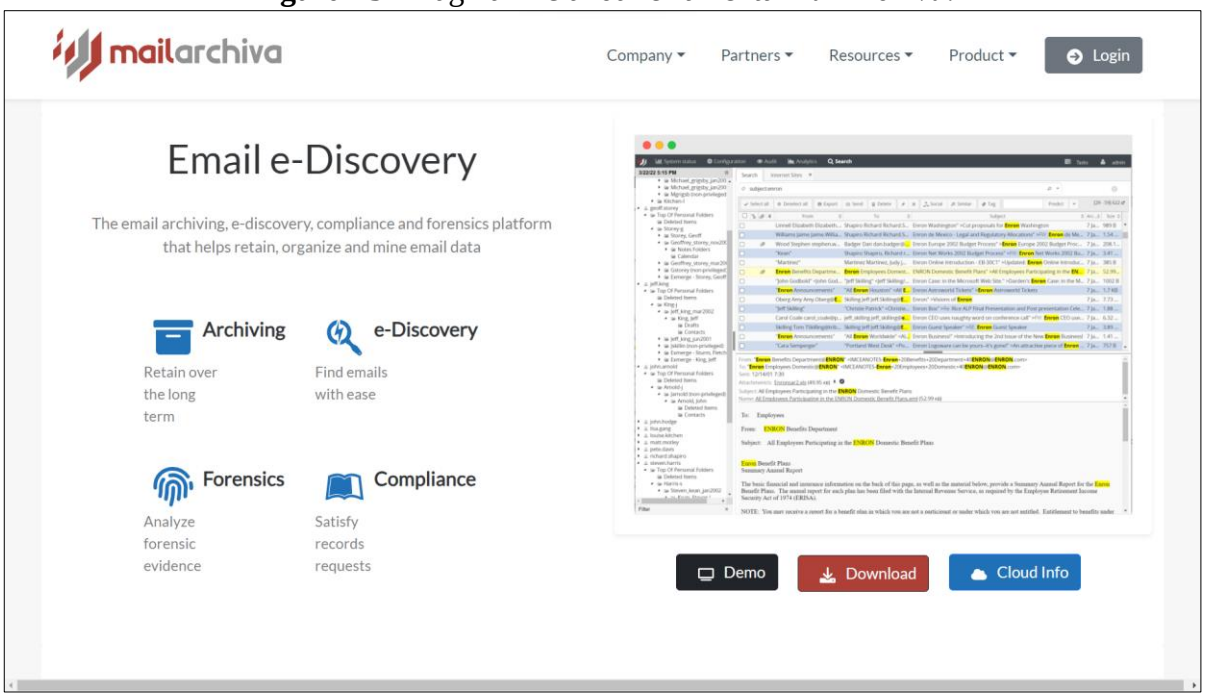
[214] Documento arquivístico é o "documento produzido (elaborado ou recebido), no curso de uma atividade prática, como instrumento ou resultado de tal atividade, e retido para ação ou referência." (CONSELHO NACIONAL DE ARQUIVOS, 2020, p. 24) ou o "[...] documento produzido, recebido ou acumulado por um órgão ou unidade no exercício de suas funções e atividades, para fins de prova, informação ou fonte de pesquisa." (UNIVERSIDADE ESTADUAL DE CAMPINAS, 2011, não paginado). Por sua vez, documento arquivístico digital diz respeito ao "[...] documento arquivístico codificado em dígitos binários, produzido, tramitado e armazenado por sistema computacional, que pode ser produzido no contexto tecnológico digital (documentos nato-digitais) ou obtido a partir de suportes analógicos (documentos digitalizados)." (UNIVERSIDADE ESTADUAL DE CAMPINAS, 2011, não paginado), e também ao "documento digital reconhecido e tratado como um documento arquivístico.", sendo que o termo documento digital se refere a "informação registrada, codificada em dígitos binários, acessível e interpretável por meio de sistema computacional." (CONSELHO NACIONAL DE ARQUIVOS, 2020, p. 25).

[215] Deve-se levar em conta que grande parte dos *softwares* identificados está em desenvolvimento ativo e os estudos citados ocorreram ao longo de vários anos, assim, pode haver inconsistências na forma como as ferramentas são descritas pelas diferenças entre versões ou, mesmo, pelas constantes atualizações e melhorias de funcionalidades.

preparação de *e-mails* para questões de requisições e conformidade legais (incluindo a triagem de informações protegidas ou restritas) como solicitações de liberdade de informação, a saber:

- *MailArchiva* – *software* de arquivamento, *e-Discovery*[216] e análise de evidências de *e-mail*, desenvolvido pela *Stimulus Software*, que ajuda reter, organizar e extrair dados de arquivos grandes de *e-mail*.

Figura 45 – Página inicial da ferramenta *MailArchiva*.



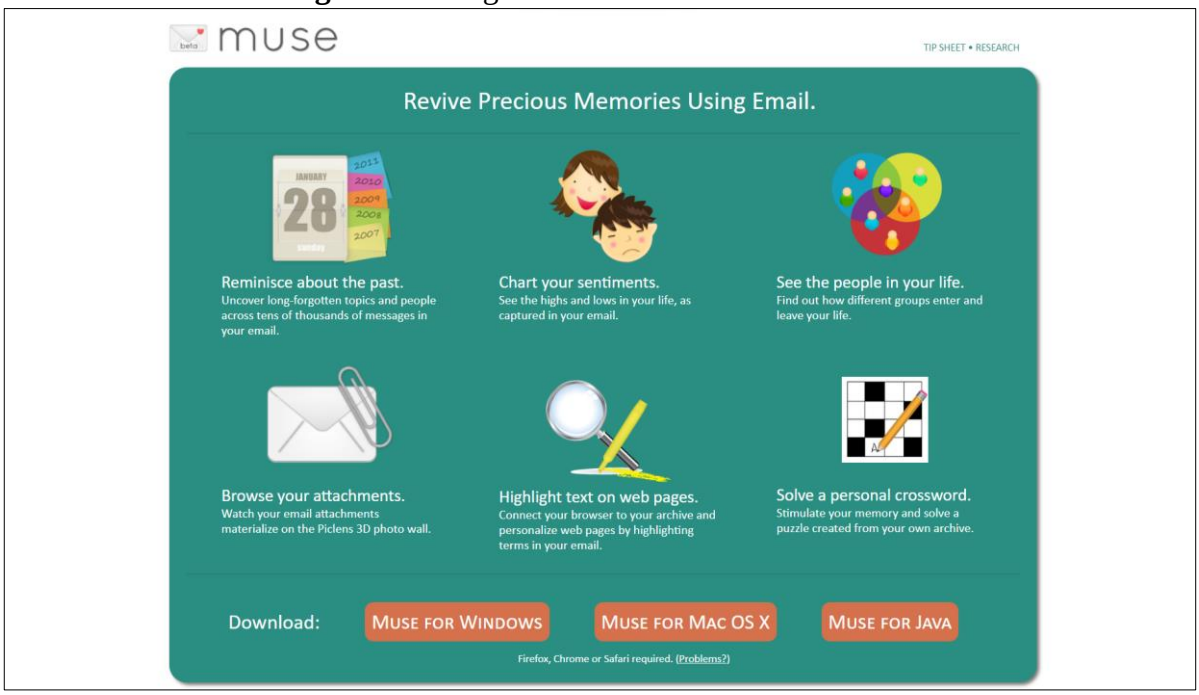
Fonte: *Stimulus Software* ([2023]).

Esta solução escalonável possibilita, por exemplo, pesquisar *e-mails* (e dentro do conteúdo do anexo) com facilidade e precisão no *Outlook* ou usando um navegador da *Web* (*browser*); gerar relatórios dos resultados da pesquisa; analisar a autenticidade dos dados de *e-mail*; compactar e desduplicar dados de *e-mail* e anexos; garantir o cumprimento da legislação; e exportar em massa dados para vários formatos populares. O *MailArchiva* é gratuito só para menos de dez caixas de *e-mail* (*mailboxes*).

- *Memories USing Email* (MUSE) – *software* de mineração de dados de coleções de *e-mail*, criado pela *Stanford University*, que ajuda os usuários a explorar, analisar, extrair e visualizar seus próprios arquivos de *e-mail* (com até cinquenta mil mensagens) usando facetas de navegação, como sentimentos, pessoas, grupos de contatos (família, colegas etc.), meses e anos.

[216] *E-discovery*, também conhecida como *electronic discovery*, concerne a quaisquer métodos de busca, pesquisa, localização e obtenção de informações eletrônicas a fim de usá-las como evidências num processo judicial, como litígios, investigações governamentais ou pedidos de acesso à informação (ELECTRONIC DISCOVERY, 2019).

Figura 46 – Página inicial da ferramenta MUSE.

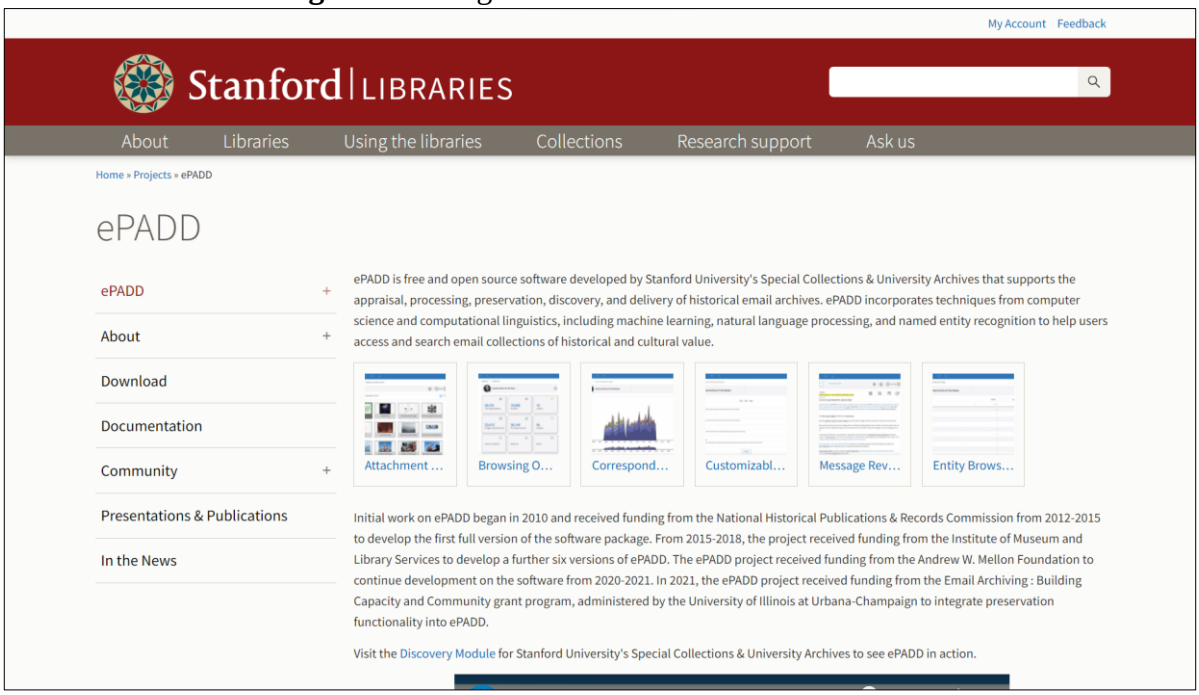


Fonte: *Stanford University* ([2023?c]).

Com um léxico personalizado abrangendo categorias de termos que expressam, por exemplo, emoções, família, saúde e eventos da vida (mortes, casamentos etc.), este sistema analisa o conteúdo do arquivo e gera “pistas” que ajudam a despertar e a redescobrir as memórias dos usuários, como a atividade de comunicação e interação com grupos de indivíduos, a ocorrência de palavras sentimentais, e as imagens de seus anexos de *e-mails*. Hoje, a *Stanford* absorveu aspectos do MUSE no *software* ePADD.

- *Email: Process, Appraise, Discover, Deliver* (ePADD) – *software* gratuito e de código aberto, desenvolvido pela *Stanford University*, que suporta processos de arquivamento em torno da avaliação e seleção por doadores ou curadores, processamento, descoberta e entrega/acesso de arquivos de *e-mail*.

Figura 47 – Página inicial da ferramenta ePADD.

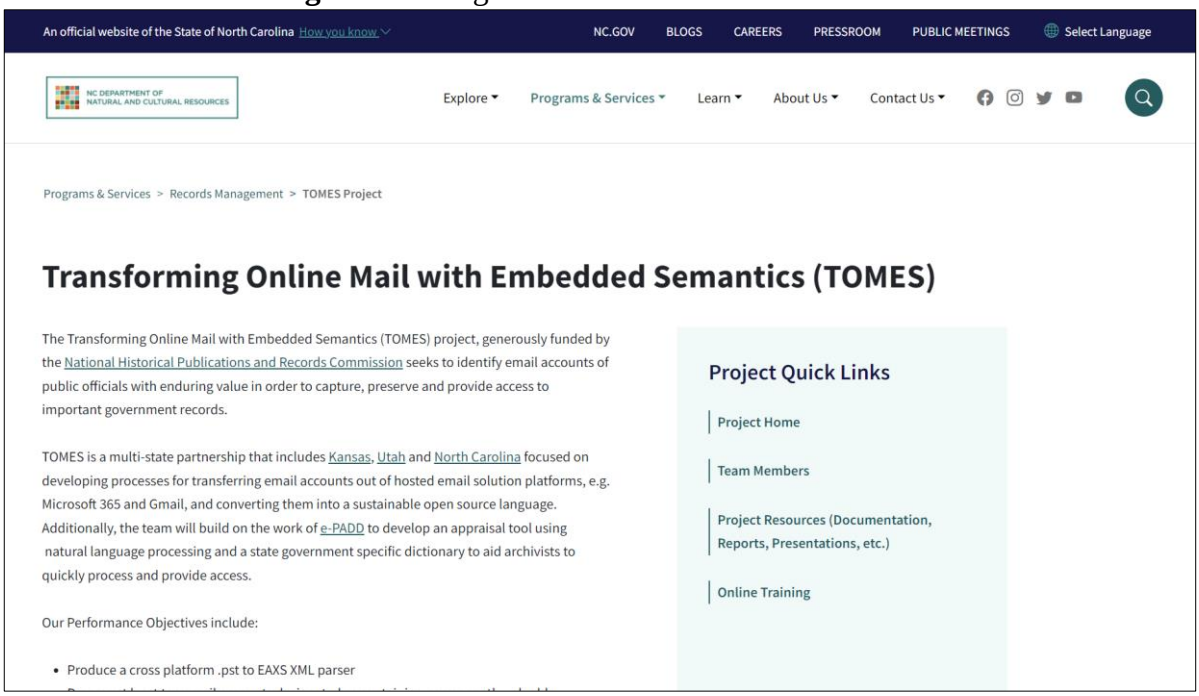


Fonte: *Stanford Libraries* ([c2022c]).

Este sistema, por exemplo, permite a navegação de todas as imagens anexas em um único lugar; o agrupamento de vários endereços de *e-mail* relativas às mesmas pessoas; anotações para as mensagens; e o fornecimento de modelos para criar pesquisas complexas na forma de léxico (termos predefinidos, mas personalizáveis). O ePADD não permite a saída (*output*) de muitos tipos de metadados.

- Ferramenta *Transforming Online Mail with Embedded Semantics* (TOMES) – *software* de código aberto, elaborado pelo projeto TOMES, que ajuda os arquivistas a processar contas de *e-mail* grandes e proporcionar acesso.

Figura 48 – Página inicial da ferramenta TOMES.



Fonte: *North Carolina Department of Natural and Cultural Resources* ([2023?]).

Usando a tecnologia de processamento de linguagem natural (*Natural Language Processing* – NLP) para "etiquetar" (*tag*) as informações especificadas em uma conta de *e-mail*, e dicionários específicos em XML do governo norte-americano, a ferramenta TOMES permite identificar mais facilmente tópicos de interesse, entidades nomeadas e, ainda, informações pessoais, confidenciais e sensíveis em *e-mails* que contêm registros públicos a serem tornados acessíveis. Os recursos de assistência de NLP do TOMES podem exigir conhecimentos especializados.

- *Archivematica* – sistema de código aberto, desenvolvido pela *Artefactual Systems*, que cede um conjunto integrado de ferramentas de *software* para processar objetos digitais (inclusive formatos de *e-mail*) desde a ingestão até o acesso segundo a norma OAIS e implantar planos de preservação.

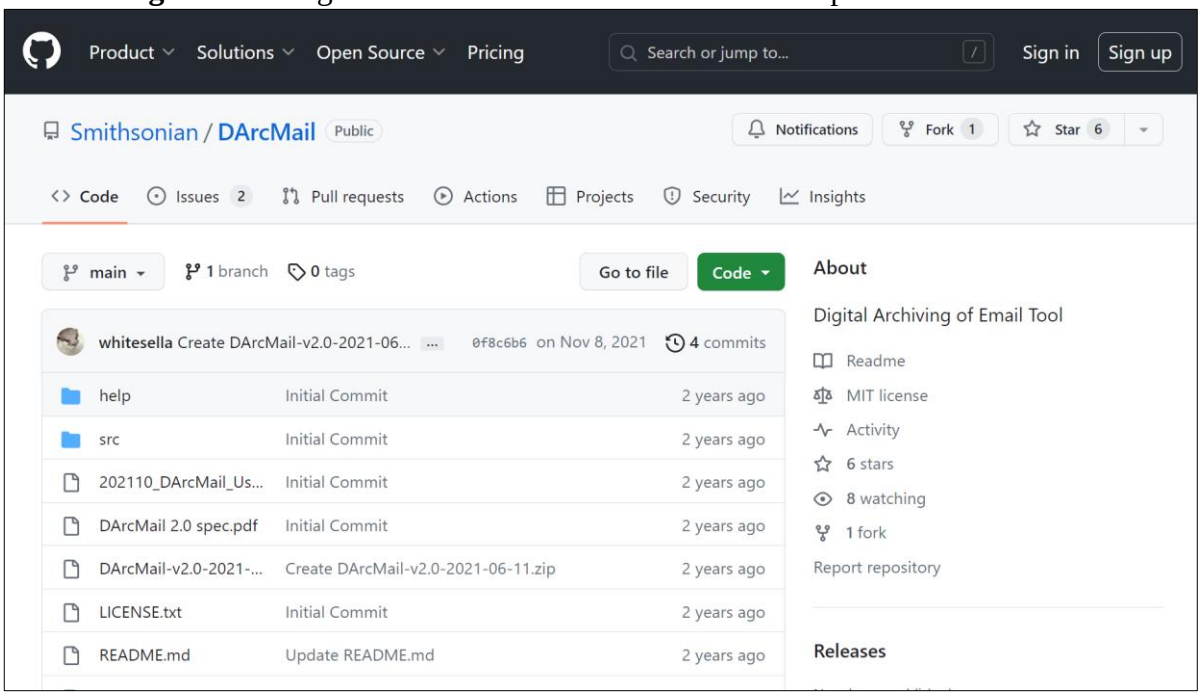
Figura 49 – Página inicial da ferramenta *Archivematica*.

Fonte: *Artefactual Systems* ([2023?]).

Empregando o METS, o PREMIS etc. para registrar e rastrear metadados, os usuários podem monitorar e controlar a ingestão e preservação de micro serviços via um painel de controle baseado na *Web*. O sistema possui falha em separar e preservar separadamente anexos, permite a normalização para um pequeno número de formatos nativos de *e-mail*, e carece de funcionalidades de análise de dados.

- *Digital Archives of Email* (DArcMail) – *software* de preservação de *e-mail* de código aberto, criado pelos *Smithsonian Institution Archives* (SIA), para realizar a preservação, processamento e acesso de contas de *e-mail*. A implementação do DArcMail é gratuita e independente de plataformas operacionais.

Figura 50 – Página inicial da ferramenta DArcMail na plataforma *GitHub*.

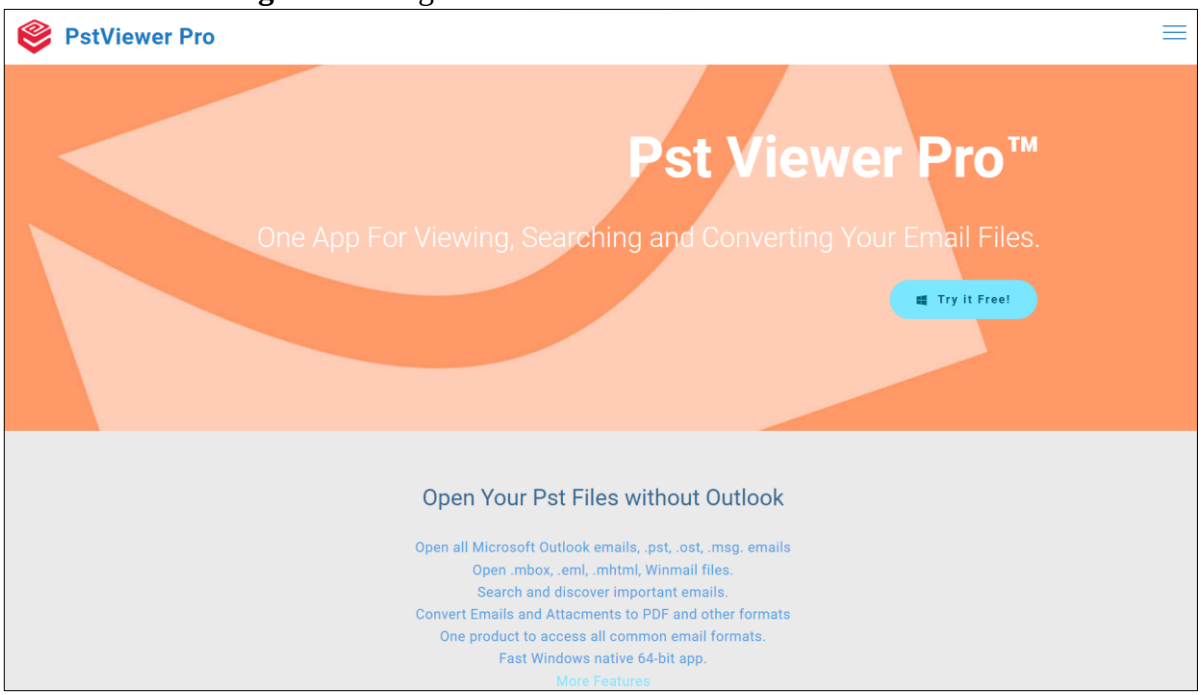


Fonte: *Smithsonian Institution Archives* (2021).

Esta aplicação permite, por exemplo, que os usuários interajam com *e-mails* de forma individual, em grupo etc.; as mensagens de *e-mail* podem ser buscadas, filtradas, rastreadas e visualizadas; e os anexos podem ser pesquisados, visualizados e separados dos *e-mails*. Igualmente ao ePADD, o DArcMail não suporta manter um histórico completo de processamento dos conteúdos de *e-mails*.

- PST *Viewer Pro* – *software* sob licença criado pela *Encryptomatic* LLC para visualizar, pesquisar, descobrir e converter arquivos de *e-mail*.

Figura 51– Página inicial da ferramenta PST *Viewer Pro*.

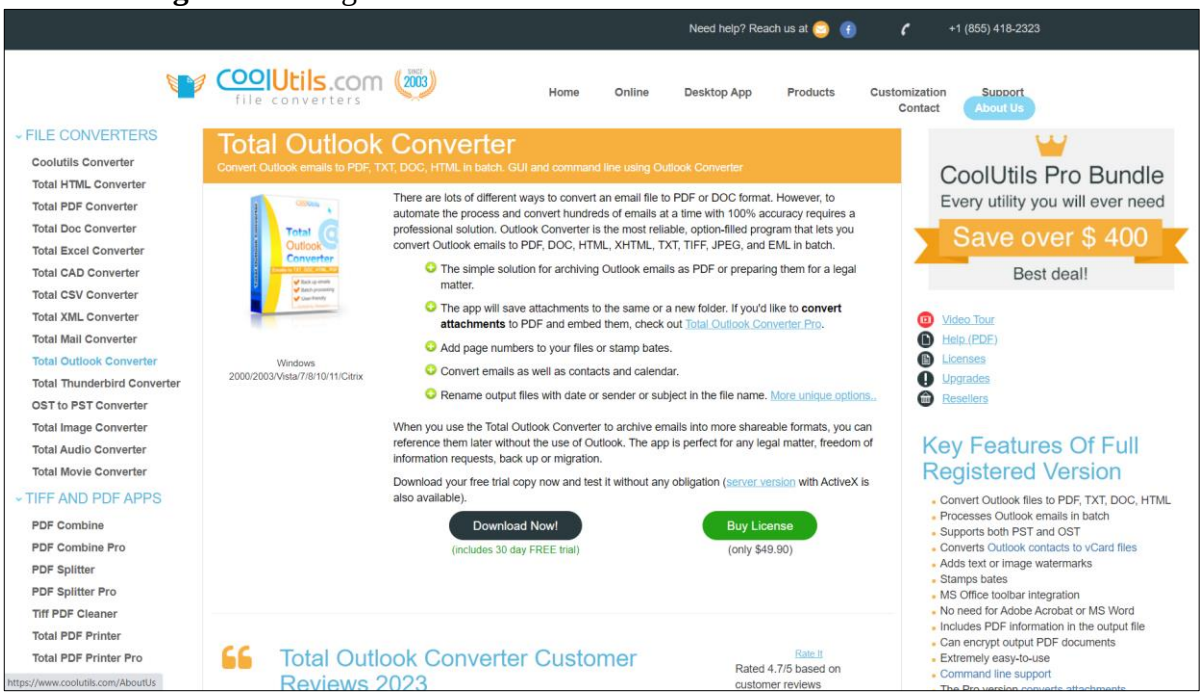


Fonte: *Encryptomatic* LLC ([c2023?]).

Tal solução permite, por exemplo, abrir mensagens em formatos do *Microsoft Outlook* ignorando a proteção por senha nos arquivos; ler *e-mails* armazenados em um estado somente leitura; converter/exportar *e-mails* e anexos para/em *Portable Document Format* (PDF) e outros formatos comuns de *e-mail*; duplicar *e-mails* selecionados; processar em massa milhares de *e-mails* para operações de impressão ou exportação; exibir arquivos formatados como notas, tarefas, calendário e contatos do *Outlook*; e suporte para vários idiomas (incluindo o português).

- *Total Outlook Converter Pro* – *software* sob licença desenvolvido pela *CoolUtils* para converter os *e-mails* (e contatos, calendário etc.) do *Outlook* para PDF e outros tipos de formatos.

Figura 52 – Página inicial da ferramenta *Total Outlook Converter Pro*.



Fonte: *CoolUtils* (c2022).

> Esta solução permite, por exemplo, pesquisar *e-mails* por remetente, data ou palavras-chave e filtrá-los por tamanho; controlar totalmente as permissões para abrir, visualizar, imprimir ou editar as cópias em PDF dos *e-mails;* e adicionar data, hora etc. a documentos. Porém, o aplicativo pode vir a travar quando atua com arquivos *Personal Folders File* (ou *Personal Storage Table* – PST)[217] grandes, como em *Library of Virginia* (c2016) no projeto *Kaine Email* que optou por uma combinação de suas funcionalidades com o PST *Viewer Pro*[218].

Iniciativas em universidades, bibliotecas nacionais e estaduais, e arquivos estaduais

A partir de Baker, Butler e Green (2012), Boudrez (2006), *British Library* (2011), Croft (2016), *Council on Library and Information Resources* (2018), *Library of Virginia* (c2016), *North Carolina Department of Natural and Cultural Resources* ([2018]), Prom (c2011, c2019) e Schneider *et al*. (2019), verificamos diferentes experiências no mundo de preservação digital de coleções de *e-mail*, que se baseiam em requisitos legais ou de doação (sobretudo, acerca de questões de privacidade), em políticas institucionais de gerência e acesso seguro deste conteúdo em vários formatos de origem e no uso de ferramentas de processamento e arquivamento, como:

[217] Disponível em: https://docs.microsoft.com/en-us/openspecs/office_file_formats/ms-pst/141923d5-15ab-4ef1-a524-6dce75aae546?redirectedfrom=MSDN. Acesso em: 7 jun. 2023.
[218] Disponível em: https://www.encryptomatic.com/pstviewer/. Acesso em: 7 jun. 2023.

- *Digitale archivering in/voor Vlaamse instellingen en diensten* (DAVID)[219] – como um programa de investigação desenvolvido de 2000 a 2003 entre os *Antwerp City Archives* (*FelixArchief*)[220] e o *Interdisciplinary Centre for Law and IT da Katholieke Universiteit Leuven* (K.U. *Leuven*)[221] na Bélgica, o projeto DAVID objetivou criar um manual sobre arquivamento eletrônico.

Figura 53 – Página inicial do projeto DAVID.

Fonte: *Digitale archivering in/voor Vlaamse instellingen en diensten* (2005).

Através do exame dos requisitos jurídicos e arquivísticos para preservação de *e-mail* e da indicação de estratégias de arquivamento possíveis, o projeto elaborou uma solução modelo. Surgido do DAVID e compondo um centro de saber em arquivamento digital na Holanda, o *expertisecentrum* DAVID *vzw* (eDAVID)[222] passou a fornecer acesso aos resultados do projeto, e em 2006 fez pesquisas na área de requisitos de qualidade para patrimônio 'nato digital' oferecendo funções de depósito para *e-mails*.

[219] Disponível em: http://www.edavid.be/davidproject/eng/index.htm. Acesso em: 7 jun. 2023.
[220] Disponível em: https://felixarchief.antwerpen.be/. Acesso em: 7 jun. 2023.
[221] Disponível em: https://www.kuleuven.be/english/about-kuleuven/. Acesso em: 7 jun. 2023.
[222] Disponível em: http://www.edavid.be/index.php. Acesso em: 7 jun. 2023.

Figura 54 – Página inicial do eDAVID.

Fonte: *Expertisecentrum* DAVID *vzw* (2014).

- *futureArch project*[223] – como biblioteca central de pesquisa da Universidade de *Oxford* no Reino Unido, a *Bodleian Library*[224] criou em 2008 o projeto *futureArch* que propôs elaborar métodos para tratar sistematicamente coleções híbridas (analógicas e digitais). Quanto aos arquivos de *e-mails*, estes foram coletados de pastas locais ou de um servidor de *e-mail* com a ajuda do doador/herdeiro e lidos e/ou migrados para o *Internet Message Format* (EML, abreviação de *electronic mail* ou *e-mail)*, especificado no RFC 5322[225], sendo que incluíram, por exemplo, registros de uma conta de *e-mail* de um ex-membro do Parlamento (no formato *Exchange/Outlook*) e o *e-mail* profissional de um acadêmico (no formato *CompuServe*[226]). Dos desafios ilustrados no projeto, esteve a verificação da migração para arquivos grandes, a migração a partir de formatos obsoletos, e a separação de *e-mails* pessoais/privados da correspondência profissional (um requisito de doação).
- *Social and Public Health Sciences Unit* (SPHSU)[227] – sendo uma unidade do *Medical Research Council*[228] e do *Chief Scientist Office*[229] sediada na Universidade de *Glasgow*

[223] Disponível em: http://futurearchives.blogspot.com/. Acesso em: 7 jun. 2023.
[224] Disponível em: https://www.bodleian.ox.ac.uk/home#/. Acesso em: 7 jun. 2023.
[225] Disponível em: https://datatracker.ietf.org/doc/html/rfc5322. Acesso em: 7 jun. 2023.
[226] Disponível em: https://www.compuserve.com/. Acesso em: 7 jun. 2023.
[227] Disponível em: https://www.gla.ac.uk/researchinstitutes/healthwellbeing/research/mrccsosocialandpublichealthsciencesunit/. Acesso em: 7 jun. 2023.
[228] Disponível em: https://mrc.ukri.org/about/. Acesso em: 7 jun. 2023.
[229] Disponível em: https://www.cso.scot.nhs.uk/about/. Acesso em: 7 jun. 2023.

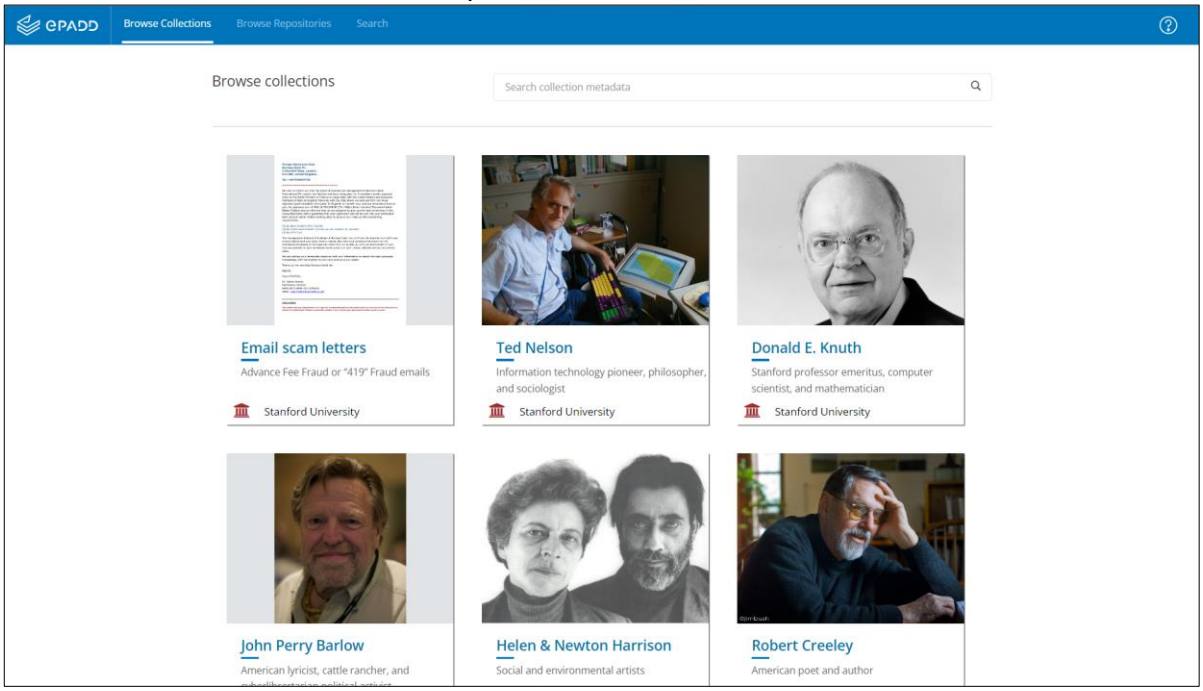
no Reino Unido e que faz pesquisas sobre os efeitos de fatores ambientais e sociais no conforto das pessoas, a SPHSU desde 2007 usa o *software MailArchiva* para reproduzir cópias dos *e-mails* trocados por membros e parceiros de pesquisa, pois o seu servidor de envio/recepção de mensagens havia atingido o limite de armazenagem e recuperação permitindo perdas de documentação sobre pesquisas nos *e-mails*. Através de diretrizes políticas, a SPHSU mantém os arquivos completos criados pelo *MailArchiva* acessíveis aos atuais e ex-funcionários como um registro das ações corporativas, e as mensagens podem ser buscadas e recuperadas num navegador e salvadas em EML fora do sistema.

- *Carcanet Press Email Preservation Project*[230], *University of Manchester Library* (UML)[231] – financiado pelo JISC, este projeto feito em 2012 visou capturar e preservar o arquivo de *e-mail Carcanet Press* da biblioteca da Universidade de *Manchester* no Reino Unido que inclui mensagens com poetas, críticos etc. entre 2001 a 2012. O projeto criou um modelo de dados e um perfil de metadados para a ingestão dos *e-mails* em um sistema baseado no *software Flexible Extensible Digital Object Repository Architecture* – Fedora (o *Manchester eScholar*[232]), apoiando-se no OAIS. Devido as leis de proteção de dados pessoais e de direitos autorais do Reino Unido (isto é, a *Data Protection Act* 2018[233] e a *Copyright, Designs and Patents Act* 1988[234]), foi definido no contrato de depósito do arquivo que ele não seria então concedido a terceiros. Desta maneira, o enfoque do projeto foi a preservação ao invés de acesso, porém houve a exploração de maneiras pelas quais os pesquisadores possam acessar e usar esse arquivo no futuro[235].
- *Special Collections & University Archives*[236], *Stanford University* – o Departamento de Coleções Especiais e Arquivos Universitários da SUL adquiriu em 1993 o acervo do poeta americano Robert Creeley e sua coleção de *e-mails* foi depositada entre 1999 e

[230] Disponível em: https://www.manchester.ac.uk/discover/news/carcanet-press-email-preservation-project/. Acesso em: 7 jun. 2023.
[231] Disponível em: https://www.library.manchester.ac.uk/about/. Acesso em: 7 jun. 2023.
[232] Disponível em: https://www.escholar.manchester.ac.uk/search/. Acesso em: 7 jun. 2023.
[233] Sendo a implantação do *General Data Protection Regulation* (GDPR) no Reino Unido, a Lei de Proteção de Dados (https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted), de 2018, controla como as informações pessoais dos cidadãos britânicos são utilizadas por organizações, empresas ou pelo governo. Todos os responsáveis pelo uso de dados pessoais devem seguir regras rígidas chamadas de "princípios de proteção de dados" (do inglês *data protection principle*s), certificando-se que as informações sejam, por exemplo, usadas de forma justa, legal e transparente e para propósitos explícitos, tratadas de modo a garantir a segurança devida, incluindo proteção contra processamento, acesso, perda, destruição ou danos ilegais ou não autorizados etc.; e ainda esta lei confere o direito dos cidadãos de descobrirem quais informações o governo e demais organizações armazenam sobre eles, incluindo, por exemplo, os direitos de serem informados sobre como seus dados estão sendo usados, de terem dados apagados, e de interromperem ou limitarem o processamento dos seus dados, dentre outros (UK GOVERNMENT, [c2022]).
[234] Disponível em: https://www.legislation.gov.uk/ukpga/1988/48/contents. Acesso em: 7 jun. 2023.
[235] Disponível em: https://www.escholar.manchester.ac.uk/uk-ac-man-scw:226625. Acesso em: 7 jun. 2023.
[236] Disponível em: https://library.stanford.edu/spc/more-about-us. Acesso em: 7 jun. 2023.

2011 em disquetes, discos ópticos e rígidos. Para adquirir, avaliar, processar e fornecer acesso a coleção de *e-mails* de Creeley (https://t.ly/nFQG) e de outras figuras[237] a SUL implementou o ePADD[238] em 2018.

Figura 55 – Página inicial das coleções de *e-mails* das bibliotecas da Universidade de *Stanford* nos Estados Unidos.



Fonte: *Stanford University* ([2023?a]).

O *software* se provou útil aos desafios de identificar e restringir mensagens com conteúdo sensível, e de manter controle das restrições para fornecer um registro do conteúdo que não pode ser disposto e para documentar quando, se acaso, ele puder ser divulgado ao público, sendo que a coleção pode ser navegada por pesquisadores interessados em salas de leitura, e os metadados estão disponíveis *online*.

- Stuart A. Rose *Manuscript, Archives, and Rare Book Library*[239], *Emory University*[240] – a Biblioteca Rose da Universidade *Emory* nos Estados Unidos obteve em 2006 o acervo do autor britânico Salman Rushdie que incluiu cópias locais de *e-mails* armazenados em computadores *Performa* 5400. Um dos computadores foi processado e os *e-mails* estão disponíveis para visualização em salas de leitura da biblioteca após terem sido revisados para identificação de conteúdos restritos conforme o acordo de doação do autor. O resto

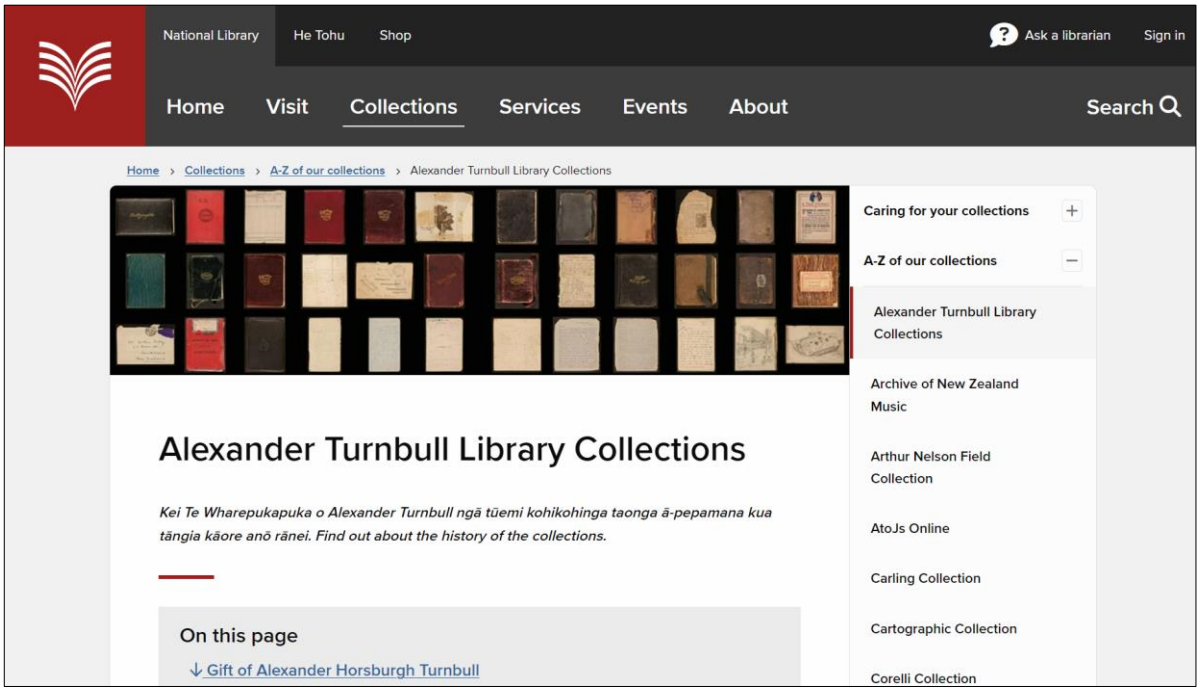
[237] Disponível em: https://epadd.stanford.edu/epadd/collections. Acesso em: 7 jun. 2023.
[238] Disponível em: https://library.stanford.edu/spc/university-archives/transferring-materials/archiving-email. Acesso em: 7 jun. 2023.
[239] Disponível em: https://rose.library.emory.edu/about/index.html. Acesso em: 7 jun. 2023.
[240] Disponível em: https://www.emory.edu/home/explore/about-emory.html. Acesso em: 7 jun. 2023.

dos *e-mails* não foi processado e está inacessível aos pesquisadores devido as restrições do doador e as preocupações com a privacidade. Assim, o uso do ePADD em 2018 foi explorado pela biblioteca para apoiar a descoberta e a pesquisa da coleção, onde só datas das mensagens, nomes de correspondentes etc. extraídos de *e-mails* seriam consultados.

- *Harry Ransom Center*[241], *University of Texas at Austin*[242] – como biblioteca de pesquisa e museu de humanidades da Universidade do *Texas* em *Austin* nos Estados Unidos, o Centro Harry Ransom adquiriu em 2014 o acervo do escritor britânico Ian McEwan que incluiu cerca de oitenta mil mensagens de *e-mail* do *Apple Mail* entre 1997 e 2014. A coleção de *e-mails* começou a ser processada em 2017, porém essa tarefa se mostrou exaustiva dada a previsão de refinar milhares de nomes de correspondentes. Assim, em consulta com desenvolvedores do ePADD, o Centro passou a usar o programa para gerir a coleção, sobretudo, na avaliação, identificação e restrição em lote de dados sensíveis. Em 2019, com base na criação de uma política de acesso e uso de correspondência por *e-mail*, a coleção foi disponibilizada para consulta em salas de leitura aos pesquisadores.
- Alexander Turnbull *Library*[243], *National Library of New Zealand* – como uma divisão da Biblioteca Nacional da Nova Zelândia, a Biblioteca Alexander Turnbull adquiriu em 2013 o acervo do autor neozelandês Ian Wedde que incluiu cópias de seus *e-mails* entre 2003 e 2005 feitos de arquivos PST do *Outlook* e mais de duas mil mensagens enviadas e recebidas.

[241] Disponível em: https://www.hrc.utexas.edu/about/. Acesso em: 7 jun. 2023.
[242] Disponível em: https://www.utexas.edu/about. Acesso em: 7 jun. 2023.
[243] Disponível em: https://natlib.govt.nz/collections/a-z/alexander-turnbull-library-collections. Acesso em: 7 jun. 2023.

Figura 56 – Página inicial das coleções da Biblioteca Alexander Turnbull vinculada a Biblioteca Nacional da Nova Zelândia.



Fonte: *National Library of New Zealand* ([2023a]).

Os arquivos PST foram convertidos para o formato aberto MBOX (às vezes conhecido como formato *Berkeley*, especificado no RFC 4155[244])[245] e, visto à natureza discreta e gerenciável desses arquivos, a biblioteca os usou como teste piloto do *software* ePADD em 2016. O programa mostrou ser útil no processamento e gerência da coleção de *e-mails*, sobretudo, em identificar dados pessoais/sensíveis e conexões de arquivos com autores presentes nas coleções da biblioteca, e para tomar decisões fiáveis do que deve ser removido, restrito por um tempo e pode ser acessado por pesquisadores.

- *British Library*[246] – como a biblioteca nacional do Reino Unido, a Biblioteca Britânica adquiriu em 2011 o acervo da poetisa britânica Wendy Cope que incluiu quarenta mil *e-mails* pessoais e profissionais mantido ainda não catalogado. Apesar dos controles de proteção de dados na biblioteca, a autora preocupou-se com os efeitos da disposição ao público desta ampla coleção de material íntimo, o que fez com que a biblioteca passasse ajudá-la na revisão de *e-mails* do *Outlook* e *BT Internet*[247] para restrição de informações sensíveis. Porém, esse processo se revelou moroso e desgastante, pois nestes serviços

[244] Disponível em: https://datatracker.ietf.org/doc/html/rfc4155. Acesso em: 7 jun. 2023.
[245] Disponível em: https://www.loc.gov/preservation/digital/formats/fdd/fdd000383.shtml. Acesso em: 7 jun. 2023.
[246] Disponível em: https://www.bl.uk/about-us. Acesso em: 7 jun. 2023.
[247] Disponível em: https://home.bt.com/login/loginform?TARGET=SMhttps%3A%2F%2Fsignin1.bt.com%2Fbtmail%2Fsecure%2Femaillogin. Acesso em: 7 jun. 2023.

de *e-mail* a funcionalidade de pesquisa é otimizada para precisão e recuperação rápida de mensagens recentes, com opções limitadas de navegação e visualização. Assim, em 2018 a biblioteca passou a explorar a adoção das funcionalidades do *software* ePADD.

- *Kaine Email Project*[248] – financiado por meio do programa federal norte-americano da *Library Services and Technology Act* (LSTA)[249] gerido pelo *Institute of Museum and Library Services* (IMLS)[250], o projeto *Kaine Email* da Biblioteca da Virgínia foi criado em 2010 para tornar acessíveis os registros de *e-mail* da administração do ex-governador Timothy M. Kaine (2006-2010) conforme a legislação do Estado da Virginia (Código 2.2-126, *Disposition of official correspondence*[251]).

Figura 57 – Página inicial do *Kaine Email Project*.

Fonte: *Library of Virginia* (c2016).

Ao lidar com cerca de 1,5 milhão de mensagens de *e-mails* e anexos, a biblioteca teve que adotar novas ferramentas, como o PST *Viewer Pro* e o *Total Outlook Converter Pro*[252], e processos para disponibilizar as mensagens. Com registros públicos do gabinete do governador e de seus secretários, a coleção de *e-mails* pode ser pesquisada e vista *online*[253] pelos usuários da biblioteca.

[248] Disponível em: https://www.virginiamemory.com/collections/kaine/. Acesso em: 7 jun. 2023.
[249] Disponível em: https://www.ala.org/advocacy/fund-libraries/LSTA. Acesso em: 7 jun. 2023.
[250] Disponível em: https://www.imls.gov/. Acesso em: 7 jun. 2023.
[251] Disponível em: https://law.lis.virginia.gov/vacode/title2.2/chapter1/section2.2-126/#. Acesso em: 7 jun. 2023.
[252] Disponível em: https://www.coolutils.com/TotalOutlookConverterPro. Acesso em: 7 jun. 2023.
[253] Disponível em: https://www.virginiamemory.com/collections/kaine/search-the-collection. Acesso em: 7 jun. 2023.

- TOMES *project*[254] – financiado pelo programa federal norte-americano de subsídios da *National Historical Publications and Records Commission* (NHPRC)[255] e executado de 2015 a 2018, o projeto TOMES objetivou identificar contas de *e-mail* de funcionários públicos cujos os *e-mails* tivessem registros de valor permanente, e capturar, preservar e fornecer acesso a esses registros do governo. Foi uma parceria multiestadual entre os *Utah State Archives*[256], os *State Archives of North Carolina*[257] e a *Kansas State Historical Society*[258] centrada na criação de processos para transferir contas de *e-mail* de arquivo em sistemas proprietários, como *Microsoft* 365[259] e *Gmail*, e convertê-las numa linguagem de código aberto sustentável. Com base no trabalho do ePAAD na SUL, o resultado do projeto foi o *software* TOMES.

Primeiramente, na análise das iniciativas levantadas de preservação digital de coleções de *e-mail*, é fato que os *e-mails* como registros eletrônicos públicos ou privados, produzidos de modo dependente em serviços/clientes de *webmail*, podem documentar informações pessoais e profissionais além de dados administrativos e de pesquisa científica. A sua preservação permite que conteúdos com valor histórico sejam acessados no futuro para uso na complementação de coleções físicas, na conformidade legal e como fontes informais sobre pessoas, locais, assuntos, eventos, publicações, governos e organizações incluídas nos *corpus* das mensagens, em acordo com Boudrez (2006), *Library of Virginia* (c2016), Prom (c2011, c2019) e *Stanford Libraries* ([c2022a]). Ademais, embora as ferramentas existentes sejam úteis em restringir dados pessoais nos arquivos de *e-mails*, elas não podem substituir a atenção do arquivista pois, para Schneider *et al*. (2019) no caso do Centro Ransom, localizar itens restritos necessita da criação de léxicos escalonáveis e personalizados baseados no perfil único do criador dos arquivos, incluindo tanto a linguagem formal como o tom coloquial das mensagens para tornar as buscas mais eficazes.

Em segundo lugar, na coleta, arquivo e processamento dos arquivos grandes de *e-mails* temos não só a necessidade do uso de tecnologias para gerir a coleção (sobretudo, na sinalização e na restrição em massa de mensagens semiestruturadas que incluam dados confidenciais e/ou sensíveis para que estas não sejam divulgadas quando o acesso for disponibilizado ao público), como os *softwares* ePADD e *MailArchiva*, mas também de estratégias de preservação digital, em especial, a migração de formatos de origem obsoletos (por exemplo, do *CompuServe*) e de

[254] Disponível em: https://www.ncdcr.gov/resources/records-management/tomes. Acesso em: 7 jun. 2023.
[255] Disponível em: https://www.archives.gov/nhprc/apply/program.html. Acesso em: 7 jun. 2023.
[256] Disponível em: https://archives.utah.gov/. Acesso em: 19 out. 2021
[257] Disponível em: https://archives.ncdcr.gov/about. Acesso em: 7 jun. 2023.
[258] Disponível em: https://www.kshs.org/p/about/19383. Acesso em: 7 jun. 2023.
[259] Disponível em: https://www.office.com/. Acesso em: 7 jun. 2023.

sistemas proprietários (por exemplo, EML e PST no *Outlook*) para formatos de destino abertos (MBOX etc.) e em outros formatos mais adequados para preservação[260]. Isto incide, em algum grau, na estratégia de preservação digital de investimentos em infraestrutura tecnológica para a sustentação do aumento das operações ao lidar com grandes volumes de dados onde, conforme *Digital Preservation Coalition* (2015c), a análise, a indexação para busca e acesso, a verificação da integridade etc. dos dados exigem mais armazenagem como um maior poder computacional.

Por fim, dentre os fatores que afetam o arquivamento de *e-mails*, destacamos as leis de proteção de dados pessoais (por exemplo, a *Data Protection Act* do Reino Unido e o GDPR[261] da União Europeia) que podem limitar o tratamento e a disponibilização deste material digital, pois ele pode conter informações privadas protegidas por códigos ou regulamentos e relativas a indivíduos dos quais a sua divulgação indevida causará implicações legais. Como explicitado nos casos do Centro Harry Ransom, da Biblioteca Rose e dos projetos *Carcanet Press* e Kaine *Email* em Baker, Butler e Green (2012), *Library of Virginia* (c2016) e Schneider *et al*. (2019), é preciso ter um equilíbrio entre a conformidade legal e o acesso aos *e-mails* com a exploração de formas por quais os pesquisadores possam utilizá-los, como a criação de uma política de uso da coleção que restrinja impressões, *downloads*, capturas de tela etc. dos materiais protegidos e que declare as responsabilidades e exija a comunicação de intenção de seu uso pelos usuários;

---

[260] Como exemplos de diretrizes para a identificação dos formatos de arquivo de *e-mails* preferíveis na preservação digital de longo prazo, podemos citar: as "*Guidelines on File Formats for Transferring Information Resources of Enduring Value*" da *Library and Archives Canada* (2015), que recomenda os formatos EML e MBOX; os "*Born-digital file format standards*" da *National Archives of Australia* ([2022?]), o qual indica os formatos EML, MBOX, *Microsoft Outlook* PST e *Microsoft Outlook Item* (MSG); a "*Sustainability of Digital Formats: Planning for Library of Congress Collections*" da *Library of Congress* (2015), que identifica os formatos MBOX e MSG como sendo promissores para a sustentabilidade em longo prazo; os "*Recommended File Formats for Digital Preservation*" da *Duke University Libraries* ([2022?]), o qual recomenda os formatos EML e MBOX, aceita o esquema XML da conta de *e-mail* e não recomenda os formatos MSG, PST e *Outlook for Mac* (OLM); a "*UW Libraries List of Preferred File Formats*" das *University of Washington Libraries* ([c2022?]), que defini uma máxima confiança aos formatos EML e MBOX, uma confiança média aos formatos PST e MSG, e a mais baixa confiança a todos os outros formatos existentes; os "*Recommended Preservation Formats for Electronic Records*" da *Smithsonian Institution Archives* ([2022?]), que prefere o formato XML de preservação de *e-mail*; dentre outros.

[261] O GDPR, ou *Regulation (EU)* 2016/679 (http://data.europa.eu/eli/reg/2016/679/2016-05-04), estabelece regras relativas à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e regras relativas à livre circulação desses dados, constituindo um passo crucial para fortalecer os direitos fundamentais dos indivíduos na era digital e facilitar as atividades comerciais e negócios através do esclarecimento das normas para empresas e órgãos públicos no mercado único digital. O pacote de medidas de proteção de dados, adotado em 2016, teve o objetivo de tornar os Estados-Membros da União Europeia aptos para a era digital incluindo na legislação do bloco não somente o GDPR, mas também a *Data Protection Law Enforcement Directive* (*Directive – EU –* 2016/680, http://data.europa.eu/eli/dir/2016/680/oj) relativa à proteção das pessoas singulares no que se refere ao tratamento de dados pessoais pelas autoridades competentes para efeitos de prevenção, investigação, detecção e repressão de infrações penais ou execução de sanções penais. Além disto, a Carta dos Direitos Fundamentais da União Europeia (do inglês *Charter of Fundamental Rights of the European Union*, http://data.europa.eu/eli/treaty/char_2012/oj) estipula no artigo 8º que todos os cidadãos dos Estados-Membros têm o direito à proteção dos seus dados pessoais, acesso aos dados que tenham sido recolhidos a seu respeito, e retificá-los (EUROPEAN COMMISSION, [2022]).

unido a adoção dos recursos do ePADD e demais soluções que permitam a busca, navegação e visualização pública *online* dos *e-mails* processados (sem conteúdos restritos pela privacidade).

3 PRESERVAÇÃO DIGITAL: desafios, requisitos, estratégias e produção científica[262]

As sociedades modernas são grandes produtoras e consumidoras de informação, consistindo, esta, num bem vital para o seu desenvolvimento cultural, econômico, político e de conhecimento. O uso das TIC associado a *World Wide Web* proposta por Tim Berners-Lee em 1989 e a *Internet*, permitiu uma explosão informacional mundial com a rápida produção, disseminação e aquisição de recursos de informação digital, seja nas esferas públicas ou privadas. Contudo, a dinamicidade e a efemeridade dos ambientes digitais impõem perdas rápidas e definitivas de registros importantes disponíveis *online*, retratando o desafio global do século XXI de garantir a preservação e o acesso contínuo à uma memória pessoal, corporativa e cultural digital.

Em reconhecimento a existência desses riscos, organizações de todo o mundo – em especial, instituições de patrimônio cultural e universidades – têm implementado técnicas e ferramentas para criação, gestão e preservação de materiais digitais acessíveis ao longo do tempo. No caso brasileiro, encontram-se políticas de preservação digital planejadas por IES, tais como Universidade Estadual de Campinas (2011) e Universidade Estadual Paulista (2017). De modo similar estas universidades integram, também, a Rede Cariniana, que foi criada em 2012 pelo IBICT no intuito de assegurar o acesso contínuo de documentos eletrônicos nacionais.

A preservação digital vêm sendo um tema de estudo da Ciência da Informação. Trata-se de um desafio complexo, inevitável e da atualidade nas publicações nacionais e internacionais da área, necessitando de análises e soluções inter/multidisciplinares. Como definição de preservação digital por longo prazo, Grácio, Fadel e Valentim (2013, p. 113) interpretam que a preservação digital se refere à "[...] um processo de gestão organizacional que abrange várias atividades necessárias para garantir que um objeto digital possa ser acessado, recuperado e utilizado no futuro, a partir das TIC existentes na época e com garantias de autenticidade", julgando o conceito de autenticidade de um recurso/objeto digital[263] ligado à salvaguarda do conteúdo informacional original de sua produção.

---

[262] Parte do texto original deste capítulo de Tese de Doutorado foi publicado como artigo na Revista Digital de Biblioteconomia e Ciência da Informação (RDBCI) vinculada ao Sistema de Bibliotecas da Universidade Estadual de Campinas (UNICAMP), volume 18, de 2020, sob o título de "Preservação digital: desafios, requisitos, estratégias e produção científica".

[263] A partir de Márdero Arellano (2008) e Santos e Flores (2015), um objeto digital é qualquer tipo de arquivo em meio digital, o qual é representado em cadeias de *bits* (*bitstream*) e formado por estrutura lógica, conteúdo e estrutura de apresentação.

Diante disso, este capítulo tem por objetivo proporcionar uma visão ampla e reflexiva das principais questões da preservação digital, em que subsidie a compreensão de tendências e políticas acerca do tema como competência exigida aos profissionais em unidades de informação (BOERES, 2017) e, ainda, aprofunde as discussões sobre as necessidades de preservação digital para a geração de novos conhecimentos e efetivação de estudos e ações futuras neste domínio. Consideramos que o entendimento de atualidades e interesses da comunidade científica de preservação digital permite a sinergia das abordagens estratégicas, institucionais e tecnológicas contemporâneas aos problemas vigentes e vindouros da gerência da preservação e acesso por longo prazo de informações digitais.

Para uma ideal estruturação do presente capítulo, dispomos com a descrição da metodologia adotada, seguido da exibição e debate dos resultados obtidos, que incluem os problemas, requisitos e estratégias de preservação digital descritos na literatura clássica e recente especializada, os dados recuperados em bases de dados científicas e as considerações finais sobre a pesquisa desenvolvida.

Adotamos uma pesquisa quanti-qualitativa, de abordagem exploratória-descritiva (SILVA; MENEZES, 2005), pautada na revisão da literatura específica nacional e internacional dos últimos vinte um anos sobre preservação digital; e na coleta e análise dos dados, disponíveis em bases de dados escolhidas, relativos à produção científica recente desse tema de investigação. As discussões do trabalho fornecem uma síntese da concepção atual de questões da preservação digital, partindo-se das principais dificuldades discernidas, dos critérios deste processo reconhecidos e das soluções estratégicas, políticas e tecnológicas exploradas pela comunidade de preservação digital, retratando o que tem sido articulado até o momento no arquivamento de conteúdos digitais em longo prazo.

Quanto aos procedimentos metodológicos, adotou-se o método bibliográfico (MARCONI; LAKATOS, 2017; SEVERINO, 2016) no qual foram levantadas publicações científicas do assunto ‘preservação digital’ na *Scopus* (*Elsevier*)[264], maior base de dados de resumos e citações da literatura com revisão por pares, e na *Web of Science* Coleção Principal (*Clarivate Analytics*)[265], base de dados de publicações da mais alta qualidade e com maior impacto no mundo. Disponíveis via Portal de Periódicos CAPES, ambas as bases de abrangência mundial, multidisciplinar e de referenciais com resumos e dados de citações oferecem também ferramentas bibliométricas para acompanhar, analisar e/ou visualizar pesquisas.

[264] Disponível em: https://www.elsevier.com/pt-br/solutions/scopus. Acesso em: 22 maio 2023.
[265] Disponível em: https://clarivate.com/webofsciencegroup/solutions/web-of-science/. Acesso em: 22 maio 2023.

Como limitadores no levantamento bibliográfico feito em 20 de março de 2020[266], usamos a pesquisa avançada (*Advanced*) com o termo "*digital preservation*" nos campos de título, resumo e palavras-chave (*Title*, *Abstract*, *Keywords*/Tópico), definimos as opções de filtro por atas/anais de conferências (*conference/proceedings paper*) e artigos de periódicos (*article*), publicados em língua portuguesa, inglesa e espanhola nos últimos cinco anos (2015-2019). Na exploração e análise dos resultados, através do programa *Microsoft Excel* 2016 ilustramos com quantitativos da produção científica recente sobre preservação digital, destacando os pesquisadores mais produtivos, além das instituições, dos países e das áreas de pesquisa com o maior número de publicações neste período.

3.1 Preservação digital: o desafio para as gerações presentes e futuras

As principais dificuldades da preservação digital advêm das especificidades dos objetos que procura salvaguardar ao longo do tempo. Estes objetos digitais, nascidos digitais ou digitalizados, são suscetíveis às constantes alterações e a efemeridade dos meios onde são criados, transportados ou armazenados bem como a alta dependência de tecnologias de *hardware*, de *software* e de suporte para a sua reprodução que se tornam obsoletas com rapidez ou são danificadas fisicamente. Assim, essas particularidades resultam em refletir as questões de fidedignidade, autenticidade e integridade dos materiais digitais[267] no seu gerenciamento, arquivamento e acesso utilizável por longo período.

Diferentemente da preservação dos meios não digitais, os objetos digitais são facilmente capazes de dissociarem em seus elementos individuais dificultando mantê-los inteiramente. Pode-se, por exemplo, manter o conteúdo de um documento eletrônico, mas perder/adulterar seu leiaute pelas migrações contínuas ou, ainda, preservar a presença física (o arquivo de dados) de um objeto, porém deixar de manter sua capacidade de interpretação (BULLOCK, 1999; THOMAZ; SOARES, 2004). Isto posto, segundo Márdero Arellano (2008, 2012), Baggio e Flores (2012) e Innarelli (2009, 2014), a preservação dos documentos digitais e eletrônicos

[266] Há de se levar em conta a possibilidade de haver pequenas variações no número total de documentos recuperados nas bases de dados caso novas buscas sejam feitas novamente em acordo com os procedimentos metodológicos explicitados no trabalho, visto que é provável que estudos publicados no período definido na metodologia poderão ser indexados pelas respectivas bases num período posterior a data de realização dos levantamentos bibliográficos.

[267] Através de Arquivo Nacional (2016) e Barbedo, Corujo e Sant'ana (2011), um documento digital autêntico constitui aquele que comprovamos ser o que significa ser e é isento de alterações não autorizadas; por sua vez, a integridade de um documento digital remete a guarda de plenitude e fixidade, onde são cruciais informações a serem registradas em metadados para identificação da proveniência e do contexto de criação e manutenção do documento ao longo do tempo.

requisita esforços específicos para manter as propriedades originais e a capacidade de servirem de registro e fonte de informação, haja vista suas fragilidades sobre a complexidade, os custos, a obsolescência tecnológica e a degradação física.

Agregam-se a estes problemas fundamentais da preservação digital, demais dificuldades de natureza gerencial, técnica, jurídica, política, econômica e social, dentre as quais ressaltamos:

- A necessidade de diferentes abordagens de preservação em distintas escalas; além de ciclos recorrentes de manutenção do acesso contínuo aos recursos digitais, mediante a utilização de um conjunto variável de ferramentas (NATIONAL LIBRARY OF AUSTRALIA, 2013).
- O compromisso institucional de não perda das propriedades significativas do objeto digital custodiado e preservado por longo prazo, assegurando sua recuperação, inteligibilidade e autenticidade, para servir de fonte de prova e informação (ARQUIVO NACIONAL, 2016).
- A ampla variedade de padrões e formatos de arquivos digitais, tornando-se fundamental o reconhecimento e o uso dos formatos mais apropriados e sustentáveis para o arquivamento de longo prazo das informações digitais (LIBRARY AND ARCHIVES CANADA, 2015).
- Os direitos de propriedade intelectual e as demais obrigações legais a serem cumpridas, que interferem na cópia, armazenamento, alteração e utilização do conteúdo de recursos digitais para fins de preservação a longo prazo (NATIONAL MUSEUM OF AUSTRALIA, 2012).
- A rápida e contínua obsolescência tecnológica ao nível do *hardware*, *software*, formatos e suportes de armazenamento, acrescido das ameaças de danos físicos nos arquivos e ao nível dos componentes de *hardware* e dos suportes (BARBEDO; CORUJO; SANT'ANA, 2011).
- O caráter dinâmico e efêmero da *Web* somado as experiências personalizadas de navegação, que trazem dilemas de validação da autenticidade e integridade na renderização de *websites* grandes e complexos arquivados em larga escala e por longo período (PENNOCK, c2013).
- A alta demanda pelo recrutamento de profissionais com experiência e habilidades práticas continuamente atualizadas em preservação digital, a fim de compor equipes distribuídas ou multidisciplinares nas organizações (DIGITAL PRESERVATION COALITION, c2015).

- O armazenamento em nuvem e os contratos de serviço que requerem gestão cuidadosa para atender o arquivamento, ensejando flexibilidade, baixos custos e dados seguros e acessíveis além da vida útil de tecnologias/provedores atuais (THE NATIONAL ARCHIVES, c2015).
- A exigência contínua de recursos financeiros para investimentos assíduos em infraestrutura organizacional e tecnológica e na capacitação de pessoal, com intuito de manter acessíveis os objetos digitais ao longo do tempo (UNIVERSIDADE ESTADUAL PAULISTA, 2017).

À vista dos desafios indicados, ainda que os objetos digitais incluam conteúdos dinâmicos, multimídia, funcionalidades e vantagens de transmissão, replicação e edição em ambientes digitais, suas complexidades trazem obstáculos para a preservação e a acessibilidade em longo prazo. Das necessidades de preservação digital, são evidentes os amplos investimentos, as exigências jurídicas e as garantias de localização, contexto, autenticidade e integridade dos conteúdos digitais. Portanto, torna-se vital reconhecer os requisitos, as estratégias e os recursos tecnológicos existentes, descritos na literatura especializada, a serem considerados para esforços efetivos e exequíveis neste domínio.

3.2 Os requisitos para a preservação digital, com base no modelo de referência OAIS

A preservação digital exige o cumprimento de um conjunto mínimo de requisitos funcionais e não funcionais, com o objetivo de alcançar os resultados almejados. Das abordagens existentes, Bullock (1999), Formenton, Gracioso e Castro (2015), Thomaz (2004) e Thomaz e Soares (2004) identificam nove requisitos a serem analisados para a preservação de objetos digitais a longo prazo, enquanto que Innarelli (2009, 2014) cita os princípios da preservação digital em dez mandamentos passíveis de interpretação e de aplicação conforme a realidade e a estrutura da organização; ambas propostas respaldadas no modelo de referência OAIS.

A partir das considerações apontadas pelos autores mencionados, idealizamos um conjunto de cinco requisitos básicos para a preservação digital, que podem ser assim compreendidos:

- Manter uma política de preservação – elaboração, implantação e manutenção de diretrizes, objetivos e métodos institucionais para o arquivamento das coleções digitais, abrangendo a delimitação clara dos tipos de informações ou de quais elementos do

objeto digital serão selecionados, visto a natureza multimídia, hipertextual e dinâmica dos conteúdos digitais.

- Garantir a fidedignidade, a autenticidade e a integridade – confiança de que o objeto digital acessado é justamente aquele que se busca, onde as prováveis alterações ou deslocamentos advindos de medidas de preservação pelas quais foi submetido (as migrações contínuas de suportes, formatos e versões, por exemplo), mantiveram a sua identificação e localização inequívoca e o seu conteúdo com leiaute e funcionalidades originais ao longo do tempo.
- Manter o contexto – salvaguarda das dependências de *hardware* e *software* (de preferência compatíveis com os padrões internacionais e de arquitetura aberta, assegurando autonomia sobre os desenvolvedores para o acesso às informações no arquivamento), das razões para a produção do objeto digital, seus modos de distribuição e relações com demais objetos.
- Manter a proveniência – identificação da origem ou fonte do objeto digital, sua cadeia de custódia e o detalhamento do histórico de alterações ocorridas, através dos metadados para preservação digital[268], com o intuito de comprovar ou garantir a autenticidade e a integridade do objeto digital e apoiar a sua reconstituição, consistência e persistência por longo prazo.
- Manter a recuperação – implantação e revisão assídua de uma política de *backup* ou cópia de segurança, que preze a replicação do objeto digital (e seus metadados) em local físico separado e o uso combinado de diferentes tipos de tecnologias de armazenamento, com o propósito de assegurar o acesso e a restauração confiável, integra e segura dos dados.

Em síntese, os cinco requisitos indicados a serem discutidos e adaptados pelas organizações comprometidas com a manutenção em longo prazo de informações digitais, objetivam, sobretudo, salvaguardar os objetos digitais e a capacidade de acesso contínuo e utilizável aos seus conteúdos, refletindo assim os pressupostos das estratégias de preservação digital. Embora cada requisito tenha as suas particularidades, todos eles estão intrinsecamente

[268] Sobre a descrição de documentos eletrônicos, ressaltamos o formato XML como um padrão aberto para produção, armazenamento e transferência de documentos por meio eletrônico, independente das plataformas operacionais e dos fabricantes de *software*, compreensível por diversas aplicações e autoexplicativo. Na preservação digital, a linguagem XML é considerada um tipo particular de migração, o qual enriquece informação sobre estruturas e significado, garante o encapsulamento dos metadados e das informações exigidas para interpretação dos objetos digitais originais e beneficia a interoperabilidade entre recursos de distintas áreas (MÁRDERO ARELLANO, 2008).

associados e detêm como componentes-chave a construção da política de preservação e a elaboração dos metadados para preservação.

Dado o caráter geral, os requisitos descritos ainda são usados no trabalho como critérios de referência para conceber a preservação digital. No entanto, outros aspectos devem ser analisados na implementação deste processo, incorporando-os numa política de preservação digital, tais como a identificação das necessidades informacionais da comunidade usuária; a disposição de recursos financeiros, humanos e tecnológicos; os direitos e deveres éticos, morais e legais das várias partes envolvidas; e a aplicação de métodos e tecnologias para preservação das propriedades originais dos objetos digitais no qual sustentam a validade de sua confiabilidade, autenticidade e integridade.

O OAIS ou Sistema Aberto de Arquivamento de Informação (SAAI)[269], esquema conceitual que normaliza um sistema de repositório direcionado para a preservação e o acesso à informação digital em longo prazo, da ISO e do *Consultative Committee for Space Data Systems* (CCSDS), define um modelo funcional e de informação, o qual especifica as operações a serem feitas pelo sistema e as informações registradas por metadados requeridas para a representação dos materiais mantidos e o arquivamento digital de longo prazo.

Para solucionar os atuais problemas e desafios impostos pelo ambiente digital à preservação por longo prazo de objetos digitais e, juntamente, usufruir dos avanços científicos e das modernas aplicações ou ferramentas tecnológicas que disponibilizam, várias estratégias vêm sendo propostas para a preservação digital. É preciso entendê-las, explorando as suas capacidades e limitações.

3.3 Estratégias: soluções para os desafios da preservação digital

Bullock (1999) reuni as estratégias de preservação digital em dois tipos: as estratégias para tentar solucionar o problema da obsolescência tecnológica, abrangendo a migração, a transferência para suportes analógicos, a emulação e a preservação de tecnologia; e as estratégias para "assumir o controle", compreendendo a adoção de padrões e de diretrizes, a documentação e a descrição dos recursos (metadados), a construção de parcerias e estabelecimento de infraestrutura. Pearson e Del Pozo (2009) ainda optam por ordenar as estratégias em ações de preservação primárias, que alteram diretamente os materiais digitais a

[269] ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **ABNT NBR 15472**: Sistemas espaciais de dados e informações: Modelo de referência para um sistema aberto de arquivamento de informação (SAAI), Rio de Janeiro: ABNT, 2007.

serem preservados, como a migração; e as ações de preservação secundárias, o qual mudam a maneira como o material é acessado e como este acesso é preservado no decorrer do tempo, incluindo a emulação e a coleta e manutenção de um "museu tecnológico"[270]. Por outro lado, Long (2009) divide as estratégias em termos de maiores chances de prover ao menos soluções parciais de preservação, isto é, as metodologias não sustentáveis em longo prazo, tal como os museus tecnológicos; e as que são sustentáveis por longo prazo, como a migração e a emulação.

Em Santos e Flores (2017a) as estratégias para preservação digital também estão organizadas por suas prioridades na preservação dos três níveis dos objetos digitais: nível físico (a integridade do *hardware* e do suporte), integrando o refrescamento ou recopia de conteúdos em suportes mais atuais; nível lógico (o *software* e a integridade da cadeia de *bits* original), englobando a emulação, a preservação de tecnologia e o encapsulamento ou reunião de tudo que seja requerido para o acesso e interpretação dos objetos; e nível conceitual (a representação visual do conteúdo interpretado por humanos), incluindo a migração. Conforme Márdero Arellano (2008), Formenton, Gracioso e Castro (2015), Santos e Flores (2015, 2018), Thomaz (2004) e Thomaz e Soares (2004) as diferentes estratégias de preservação digital podem ser usualmente agrupadas em: estratégias estruturais, que consistem nos investimentos iniciais advindos das instituições, a fim de construir um ambiente adequado para a preservação digital; e estratégias operacionais, o qual constituem as medidas reais de preservação física, lógica ou conceitual dos objetos digitais a serem executadas pelas respectivas organizações.

Baseando-se nesta categorização genérica, o trabalho reuni as principais estratégias de preservação digital em estratégias estruturais e operacionais, segundo apresentado na Figura 58.

[270] Independentemente da preservação de tecnologia e os museus tecnológicos terem designações distintas, neste trabalho julgaremos que ambos se tratam da mesma estratégia de preservação digital. Através de Long (2009) e Pearson e Del Pozo (2009), uma lista de prós e contras poderá ser obtida para consulta e investigação acerca da coleta e manutenção dependente do contexto tecnológico original de criação e uso de materiais digitais a serem preservados em curto prazo.

Figura 58 – Principais estratégias de preservação digital.

Fonte: Elaborado pelo autor.

A seguir discutiremos cada uma das estratégias para a preservação digital, indicadas na Figura 58, que são atualmente mais adotadas, difundidas ou relatadas na literatura especializada.

3.4 Estratégias estruturais

a) Adoção de padrões abertos

A adoção e a conformidade com padrões abertos, seja para criação ou gestão de documentos digitais, propicia a redução dos efeitos da obsolescência tecnológica na preservação digital. Santos e Flores (2015, 2017a) e Schäfer e Constante (2012), considerando a migração dos padrões em caso de obsolescência e a restrição de formatos para armazenar os dados[271] inferem que os padrões abertos e não proprietários possibilitam a reconstrução do *software* interpretador e a distribuição irrestrita, que assegura a não dependência por atualizações dos desenvolvedores para o acesso as informações e a reprogramação e adaptação do *software* em consenso com a política de preservação. Os padrões abertos para preservação digital e arquivamento são fixados por órgãos oficiais de normalização e consórcios internacionais, como a *National Digital Stewardship Alliance* (NDSA)[272], o *World Wide Web Consortium* (W3C)[273], a ISO[274] e o IIPC[275].

b) Documentos de políticas e estratégias institucionais

O conjunto de documentos criados ao redor do mundo[276], através da iniciativa de instituições envolvidas com a preservação digital e de pesquisadores do tema, cedem diretrizes para o acertado desenvolvimento e implementação de políticas e estratégias de gestão de

---

[271] Sobre exemplos de diretrizes para identificação dos formatos de arquivo digital confiáveis para preservação em longo prazo, podemos indicar as "*Guidelines on File Formats for Transferring Information Resources of Enduring Value*" da *Library and Archives Canada* (2015); e os "*Long-term File Formats*" dos *National Archives of Australia* (2019).

[272] A NDSA, lançada em 2010, é um consórcio de organizações envolvidas na preservação a longo prazo da informação digital, o qual as suas atividades ocorrem através de grupos de interesse e trabalho, como pesquisas sobre arquivamento da *Web* pelo *Web Archiving Survey Working Group*. Disponível em: https://ndsa.org/about/. Acesso em: 22 maio 2023.

[273] O W3C, fundado em 1994, é uma comunidade internacional onde as organizações membros (o Núcleo de Informação e Coordenação do Ponto BR - NIC.br do Comitê Gestor da *Internet* no Brasil, por exemplo), uma equipe e o público trabalham para desenvolver padrões *Web*. Disponível em: https://www.w3.org/Consortium/. Acesso em: 22 maio 2023.

[274] Instaurada em 1947, a ISO é uma rede global de órgãos de normalização nacionais de 164 países (apenas um membro por país), tendo o Brasil a participação da ABNT, que define a norma internacional ISO 14721:2012 *Space Data and Information Transfer Systems* - OAIS. Disponível em: https://www.iso.org/about-us.html. Acesso em: 22 maio 2023.

[275] Fundado em 2003 na BnF, o IIPC possui como membros organizações de mais de trinta e cinco países, abarcando bibliotecas e arquivos nacionais, universitários e regionais (*Internet Archive*, Arquivo.pt, Biblioteca Nacional da Austrália, Biblioteca Britânica e *Library of Congress*, por exemplo), o qual participam de projetos e grupos de trabalho que focam aspectos do arquivamento da *Web*, incluindo desenvolvimento de coleções, preservação dos arquivos da *Web* etc. Disponível em: https://netpreserve.org/about-us/. Acesso em: 22 maio 2023.

[276] Podem-se citar, como exemplo, as "Recomendações para a Produção de Planos de Preservação Digital" da Direção-Geral de Arquivos de Portugal (BARBEDO; CORUJO; SANT'ANA, 2011); e a "Política de Preservação Digital do Programa Permanente de Preservação e Acesso a Documentos Arquivísticos Digitais" do Arquivo Nacional (2016).

materiais digitais. Uma política de preservação digital em IES, segundo Grácio (c2012) e Grácio, Fadel e Valentim (2013), respalda-se em elementos distribuídos em três categorias: organizacional, incluindo elementos de gestão para fixação e estabilização institucional sobre a política e as medidas de preservação; legal, abrangendo normas institucionais e a legislação vigente a nível nacional e internacional; e técnico, envolvendo elementos sobre os fluxos, os processos e as medidas de preservação. Ademais, certos aspectos devem ser analisados na criação dessas políticas e estratégias, como entender e incorporar o contexto organizacional que irão existir (DIGITAL PRESERVATION COALITION, c2015).

c) Orçamentos e custos da preservação digital

O cálculo dos custos da preservação digital é uma atividade complexa e primordial para a determinação de práticas rentáveis e a justificação dos investimentos de recursos e despesas. Entre as questões de impacto para os custos a serem analisadas estão os recursos humanos (especialistas e equipe multidisciplinar de profissionais); a implantação, a operação e a manutenção das medidas de preservação; os recursos materiais; e a missão e os objetivos institucionais, com a inserção do tipo e do volume das coleções, dos níveis de preservação e de acesso definidos e o prazo proposto as ações (BARBEDO; CORUJO; SANT'ANA, 2011; DIGITAL PRESERVATION COALITION, c2015). Há ainda incertezas dos custos da preservação digital, onde a maioria dos estudos apontam para altas despesas envolvidas, como Boeres (2017), e alguns indicam o baixo custo deste processo em comparação a preservação de acervos tradicionais, como Andrade, Borges e Jambeiro (2006).

d) Seleção para preservação digital e conformidade legal

Os volumes crescentes de informações criadas em ambientes digitais com suas restrições, dependências e relevância variável, tornam indispensável a seletividade daquilo em que o acesso será mantido para um certo fim e período. Conforme Boeres e Márdero Arellano (2005) e Grácio (c2012), a preservação eterna de tudo e para todos é inviável e nulo, assim uma política de preservação digital em IES deverá definir critérios de seleção com base nos objetivos institucionais, nas necessidades acadêmicas e da sociedade e em termos de custo-benefício da preservação por longo prazo. Outra dificuldade colocada por *Digital Preservation Coalition* (c2015), envolve a legalidade na coleta, preservação e acesso aos materiais digitais geridos e armazenados, visto os deveres legais sobre a proteção/disposição de dados, os direitos autorais

e os contratos de serviço, além da lei[277] que muitas vezes também está atrasada quanto a mudança tecnológica e as demandas de preservação digital.

e) Treinamento e desenvolvimento de pessoal

Como levantado por Farias, Araújo e Evangelista (2018), as instituições da Rede Cariniana colocam a carência de recursos humanos, de pessoal especializado e de orçamento específico como fatores preponderantes na adoção das estratégias de preservação digital. Nesta perspectiva, é crucial um programa de treinamento de pessoal e de desenvolvimento profissional contínuo, que considere as diferentes habilidades para preservação[278] e a definição clara das funções e responsabilidades das diversas partes interessadas, com garantias de sua implementação, aceitação e aprimoramento. Das alternativas de treinamento, desenvolvimento e aprendizado das equipes estão o compartilhamento de informações e troca de pessoal com organizações similares, além de cursos de curta duração ou programas completos teóricos e práticos acerca da preservação digital, disponíveis presencialmente ou, ainda, na modalidade a distância e *online* (DIGITAL PRESERVATION COALITION, c2015).

f) Metadados para preservação digital

A adoção efetiva de padrões ou esquemas de metadados[279] será essencial para a garantia da preservação digital, de modo a apoiar a gerência do arquivamento e manutenção do acesso contínuo de objetos digitais em ambientes informacionais, como repositórios e serviços de nuvem. Conforme Formenton *et al.* (2017) e Lavoie e Gartner (c2013), estes esquemas determinarão a identidade, a representação, a coesão e a persistência do objeto no repositório, com garantias de fidedignidade, autenticidade e integridade, além de definir a

[277] No Brasil, destacamos a Lei nº 12.527, de 2011, que regula o direito de acesso dos cidadãos a informações produzidas ou sob a guarda de entidades públicas, como as IES; a Lei nº 9.610, de 1998, que defini os direitos autorais no país; a Lei nº 13.853, de 2019, ou LGPD, para dispor sobre a proteção de dados pessoais; o Decreto nº 10.278, de 2020, que estabelece a técnica e os requisitos para a digitalização de documentos públicos ou privados; a resolução do Conselho Nacional de Arquivos (CONARQ) nº 43, de 2015, que estabelece diretrizes para a implementação de repositórios arquivísticos digitais confiáveis; e a Lei nº 12.682, de 2012, que regula a elaboração e o arquivamento de documentos em meios eletromagnéticos.

[278] Estas questões são notadamente discutidas em Boeres (2017) que, através das indicações de especialistas brasileiros e estrangeiros, identifica um conjunto de competências requeridas para equipes de profissionais de preservação digital em unidades de informação, entre as quais estão: o saber em gestão de dados digitais por longo prazo; e a compreensão acerca de confiabilidade, autenticidade e integridade dos registros, de tendências e de políticas de preservação digital.

[279] Dos exemplos de padrões de metadados para preservação digital, estão: o MODS (http://www.loc.gov/standards/mods/) como esquema de metadados descritivos para serviços de biblioteca; o METS (http://www.loc.gov/standards/mets/) para codificação de metadados descritivos, administrativos e estruturais sobre objetos numa biblioteca digital; e o PREMIS (https://www.loc.gov/standards/premis/) para codificação, gerenciamento e intercâmbio de metadados de preservação entre sistemas de repositórios para preservação digital de longo prazo.

interoperabilidade[280] entre sistemas. Sobre padrões de metadados de preservação digital nos serviços em nuvem, Castro e Silveira (2018) inferem que a temática se situa ainda pouco explorada na área da Ciência da Informação, seja a nível nacional ou internacional, mesmo com o uso crescente destes ambientes digitais no armazenamento de objetos digitais como estratégia na preservação de longo período para conter a obsolescência tecnológica.

g) Investimento e montagem de infraestrutura tecnológica

Uma factual preservação digital exige investimentos notáveis em infraestrutura tecnológica para sustentação dos fluxos, processos e atividades de arquivamento dos materiais digitais. Tendo em vista o tratamento de quantias crescentes de dados, através de Boeres e Márdero Arellano (2005) e *Digital Preservation Coalition* (c2015), algumas instalações devem ser presumidas: os serviços em nuvem e os sistemas de armazenamento corporativo para a replicação dos dados mantidos; os sistemas de repositório digital fiáveis para armazenar, gerir e acessar a produção intelectual das IES no decorrer do tempo; e uma computação de alto desempenho para manipular dados grandes, sejam de pesquisa ou arquivos da *Web*. Nesta infraestrutura é cabível dois ambientes tecnológicos pois, para Grácio (c2012), Grácio, Fadel e Valentim (2013) e Universidade Estadual Paulista (2017), deverá atender a preservação dos objetos (e seus metadados) e viabilizar o seu ideal acesso, busca e recuperação.

h) Formação de redes de colaboração

O sucesso na superação dos desafios da preservação digital requer uma maior colaboração entre organizações, equipes de profissionais e criadores de objetos digitais a serem mantidos[281]. De fato, segundo Grácio (c2012) e Grácio, Fadel e Valentim (2013), as iniciativas colaborativas incluem os atrativos da troca de saberes e de experiências como da padronização de estratégias institucionais em suporte à interoperabilidade dos objetos digitais entre sistemas. Todavia, estas medidas impõem flexibilidade no interior das estruturas organizacionais e implicam conflitos potenciais que podem se manifestar sob a forma de múltiplas agendas, prazos ou mecanismos de financiamento. Assim, julgando a preservação digital como um esforço global, vital e exequível, a chave da construção e manutenção das colaborações está no

[280] Interoperabilidade é "[...] a capacidade de vários sistemas com diferentes plataformas de *hardware* e *software*, estruturas de dados e interfaces, de trocar dados com perda mínima de conteúdo e funcionalidade." (NATIONAL INFORMATION STANDARDS ORGANIZATION, 2004, p. 2, tradução nossa).

[281] Por exemplo, o IBICT na criação da Rede Cariniana aderiu a Aliança Internacional LOCKSS para a preservação das publicações nacionais de acesso livre incluídas no OJS e no *software DSpace*, ensejando a troca de experiências com instituições ao redor do mundo unidas em demais redes colaborativas (MÁRDERO ARELLANO, 2012).

diálogo eficaz entre as partes interessadas, através do uso de termos e de linguagem inteligível por todos (DIGITAL PRESERVATION COALITION, c2015).

3.5 Estratégias operacionais

a) Definição do meio de armazenamento

A preservação digital depende da escolha apropriada do meio de armazenamento de dados a longo prazo. Nas últimas décadas, o uso de mídias magnéticas e ópticas julgava critérios técnicos para a sua avaliação contínua, como os fatores citados em Brown (2008) e Thomaz e Soares (2004): capacidade (de armazenamento, leitura etc.); obsolescência; padronização; viabilidade (de reter a integridade probatória etc.); custo; e outros. Porém, hoje fez-se mais comum o uso de sistemas de armazenamento local e/ou em nuvem resilientes (o *DSpace*[282], o Fedora e o LOCKSS[283], por exemplo) apoiado nos princípios de redundância e variedade, fixidez, monitoração e reparo etc. para volumes crescentes de material digital a serem acessados, preservados e recuperados com facilidade, rapidez e maior precisão (DIGITAL PRESERVATION COALITION, c2015; SANTOS; FLORES, 2017a).

b) Migração

A migração é uma das estratégias mais adotadas na preservação de longo prazo. Através de Baggio e Flores (2012), Barbedo, Corujo e Sant'ana (2011), Hedstrom (2001) e Pearson e Del Pozo (2009) consiste em transferir objetos digitais de plataformas tecnológicas em vias de obsolescência, degradação física ou descontinuidade para outras mais novas, estáveis e padronizadas, assegurando a atualização/conversão de versões, formatos e suportes, a compatibilidade com tecnologias atuais e o acesso às informações. Contudo, para os autores as variações de migração intervêm na estrutura e conteúdo original do objeto, exigindo um plano contínuo e a apreensão, controle e documentação eficaz das alterações. Em Santos e Flores (2017a, 2018) e Schäfer e Constante (2012) as constantes migrações podem ainda refletir em

[282] O projeto DSpace é um *software* de código aberto desenvolvido pelo Instituto de Tecnologia de *Massachusetts* nos Estados Unidos e amplamente usado por organizações acadêmicas na criação de repositórios para preservação e acesso aberto a todo tipo de conteúdo digital. Disponível em: https://duraspace.org/dspace/about/. Acesso em: 22 maio 2023.

[283] Como uma iniciativa de preservação digital em operação desde 1999 sob os auspícios da Universidade *Stanford* nos Estados Unidos, o Programa LOCKSS oferece um *software* de repositório confiável e de código aberto elaborado para preservar publicações acadêmicas. Disponível em: https://www.lockss.org/about/why-lockss. Acesso em: 22 maio 2023.

perda/adição de dados, falhas de representação fiel dos objetos complexos e incompatibilidades entre formatos de origem e destino, requerendo outras estratégias.

c) Transferência para suportes analógicos

A transferência para suportes analógicos de longa duração é um método a ser utilizado em última alternativa para o arquivamento das informações digitais. De acordo com Bullock (1999), Hedstrom (2001) e Rothenberg (1999) a cópia impressa em papel fixa os objetos digitais simples como um todo, mantêm o conteúdo e, de certo modo, o leiaute, sendo uma ação paliativa, operosa e aplicável na ausência de uma infraestrutura tecnológica. Não obstante, os autores inferem que a impressão resulta na perda de funcionalidade interativa ou dinâmica e da forma original de objetos complexos restringindo a prática desta abordagem. Por isso, outra opção seria a estratégia híbrida de uso de cópias em microfilme como substitutos arquivísticos e da geração de cópias digitais pois, consoante Schäfer e Constante (2012) e Thomaz e Soares (2004), possibilita a reformatação dos documentos produzidos originalmente em papel e a melhoria da funcionalidade e da acessibilidade.

d) Emulação

A emulação propõe a criação e o uso de um emulador moderno que substitua e reproduza o comportamento de tecnologias de *hardware* e de *software* antigas e/ou obsoletas. Como observado em Barbedo, Corujo e Sant'ana (2011), Long (2009) e Rothenberg (1999), este método mantêm o conteúdo e a visualização dos objetos digitais em seu formato nativo com leiaute e funcionalidade original, sendo livre da manutenção de plataformas e sistemas específicos e, também, útil quando há interesse na preservação do contexto tecnológico original dos objetos. Porém, os autores expõem que a emulação esta suscetível aos riscos da dependência e obsolescência dos emuladores e supõe limitações com o tempo na capacidade de representação fidedigna dos materiais. Assim, conforme Santos e Flores (2015), a emulação se aplica junto com encapsulamento e em troca da conservação de tecnologia, tendo função em curto e médio prazo através de altos recursos técnicos e financeiros.

e) Conservação de tecnologia

A conservação de tecnologia trata-se de um método interino e de curto prazo. Baseado em Márdero Arellano (2008), Bullock (1999) e Santos e Flores (2017a) envolve a manutenção do *hardware* e do *software* original de criação ou acesso aos objetos digitais a fim de disponibilizá-lo para uso. Para os autores esta estratégia mantêm o conteúdo e a visualização

dos materiais digitais em seu formato nativo com leiaute e funcionalidade original, porém a criação de "museus" não reduz os efeitos da obsolescência tecnológica e requer condições de custo, espaço e suporte técnico de difícil operação, tornando o seu uso útil para objetos valiosos[284] em formato proprietário e em *software* obsoleto. Um dos papéis possíveis dos "museus de computadores" na preservação digital, segundo Rothenberg (1999), pode estar em efetuar esforços heroicos de recuperação de dados legíveis em mídias antigas e na verificação de emuladores cotejando seu comportamento com o de máquinas obsoletas salvas.

f) Arqueologia digital

A arqueologia digital é um método dispendioso e parcial de preservação. Como analisado em Baggio e Flores (2012), Hedstrom (2001) e Schäfer e Constante (2012) consiste no resgate de materiais digitais inacessíveis, seja pela obsolescência tecnológica e/ou pela degradação física do suporte, os quais não foram atendidos por outras estratégias ou ficaram carecidos de qualquer ação de preservação. Para os autores a arqueologia digital é indicada somente para situações em que a relevância das informações legitime os elevados custos do procedimento, uma vez que não existem garantias de recuperação, restauração e interpretação da plenitude dos dados que, dessa maneira, comprometem a definição da identidade, da integridade e do contexto do material recuperado. A título de exemplo, Galrão (2017) traz um caso prático de dados digitais submetidos à arqueologia digital, onde se constatou que apenas uma parte da documentação pôde ser somente interpretada.

g) Arquivamento da *Web*

O arquivamento da *Web* alude o processo de seleção e coleta, armazenamento, recuperação, acesso e preservação em longo prazo de conteúdos na *Web*. Através de Costa, Gomes e Silva (2017) e Pennock (c2013), a execução destes métodos inclui se defrontar com algumas dificuldades, como o amplo volume de informações que são perdidas ou ficam indisponíveis rapidamente em sua forma original pela dinâmica da *Web* e as condições de legalidade e de licenças/permissões do proprietário do conteúdo. Para apoiar o avanço e a preservação da *Web*, iniciativas vêm sendo criadas no mundo (por exemplo, o *Internet*

[284] Embora seja tida em declínio, a conservação de tecnologia ainda é utilizada por algumas organizações. Por exemplo, a Biblioteca Nacional da Austrália tem atuado na coleta de doações de *hardwares* e *softwares* obsoletos para comporem a sua coleção, a fim de apoiar a recuperação de dados valiosos em formatos digitais desatualizados com a aplicação de emuladores para a criação de ambientes virtuais que possibilitam o uso de tecnologias ultrapassadas (THORPE, 2015).

Archive[285] e o Arquivo.pt[286]) como ferramentas e técnicas são desenvolvidas por consórcios, frisando o W3C[287] e o IIPC. Na esfera nacional, de acordo com Rockembach e Pavão (2018), o tema é atual na área da Ciência da Informação e não há ainda iniciativas sistematizadas[288], implicando a carência de uma memória cultural da *Web* brasileira às gerações presentes e futuras.

À vista dos métodos discutidos, observamos que não há soluções plenamente satisfatórias e definitivas quando aplicadas isoladamente, sendo necessário uma combinação de estratégias para a preservação digital. Não obstante, as estratégias operacionais toleram alterações nas propriedades originais dos objetos digitais que poderão suceder em perda significativa de dados no decorrer do tempo aos usuários finais ou, ainda, preveem restrições de acesso, uso e preservação de conteúdos atreladas a condições legais e a disposição de recursos financeiros, humanos e tecnológicos; o qual contrariam os princípios da preservação de longo prazo. Estas questões são potencializadas com a rápida obsolescência tecnológica e a crescente complexidade e interatividade dos objetos digitais, refletindo a urgência de maiores estudos, políticas e tecnologias padronizadas e em colaboração.

Assim, as estratégias estruturais (sobretudo, orçamento e custos, metadados de preservação e formação de redes de relações) podem vir a mitigar os problemas supracitados. O uso de modelos para o cálculo de custos, como Boté, Fernandez-Feijoo e Ruiz (2013) e Willer *et al*. (2008), ajudam nas decisões sobre investimentos em preservação e o registro das alterações por metadados trazem garantias de que as perdas ocorridas não afetaram a confiabilidade, autenticidade e integridade dos materiais (e suas propriedades significativas); mas, são métodos carentes de estudos na Ciência da Informação nacional. Ademais, as parcerias

---

[285] Criado em 1996, o *Internet Archive* é uma organização sem fins lucrativos que fornece acesso gratuito e universal a uma biblioteca digital de páginas *Web*, livros, textos, vídeos, músicas, imagens, *softwares* etc., através do seu *website* oficial e pela ferramenta *Wayback Machine*. Disponível em: https://archive.org/about/. Acesso em: 22 maio 2023.

[286] O arquivo da *Web* portuguesa, iniciado em 2008, consiste numa infraestrutura que propicia a pesquisa e o acesso a páginas *Web* de Portugal arquivadas desde 1996, objetivando a preservação da informação publicada na *Web* para fins de investigação. Disponível em: https://sobre.arquivo.pt/pt/ajuda/o-que-e-o-arquivo-pt/. Acesso em: 22 maio 2023.

[287] Liderado por Tim Berners-Lee e fundado em 1994, o W3C é "[...] um consórcio internacional que desenvolve protocolos de consenso e especificações para garantir a interoperabilidade da *World Wide Web*." (NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004, p. 16, tradução nossa), sendo que os órgãos membros (a *Library of Congress* e a *Online Computer Library Center* – OCLC, por exemplo) criam padrões com a missão de crescimento da *Web* a longo prazo. Disponível em: https://www.w3.org/Consortium/. Acesso em: 22 maio 2023.

[288] Apesar de não haver iniciativas organizadas de arquivamento da *Web* no Brasil, nos últimos anos, grupos de pesquisa têm buscado investigar o tema, tais como o NUAWEB, da Faculdade de Biblioteconomia e Comunicação (FABICO) da UFRGS (http://dgp.cnpq.br/dgp/espelhogrupo/1769372358627653); e o grupo "Estudos e Práticas de Preservação Digital" ou Rede DRÍADE (http://dgp.cnpq.br/dgp/espelhogrupo/3997875180380796), da Rede Cariniana do IBICT, que possui grupos de estudo sobre arquivamento de *e-mail*, preservação de dados de pesquisa, de teses e dissertações e/ou de periódicos eletrônicos etc. (https://cariniana.ibict.br/?page_id=341).

colaborativas cedem respostas a falta de subsídios e a superação de entraves onde, através de Márdero Arellano (2012) e Farias, Araújo e Evangelista (2018), pode ser evidenciado pelo avanço expressivo da preservação digital na Rede Cariniana com o LOCKSS.

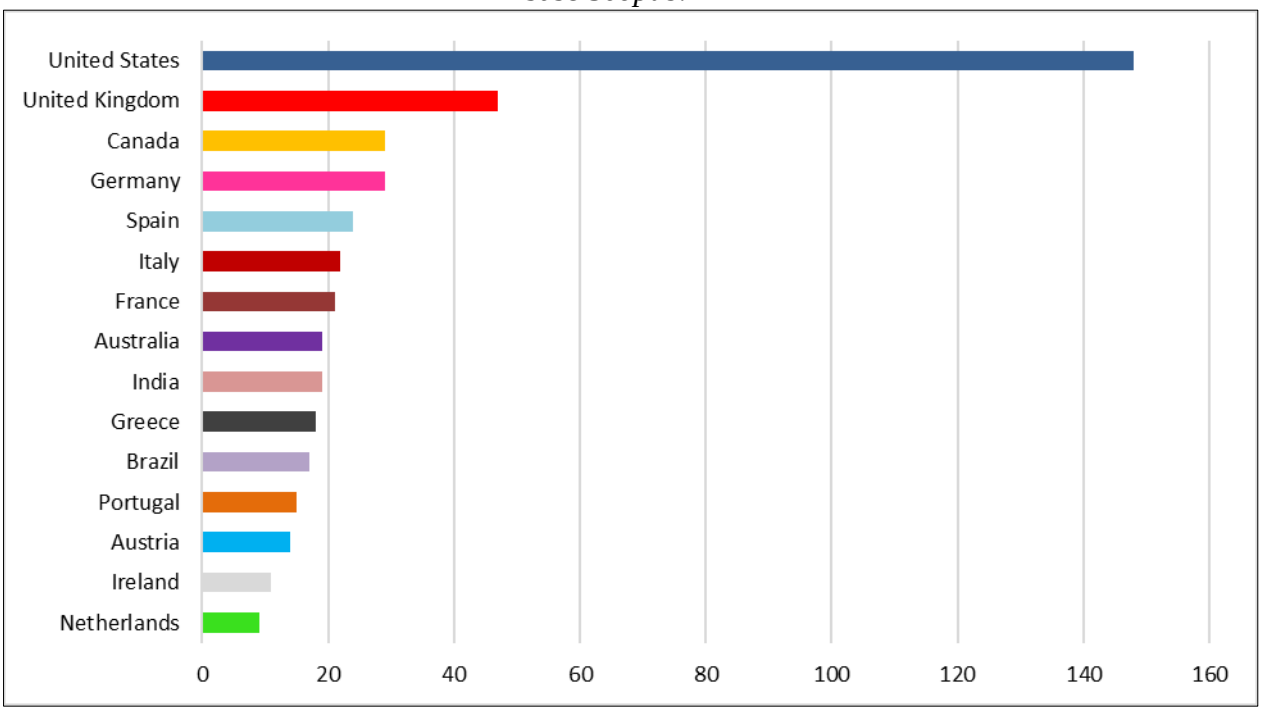
Considerando as publicações nas bases *Scopus* e *Web of Science*, a seguir apresentamos os somatórios da produção científica nos últimos anos sobre preservação digital, realçando os autores mais produtivos como as instituições, países e áreas de pesquisa com maior número de publicações.

3.6 Análise das publicações em preservação digital nas bases *Scopus* e *Web of Science*

No levantamento bibliográfico da base *Scopus*, através dos procedimentos metodológicos descritos, foram recuperados 274 artigos (56,5%) e 211 anais de conferências (43,5%), totalizando 485 documentos. Quanto aos autores que publicaram no período definido, os mais produtivos foram Efstratios Kontopoulos do Centro de Pesquisa e Tecnologia *Hellas* (do inglês *Centre for Research and Technology-Hellas* – CERTH)[289] na Grécia e Michael L. Nelson da Universidade de *Old Dominion* nos Estados Unidos, com 9 publicações cada. As instituições que detêm maior produção científica são as Universidades de *Illinois* em *Urbana-Champaign* e de *Old Dominion* nos Estados Unidos e a Universidade de Toronto no Canadá[290], além disso, os países que mais publicaram foram o Reino Unido e, especialmente, os Estados Unidos como a Figura 59.

[289] Disponível em: https://www.certh.gr/root.en.aspx. Acesso em: 23 maio 2023.
[290] "Instituições que detêm maior produção científica" remete as afiliações institucionais ou as proveniências dessas afiliações dos autores das publicações, no período de 2015 a 2019, do levantamento bibliográfico na base *Scopus*.

Figura 59 – Número de publicações em "*digital preservation*" por país[291] (2015-2019) na base *Scopus*.



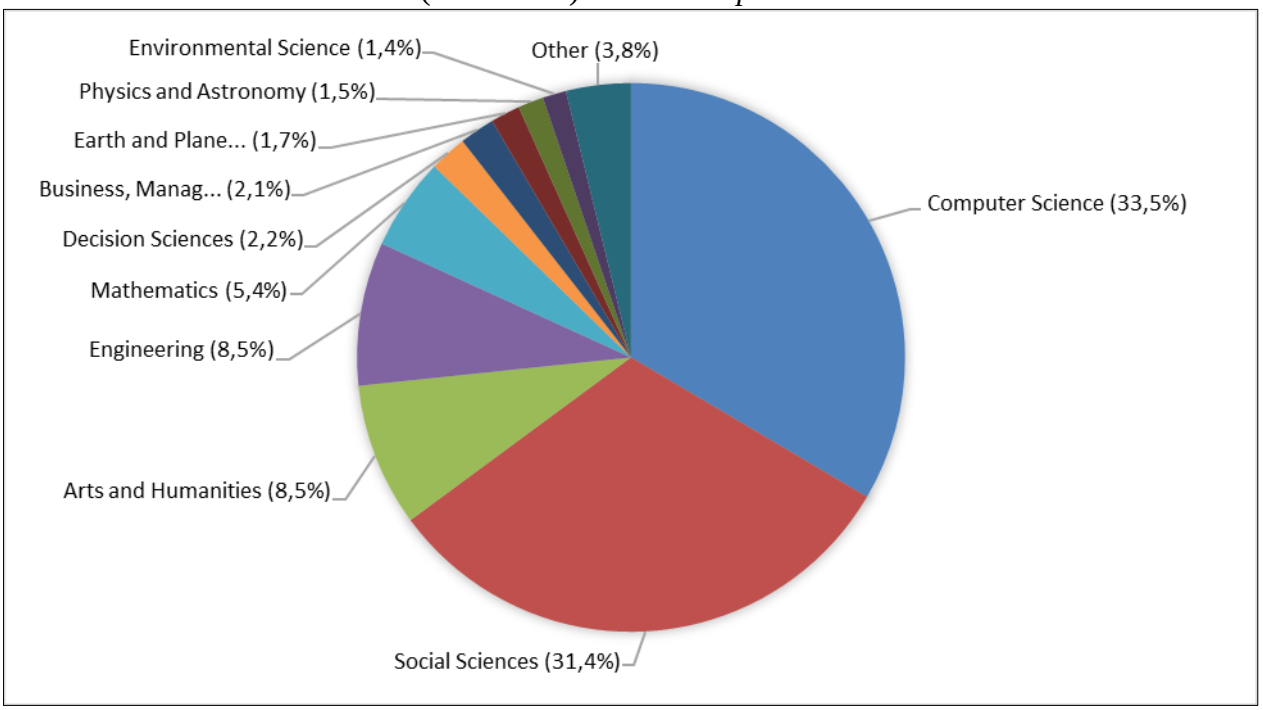
Fonte: Elaborado pelo autor.

A Universidade de *Old Dominion*, membro do IIPC mediante Departamento de Ciência da Computação, ainda coordena junto ao Laboratório Nacional de *Los Alamos* dos Estados Unidos o *Memento Framework* (RFC 7089)[292] e a ferramenta *Time Travel*[293] para busca de versões anteriores de *websites* armazenados em arquivos da *Web*, como Arquivo.pt e *Internet Archive*. Também parte das publicações da Universidade de Toronto afiliam-se a Faculdade de Informação que dispõe do Instituto de Curadoria Digital[294], uma unidade interdisciplinar de pesquisadores para investigação da preservação de recursos digitais. Das publicações do Reino Unido, se destaca as Universidades de *Glasgow*, de *Edinburgh* e de *Cambridge*, membros da DPC[295], uma organização associativa sem fins lucrativos fundada em 2002 e dedicada a preservação digital.

Reforçando a pesquisa de Rockembach e Pavão (2018), a multi e interdisciplinaridade da preservação digital também manifesta nos documentos recuperados, segundo exposto na Figura 60.

[291] Publicações "por país" referem-se as afiliações institucionais ou as proveniências dessas afiliações dos autores das publicações, no período de 2015 a 2019, do levantamento bibliográfico na base *Scopus*.
[292] Disponível em: https://datatracker.ietf.org/doc/html/rfc7089. Acesso em: 23 maio 2023.
[293] Disponível em: http://timetravel.mementoweb.org/. Acesso em: 23 maio 2023.
[294] Disponível em: http://dci.ischool.utoronto.ca/about-the-dci/. Acesso em: 23 maio 2023.
[295] Disponível em: https://www.dpconline.org/about. Acesso em: 23 maio 2023.

Figura 60 – Distribuição das publicações em "*digital preservation*" por área de pesquisa (2015-2019) na base *Scopus*.



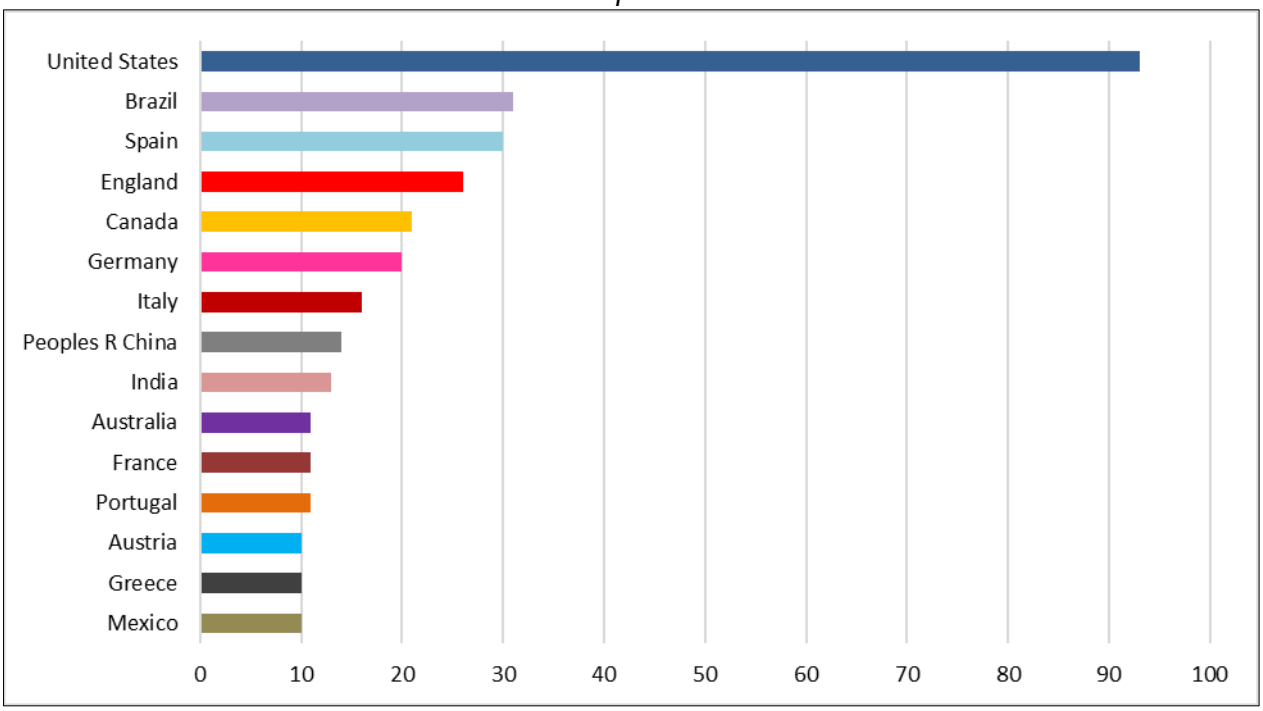
Fonte: Elaborado pelo autor.

Em vista disto, verificamos que as duas grandes áreas de pesquisa que estudam e publicam sobre o tema de preservação digital agrupam-se na Ciência da Computação e nas Ciências Sociais, junto de Artes e Humanidades, Engenharia, Matemática, *Decision Sciences*, Negócios, Gestão e Contabilidade, *Earth and Planetary Sciences*, Física e Astronomia, Ciência Ambiental e outras. Dos termos utilizados nas palavras-chave dos 485 documentos recuperados, foram 241 registros para "*Digital Storage*", 25 registros para "*Information Management*", 23 registros para "*Metadata*", 17 registros para "*Web Archiving*", 15 registros para "*Digital Repository*", 12 registros para "OAIS" e 8 registros para "*Collaboration*", reproduzindo parte dos aspectos do tema tratados neste trabalho.

Por sua vez, no levantamento bibliográfico da base *Web of Science*, através da metodologia descrita, foram recuperados 363 documentos. Sobre aos autores que publicaram no período, os mais produtivos foram Michael L. Nelson e Michele C. Weigle, da Universidade de *Old Dominion*, com 8 e 6 publicações cada, respectivamente; e Daniel Flores, da Universidade Federal de Santa Maria (UFSM) no Brasil, com 7 publicações. As duas grandes áreas de pesquisa que publicam acerca do tema concentram-se na Biblioteconomia e Ciência da Informação e na Ciência da Computação. As instituições que detêm maior produção

científica são a Universidade de *Old Dominion* e a UFSM[296], e os países que mais publicaram foram o Brasil e, mormente, os Estados Unidos como a Figura 61.

Figura 61 – Número de publicações em "*digital preservation*" por país[297] (2015-2019) na base *Web of Science*.



Fonte: Elaborado pelo autor.

As publicações dos Estados Unidos afiliam-se, sobretudo, a Universidade de *Oklahoma* em que mediante a Biblioteca de Direito Donald E. Pray participa da *Legal Information Preservation Alliance* (LIPA)[298], um consórcio de bibliotecas criado em 2003 e dirigido a projetos de preservação de informações jurídicas impressas e eletrônicas. Por fim, nas publicações do Brasil, a sua maioria associa-se a UFSM e o IBICT, acompanhado da UnB, Universidade de São Paulo (USP), da Universidade Federal da Paraíba (UFPB), da Universidade Federal de Santa Catarina (UFSC), da Universidade Federal de Goiás (UFG), dentre outras, instituições estas que configuram parte dos parceiros integrais da Rede Cariniana[299] coordenada pelo próprio IBICT.

Diante dos resultados apresentados, vemos que nos últimos anos a produção científica sobre preservação digital se manteve em geral nos países da Europa e da América do Norte,

[296] "Instituições que detêm maior produção científica" remete as afiliações institucionais ou as proveniências dessas afiliações dos autores das publicações, no período de 2015 a 2019, do levantamento bibliográfico na base *Web of Science*.
[297] Publicações "por país" referem-se as afiliações institucionais ou as proveniências dessas afiliações dos autores das publicações, no período de 2015 a 2019, do levantamento bibliográfico na base *Web of Science*.
[298] Disponível em: https://www.lipalliance.org/history-of-lipa. Acesso em: 23 maio 2023.
[299] Disponível em: https://cariniana.ibict.br/?page_id=222. Acesso em: 23 maio 2023.

considerando as limitações da representação de periódicos, áreas e países das bases tratadas no presente trabalho. Ainda assim, países como o Brasil que, por sua vez, detêm as primeiras publicações sobre o tema no ano de 2007 e 2003 na *Scopus* e na *Web of Science*, nesta ordem, destaca-se pelas quantias notáveis de estudos nas Ciências Sociais (em especial, Biblioteconomia e Ciência da Informação). Reforçando Boeres (2017), os dados indicam uma maduração do tema (e sua relevância) no Brasil ante a maior solidez deste no exterior, condizendo com o crescimento de iniciativas nacionais onde, por exemplo, o uso do LOCKSS ensejou a aplicação de saberes teóricos em preservação digital (MÁRDERO ARELLANO, 2012).

3.7 Atualização do estado da arte da produção científica em preservação digital

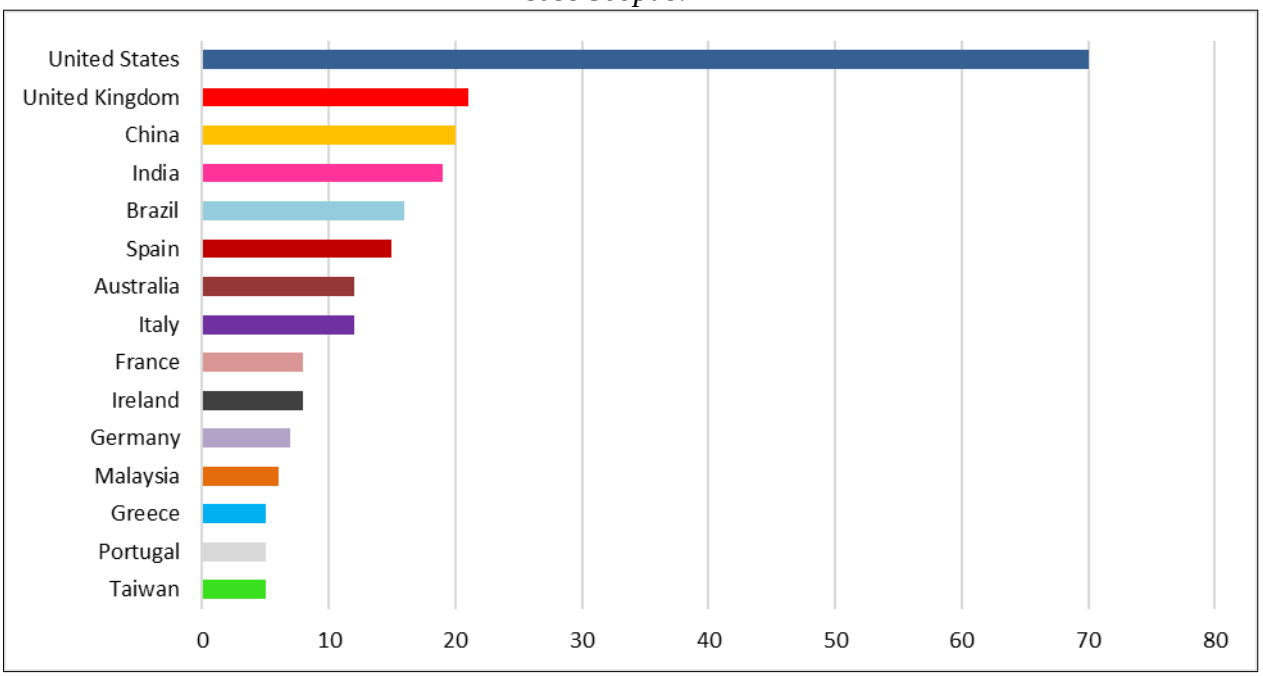
Para fins de atualização dos dados exibidos anteriormente no trabalho, baseando-se nos mesmos procedimentos metodológicos descritos (mas com a definição de filtro por documentos publicados de 2020 a 2022), foram realizados mais uma vez levantamentos bibliográficos nas bases *Scopus* e *Web of Science* em 30 de abril de 2023[300].

No levantamento da base *Scopus*, foram recuperados 212 artigos (75,2%) e 70 anais de conferências (24,8%), totalizando 282 documentos. Quanto aos autores que publicaram no período definido, os mais produtivos foram Amelia Acker da Universidade do *Texas* em *Austin* nos Estados Unidos e Rebecca D. Frank da Universidade *Humboldt* de Berlim na Alemanha e da Universidade do *Tennessee* nos Estados Unidos, com 4 publicações cada. As instituições que detêm maior produção científica são a Universidade do *Texas* em *Austin* e a Universidade Federal de Minas Gerais (UFMG) no Brasil[301]; além disto, os países que mais publicaram foram o Reino Unido e, principalmente, os Estados Unidos, conforme a Figura 62.

[300] Há de se levar em conta a possibilidade de haver pequenas variações no número total de documentos recuperados nas bases de dados caso novas buscas sejam feitas novamente em acordo com os procedimentos metodológicos explicitados no trabalho, visto que é provável que estudos publicados no período definido na metodologia poderão ser indexados pelas respectivas bases num período posterior a data de realização dos levantamentos bibliográficos.

[301] "Instituições que detêm maior produção científica" remete as afiliações institucionais ou as proveniências dessas afiliações dos autores das publicações, no período de 2020 a 2022, do levantamento bibliográfico na base *Scopus*.

Figura 62 – Número de publicações em "*digital preservation*" por país[302] (2020-2022) na base *Scopus*.



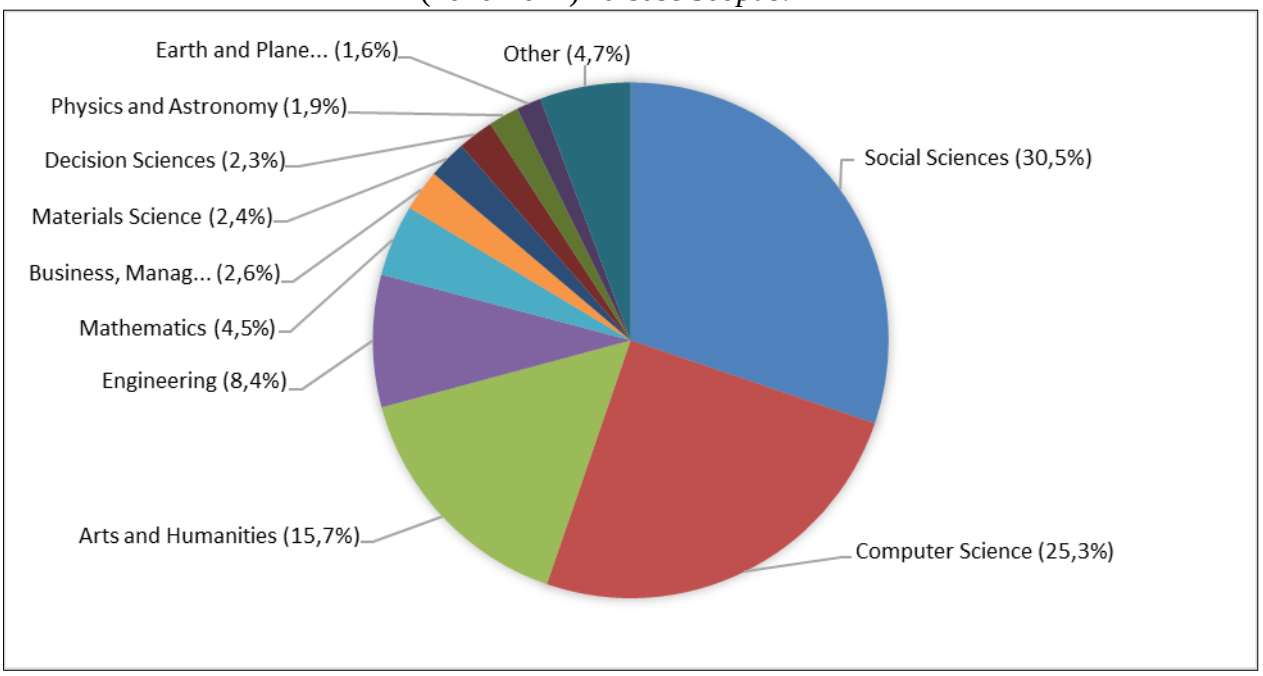
Fonte: Elaborado pelo autor.

A Universidade do *Tennessee* e a Universidade do *Texas* em *Austin* são ambos membros da NDSA[303] tendo como foco de preservação digital, nesta ordem, dados geoespaciais e inovação, e imagem/texto, ferramentas/infraestrutura, padrões/melhores práticas e armazenamento distribuído.

Novamente, a multi e interdisciplinaridade da preservação digital da mesma forma se manifesta nos documentos recuperados, segundo exposto no Figura 63.

[302] Publicações "por país" referem-se as afiliações institucionais ou as proveniências dessas afiliações dos autores das publicações, no período de 2020 a 2022, do levantamento bibliográfico na base *Scopus*.
[303] Disponível em: https://ndsa.org/membership/members/. Acesso em: 23 maio 2023.

Figura 63 – Distribuição das publicações em "*digital preservation*" por área de pesquisa (2020-2022) na base *Scopus*.



Fonte: Elaborado pelo autor.

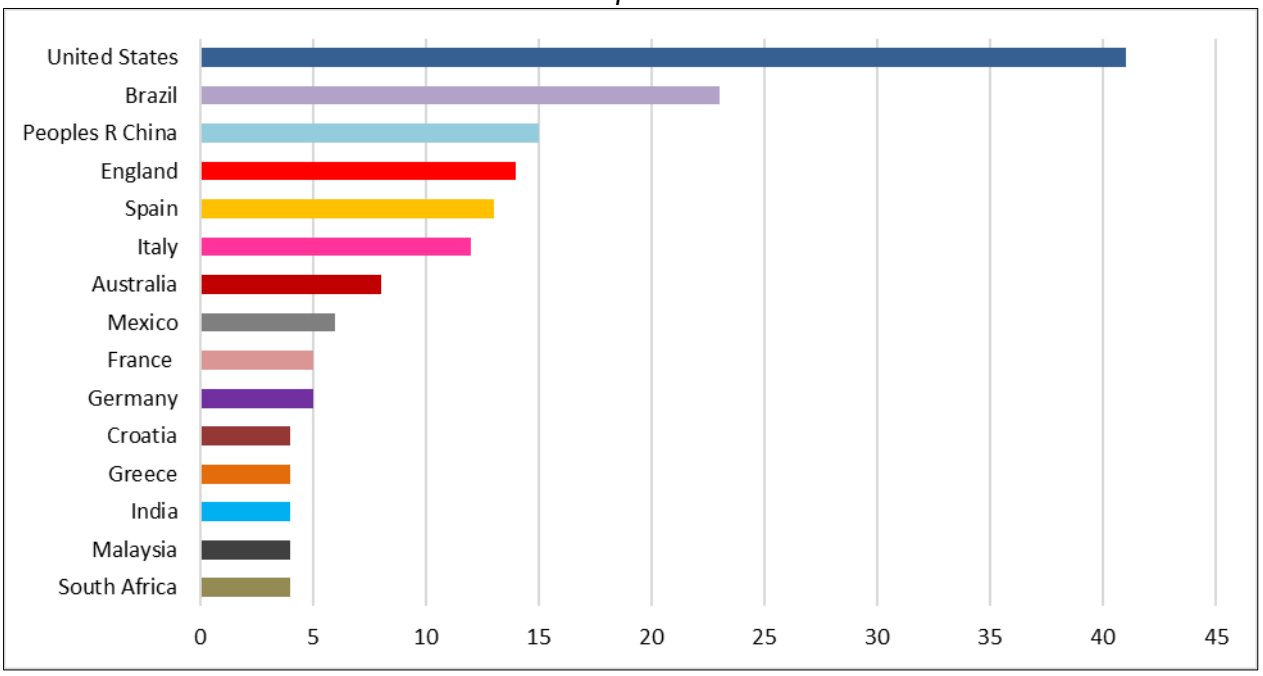
Deste modo, constatamos que as duas grandes áreas de pesquisa que estudam e publicam sobre o tema de preservação digital agrupam-se nas Ciências Sociais e Ciência da Computação (oposto ao que foi verificado na Figura 60), acompanhado de Artes e Humanidades, Engenharia, Matemática, Negócios, Gestão e Contabilidade, *Materials Science*, *Decision Sciences*, Física e Astronomia, *Earth and Planetary Sciences*, e outras. Dos termos utilizados nas palavras-chave dos 282 documentos recuperados, foram 85 registros para "*Digital Storage*", 12 registros para "*Metadata*", 11 registros para "*Web Archiving*" e "*Information Management*", 6 registros para "OAIS", 3 registros para "*Documentation*", "*Digital Preservation Strategies*" e "*Standards*", e 2 registros para "*Collaboration*", reproduzindo igualmente parte dos aspectos do tema tratados neste trabalho.

Adicionalmente, por meio dos procedimentos metodológicos descritos (porém com a definição de filtro por documentos publicados somente no ano de 2023), na base *Scopus* foram recuperados até o mês de abril de 2023 exatamente 8 artigos (72,7%) e 3 anais de conferências (27,3%), totalizando 11 documentos.

No levantamento da base *Web of Science*, foram recuperados 166 artigos (87%) e 25 artigos de conferência (13%), totalizando 191 documentos. Quanto aos autores que publicaram no período definido (2020-2022), os mais produtivos foram outra vez Rebecca D. Frank e Amelia Acker, com 4 e 3 publicações cada, respectivamente. As duas grandes áreas de pesquisa que publicam sobre o tema centram-se na Biblioteconomia e Ciência da Informação e na

Ciência da Computação (igual ao levantamento bibliográfico no período de 2015 a 2019). As instituições que detêm maior produção científica são a Universidade do *Texas* nos Estados Unidos, o *Consiglio Nazionale delle Ricerche*[304] na Itália, a UFPB, a UFMG e a UFRGS no Brasil[305]; ademais, os países que mais publicaram continuaram sendo o Brasil e, sobretudo, os Estados Unidos, conforme a Figura 64.

Figura 64 – Número de publicações em "*digital preservation*" por país[306] (2020-2022) na base *Web of Science*.



Fonte: Elaborado pelo autor.

As publicações dos Estados Unidos afiliam-se, principalmente, a Universidade do *Texas* em *Austin* e a Universidade de *Indiana*. Esta última também é membro da NDSA, através das suas bibliotecas, tendo como foco de preservação digital ferramentas/infraestrutura.

Por último, nas publicações do Brasil, a sua maioria associa-se a UFMG, a UFPB, a UFRGS e a UFSCar, dentre outras.

Em adição, por meio dos procedimentos metodológicos descritos (mas com a definição de filtro por documentos publicados apenas no ano de 2023), na base *Web of Science* foram recuperados até o mês de abril de 2023 exatamente 5 artigos e 1 artigo de conferência, totalizando 6 documentos.

[304] Disponível em: https://www.cnr.it/. Acesso em: 23 maio 2023.
[305] "Instituições que detêm maior produção científica" remete as afiliações institucionais ou as proveniências dessas afiliações dos autores das publicações, no período de 2020 a 2022, do levantamento bibliográfico na base *Web of Science*.
[306] Publicações "por país" referem-se as afiliações institucionais ou as proveniências dessas afiliações dos autores das publicações, no período de 2020 a 2022, do levantamento bibliográfico na base *Web of Science*.

Ao fornecer uma visão ampla e reflexiva das principais questões da preservação digital, a pesquisa feita mostra um quadro atual de desafios, experiências e oportunidades entre instituições e profissionais envolvidos na criação, aquisição e gerenciamento de materiais digitais. Abrangendo diversos tipos de objetos digitais – textos, imagens, áudios, vídeos, *softwares*, jogos, conteúdos em mídias sociais, páginas da *Web* etc. – as estratégias para preservação digital, em resguardo a uma possibilidade de as gerações futuras terem pouco ou nenhum registro do século XXI (tido por Vint Cerf, um dos fundadores da *Internet*, como uma "idade das trevas digital")[307], ainda apresentam um caráter relativamente insuficiente de conhecimentos e ensaios práticos na comunidade científica.

Apesar do destaque da Rede Cariniana em publicações obtidas nas bases de dados tratadas no trabalho, a maioria das instituições parceiras da Rede não têm uma estimativa de crescimento das atividades e nem prospectiva de implementação de estratégias nos próximos anos em razão da falta de pessoal especializado, de orçamentos próprios e de apoio tecnológico (FARIAS; ARAÚJO; EVANGELISTA, 2018). Unido a esta realidade, mesmo na condição de estar entre os países líderes no número de usuários em mídias sociais, como *Twitter*[308] e *Facebook*[309], ou também de possuir em 2019 aproximadamente 149 milhões de usuários da *Internet*[310], no Brasil o arquivamento da *Web* é um dos temas mais em ascensão e carente de ações efetivas e de estudos na Ciência da Informação. Além do mais, é pertinente indicar que no âmbito internacional os temas de preservação digital e, também, de arquivamento da *Web* (ROCKEMBACH; PAVÃO, 2018), têm abrangência interdisciplinar, o que fundamenta as suas investigações no campo dos estudos CTS através deste trabalho no PPGCTS, porém no Brasil, eles se manifestam diretamente com o campo da Ciência da Informação e suas áreas afins, justificando igualmente o recorte disciplinar assumido no trabalho.

Não obstante, outros temas relacionados à preservação digital constatados na pesquisa são hoje igualmente pouco analisados por esta grande área de pesquisa no panorama nacional, ou seja: o planejamento orçamentário e o cálculo dos custos aproximados para preservação, que trazem um entendimento sobre a viabilidade de investimentos e efetivação de cooperações ou terceirização; e os padrões de metadados de preservação, estruturas formais de descrição de recursos, que registram informações contextuais e de proveniência para que os conteúdos

---

[307] Disponível em: https://www.bbc.com/news/science-environment-31450389. Acesso em: 23 maio 2023.
[308] Disponível em: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/. Acesso em: 15 mar. 2020.
[309] Disponível em: https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/. Acesso em: 15 mar. 2020.
[310] Disponível em: https://www.statista.com/statistics/262966/number-of-internet-users-in-selected-countries/. Acesso em: 15 mar. 2020.

digitais sejam interpretados no presente e futuro. Ambos os temas são vitais na preservação digital de larga escala em arquivos da *Web*, sistemas etc., dado a alta produção de objetos complexos, heterogêneos e com dependências.

Em vista disso, com base nas discussões deste trabalho, propomos que além da urgência de maiores estudos nacionais da Ciência da Informação sobre preservação digital, as novas pesquisas poderiam focar as dificuldades e os tópicos pouco explorados identificados, prevendo a ampliação de nossos saberes e experiências nesse domínio. Há que cuidar da conscientização da importância do processo aos que produzem, usam e acessam objetos digitais e do dever das partes interessadas em todo o ciclo de vida do objeto, como da definição de parâmetros para avaliação dos efeitos das abordagens de preservação vigentes (e o uso de emuladores, tecnologias de arquivamento etc.), no intuito de não cometermos um distanciamento dos princípios da preservação digital de longo prazo.

4 PADRÕES DE METADADOS NO ARQUIVAMENTO DA *WEB*: recursos tecnológicos para a garantia da preservação digital de *websites* arquivados[311]

Inventada por Tim Berners-Lee em 1989, a *World Wide Web* é um registro exclusivo da vida no século XXI e recurso de informação único que hospeda milhões de *sites* onde conectam diferentes comunidades e indivíduos no mundo (PENNOCK, c2013). Porém, como *Library of Congress* ([2022a]), Masanès (c2006b), Pennock (c2013) e Rockembach e Pavão (2018), a própria velocidade com que se este ambiente se desenvolve e as informações são publicadas e movem-se ao esquecimento rapidamente com a *Internet* e o uso intensivo de tecnologias, somado aos *websites*[312] ao vivo em si (e suas páginas da *Web* que testemunham eventos atuais, organizações, reações públicas, ou informações governamentais, culturais e acadêmicas acerca de uma ampla variedade de assuntos em múltiplas mídias) os quais são criados com rapidez e os seus URLs e conteúdos mudam regularmente e, por vez, somem completamente constituindo objetos digitais complexos, dinâmicos e efêmeros, representam uma ameaça muito real à nossa memória individual, organizacional, de fatos, ou cultural digital, seu legado técnico, evolução e história social. Em admissão desta ameaça, através de Costa, Gomes e Silva (2017), organizações do mundo todo – sobretudo, universidades e instituições de patrimônio cultural, como bibliotecas, arquivos e museus – tem investido em políticas, métodos e tecnologias para coletar, preservar ao longo do tempo e tornar acessíveis cópias arquivadas do conteúdo da *Web*.

Nas últimas décadas, a preservação digital fez-se uma temática de estudo que se firmou na Ciência da Informação, tal como já mencionado anteriormente. É um problema emergente, coletivo e atual nas publicações nacionais e internacionais do campo, exigindo análises inter e multidisciplinares e soluções sustentáveis, integradas e colaborativas. Para maior clareza do conceito, a preservação digital é definida por Hedstrom (1998, p. 190, tradução nossa) como sendo "[...] o planejamento, alocação de recursos e aplicação de métodos de preservação e tecnologias necessárias para garantir que informações digitais de valor contínuo permaneçam acessíveis e utilizáveis". Além da emulação tecnológica, ou da migração de dados onde a

[311] Parte do texto original deste capítulo de Tese de Doutorado foi publicado como artigo na RDBCI vinculada ao Sistema de Bibliotecas da UNICAMP, volume 20, de 2022, sob o título de "Padrões de metadados no arquivamento da *Web*: recursos tecnológicos para a garantia da preservação digital de *websites* arquivados".

[312] Como *Premis Editorial Committee* (2015) e *Web Page* (2022d), um *site* é uma coleção de páginas *Web* interligadas no mesmo local na *Internet*, ou melhor, documentos *HyperText Markup Language* (HTML)/*Extensible Hypertext Markup Language* (XHTML) (onde as extensões de arquivo são .htm e .html) acessíveis via o *Hypertext Transfer Protocol* (HTTP) na *Internet* e com *links* de hipertexto para a navegação de uma página ou seção para outra, assim o conjunto de todos os *sites* acessíveis ao público e existentes engloba a *Web*; ademais, acessadas por um URL *root* comum ou *homepage*, as páginas de um *site* utilizam arquivos gráficos associados (que podem fazer-se *links* clicáveis) para fornecer ilustração como, também, os seus URLs as organizam em uma hierarquia, apesar de que os *hyperlinks* entre eles controlem como o leitor percepciona a estrutura geral do *website*.

XML[313] é tida, segundo Márdero Arellano (2008), como um tipo especial de migração, uma das estratégias de preservação digital se refere à adoção efetiva de padrões de metadados de forma a apoiar a interpretação, a gerência e a preservação por longo prazo de objetos digitais em meios informacionais, como repositórios.

Outra estratégia notável abrange a preservação digital do conteúdo de *websites*. Sendo um tema novo e carente de pesquisas e iniciativas sistematizadas no Brasil (ROCKEMBACH; PAVÃO, 2018), o arquivamento da *Web* (*web archiving*) envolve cinco etapas, descritas em Kim e Lee (2007) e Masanés (c2006b), a saber: seleção (incluindo as fases de preparação – definir o objetivo da coleta, política de captura e ferramentas –, de descoberta – fixar os pontos de entrada para a captura, como a frequência e o escopo desta, e de filtragem para reduzir o espaço aberto pela fase anterior aos limites na política de seleção), captura, arquivamento, acesso e revisão de qualidade; podendo este processo ser extensivo (não seletivo), intensivo (seletivo), centrado no tópico (temático) e/ou no domínio de *sites*. Unido aos arquivos da *Web* surgidos, ferramentas, padrões e técnicas para o avanço e a preservação da *Web* são criadas por consórcios: o W3C mais o IIPC.

Integrados em páginas *Web* na forma de *links* de uma página para outras e registros do comportamento do usuário (RILEY, c2017), metadados (e padrões de metadados) têm a função de descrever unicamente um recurso informacional em ambientes digitais, multidimensionando suas formas de acesso e uso, assegurando sua representação e recuperação pelo usuário. Como exemplo, no domínio *Web*[314], o principal padrão é o DC; no domínio arquivístico e museológico[315], há o EAD, e o *Visual Resources Association* (VRA) *Core* além do *Categories*

---

[313] XML remete "[...] uma forma estendida de HTML que permite conjuntos de *tags* definidos localmente e o fácil intercâmbio de informações estruturadas." (NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004, p. 2-3, tradução nossa) consistindo, segundo Baca (c2016), numa linguagem de marcação simples e flexível para publicação e intercâmbio de ampla variedade de dados na *Web*. Publicado em 1998, a XML foi desenvolvida e é oficialmente uma recomendação do W3C. Disponível em: https://www.w3.org/XML/. Acesso em: 24 maio 2023.

[314] O domínio *Web*, em conformidade com Alves (2010), alude o universo referente aos diversos tipos de ambientes digitais de informação e os vários tipos de recursos disponíveis por diferentes áreas do saber. "As representações, entretanto, geradas nesses domínios, serão diferentes das representações geradas no domínio bibliográfico e correspondentes às necessidades informacionais próprias de cada um deles." (ALVES; SANTOS, 2013, p. 16).

[315] Como o domínio bibliográfico, nos domínios arquivístico e museológico os padrões de metadados são complexos e muito estruturados com esquemas detalhados, formais e criados ante seus princípios, normas e códigos próprios (ALVES; SANTOS, 2013). Conforme Alves (2017) no domínio arquivístico os metadados e padrões de metadados estabeleceram-se em princípios e teorias arquivísticas como na descrição e nos processos de gestão arquivística de documentos; já no domínio museológico estas estruturas instituíram-se em contexto internacional, sendo que neste domínio elaboraram-se instrumentos (padrões de conteúdo etc.) para a interoperabilidade de dados do patrimônio cultural; os instrumentos gerados em domínios específicos (bibliográfico, arquivístico etc.) devem ser sabidos para criar representações ideais e padronizadas que concedam a recuperação eficaz de recursos informacionais da *Web*.

for the Description of Works of Art (CDWA)[316]; e no domínio bibliográfico[317], destacamos o MARC[318] e o MODS. Consoante Formenton *et al*. (2017) os metadados ainda definem a garantia da preservação digital de um recurso/objeto digital (por exemplo, *sites* arquivados), através de padrões de metadados específicos, como o PREMIS em conjunto ao METS.

Os padrões de metadados, sejam eles para descrição, gerência ou preservação de objetos digitais, são recursos tecnológicos-chave na interoperabilidade[319]. Esta função está assegurada por práticas e padrões de descrição que se traduz nas sintaxes para codificação de dados, como a XML e a *Standard Generalized Markup Language* (SGML)[320] *Document Type Definition* (DTD)[321], além dos padrões de conteúdo (regras e códigos de catalogação), como CCO[322],

[316] CDWA refere-se a "[...] um conjunto de elementos de metadados para descrever obras de arte." (NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004, p. 15, tradução nossa), sendo que o CDWA *Lite* 1.1 alude um esquema XML para registros centrais de arte, arquitetura e cultura material criado para funcionar com o OAI-PMH – protocolo aplicado para coletar registros de metadados dispostos em repositórios por organizações (ou provedores de dados, *data providers*) – cujo os seus elementos baseiam num subconjunto do conjunto completo de elementos CDWA (BACA, c2016). Disponível em: http://www.getty.edu/research/publications/electronic_publications/cdwa/index.html. Acesso em: 24 maio 2023.

[317] O termo domínio bibliográfico é fundamentado na definição da *International Federation of Library Associations and Institutions* (IFLA) para Universo Bibliográfico que é o domínio das coleções de bibliotecas, arquivos, museus e outras comunidades de informação (INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 1998). Desta maneira, o significado aqui empregado para domínio bibliográfico "[...] designa o campo de estudo voltado para o tratamento descritivo da informação (processo de representação do recurso) em bibliotecas." (ALVES; SANTOS, 2013, p. 16), além do mais "[...] os metadados e padrões de metadados deste domínio foram criados com base nos princípios, códigos e regras de catalogação." (ALVES, 2017, p. 101).

[318] Como um conjunto de estruturas de dados padronizadas para descrição de materiais bibliográficos que promove a catalogação cooperativa e a troca de dados em sistemas informacionais (BACA, c2016), o formato MARC 21 constitui "[...] uma formatação, estrutura de registro, e padrão de codificação para registros de catalogação bibliográfica eletrônica desenvolvido pela *Library of Congress*.", do qual o algarismo "21" diz respeito "[...] à versão do MARC emitida em 1998 [...]" (NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004, p. 16, tradução nossa). Disponível em: https://www.loc.gov/marc/. Acesso em: 24 maio 2023.

[319] Interoperabilidade é "[...] a capacidade de vários sistemas com diferentes plataformas de *hardware* e *software*, estruturas de dados e interfaces, de trocar dados com perda mínima de conteúdo e funcionalidade." (NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004, p. 2, tradução nossa).

[320] SGML é "[...] uma linguagem usada para a marcação de documentos eletrônicos com *tags* que definem a relação entre o conteúdo e a estrutura." (NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004, p. 16, tradução nossa) que constitui, de acordo com Márdero Arellano (2008, p. 354), "[...] a base para a criação de todas as linguagens de marcação [...]" definida oficialmente pela norma ISO 8879:1986 *Information processing – Text and office systems* – SGML. Disponível em: https://www.iso.org/standard/16387.html. Acesso em: 24 maio 2023.

[321] DTD é "[...] uma descrição formal em SGML ou sintaxe XML da estrutura (elementos, atributos e entidades) a ser utilizada para descrever o tipo de documento especificado." (NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004, p. 15, tradução nossa).

[322] Como um padrão de conteúdo para descrição de obras de arte, arquitetura e cultura material (BACA, c2016), o CCO diz respeito a um manual para descrever, documentar e catalogar artefatos culturais e mídias visuais que os representam, podendo ser adotado para criar metadados compartilháveis, melhorar a descoberta e o acesso a obras culturais etc. Disponível em: https://vraweb.org/resources/cataloging-cultural-objects/. Acesso em: 24 maio 2023.

General International Standard Archival Description (ISAD(G))[323], AACR2[324] e RDA[325], e dos padrões de valor de dados (vocabulários, tesauros e listas controladas[326]), como *United States* (US) LCSH[327] e *Getty* TGN[328]; as quais são recomendadas por consórcios, órgãos de normalização e/ou líderes de comunidades como, por exemplo, a OCLC[329], a ISO, a NISO e o W3C.

Diante da carência de estudos nacionais de arquivamento da *Web* no campo CTS de um modo geral constatado pela busca deste tema junto as principais revistas associadas à área, e em específico no campo da Ciência da Informação e áreas afins (Arquivologia, Biblioteconomia e Museologia) verificada por levantamento bibliográfico e revisão da literatura produzida neste assunto, como já mencionado, os quais investiguem, sistematizem e analisem em profundidade os metadados e as características de padrões de metadados aplicáveis para a preservação de conteúdos da *Web* em sistemas de arquivamento digital, é que se constatou a necessidade de identificar e determinar quais padrões e esquemas de metadados poderiam ser julgados pelas organizações – acima de tudo, as instituições de patrimônio cultural e as universidades –, que

---

[323] Desenvolvida pelo *International Council on Archives* (ICA), a ISAD(G) (https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition) é um modelo de conteúdo que fornece diretrizes para a criação de descrições de materiais de arquivo respeitando-se às fontes dentro de uma descrição multinível. Hoje, encontra-se na segunda versão lançada em 2000. Disponível em: https://www2.archivists.org/groups/standards-committee/international-standard-archival-description-general-isadg. Acesso em: 24 maio 2023.

[324] O AACR consiste em "[...] um conjunto padrão de regras para catalogação de materiais de biblioteca.", onde o número "2" "[...] refere-se à segunda edição." (NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004, p. 15, tradução nossa) e, em conformidade com Baca (c2016), é um padrão de conteúdo de dados para descrição de materiais no domínio bibliográfico. Disponível em: http://www.aacr2.org/. Acesso em: 24 maio 2023.

[325] De modo que é um padrão de catalogação para bibliotecas que, atualmente, substitui o AACR2 (BACA, c2016), o RDA (https://access.rdatoolkit.org/) foi desenvolvido pelo RDA *Steering Committee* (RSC) e consisti em um pacote de elementos de dados, diretrizes e instruções para criação de metadados de recursos de patrimônio cultural e biblioteca, que é devidamente estruturado segundo modelos internacionais para aplicações de dados vinculados (*linked data*) centrados no usuário. Disponível em: http://rda-rsc.org/content/about-rda. Acesso em: 24 maio 2023.

[326] Vocabulário controlado (*controlled vocabulary*) concerne um arranjo organizado de palavras e frases servidas à indexação e/ou recuperação de conteúdos mediante navegação ou pesquisa, sendo que normalmente compõem-se de termos preferidos e variantes, com um escopo restrito ou que descreve um domínio particular (BACA, c2016).

[327] A LCSH (https://id.loc.gov/authorities/subjects.html) é uma lista de palavras e frases (ou melhor, cabeçalhos) que são empregadas para indicar os tópicos dos recursos de bibliotecas, trazendo consistência às coleções delas com a categorização dos temas em arranjos lógicos e com o controle de sinônimos, grafias variantes e homógrafos. É continuamente atualizada pela Biblioteca do Congresso norte-americano desde a sua primeira edição publicada em 1914. Disponível em: https://www.loc.gov/aba/cataloging/subject/lcsh-process.html. Acesso em: 24 maio 2023.

[328] O TGN (https://www.getty.edu/research/tools/vocabularies/tgn/) é um vocabulário do *Getty Research Institute* que se centra em lugares relevantes para a arte, arquitetura e disciplinas afins, documentando nomes, relações, tipos de localidades, coordenadas para cidades atuais e históricas, nações, impérios, sítios arqueológicos, aspectos físicos, entre outros. Disponível em: https://www.getty.edu/research/tools/vocabularies/. Acesso em: 24 maio 2023.

[329] Fundada em 1967, a OCLC é uma comunidade de bibliotecas sem fins lucrativos dirigida a milhares de membros em mais de cem países, como a Universidade Estadual Paulista (UNESP), a UNICAMP e o IBICT no Brasil, que proporciona serviços de compartilhamento de recursos, pesquisa original e programas comunitários para os seus membros e a comunidade de bibliotecas em geral. Disponível em: https://www.oclc.org/en/about.html. Acesso em: 24 maio 2023.

estão desenvolvendo seus sistemas, para que estas pudessem atender a preservação digital em arquivos da *Web*. Objetiva-se também, mais notadamente examinar em que grau os padrões e esquemas de metadados no âmbito da preservação digital e do arquivamento da *Web* têm sido discutidos pela Ciência da Informação e áreas afins, apontando os elementos de metadados que poderiam ser úteis as demandas de estruturação dos sistemas de arquivos da *Web* de forma mais apta à preservação da informação publicada na *Web* para fins históricos, culturais e de pesquisa.

Para isto, faz-se uma pesquisa qualitativa, de abordagem exploratória e descritiva (GIL, 2010; SILVA; MENEZES, 2005), que adota o método bibliográfico (MARCONI; LAKATOS, 2017; SEVERINO, 2016) a partir de um levantamento assistemático de dados em várias fontes de informação (com o uso de múltiplas palavras-chave e em diversos momentos) e uma revisão (narrativa) da literatura nacional e internacional, dos últimos vinte anos, dirigida e referente aos padrões de metadados aplicados à preservação digital e o arquivamento da *Web*. De acordo com Witter (1990, p. 7) "o levantamento bibliográfico assistemático é feito muitas vezes sem muita regularidade e sem alvos claramente estabelecidos.". Por sua vez, as revisões narrativas compõem de análise da literatura na interpretação e análise crítico pessoal do autor, geralmente não especificando as fontes usadas, os métodos para busca de referências e os critérios adotados na avaliação e na seleção de estudos, segundo Bernardo, Nobre e Jatene (2004) e Rother (2007).

Assim, através da análise do conteúdo da revisão de produções científicas como artigos de periódicos, anais de eventos, teses e livros buscados no *Google Scholar*[330] e na SciELO e nas bases *Scopus*[331] e *ScienceDirect*[332] (*Elsevier*), *Emerald Insight*[333] (*Emerald Publishing*), *Web of Science*[334] (*Clarivate Analytics*), *Library & Information Science Abstracts* (LISA)[335] (*ProQuest*) e *Library, Information Science & Technology Abstracts with Full Text* (LISTA)[336] e *Information Science & Technology Abstracts* (ISTA)[337] (EBSCO) disponíveis mediante Portal de Periódicos CAPES somado a *sites*, relatórios e guias de consórcios e órgãos de normalização/líderes de comunidades, foram reconhecidas e sistematizadas uma definição, categorização e funções dos metadados; o conceito de metadados de preservação e as informações registradas por metadados que apoiam o gerenciamento da preservação digital e o

[330] Disponível em: https://scholar.google.com.br/. Acesso em: 24 maio 2023.
[331] Disponível em: https://www.elsevier.com/pt-br/solutions/scopus. Acesso em: 24 maio 2023.
[332] Disponível em: https://www.sciencedirect.com/. Acesso em: 24 maio 2023.
[333] Disponível em: https://www.emerald.com/insight/. Acesso em: 24 maio 2023.
[334] Disponível em: https://clarivate.com/webofsciencegroup/solutions/web-of-science/; Acesso em: 24 maio 2023.
[335] Disponível em: https://about.proquest.com/products-services/lisa-set-c.html. Acesso em: 24 maio 2023.
[336] Disponível em: https://www.ebsco.com/products/research-databases/library-information-science-and-technology-abstracts. Acesso em: 24 maio 2023.
[337] Disponível em: https://www.ebsco.com/products/research-databases/information-science-technology-abstracts. Acesso em: 24 maio 2023.

arquivamento da *Web*; e os principais padrões de metadados usados na descrição e preservação digital de conteúdos *Web* arquivados.

Portanto, o presente capítulo se dispõe a expor os resultados e as análises dos conteúdos coletados, sendo que o produto deste mapeamento previu colaborar em prováveis delimitações de diretrizes e políticas as quais serão empregadas por instituições interessadas e/ou envolvidas com a captura, a retenção e o acesso permanente de um *website* ou coleção de *sites* arquivados.

4.1 Definição, categorização e funções dos metadados

Como informações criadas, guardadas e partilhadas para descrever coisas, os metadados nos permitem interagir com elas para obter o conhecimento que necessitamos. Difundidos nos sistemas informacionais, os metadados vêm de várias formas que nos mostram como os mesmos são todos estruturados até certo ponto, coletados para servirem um objetivo útil e dispostos em categorias conhecidas. Na definição ampla e clássica de que metadados significam "dados sobre dados" é de se esperar que os metadados podem ser encontrados em qualquer lugar, e realmente são (RILEY, c2017). Entretanto, esta definição literal e minimalista do termo metadados não é satisfatória visto que, pautando-se em Alves (2017) e Sayão (2010), se faz inexpressiva e rasa ante a complexidade das funções designadas a eles nos contextos atuais da gestão da informação bem como pelo fato de ser preciso entendê-los no domínio de aplicação onde estejam inseridos.

Neste trabalho adotaremos a definição de metadados de Alves (2010, p. 47), por julgá-la aplicável ao domínio da *Web* e em domínios específicos, como o domínio bibliográfico; além de atender aos propósitos da presente investigação e fundamentar-se na construção padronizada e consistente de representações unívocas dos recursos informacionais em diferentes ambientes digitais estruturados. Desta maneira, os metadados (*metadata*) podem ser conceituados como:

> [...] atributos que representam uma entidade (objeto do mundo real) em um sistema de informação. [...] são elementos descritivos ou atributos referenciais codificados que representam características próprias ou atribuídas às entidades [...] com o intuito de identificar de forma única uma entidade (recurso informacional) para posterior recuperação.

Para a autora a existência dos metadados dá-se através da sua codificação em estruturas de descrição padronizadas chamadas de padrões de metadados (*metadata statement*), sendo que o conjunto de metadados ou de elementos de metadados (*element sets*) integrará o esquema de metadados (*metadata schema*) do formato ou padrão de metadados. Assim, o esquema (*schema*) tange "[...] um conjunto de elementos de metadados e regras para sua utilização." (NATIONAL

INFORMATION STANDARDS ORGANIZATION, c2004, p. 16, tradução nossa) e, conforme Castro (2012) e Zeng e Qin (2008), o elemento de metadado (*metadata element*) corresponde a um termo formalmente definido para descrever uma das propriedades (ou atributos) do recurso de certo tipo ou com um propósito em particular, tal como 'o formato' de um arquivo eletrônico.

Além do conjunto de metadados (ou elementos prescritos, que são especificados através de declarações – *statements*), o esquema de metadados é composto pelo espaço de valor (*value spaces*), isto é, o conjunto de valores e regras de especificação para cada elemento e posição na estrutura descritiva que são definidos por padrões externos ao esquema, como uma sintaxe para exprimir os valores nos elementos e esquemas de codificação que fixam regras de codificação, sintaxe dos dados e formas/valores aceitos. Tais componentes indicarão os aspectos estruturais (disposição dos atributos e relações entre elementos), de sintaxe (codificação dos elementos e ordem lógica dos valores) e semânticos (significado do atributo, ordem lógica de codificação etc.) para a definição do esquema de metadados do padrão (ALVES, 2010; ZENG; QIN, 2008).

Para entender melhor a concepção de metadados, é útil separar metadados em categorias distintas que refletem aspectos-chave da funcionalidade deles num sistema. Os tipos principais de metadados existentes são utilizados sob as particularidades do domínio (e as funções a serem feitas), as demandas dos usuários e os tipos de recursos/entidades para representação (ALVES, 2017; GILLILAND, c2016; NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004). A partir de *Digital Preservation Coalition* ([2018a]), Riley (c2017) e Sayão (2010) são consideradas várias categorias funcionais de tipos de metadados, sendo assim compreendidas:

- Metadados descritivos – detalham um recurso digital para localização, identificação ou compreensão. Podem incluir propriedades ou elementos, tais como título, autor, assunto, gênero etc., onde os usos primários são descoberta, apresentação e interoperabilidade.
- Metadados estruturais – explicitam a estrutura interna de um arquivo digital e as relações hierárquicas de partes integrantes de recursos entre si. Podem conter propriedades, como ordem, lugar na hierarquia etc., em que os usos primários são navegação e apresentação.
- Metadados administrativos – fornecem informações que apoiam a gerência do ciclo de vida (criação, seleção, descrição etc.) dos recursos informacionais, subdividindo-se em:
    - Metadados técnicos – indicam os aspectos e dependências técnicas de um arquivo digital para decodificá-lo e renderizá-lo. Podem incluir propriedades, como tipo e tamanho de arquivo, data/hora de criação, esquema de compressão etc., onde os usos primários são interoperabilidade, gerência de objetos digitais e preservação.
    - Metadados de preservação – incluem informações (por exemplo, as dependências de *hardware* e *software*) exigidas ao gerenciamento de um arquivo digital a longo

prazo. Podem conter propriedades, como soma de verificação (*checksum*)[338], evento de preservação etc., em que os usos primários refletem a dos metadados técnicos.

- Metadados de direitos – documentam informações para apoio à gestão dos direitos de propriedade intelectual associados a um conteúdo. Podem incluir propriedades, tais como status dos direitos autorais, termos de licença, detentor dos direitos etc., onde os usos primários são interoperabilidade e gerenciamento de objetos digitais.

- Linguagens de marcação – incluem metadados e sinalizadores para outros recursos estruturais ou semânticos dentro do conteúdo. Podem conter propriedades, como nome, parágrafo, lista, data etc., no qual os usos primários são navegação e interoperabilidade.

À vista disso, uma razão notável para criar metadados descritivos é facilitar a descoberta de recursos informacionais relevantes no domínio *Web* ou em domínios específicos; em adição, os metadados podem auxiliar a organizar recursos eletrônicos, promover a interoperabilidade, apoiar o arquivamento e a preservação além de outras atividades comuns a serem realizadas em um sistema de informação digital que, de acordo com Gilliland (c2016) e *National Information Standards Organization* (c2004) retratam algumas das funções primárias dos metadados, como:

- Criação, multiversionamento, reutilização e recontextualização de objetos – os objetos de informação entram em um sistema sendo criados digitalmente ou convertidos em um formato digital. Múltiplas versões do mesmo objeto podem ser criadas para preservação, pesquisa, difusão, exposição etc. Determinados metadados descritivos e administrativos podem e devem ser inseridos pelo criador ou equipamento digitalizador, principalmente, se a reutilização estiver prevista, como num sistema de gerenciamento de ativos digitais.
- Organização e descrição – os metadados descrevem e ordenam tanto os objetos originais em um repositório como os objetos relacionados aos originais. Os objetos de informação são organizados automaticamente ou de forma manual na estrutura do sistema e podem conter descrições geradas pelo criador original. Ademais, metadados adicionais podem

[338] *Checksum* é "um código de comprimento fixo gerado a partir de um objeto digital com o objetivo de detectar erros durante a transmissão e armazenamento" (LAVOIE; GARTNER, c2013, p. 31, tradução nossa). Um resumo da mensagem (*message digest*, chamado informalmente de soma de verificação – *checksum*) é o fruto da aplicação de uma função *hash* criptográfica (algoritmo) unidirecional numa sequência de *bits* (mensagem), como um arquivo ou fluxo de *bits* (*bitstream*). Para apuração de que um arquivo não foi alterado durante um período (verificação de fixidez), um método é calcular um *message digest* num ponto no tempo e recalcular o *message digest* num ponto posterior; se os resumos forem idênticos, o objeto não foi alterado (PREMIS EDITORIAL COMMITTEE, 2015).

ser gerados por profissionais da informação mediante processos de registro, catalogação e indexação, ou via folksonomias[339] e outras formas de metadados cedidos pelo usuário.

- Interoperabilidade – a descrição de um recurso com metadados possibilita que ele seja entendido por humanos e máquinas para facilitar a interoperabilidade. Usando esquemas predefinidos, protocolos de transferência partilhados e mapeamentos (*crosswalks*) entre esquemas[340], os recursos na rede (*network*) podem ser buscados com facilidade. Segundo Baca (c2016) a interoperabilidade é a capacidade de distintos sistemas de trabalharem em conjunto, sobretudo, na interpretação exata da semântica dos dados e funcionalidade.
- Identificação e validação – os esquemas incluem como elementos números padrão para a identificação única do objeto digital que os metadados se refiram. Além dos elementos reais que remetem para o objeto, os metadados podem ser reunidos para atuarem como conjunto de dados de identificação, distinguindo um objeto de outros para validação. A localização dos objetos dá-se usando sistemas de identificadores[341] persistentes, como

[339] Folksonomia (*folksonomy*) é um conjunto de conceitos, representados por termos e nomes (chamados etiquetas – *tags*), que derivam da etiquetagem ou marcação social (*social tagging*). Os autores da folksonomia são os usuários eventuais do conteúdo ao invés de profissionais indexadores que seguem protocolos normalizados e vocabulários controlados padronizados. Uma folksonomia difere de uma taxonomia por não estar estruturada hierarquicamente, sendo que esta última pode ser utilizada como um vocabulário controlado e consiste em uma classificação ordenada que exprime as relações (comumente, hierárquicas) entre os objetos que estão sendo classificados (BACA, c2016).

[340] Independente do sistema usado, o protocolo ajuda no acesso concomitante à catálogos de outras instituições como compartilha registros bibliográficos e dispõe interface única para várias fontes; e o mapeamento (*crosswalk*) é uma ferramenta que propicia aos sistemas de fato converter dados de um padrão de metadados para outro, contribuindo para a interoperabilidade entre esquemas em ambientes informacionais digitais (CASTRO, 2012). De acordo com Baca (c2016), o protocolo é uma especificação que descreve como os computadores se comunicam entre si, por exemplo, o OAI-PMH remete um protocolo da *Open Archives Initiative* (OAI) para coletar registros de metadados de organizações ou provedores de dados (*data provider*) que dispõem registros de metadados em um repositório – servidores especificamente configurados – para coleta por provedores de serviço (*service providers*).

[341] Identificador é um conjunto de caracteres apontados para identificar de maneira inequívoca um recurso. Entende-se por identificador persistente como um único identificador ininterruptamente vinculado a um objeto digital que, enquanto gerido, continuamente fornecerá acesso permanente ao objeto apesar de mudanças de local (MÁRDERO ARELLANO, 2008). Além do DOI e PURL, outros exemplos de identificadores são o URI no qual, segundo Baca (c2016), é uma cadeia curta de caracteres que identifica de forma única um recurso, classificando-se em: *Uniform Resource Name* (URN), que consiste em um identificador único e independente das mudanças de local de um arquivo na *Internet* (por exemplo, urn:issn:0167-6423 é o URN para a revista *Science of Computer Programming*); e URL, um endereço da *Internet* que comunica aos usuários como e onde encontrar um certo arquivo na *Web* (por exemplo, http://www.getty.edu/publications/intrometadata/ e https://repositorio.ufscar.br/handle/ufscar/7221).

o *Digital Object Identifier* (DOI)[342] e um *Persistent Uniform Resource Locator* (PURL)[343].

- Busca, descoberta e recuperação – a exigência de metadados descritivos de qualidade[344] decorre em permitir que os recursos sejam encontrados por critérios pertinentes; reunir, distinguir e identificar os recursos; além de capacitar os usuários a localizar e recuperar metadados e objetos de informação relevantes armazenados localmente e virtualmente distribuídos. Os sistemas constroem e mantêm metadados[345] que rastreiam algoritmos[346] de recuperação, transações de usuário e ação do sistema no arquivamento e recuperação.
- Utilização e preservação – os objetos de informação em ambientes digitais podem estar submetidos a vários tipos de usos ao longo de suas vidas, período os quais eles também são reproduzidos e alterados. Assim, metadados sobre comentários do usuário, controle de versão e rastreamento de direitos podem ser criados. A preservação e o arquivamento exigem elementos específicos para rastrear a "linhagem" do objeto digital (de onde ele veio e como ele mudou no tempo, por exemplo), detalhar seus aspectos, ou registrar o

[342] DOI é "[...] um identificador único atribuído a objetos eletrônicos de propriedade intelectual que pode ser definido para localizar o objeto na *Internet*." (NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004, p. 15, tradução nossa) e, para Márdero Arellano (2008, p. 352) um "[...] sistema de identificador de objetos digitais *on-line* para gerenciar a propriedade intelectual e o uso comercial dos objetos materiais digitais.". Introduzido em 2000, o sistema DOI foi criado pela *International DOI Foundatione* (IDF), organização sem fins lucrativos que é a autoridade de registro da ISO 26324 para o DOI. Disponível em: https://www.doi.org/. Acesso em: 25 maio 2023.

[343] PURL (*Persistent* URL) é "[...] é um sistema de nomeação e resolução desenvolvido pela OCLC utilizando um serviço de redirecionamento intermediário para localizar o URL de um recurso." (NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004, p. 16, tradução nossa) e, para Márdero Arellano (2008, p. 144), "[...] está estruturado na conexão de URL que apontam para um serviço de resolução (*service resolver*) mantido para redirecionar um endereço ou *link* de páginas *Web* que não esteja funcionando para o endereço ativo". Hoje, é uma iniciativa do *Internet Archive*, tratando-se de um URL persistente que cede um endereço permanente para acessar recursos na *Web*, onde ao recuperar um PURL redireciona-se para a localização atual do recurso, podendo ele ser atualizado se o recurso for movido. Disponível em: https://purl.prod.archive.org/help. Acesso em: 25 maio 2023.

[344] A NISO, através da *Framework of Guidance for Building Good Digital Collections*, articula princípios aplicáveis a metadados de qualidade (*good metadata*), que indicam requisitos à construção de metadados, como: a adequação aos materiais de uma coleção, aos usuários da coleção e ao uso do objeto digital; o suporte à interoperabilidade; o uso de vocabulários controlados padronizados; o apoio a gerência a longo prazo de objetos em coleções; a detenção de confiabilidade e verificação, sendo que os registros como objetos em si devem ter qualidades de arquivabilidade, identificação única e persistência; etc. (NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004).

[345] Sobre princípios comuns para construção de metadados, Alves (2010) e Zeng e Qin (2008) assentes nas definições da *Dublin Core Metadata Initiative* (DCMI) ressaltam três princípios básicos que são requisitos de um sistema de informação, a saber: simplicidade, que alude os metadados para manutenção de um conjunto mínimo de elementos descritivos em suporte a implementação que devem ser flexíveis permitindo que sejam incluídos novos metadados para acudir certas demandas de descrição e de um domínio particular; extensibilidade, que remete a capacidade do esquema conceder um conjunto de elementos descritivos que unifique variados padrões de descrição ou, também, a capacidade de intercâmbio entre registros de metadados de esquemas simples para outros mais complexos; e a interoperabilidade, o qual refere-se a "[...] interação de documentos digitais entre diferentes sistemas tecnológicos; configuração de todos os arquivos, padronizando-os de forma global." (MÁRDERO ARELLANO, 2008, p. 353).

[346] Algoritmo é uma fórmula/processo para solucionar um problema ou executar uma tarefa, que consiste em um conjunto de etapas numa ordem particular, tal como as instruções em um programa de computador (BACA, c2016).

seu comportamento para imitá-lo em tecnologias futuras, como um modo de garantir a sobrevivência e o acesso contínuo e utilizável dos recursos digitais no presente e futuro.

- Disposição – os metadados são componentes-chave na documentação da disposição (por exemplo, adesão, cancelamento) de objetos e itens originais num repositório ou coleção, assim como dos objetos de informação associados a esses originais. Neste seguimento, os objetos que estão inativos ou então que não são mais exigidos podem ser descartados.

Para os objetivos deste trabalho destacaremos a categoria de metadados de preservação devido a serem essenciais para a obtenção de uma efetiva gerência e preservação a longo prazo dos arquivos digitais e eletrônicos; e a classe de metadados descritivos, a faceta mais conhecida dos metadados (SAYÃO, 2010) que, respaldando-se nas considerações do grupo de trabalho internacional *Web Archiving Metadata* (WAM) da OCLC *Research*[347] pelos estudos de Dooley e Bowers (c2018), Samouelian e Dooley (c2018) e Venlet *et al*. (c2018), serão abordados em conformidade com melhores práticas de criação de metadados descritivos coerentes e eficientes acerca de conteúdos da *Web* arquivados (ou melhor, *websites*) e para o arquivamento da *Web*.

4.2 Metadados de preservação e metadados descritivos para arquivamento da *Web*

A julgar que a preservação digital é um processo de gestão, os metadados de preservação são categorizados principalmente como metadados administrativos, porém é admissível que os esquemas de metadados de preservação incluam elementos que se estendem em várias classes, como descritivos, estruturais e administrativos. Tais metadados compõem uma parte crucial das estratégias de preservação digital e são concebidos no dicionário de dados PREMIS, um padrão internacional de fato para metadados de preservação (CHEN; REILLY, 2011; DAPPERT *et al*., 2013; SAYÃO, 2010). O *Premis Editorial Committee* (2015, p. 2, tradução nossa) define os metadados de preservação (*preservation metadata*) "[...] como a informação que um repositório usa para apoiar o processo de preservação digital". Através de Dappert e Enders (2010) e Caplan (2017) tratam-se das informações que descrevem um recurso digital no repositório para garantir o seu acesso e uso a longo prazo. Em Márdero Arellano (2008) são aqueles alusivos ao conteúdo do recurso, seu contexto e estrutura de criação, além das alterações feitas em seu ciclo de vida.

[347] A OCLC *Research* (https://www.oclc.org/research/home.html) estabeleceu o WAM com o fim de desenvolver recomendações para metadados descritivos. O resultado foram as três publicações supracitadas que compreendem orientações para ajudar as instituições a melhorar a consistência e a eficiência de suas práticas de metadados, uma revisão das ferramentas de coleta da *Web* e uma revisão de literatura das necessidades do usuário. Disponível em: https://www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata.html. Acesso em: 25 maio 2023.

Definidos assim, nota-se que os metadados de preservação são construídos para cumprir uma ampla série de funções diferentes, todavia relacionadas (SAYÃO, 2010). Esses metadados suportam os distintos requisitos da preservação digital que, consoante Lavoie e Gartner (c2013) e *Premis Editorial Committee* (2015), se refletem em manter a disponibilidade; a renderização (tornar o objeto perceptível para um usuário via reprodução – para materiais visuais –, exibição – para materiais de áudio –, ou de outros meios próprios ao seu formato); a compreensibilidade; a identidade; a persistência; a autenticidade (qualidade de que o objeto é o que ele pretende ser, onde a integridade do seu conteúdo e origem pode ser verificada); e a viabilidade (propriedade de ser legível pela mídia de armazenamento) de objetos digitais por longos períodos de tempo.

Isto posto, *Digital Preservation Coalition* ([2018a]) e Gilliland (c2016) deduzem certas razões dos metadados serem importantes para a preservação digital de longo prazo o qual, em conjunto com as considerações dos autores supracitados, podem ser descritas da seguinte forma:

- Tomada de decisões – informações vinculadas a um objeto digital, como o histórico das alterações feitas nele, o *software* requerido para abri-lo ou quanto tempo ele precisa ser mantido, ajudam os profissionais da informação a tomar decisões sobre como e por que preservá-lo. Os metadados ainda contém detalhes sobre direitos e propriedade para que os usuários do registro digital avaliem o que pode ser copiado, partilhado e modificado.
- Questões legais – os metadados permitem que os sistemas rastreiem níveis de direitos, licenças e informações de reprodução existentes para os itens originais, os seus objetos associados e as múltiplas versões destes. Além do mais, metadados documentam outros requisitos legais ou de doadores impostos aos objetos originais e seus substitutos, como questões de privacidade, restrições às reproduções e interesses proprietários/comerciais.
- Controle de versão – metadados devem ser capazes de elucidar as distinções nas versões de um objeto informacional. Eles devem ainda ajudar os usuários a distinguir e rastrear mudanças em: versões analógicas e digitalizadas originais, indicando qualquer alteração acidental ou decidida ocorrida no processo de digitalização; materiais originais nascidos digitais e versões atualizadas ou revisadas como, por exemplo, *websites*; dentre outros.
- Persistência – a documentação por metadados de como o objeto de informação foi criado e mantido, como se comporta e liga-se com outros objetos será crucial à sua existência independente do sistema atual usado para armazená-lo e recuperá-lo. Para que os objetos digitais permaneçam acessíveis e inteligíveis no tempo, também será essencial preservar e migrar esses metadados não tornando-os desconectados dos objetos a qual descrevem.
- Contexto para significado – os metadados fornecem informações de contexto requeridas para que futuros usuários entendam o significado do conteúdo de um registro, exercendo

um papel vital na documentação de relações e na indicação da autenticidade, integridade estrutural/processual e grau de completude dos objetos. Eles podem conter informações acerca da razão pela qual um registro foi criado, quem o criou e por que foi preservado.

- Usabilidade – sem a informação certa, um registro digital pode perder a sua estrutura ou significado. Os metadados garantem que futuros usuários serão capazes de renderizar e interpretar um arquivo, por exemplo, com o *software* correto. Na dissociação, isto é, se dois ou mais arquivos dependem um do outro por significado ou estrutura e um se separa ou se perde, os metadados dessa relação ajudarão reuni-los para não se tornarem inúteis.

Estas justificações expressam determinadas informações descritivas, administrativas e estruturais a serem incorporadas pelos metadados de preservação. Neste seguimento, agrupando as ponderações de Caplan (2017), Dappert *et al*. (2013), Dappert e Enders (2010), Formenton *et al*. (2017), Lavoie e Gartner (c2013), *National Library of New Zealand* (2003) e Sayão (2010) identificamos um conjunto de informações e funções interrelacionadas, que apoiam a gestão da preservação digital, abrangidas na captura, criação e manutenção de metadados de preservação:

- Registro do histórico do objeto durante o seu ciclo de vida. Pode conter a documentação das mudanças na cadeia de custódia e propriedade, das circunstâncias de criação (a data de criação, o nome e a versão do aplicativo de criação etc.) e das alterações autorizadas.
- Registro das ações tomadas para preservar o objeto no decorrer do tempo. Pode incluir a descrição dos aspectos do processo de preservação digital utilizado, como migração, emulação, encapsulamento, arquivamento da *Web* etc., e a documentação dos efeitos do processo sobre o conteúdo, a apresentação, a usabilidade e as funcionalidades do objeto.
- Registro de informações que definam e validem a autenticidade do objeto, isto é, que o objeto é de fato o que diz ser e que não teve alterações intencionais ou involuntárias não documentadas. Pode incluir a documentação de informações de fixidez e integridade – assinatura digital, *checksum* etc. –, e informação técnica específica do tipo de conteúdo, como largura de imagem para uma imagem e tempo decorrido para um arquivo de áudio.
- Registro de permissões e direitos de propriedade intelectual que validem (ou restrinjam) as ações de preservação e de difusão, acesso e uso do objeto. Pode conter a descrição da natureza do direito, do seu status e da fonte onde este é conferido (por exemplo, termos de licenças/estatutos, direitos autorais, permissões especiais, política institucional etc.).
- Registro das dependências técnicas necessárias para acessar, renderizar – ou apresentar, executar etc. – e usar o objeto. Pode incluir a descrição do formato de arquivo do objeto e o *software* aplicativo, *hardware* e sistema operacional exigidos para torná-lo utilizável.

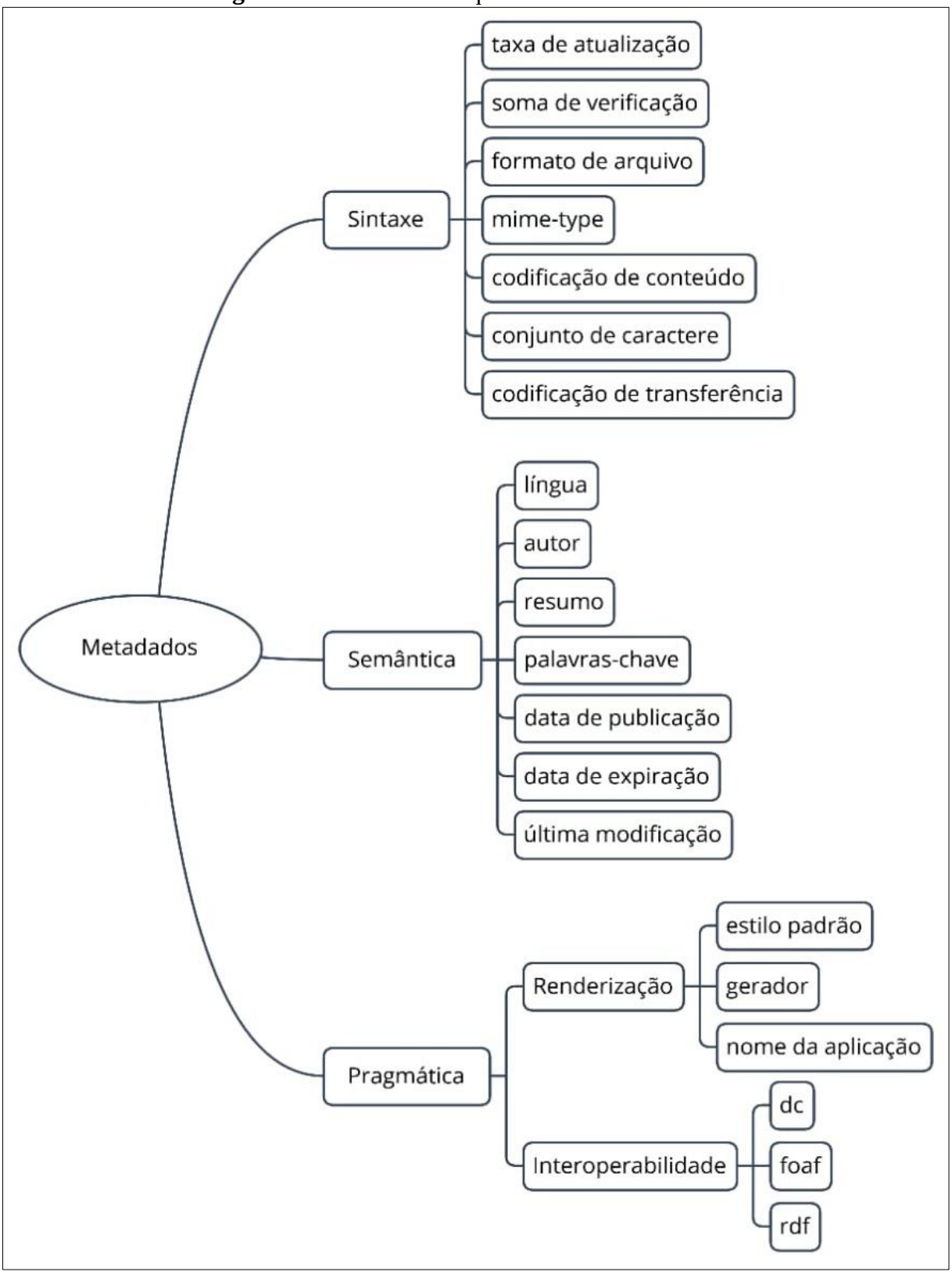
- Registro de informações que estabeleçam as propriedades significativas do objeto, isto é, características do objeto original e do ambiente que devem ser mantidas por ações de preservação para uma comunidade de usuários (por exemplo, as imagens em uma página da *Web*), orientando decisões sobre quais ações de preservação devem ser selecionadas.
- Registro das relações estruturais físicas e lógicas do objeto (por exemplo, qual imagem está integrada em qual *website* e qual página segue qual em um livro digitalizado); além de informações sobre seu meio de armazenamento. Pode conter a documentação do tipo e idade da mídia como das datas em que os arquivos foram atualizados pela última vez.
- Registro de informações sobre agentes – pessoas, organizações, *software* aplicativo ou *hardware* – com funções nos direitos, nos ambientes computacionais de renderização e nas ações que afetam o objeto. Pode incluir a documentação da função do agente (autor, detentor dos direitos, depositante etc.) e da versão do agente (para *software* e *hardware*).
- Registro de informações sobre inibidores, ou seja, qualquer recurso do objeto dirigido a inibir o acesso, a utilização, a cópia, a disseminação ou a migração como, por exemplo, criptografia e proteção por senha. Pode conter a descrição do tipo de inibidor, do alvo (as ações inibidas) e da chave (senha ou demais mecanismos para contornar o inibidor).

Apesar de haver pouco trabalho que reúna e sintetize experiências de implementação de metadados de preservação para o acúmulo e a consolidação de melhores práticas na preservação digital ou, ainda, que avalie os custos inclusos na coleta e gerência de metadados de preservação e os benefícios práticos de recair nestes custos (LAVOIE; GARTNER, c2013), os metadados de preservação são um componente-chave de todo o processo de arquivamento digital em longo prazo. Tais metadados documentam informações de conteúdo e de proveniência, autenticidade, fixidez, referência, contexto, direitos etc. alinhadas ao modelo de informação do OAIS e seus três pacotes de informação (isto é, Pacote de Submissão de Informação – PSI, Pacote de Arquivamento de Informação – PAI e Pacote de Disseminação de Informação – PDI)[348], os quais asseguram que recursos/objetos digitais sejam mantidos, retidos, identificados, acessados, decifrados, renderizados e usados de forma coesa e precisa no tempo.

[348] O OAIS, do CCSDS e da ISO, é um "[...] esquema conceitual que disciplina e orienta um sistema para a preservação e manutenção de acesso à informação digital por longo prazo [...]" (MÁRDERO ARELLANO, 2008, p. 353). No esquema há um modelo de informação para a inclusão dos metadados exigidos na preservação e acesso de informação digital por longo prazo onde, com base em Lavoie e Gartner (c2013) e Márdero Arellano (2008), são observados três pacotes de informação, a saber: o PSI, pacote expedido do produtor para o arquivo OAIS; o PAI, pacote armazenado, gerido e protegido no sistema, incluindo o material que será mantido e a informação para representá-lo e preservá-lo; e o PDI, pacote transportado do sistema para um cliente/usuário em resposta a uma solicitação. Estes pacotes são reservatórios que encapsulam o recurso a ser preservado e os vários metadados (informação de proveniência, contexto etc.) de apoio a gerência da preservação. O OAIS está especificado na norma ISO 14721:2012 – *Space data and information transfer systems* – OAIS – *Reference model*. Disponível em: https://www.iso.org/standard/57284.html. Acesso em: 25 maio 2023.

Como indicado por Banos *et al*. (c2013), o uso de metadados também constitui uma das principais facetas de arquivabilidade (*archivability facets*) de *websites*, ou melhor, fatores que devem ser levados em conta para calcular a extensão em que o *site* satisfaz às condições para a transferência segura de seu conteúdo a um arquivo da *Web* para intuitos de preservação. Com base em um modelo geral de perspectiva compartilhada em várias disciplinas de informação – Filosofia, Linguística, Ciências da Computação etc. –, os autores consideram os metadados em três níveis (resumidos e demostrados na Figura 65) para medição da capacidade de arquivamento de um *site* (arquivabilidade), a saber: sintaxe (por exemplo, como isso é expresso); semântica (por exemplo, sobre o que é isso); e pragmática (por exemplo, o que você pode fazer com isso).

Figura 65 – Facetas de arquivabilidade: Metadados.



Fonte: Adaptado de Banos *et al*. (2013).

Banos *et al*. (c2013) explicam que metadados de codificação de conteúdo e transferência podem ser inseridos pelo servidor em cabeçalhos *Hypertext Transfer Protocol* (HTTP)[349]; além disso, metadados de renderização, como nome do aplicativo, a linguagem do usuário final para compreender o conteúdo e, ainda, informações descritivas, como autor e palavras-chave, que ajudam a entender como o conteúdo é classificado podem ser incluídos no atributo e nos valores do elemento *HyperText Markup Language* (HTML)[350]. Para os autores a utilização de outros metadados e esquemas de descrição conhecidos como, por exemplo, o DC, o *Friend of a Friend* (FOAF)[351] e o *Resource Description Framework* (RDF)[352], pode ser incorporada para propiciar uma melhor interoperabilidade; ademais, a existência de elementos de metadados selecionados é verificada para elevar a possibilidade de implementar a extração automatizada e o refinamento dos metadados na coleta e ingestão do conteúdo *Web* ou, em seguida, na gestão de repositório.

A julgar que o arquivamento da *Web* é um processo relativamente novo para as instituições de patrimônio cultural, existem poucos padrões; assim, as práticas de metadados variam muito dentre as distintas iniciativas na área, seja entre as bibliotecas nacionais ou, até, por diferenças nas abordagens de descrição entre as duas tradições de descrição bibliográfica e arquivística de recursos que não promove a interoperabilidade dos metadados, pois, por exemplo: a catalogação nas bibliotecas em geral é feita apenas no nível do título e os títulos são transcritos literalmente; de outra forma, os arquivos trabalham com descrições multiníveis de coleções para as quais os títulos são comumente concebidos (DI PRETORO; GEERAERT, 2019; DOOLEY; BOWERS, c2018). Destas práticas de metadados nas iniciativas de arquivamento da *Web*, mencionamos:

---

[349] HTTP é um protocolo padrão que permite aos usuários com navegadores *Web* (*Web browsers* –, *softwares* para visualizar e interagir com informações e arquivos de mídia na *Web*) acessar documentos HTML (BACA, c2016).
[350] Constituindo "[...] a principal linguagem de marcação para exibir páginas na *Internet* através de um navegador *Web* (PENNOCK, c2013, p. 35, tradução nossa), o HTML representa, de acordo com Baca (c2016) e *National Information Standards Organization* (c2004), um conjunto de *tags* e regras derivado da linguagem de marcação SGML e que é empregado para criar documentos de hipertexto para aplicações da *Web*, sendo oficialmente uma recomendação do W3C. Disponível em: https://html.spec.whatwg.org/multipage/. Acesso em: 25 maio de 2023.
[351] Considerando ontologia como uma especificação formal, legível por máquina, de um modelo conceitual no qual conceitos, propriedades, funções etc. são claramente definidos, o FOAF é uma ontologia que modela dados para pessoas, suas atividades e relações com outras pessoas e objetos (BACA, c2016). Para Riley (c2017) o FOAF é usado em sistemas para identificar pessoas e organizações e ceder informações básicas sobre elas. Trata-se de uma linguagem de computador que define um dicionário de termos associados a pessoas que pode ser adotado em dados estruturados, como RDF *in Attributes* (RDFa) e *Linked Data*. Disponível em: http://xmlns.com/foaf/0.1/. Acesso em: 25 maio 2023.
[352] RDF é "[...] um conjunto de especificações W3C." constituindo "[...] uma linguagem para representar metadados sobre recursos da *Web* para que possam ser trocados entre aplicações sem perda de significado." (NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004, p. 16, tradução nossa). Trata-se de um modelo padrão para troca de dados na *Web* a qual permite que dados estruturados e semiestruturados sejam mesclados, expostos e compartilhados entre diversas aplicações. Disponível em: https://www.w3.org/RDF/. Acesso em: 25 maio 2023.

- A Biblioteca Britânica, líder do arquivo da *Web* do Reino Unido[353] que coleta, preserva e dispõe acesso a *sites* no domínio da *Web* britânica, atua com vários níveis de descrição onde os possíveis elementos de metadados descritivos e administrativos gerados para os *sites* selecionados são: Dados de Direitos e Licenciamento (*rights data and licensing*); Agenda de Rastreio (*crawl schedule*); Atribuição a uma Coleção Especial (*allocation to a special collection*); Título (*title*); Assunto (*subject*); Breve Resumo (*short abstract)*; e Palavras-chave (*keywords*) (COOKE[354], 2018 *apud* DI PRETORO; GEERAERT, 2019).
- O arquivo da *Web* Suíça[355], mantido pela Biblioteca Nacional Suíça e outras bibliotecas como maneira de preservar e tornar acessível partes da *Internet* da Suíça para pesquisa, codifica os seguintes elementos de metadados descritivos no registro de *websites* para a coleção: URL; Título (*titel*); Organização (*körperschaft*); Produtor (Editor/Distribuidor) (*produzent/in – verlag/vertrieb –*), incluindo Nome (*name*), Local (*ort*), Região (*kanton*), País (*land*) e Pessoa de Contato para a Atribuição de Direitos (*ansprechpartner/in für rechtevergabe*); Língua (*sprache*); Classificação de Dewey (dewey); Palavras-chave (*schlagwort*); e Frequência de Coleta (*sammelhäufigkeit*) (SIGNORI; BÄTTIG, 2017).
- O *Nationaal Register Webarchieven*[356], projeto do *Netwerk Digitaal Erfgoed* que fornece uma visão geral pesquisável de *sites* arquivados na Holanda, os metadados são expostos em uma plataforma compartilhada onde os usuários podem usar filtros de busca, como: Organização do Arquivo (*archieforganisatie*); Nome (*naam*); URL do *Site*

[353] De forma seletiva desde 2005 e em um nível de domínio completo desde 2013, a Biblioteca Britânica coleciona *sites* adquiridos segundo o *Non-Print Legal Deposit Regulations*. As coleções de arquivos da *Web* disponíveis são localizadas no UK *Web Archive* (https://www.webarchive.org.uk/) que permite a navegação por coleções de temas de *websites* arquivados, como *sites* criados por comunidades latino-americanas no Reino Unido e por organizações do Reino Unido com *links* diretos para essas comunidades (o *BrazilianAritists.net* e o *Action For Brazil's Children Trust*, por exemplo). Disponível em: https://www.bl.uk/collection-guides/uk-web-archive. Acesso em: 25 maio 2023.

[354] I. Cooke. (*personal communication*, *February* 8, 2018).

[355] Iniciado em 2008, o arquivo da *Web* Suíça (https://www.e-helvetica.nb.admin.ch/) é a coleção de *sites* coletados e arquivados de importância patrimonial-cultural da Biblioteca Nacional da Suíça junto com outras instituições, os quais são indexados no catálogo *online Helveticat* da biblioteca nacional. Essa coleção é seletiva e não exaustiva (*sites* de domínio *.ch*, *sites* com conexão estreita à Suíça ou que cobrem eventos específicos ocorridos na Suíça), do qual as versões arquivadas de *sites* podem ser pesquisadas no sistema *e-Helvetica Access*, mas apenas acessadas nas instalações das bibliotecas da iniciativa. Disponível em: https://www.nb.admin.ch/snl/en/home/information-professionals/e-helvetica/web-archive-switzerland/faq-webarchiving.html. Acesso em: 25 maio 2023.

[356] O Registro Nacional de Arquivos da *Web* (*Nationaal Register Webarchieven*) é uma visão geral pesquisável de quais *sites* da Holanda (isto é, *sites* de domínio *.nl*, ou *.com*, *.eu*, *.org* etc.) são arquivados, onde, desde e até quando, por que, como e com qual frequência. Tendo em conta que nem todos os *sites* holandeses foram arquivados e nem todos os *sites* arquivados estão inseridos no projeto, o *site* oficial da iniciativa aponta que ela é útil as instituições de patrimônio para obter dados sobre o que já foi arquivado, de maneira que duplicações de trabalho sejam evitadas e parcerias possam ser formadas. Disponível em: https://www.registerwebarchieven.nl/. Acesso em: 25 maio 2023.

(*website* url); Período (*periode*); Intervalo (*interval*); Motivo (*reden*); Ferramenta de Arquivamento (*archiveringstool*); e Acesso (*toegang*) (NETWERK DIGITAAL ERFGOED, [2021?]).

- O Arquivo.pt[357], serviço da FCT do Ministério da Educação e Ciência de Portugal que permite a pesquisa e o acesso a páginas da *Web* portuguesa arquivadas, elucida certos metadados sobre os conteúdos de um *website* para a sua preservação, como: Descrição (*description*), texto breve descrevendo o conteúdo da página; Palavras-chave (*keywords*), expressões representativas dos principais temas da página; e *Dublin Core*, metadados conforme o esquema DC (ARQUIVO.PT, 2018).
- O *Internet Archive*[358], fundação que fornece acesso gratuito e universal a uma biblioteca digital com mais de 500 bilhões de páginas *Web* e outros conteúdos arquivados, indica metadados que têm significado especial na descrição do conteúdo dos itens do arquivo, tais como Patrocinador (*sponsor*); Scanner; Data de Digitalização (*scandate*); Contagem de Imagens (*imagecount*); e Tipo de Mídia (*mediatype*) (INTERNET ARCHIVE, 2018). Aliás, como levantado por Samouelian e Dooley (c2018), o seu serviço de arquivamento da *Web* dispõe dezesseis campos de metadados DC do qual os usuários podem escolher, bem como a capacidade de adicionar campos personalizados manualmente[359].
- No arquivo da *Web* da Biblioteca do Congresso americano[360], programa que gerencia, preserva e oferece acesso a conteúdo da *Web* arquivado, codificam-se subelementos de metadados MODS no registro de *websites* para coleções temáticas e por eventos, como:

---

[357] Iniciado oficialmente em 2008, o Arquivo.pt é uma infraestrutura que, com base em coletas exaustivas da *Web* portuguesa, objetiva a preservação da informação produzida e publicada na *Web* para intuitos de investigação. O arquivo da *Web* portuguesa possui páginas *Web* arquivadas disponíveis para consulta desde 1996; ademais, mesmo fora do seu escopo (isto é, *sites* de domínio .pt) o Arquivo.pt preservou de forma não exaustiva certas páginas *Web* de *sites* do Brasil (*sites* de domínio .br) (https://arquivo.pt/wayback/20170210103711/http://www.brasil.gov.br/, por exemplo). Disponível em: https://sobre.arquivo.pt/pt/ajuda/o-que-e-o-arquivo-pt/. Acesso em: 25 maio 2023.

[358] Começado em 1996, o *Internet Archive* é uma instituição sem fins lucrativos que oferece acesso gratuito a uma biblioteca digital de *sites* da *Internet* e outros artefatos culturais em formato digital (livros, textos etc.). No *website* oficial da organização indica-se que o *Internet Archive* mantém arquivado mais de 549 bilhões de páginas de toda a *Web* mundial salvas ao longo do tempo. Hoje, a fundação dispõe de mais de 20 anos de história da *Web* acessível pela ferramenta *Wayback Machine* (https://archive.org/web/) além de parceiros na identificação de páginas *Web* via programa *Archive-It* (https://archive-it.org/). Disponível em: https://archive.org/about/. Acesso em 25 maio 2023.

[359] Disponível em: https://support.archive-it.org/hc/en-us/articles/208332603-Add-edit-and-manage-your-metadata. Acesso em: 25 maio 2023.

[360] Iniciado em 2000, a página inicial do *site* do programa do arquivo da *Web* da Biblioteca do Congresso americano afirma que ele gerencia, preserva e cede acesso a conteúdo *Web* arquivado e selecionado para que esteja disponível à pesquisadores interessados. Os arquivos podem ser explorados e navegados (https://www.loc.gov/web-archives/) e estão organizados em coleções temáticas e baseadas em eventos, como incluem *sites* que documentam entidades internacionais e dos Estados Unidos (*sites* de governos, notícias etc.) que retratam uma série de áreas temáticas e assuntos. Disponível em: https://www.loc.gov/programs/web-archiving/for-researchers/. Acesso em: 25 maio 2023.

> Título (<*title*>), no elemento Informação de Título (<*titleInfo*>); Texto (<*text*>) para os escopos (domínio), no elemento Parte (<*part*>); Identificador (<*identifier*>) para o URL da fonte, no elemento Item Relacionado (<*relatedItem*>); Gênero (<*genre*>) e Forma (<*form*>), no elemento Descrição Física (<*physicalDescription*>); e Lugar (<*place*>) no elemento Informação de Origem (<*originInfo*>) (LIBRARY OF CONGRESS, [2021]).

Diante disso, consoante Dooley *et al*. (2017), a OCLC *Research* estabeleceu o grupo de trabalho WAM em face do desafio da falta de uma abordagem comum para criar metadados na comunidade de arquivamento da *Web*. Sob o intuito de elaborar recomendações para metadados descritivos e facilitar a descoberta de conteúdo da *Web* arquivado provendo uma ponte entre as abordagens bibliográficas e arquivísticas para a descrição, o grupo publicou três relatórios que incluem: uma revisão de literatura das necessidades de metadados descritivos tanto dos usuários finais de arquivos da *Web* como dos profissionais que criam e gerem tais metadados (VENLET *et al*., c2018); uma análise de onze ferramentas de coleta da *Web* com vista à sua funcionalidade para extração de metadados descritivos dos arquivos rastreados (SAMOUELIAN; DOOLEY, c2018); e diretrizes para ajudar instituições e pessoas a melhorarem a consistência e a eficiência de suas práticas de criação de metadados nessa área emergente (DOOLEY; BOWERS, c2018).

Neste último relatório, o grupo WAM indica um conjunto enxuto de elementos de dados com definições de conteúdo e notas de utilização (ou seja, um dicionário de dados) adequados às características únicas dos *websites* arquivados e relevantes para a descrição de materiais em bibliotecas e arquivos como nos níveis de item e coleção, o qual podem ser usados isoladamente ou junto com outros padrões de conteúdo e de estrutura de dados mais granulares[361] (DOOLEY *et al*., 2017; DOOLEY; BOWERS, c2018). Além disto, conforme Di Pretoro e Geeraert (2019), cada elemento de dados WAN contém a vantagem de breves mapeamentos (*crosswalks*)[362] para

---

[361] Granularidade (*granular*, *granularity*) refere-se ao nível de detalhe em que um objeto de informação é visualizado ou descrito (BACA, c2016), ou conforme Lavoie e Gartner (c2013, p. 31, tradução nossa), concerne "o tamanho das unidades nas quais os componentes de dados são divididos, em geral em diferentes níveis de uma hierarquia."

[362] *Crosswalk* remete "[...] um mapeamento dos elementos, semântica e sintaxe de um esquema de metadados para os de outro." (NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004, p. 11, tradução nossa). A julgar que mapeamento (*mapping*) são correspondências entre termos, campos ou nomes de elementos usadas para traduzir dados de um padrão para outro, ou como meio de combinar termos/dados para busca e recuperação, e que mapeamento de metadados (*metadata mapping*) alude a identificação formal de elementos ou grupos de elementos de metadados dentro de diferentes esquemas para intuitos de facilitar a interoperabilidade semântica, Baca (c2016) defini *crosswalk* como uma ilustração (tabela etc.) que representa o mapeamento semântico ou técnico de campos ou elementos de dados de um padrão de dados para os de outro que possua uma função ou significado semelhante.

o DC, o EAD, o MARC 21, o MODS e o *schema.org*[363] que se destinam a facilitar tais conversões. Com base em Dooley e Bowers (c2018), o conjunto de quatorze elementos de dados do dicionário de dados WAM para a descrição de *websites* ou coleções de *sites* arquivados são:

1. Colecionador (*collector*) – a instituição incumbida pela curadoria e gestão de um *site* ou coleção arquivada. Registra o órgão que seleciona o conteúdo *Web* para arquivamento, cria metadados e executa demais atividades associadas à "apropriação" de um recurso.
2. Contribuidor (*contributor*) – a entidade (organização ou pessoa) que fez contribuições significativas, mas secundárias, ao conteúdo de um *site* ou coleção arquivada. Orienta-se indicar o papel exercido por um contribuidor, como autor, colaborador, ilustrador etc.
3. Criador (*creator*) – uma organização ou pessoa com a responsabilidade principal de ter criado o conteúdo intelectual de um *site* ou coleção arquivada. O *blog* pessoal ou o *feed* do *Twitter* de um indivíduo, são exemplos de quando uma pessoa individual é o Criador.
4. Data (*date*) – uma única data ou intervalo de datas ligado a um evento no ciclo de vida de um *site* ou coleção arquivada. Indica-se aditar um texto para clarear o significado do elemento; incluir a data cujo um URL foi rastreado; e expressar as datas em ISO 8601[364].
5. Descrição (*description*) – as notas que explicam o conteúdo, contexto e aspectos de um *site* ou coleção arquivada. Registra informação de proveniência, a razão de selecionar o conteúdo, a natureza do conteúdo do *site* ao vivo que não está na versão arquivada etc.
6. Extensão (*extent*) – uma indicação do tamanho de um *website* ou coleção arquivada. É expressa como o número/quantidade aproximada de *websites* ou de dados armazenados (em *megabytes* etc.), assim como o número e/ou tipo aproximado de arquivos coletados.
7. Gênero/Forma (*genre/form*) – um termo que determina o tipo de conteúdo de um *site* ou coleção arquivada. Dos termos, estão: *website* de notícias, página do *Twitter* etc. Indica-se adotar um vocabulário, como o *Library of Congress Genre/Form Terms* (LCGFT)[365].

---

[363] *Schema.org* alude "[...] um vocabulário RDF que permite aos criadores marcar a semântica dentro do texto das páginas da *Web*, aumentando a capacidade dos sistemas de fazer coisas interessantes com este conteúdo.", no qual "[...] é gerido através de um processo de governança comunitária." e "[...] seu escopo é em grande parte descritivo." (RILEY, c2017, p. 19, tradução nossa). É uma atividade colaborativa com a missão de criar, manter e promover esquemas para dados estruturados na *Internet*, páginas *Web* etc., o qual o vocabulário pode ser usado com muitas codificações, como RDFa, *Microdata* ou *JavaScript Object Notation for Linked Data* (JSON-LD). Disponível em: https://schema.org/. Acesso em: 25 maio 2023.

[364] Disponível em: https://www.iso.org/iso-8601-date-and-time-format.html. Acesso em: 25 maio 2023.

[365] LCGFT é um tesauro da Biblioteca do Congresso americano que define o que é uma obra *versus* o que ela trata (por exemplo, o cabeçalho de assunto *Horror films* seria um termo gênero/forma para o filme que é um filme de terror e não um filme sobre filmes de terror); além disto, combina forma (uma característica de obras com um formato e/ou finalidade específica, tal como animação) e gênero (categorias de obras caracterizadas por enredos, temas, cenários, situações etc. similares, como faroeste e suspense), onde no termo *Horror films* "*films*" é a forma e "*horror*" tange o gênero. Disponível em: https://id.loc.gov/authorities/genreForms.html. Acesso em: 25 maio 2023.

8. Língua (*language*) – o(s) idioma(s) do conteúdo arquivado, incluindo os recursos visuais e de áudio com componentes linguísticos. Recomenda-se usar uma fonte controlada de nomes de línguas ou um vocabulário controlado, tal como a série de normas ISO 639[366].
9. Relação (*relation*) – as relações todo/parte entre um único *website* arquivado e qualquer coleção a qual pertença. Orienta-se incluir o título da coleção no elemento para fornecer o contexto no qual o *site* foi coletado e, se possível, um URL para o recurso relacionado.
10. Direitos (*rights*) – declarações de direitos e permissões legais outorgados pelo direito de propriedade intelectual ou demais acordos jurídicos. Registra condições que restringem o acesso dos usuários ao conteúdo arquivado, se o acesso ao conteúdo estiver aberto etc.
11. Fonte de descrição (*source of description*) – informações sobre extração ou criação dos metadados em si, como fontes de dados e data em que os dados das fontes foram obtidos. Orienta-se incluir a data que o *site* foi analisado e o local onde a informação foi retirada.
12. Assunto (*subject*) – o tópico(s) principal(s) que descreve(m) o conteúdo de um *site* ou coleção arquivada. Registra assuntos temáticos, nomes de pessoas ou organizações e de lugares geográficos. Indica-se usar um vocabulário controlado em prol da consistência.
13. Título (*title*) – o nome pelo qual um *site* ou coleção arquivada é conhecida. Geralmente, num *site* o título é transcrito do cabeçalho da página inicial e em coleções os títulos são concebidos. Indica-se incluir títulos variantes, tais como siglas, versões multilíngues etc.
14. URL (url) – o endereço na *Internet* de um *site* ou coleção arquivada. Registra os URLs, URIs e URNs, que sejam úteis aos usuários, sobretudo, URLs iniciais e de acesso; além de URL no instante de captura, de coletas da *Web* (e período) etc. Indica-se incluir um texto para explicar a sua função.

[366] Disponível em: https://www.iso.org/iso-639-language-codes.html. Acesso em: 25 maio 2023.

De outro modo, mediante a análise dos metadados de vários projetos de arquivamento da *Web*, como da Biblioteca Nacional da Austrália[367] e dos SIA[368] nos Estados Unidos, Kim e Lee (2007) sugerem metadados descritivos e administrativos para o arquivamento intensivo da *Web*. Julgando que a maioria dos metadados dos projetos revisados se baseiam em DC e que o arquivamento intensivo da *Web* exige elementos de metadados mais detalhados devido à seletividade voltada na qualidade, os autores adotaram além dos elementos básicos do DC simples, outros elementos administrativos comuns dos projetos revisados, como:

- Disponibilidade (*availability*) – como o conteúdo da *Web* pode ser obtido ou informação de contato.
- Público (*audience*) – o grupo esperado para utilizar o conteúdo da *Web*.
- Data da captura (*date captured*) – a data associada com a captura do *website* no arquivo.
- Condição de acesso (*access condition*) – a declaração do âmbito dos usos do conteúdo da *Web* e de onde pode ser fornecido.
- Título da coleção (*collection title*) – o nome do domínio *Web*[369] ou projeto particular ou, ainda, o nome da coleção de informação organizada especialmente.
- Data de metadados modificados (*date metadata modified*) – a data em que foram feitas alterações aos metadados.
- Data de validação (*date validated*) – a data onde a página *Web* foi validada como sendo devidamente codificada usando o W3C *Markup Validation Service*[370] ou outros serviços.
- Método de coleta (*collecting method*) – o método de coleta de conteúdos da *Web* como, por exemplo, automático, manual ou transferido.

---

[367] A Biblioteca Nacional da Austrália coleta e preserva para acesso a longo prazo capturas de páginas da *Web* ou instantâneos (*snapshots*) que documentam e refletem a sociedade e a cultura australiana. Tido como *Australian Web Archive*, a coleção de *websites* e documentos da *Web* da biblioteca é acessível pelo portal de pesquisa *online Trove* (https://webarchive.nla.gov.au/collection). A coleção de arquivos da *Web* é construída pelo arquivamento seletivo colaborativo da *Web* australiana, criado em 1996 e chamado de PANDORA *Archive* (http://pandora.nla.gov.au/); a coleta em massa de *sites* do Governo da Austrália, lançado em 2014 e conhecido como *Australian Government Web Archive* – AGWA (http://webarchive.nla.gov.au/gov); e coletas anuais de depósitos legais de todo o domínio *Web .au*. Disponível em: https://www.nla.gov.au/what-we-collect/archived-websites. Acesso em: 25 maio 2023.

[368] Iniciado em 2000, o arquivo *Web* dos SIA inclui o *site* da instituição e as coleções de mídia social que foram selecionados e capturados de vários modos, como sob rastreamento de *sites* (através de um *software* que percorre a *Web* de forma automática para capturar conteúdo) e capturas de contas de mídia social (*Youtube* etc.). Os *websites* e contas de mídias sociais arquivados e preservados servem como a face pública da instituição e registram a sua história, tendo informações que podem não estar disponíveis em outro lugar. Hoje, os processos do projeto usam o *Archive-It* e o serviço *Web Recorder* (https://conifer.rhizome.org/). Disponível em: https://siarchives.si.edu/what-we-do/digital-curation/web-and-social-media-archiving. Acesso em: 25 maio 2023.

[369] Nome de domínio (*domain name*) é o endereço exclusivo que identifica um *site* da *Internet* ou outro *site* da rede (*network*). Por exemplo, a fundação J. Paul Getty Trust registrou o nome de domínio "*getty.edu*" e, assim, ela passa a ser responsável por quaisquer nomes de subdomínios de *websites*, como "*www.getty.edu*" (BACA, c2016).

[370] Disponível em: https://validator.w3.org/. Acesso em: 25 maio 2023.

- Ferramenta de coleta (*collecting tool*) – os *softwares* necessários no processo de coleta de conteúdos *Web*.

Com efeito, o esquema DC demonstra ser notável para a descrição de conteúdos da *Web* arquivados, visto as semelhanças substanciais do padrão com o conjunto de elementos de dados WAN (DOOLEY; BOWERS, c2018), a adoção e adaptação dos seus elementos essenciais nos metadados de iniciativas de arquivamento intensivo da *Web* ou, ainda, a simplicidade do padrão que motiva o seu uso geral no arquivamento extensivo da *Web* (KIM; LEE, 2007). A despeito da ambiguidade envolvida no escopo dos metadados de preservação que, consoante Dappert e Enders (2010) e Lavoie e Gartner (c2013), é retratada pelas dificuldades em categorizá-los com precisão podendo estes se estenderem por todas as classes de metadados; o trabalho centrou nos metadados descritivos e de preservação que apoiam a descoberta, identificação, apresentação, interoperabilidade e preservação digital por longo período de coleções de *websites* arquivados.

Nesse sentido, a definição e, talvez, adaptação de padrões de metadados se faz uma ação necessária em políticas de preservação digital no arquivamento da *Web*, devendo-se considerar as etapas do processo de arquivamento (seleção, captura etc.), as tecnologias (robôs rastreadores etc.) e métodos de arquivamento adotados (domínio, temático etc.) e os tipos de conteúdo *Web* coletados, mantidos e disponibilizados (página *Web*, rede social etc.), bem como o atendimento às necessidades dos usuários finais, a série de informações a serem documentadas e as decisões tomadas ante questões de direitos de propriedade intelectual, privacidade, custos, qualidade etc. e um futuro de imprevisões inerentes à preservação digital das informações publicadas na *Web*.

4.3 Identificação dos padrões e esquemas de metadados para arquivamento da *Web*

A utilidade dos metadados dá-se da sua compreensibilidade por aplicações de *software* e pessoas que os usam. Conhecidos como vocabulários de metadados, conjuntos de elementos ou, também, formatos, os esquemas (*schemas*) podem ser formalmente padronizados através de organizações de normalização (ISO, NISO e W3C, por exemplo) e, em acréscimo, hospedados e mantidos por órgãos líderes da indústria ou da comunidade, como a US *Library of Congress*, que os endossa para uso em suas comunidades-alvo (RILEY, c2017). Os padrões de metadados (*metadata standards*) ajudam a tornar os metadados os mais úteis possíveis pois, conforme *Digital Preservation Coalition* [2018a], cedem diretrizes para uma formatação uniforme à medida que os esquemas são diretrizes para formatos uniformes de metadados, desta forma, tanto os padrões quanto os esquemas asseguram que os metadados para registros digitais sejam interoperáveis.

Como definição conceitual de padrões de metadados, Alves (2010, p. 47-48) interpreta que os padrões de metadados (*metadata statement*) "[...] são estruturas de descrição constituídas por um conjunto predeterminado de metadados (atributos codificados ou identificadores de uma entidade) metodologicamente construídos e padronizados [...]" que se propõem "[...] descrever uma entidade gerando uma representação unívoca e padronizada que possa ser utilizada para recuperação da mesma". Isto posto, chamados também de esquemas, os esquemas de metadados (*metadata schema*) são o conjunto de elementos de metadados (e regras para o seu uso) de um padrão criados para um propósito, como descrever um tipo particular de recurso informacional (CHAN; ZENG, c2006; NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004). Em Zeng e Qin (2008, p. 323, tradução nossa) os esquemas de metadados constituem:

> Uma especificação processável por máquina que define a estrutura, codificação de sintaxe, regras, e formatos para o conjunto de elementos de metadados em uma linguagem formal num esquema. Na literatura o termo esquema de metadados refere-se usualmente ao conjunto de elementos na sua totalidade, bem como a codificação dos elementos e a estrutura com uma linguagem de marcação.

De fato, através de Castro (2012), Chan e Zeng (c2006), *National Information Standards Organization* (c2004) e Vellucci (2000), o esquema (*schema*) é uma entidade no todo, incluindo os componentes semânticos e de conteúdo (chamados de conjunto de elementos de metadados) como a codificação dos metadados com uma sintaxe ou linguagem de marcação[371] (o formato MARC e uma XML/SGML DTD, por exemplo), que têm três partes ou características básicas:

1. Estrutura – o modelo de dados ou arquitetura utilizada para comportar os metadados e a maneira como as declarações (*statements*) dos metadados são expressas. Diz respeito à estrutura dos metadados e não com a categoria de metadados estruturais os quais trata-se da estrutura descritiva inicialmente dos recursos informacionais. Como exemplos de modelos de dados, estão a arquitetura de metadados RDF e o esquema XML[372] METS.
2. Semântica – os nomes e significados dos elementos e seus refinamentos. Não determina o espaço do conteúdo junto aos elementos. Isto cabe aos padrões de conteúdo de dados

[371] De acordo com Baca (c2016), linguagem de marcação (*markup language*) alude uma forma formal de anotar um documento com o uso de etiquetas (*tags*) de codificação inseridas para indicar a estrutura do documento ou arquivo e o conteúdo de seus elementos de dados. Como exemplos, existem as linguagens padronizadas HTML, XML e SGML, do qual esta última "[...] é um super conjunto de HTML e XML e proporciona a mais rica marcação de um documento." (NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004, p. 3, tradução nossa).

[372] Conforme Baca (c2016), esquema XML (XML *schema* –, onde *schema* é a organização, estrutura e regras para a codificação de informações) refere-se a uma definição legível por computador da estrutura, elementos e atributos permitidos numa instância válida de um documento XML. É expresso usando a XML *Schema Definition* (XSD) (https://www.w3.org/TR/xmlschema-0/), uma recomendação do W3C e linguagem baseada no formato XML, isto é, no "[...] perfil de aplicação de SGML projetado para uso em aplicações *Web*." (NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004, p. 16, tradução nossa).

(regras de conteúdo) e vocabulários controlados (padrões de valor de dados), sendo que os primeiros especificam, por exemplo, como a data será formatada junto aos elementos; e os últimos remetem a listas de palavras que termos são selecionados e seus sinônimos, delimitando a extensão de valores que poderão ser inseridos no interior de uma classe.

3. Conteúdo – as declarações ou instruções de como e quais valores devem ser atribuídos aos elementos. Refere-se a definição de regras de conteúdo para como o conteúdo deve ser formulado (por exemplo, como identificar o título principal); regras de representação para conteúdo, como padrões de representação do tempo; valores de conteúdo admitidos (isto é, se os valores devem ser retirados a partir de um vocabulário controlado, aditados por criadores de metadados sem uma lista de termos controlados, derivados do texto ou cedidos pelo autor); e regras de sintaxe para codificação dos elementos e seu conteúdo.

Deste modo, o esquema de metadados (*schema*) defini atributos e regras sob os aspectos semânticos e estruturais, consistindo de outros tipos de esquemas (os *schemes*) que determinam a sintaxe de codificação dos dados em suporte ao estabelecimento da estrutura e da semântica (significado) dos atributos e valores em um padrão de metadados (ALVES, 2010; ZENG; QIN, 2008). Julgando que os metadados são regidos por padrões e práticas para assegurar qualidade, consistência e interoperabilidade, Gilliland (c2016) traz uma tipologia de padrões de dados que organiza estes padrões em quatro categorias e fornece exemplos, o qual são assim entendidas:

- Padrões de estrutura de dados (conjuntos de elementos de metadados, schemas) – são "categorias" ou "contêineres" de dados que constituem um registro ou outro objeto de informação. A título de exemplo, temos os formatos MARC, *Bibliographic Framework* (BIBFRAME)[373], DC, CDWA, EAD, VRA *Core* e *Text Encoding Initiative* (TEI)[374].
- Padrões de valor de dados (vocabulários controlados, tesauros, listas controladas) – são os termos, nomes e outros valores usados para preencher padrões de estrutura de dados. Por exemplo, temos os vocabulários LCSH, *Library of Congress Thesaurus for Graphic*

[373] BIBFRAME é um projeto "[...] que visa desenvolver um novo modelo para codificação e compartilhamento de informações bibliográficas." como "[...] um vocabulário RDF formal." que pretende "[...] substituir o MARC 21, e manter a semântica do MARC 21, permitindo que proporções significativas dos dados existentes sejam migradas adiante." (RILEY, c2017, p. 28-29, tradução nossa). Trata-se de um modelo de dados para descrição bibliográfica que, para Baca (c2016), é estruturado sob os princípios de *Linked Data* para tornar os dados bibliográficos mais úteis dentro da comunidade de biblioteca. Disponível em: https://www.loc.gov/bibframe/. Acesso em: 26 maio 2023.

[374] TEI remete [...] uma linguagem de marcação para textos legíveis por máquina de todos os tipos, incluindo prosa, verso, textos, transcrições de apresentações de palavras faladas, dicionários, e manuscritos [...]" (RILEY, c2017, p. 36, tradução nossa), e [...] um esquema de metadados para texto eletrônico." (NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004, p. 16, tradução nossa). Também para Baca (c2016) a TEI é um esforço cooperativo internacional em elaborar diretrizes para esquemas de codificação (*encoding schemes*) padrão, como DTD TEI e TEI *Lite*, para textos literários e linguísticos. Disponível em: https://tei-c.org/. Acesso em: 26 maio 2023.

Materials (TGM)[375] e LCNAF[376]; US NLM *Medical Subject Headings* (MeSH)[377]; Getty TGN, *Art & Architecture Thesaurus* (AAT)[378] e *Union List of Artist Names* (ULAN)[379]; e *Iconclass*[380].

- Padrões de conteúdo de dados (regras e códigos de catalogação) – são diretrizes para o formato e sintaxe dos valores de dados que são usados para preencher os elementos de metadados. Dos exemplos, há as normas RDA, DACS[381], AACR2, CCO e ISBD[382].
- Padrões de intercâmbio técnico ou formato de dados (padrões de metadados expressos em formato legível por máquina) – são comumente uma manifestação de um padrão de estrutura de dados particular, codificado ou marcado (*marked up*) para o processamento legível por máquina. Como exemplos, estão os formatos MARC21 e MARCXML; DTD

[375] O TGM é uma ferramenta para indexação de materiais visuais por assunto e por gênero/formato, que tem mais de sete mil termos de assunto e seiscentos termos de gênero/formato para indexar tipos de fotografias, desenhos de *design*, impressões etc. Disponível em: https://www.loc.gov/pictures/collection/tgm/. Acesso em: 26 maio 2023.

[376] O LCNAF concede dados confiáveis para nomes de pessoas, organizações, eventos, localidades e títulos com o intuito de identificar essas entidades e, por intermédio da sua utilização, proporciona acesso uniforme aos recursos de informação bibliográficos. Disponível em: https://id.loc.gov/authorities/names.html. Acesso em: 26 maio 2023.

[377] O tesauro MeSH é um vocabulário controlado organizado de forma hierárquica e desenvolvido pela Biblioteca Nacional de Medicina americana (ou NLM), no qual pode ser usado para indexar, catalogar e buscar informações biomédicas e de saúde. Disponível em: https://www.nlm.nih.gov/mesh/meshhome.html. Acesso em: 26 maio 2023.

[378] O AAT (https://www.getty.edu/research/tools/vocabularies/aat/index.html) alude um tesauro do *Getty Research Institute* que inclui termos genéricos, datas, relações, fontes e notas para tipos de trabalho, papéis, estilos, técnicas, materiais e demais conceitos associados à arte, arquitetura e outro patrimônio cultural (por exemplo, tinta a óleo, renascimento etc.). Disponível em: https://www.getty.edu/research/tools/vocabularies/. Acesso em: 26 maio 2023.

[379] Mantidos pelo *Getty Research Institute*, AAT, TGN e ULAN são recursos estruturados usados para melhorar o acesso a informações sobre arte, materiais bibliográficos e de arquivo e outros materiais culturais, sendo que este último vocabulário inclui nomes, relações, fontes e informações biográficas para pessoas físicas e jurídicas, sejam nomeadas/anônimas. Disponível em: https://www.getty.edu/research/tools/vocabularies/. Acesso em: 26 maio 2023.

[380] *Iconclass* é um sistema de classificação para arte e iconografia utilizado na descrição e recuperação de assuntos representados em imagens (obras de arte etc.). Disponível em: http://www.iconclass.nl/. Acesso em: 26 maio 2023.

[381] DACS é um padrão de conteúdo para descrever coleções de arquivo e que adota uma abordagem multinível, isto é, uma descrição pode ter elementos de dados que descrevem o grupo todo de materiais, um subconjunto ou itens individuais (DOOLEY; BOWERS, c2018). Mantido pelo *Technical Subcommittee for* DACS (TS-DACS) da *Society of American Archivists* (SAA) *Standards Committee*, o DACS é compatível com a *International Standard Archival Authority Record for Corporate Bodies, Persons and Families* (ISAAR (CPF)) e a ISAD(G). Disponível em: https://saa-ts-dacs.github.io/. Acesso em: 26 maio 2023.

[382] ISBD é um conjunto de regras concisas, produzidas e mantidas pelo Grupo de Revisão (*Review Group*) ISBD da IFLA (https://www.ifla.org/units/isbd-rg/), que regularizam a forma e o conteúdo das descrições bibliográficas e, principalmente, visam oferecer consistência ao compartilhar informações bibliográficas. Disponível em: https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/8912. Acesso em: 26 maio 2023.

EAD XML; esquemas XML DC Simples e Qualificado, CDWA *Lite* 1.1, VRA Core 4.0 e *Lightweight Information Describing Objects* (LIDO)[383]; RDF; dentre outros.

Tais padrões de dados são tratados por Alves (2010) e Zeng e Qin (2008), que os inferem como componentes básicos dos formatos de metadados e esquemas de codificação usados neles para a construção padronizada de representações das informações. Dividindo-os em esquemas para codificação de valores e esquemas para codificação de conteúdo, as autoras entendem que padrões de valores de dados (*data value standards*) e padrões de conteúdo de dados (*data content standards*) constituem o primeiro tipo de esquema de codificação; já os esquemas de codificação de conteúdo são compostos pelos padrões para intercâmbio de dados (*data exchange standards*) e, igualmente, pelos padrões para estrutura de dados (*data structure standards*) que remetem a estrutura do conjunto de elementos de um formato representada pelas declarações (*statements*).

Um modo de padronizar os metadados pelo controle dos valores reais utilizados é o uso dos vocabulários controlados. Estes padrões de valor de dados são listas predefinidas de termos sobre um certo tópico ou tipo, como *Internet* MIME *types*[384] e gêneros *Spotify*[385], que geralmente identificam uma palavra ou frase preferida para algum conceito e, por vez, cedem mapeamentos dos outros termos para o conceito ao o preferido, como definem relações hierárquicas entre os termos (RILEY, c2017). Outro método de padronizar os valores apresentados nos metadados é o uso dos padrões de conteúdo de dados, esquemas externos, diretrizes que, para Alves (2010, 2017) e Baca (c2016), definem vocabulário, ordem, sintaxe, ou forma do conteúdo inserido nos elementos de metadados como, por exemplo, a norma ISO 8601 para o registro de data e hora.

Sobre a padronização da sintaxe, as linguagens de marcação (*markup languages*) e uma série de esquemas (*schemas*) e formatos de metadados cedem formas padronizadas de estruturar e expressar os padrões que regem os metadados para o processamento por máquina, publicação e implementação (GILLILAND, c2016). Na preservação digital o formato XML é tido um tipo de migração pois, segundo Márdero Arellano (2008), participa enriquecendo informação sobre estruturas e significado, assegura o encapsulamento dos metadados e das informações exigidas para interpretação dos objetos digitais originais e propicia a interoperabilidade entre recursos

[383] Mantido pelo grupo de trabalho LIDO – anteriormente *Data Harvesting and Interchange Working Group* – do *International Committee for Documentation* (CIDOC)/*International Council of Museums* (ICOM), o LIDO é um esquema XML segundo o CIDOC *Conceptual Reference Model* (CRM) para descrever e trocar informações sobre objetos de museu. Disponível em: http://cidoc.mini.icom.museum/working-groups/lido/. Acesso em: 4 mar. 2021.
[384] Disponível em: https://www.iana.org/assignments/media-types/media-types.xhtml. Acesso em: 26 maio 2023.
[385] Disponível em: https://www.spotify.com/br/. Acesso em: 26 maio 2023.

de distintas áreas. Já o RDF atua como o núcleo de armazenamento para descrições semânticas compiladas de outros formatos, que permite a interoperabilidade semântica[386] (CASTRO, 2012).

Da revisão e análise da literatura identificamos vários padrões e esquemas de metadados utilizados para a descrição de recursos em distintos domínios, como demonstrado no Quadro 1.

[386] Interoperabilidade semântica é a capacidade de distintos agentes, serviços ou aplicações de comunicarem dados, com garantias de precisão e salvaguarda do significado dos dados (BACA, c2016).

Quadro 1 – Alguns padrões e esquemas de metadados e seus escopos.

Padrão	Especificação
Dublin Core Metadata Element Set (DCMES)	Publicado em 1995, o DC é um esquema representado em diversas sintaxes, como XML e RDF, para fins de catalogação e de descoberta de recursos eletrônicos em ambiente *Web*. É amplamente utilizado no arquivamento da *Web*. Hoje, situa-se na versão 1.1 de 2012 (HARPER, 2010; RILEY, c2017; VENLET *et al*., c2018).
Visual Resources Association Core (VRA *Core*)	Publicado em 1996, o VRA Core é um esquema XML para registrar informações descritivas sobre obras de arte e reproduções exclusivas delas. Hoje, está na versão 4.0 de 2007 (LIMA; SANTOS; SANTARÉM SEGUNDO, 2016; RILEY, c2017).
Encoded Archival Description (EAD)	Lançado em 1998, o EAD é um esquema XML para codificação de instrumentos de pesquisa e acesso, como inventários, índices, guias, registros etc. Hoje, está na versão EAD3 1.1.1 de 2019 (LIBRARY OF CONGRESS, 2013; RILEY, c2017).
Technical Metadata for Text (*Text*MD)	Publicado no início dos anos 2000, o *Text*MD é um esquema XML que detalha metadados técnicos para objetos digitais baseados em texto. Em geral, serve como um esquema de extensão nos padrões METS e PREMIS. Atualmente, encontra-se na versão *alpha* 3.01 lançada em 2009 (LIBRARY OF CONGRESS, 2017, 2020b).
Metadata Encoding and Transmission Standard (METS)	Lançado em 2001, o METS é um documento XML para codificar metadados sobre objetos complexos em bibliotecas digitais. Pode servir de invólucro de metadados descritivos, estruturais e administrativos para conteúdo *Web*. Está na versão 1.12.1 de 2019 (ENDERS, 2010; LIBRARY OF CONGRESS, 2017; SAYÃO, 2010).
Multimedia Content Description Interface (MPEG-7)	Definido na ISO/IEC 15938 (2002), o MPEG-7 do *Moving Picture Experts Group* (MPEG) é um esquema XML para descrição de conteúdo audiovisual em ambientes multimídia (MARTÍNEZ, 2004).
Learning Object Metadata (LOM)	Especificado pela norma *Institute of Electrical and Electronics Engineers* (IEEE) 1484.12.1-2002[387], o LOM é um esquema XML que define o conjunto mínimo de atributos para gerir, localizar e avaliar objetos de aprendizagem (CHAN; ZENG, 2006; NATIONAL INFORMATION STANDARDS ORGANIZATION, c2004).
Machine Readable Cataloging (MARC) 21 XML *Schema*	Lançado em 2002, o MARC XML é um esquema para trabalhar com dados MARC em um ambiente XML. Visa ser flexível e extensível para permitir que os usuários trabalhem com dados MARC de maneiras específicas às suas necessidades. Hoje, encontra-se na versão 1.2 emitida em 2009 (LIBRARY OF CONGRESS, 2020a).
Audio/Video Technical Metadata Extension Schema (*Audio*/*Video*MD)	Lançados em 2002, o *Audio*MD e o *Video*MD são esquemas XML que detalham metadados técnicos para objetos digitais baseados em áudio e vídeo. Geralmente servem como esquemas de extensão nos padrões METS e PREMIS. Hoje, ambos situam na versão 2.0 revisada em 2011 (LIBRARY OF CONGRESS, 2011, 2017).
Metadata Object Description Schema (MODS)	Publicado em 2003, o MODS é um esquema XML derivado do padrão MARC 21 para informações bibliográficas de particular interesse para bibliotecas. Hoje, está na versão 3.7 de 2018 (LIBRARY OF CONGRESS, 2016, 2018a; RILEY, c2017).
Metadata Authority Description Schema (MADS)	Publicado em 2005, o MADS é um esquema XML para um conjunto de elementos de autoridade que provê metadados sobre agentes (pessoas, organizações), eventos e termos (tópicos, geográficos, gêneros etc.), sendo uma companhia para o MODS. Hoje, situa-se na versão 2.1 de 2016 (LIBRARY OF CONGRESS, 2018a, 2018b).
PREservation Metadata: Implementation Strategies (PREMIS)	Lançado em 2005, o esquema PREMIS XML permite implementar um conjunto básico de unidades semânticas para codificar, armazenar, gerenciar e intercambiar metadados de preservação entre os sistemas de arquivamento digital. Atualmente, encontra-se na versão 3.0 de 2016 (LAVOIE; GARTNER, c2013; RILEY, c2017).
NISO *Metadata for Images in* XML *Schema* (MIX)	Publicado em 2007, o MIX é um esquema XML que fornece um formato para troca e/ou armazenamento dos dados definidos na norma ANSI/NISO Z39.87-2006 (R2017)[388]. Atualmente, situa-se na versão 2.0 emitida em 2008 (FORMENTON *et al*., 2017; LIBRARY OF CONGRESS, 2015).
Encoded Archival Context – Corporate Bodies, Persons, and Families (EAC-CPF)	Lançado em 2010, o EAC-CPF é um esquema XML que padroniza a codificação de descrições sobre agentes (que criam/usam materiais arquivísticos) para permitir a troca, descoberta e exibição dessas informações num meio eletrônico. Hoje, está na versão 2010 revisada em 2018 (STAATSBIBLIOTHEK ZU BERLIN, 2017).

Fonte: Elaborado pelo autor.

[387] Disponível em: https://ieeexplore.ieee.org/document/1032843. Acesso em: 26 maio 2023.
[388] Disponível em: https://bit.ly/2DUU1KG. Acesso em: 26 maio 2023.

A maioria dos padrões de metadados supracitados teve as suas origens no momento em que a *Web* estava em seu começo. Na segunda metade dos anos 90 e início dos anos 2000 houve um rápido desenvolvimento de formatos para as necessidades de comunidades específicas e a codificação de objetos digitais complexos, os quais são delimitados por seus próprios conjuntos de elementos de metadados, particularidades e domínios de aplicação. A seguir são discutidos alguns dos principais padrões de metadados vigentes e indicados na literatura especializada para o arquivamento da *Web*. No entanto, não serão retratados todos os elementos de cada esquema e sim será realizada o apontamento apenas daqueles que fazem parte da análise dos resultados, ocasião onde são expostos o mapeamento e a indicação de elementos para o arquivamento da *Web*. O estudo de identificação, sistematização e análise de padrões de metadados para a preservação digital (em especial, o DC, o MODS, o EAD, o MIX, o PREMIS e o METS) pode ser examinado em Formenton (2015), o qual fornece outras informações adicionais associadas a este capítulo.

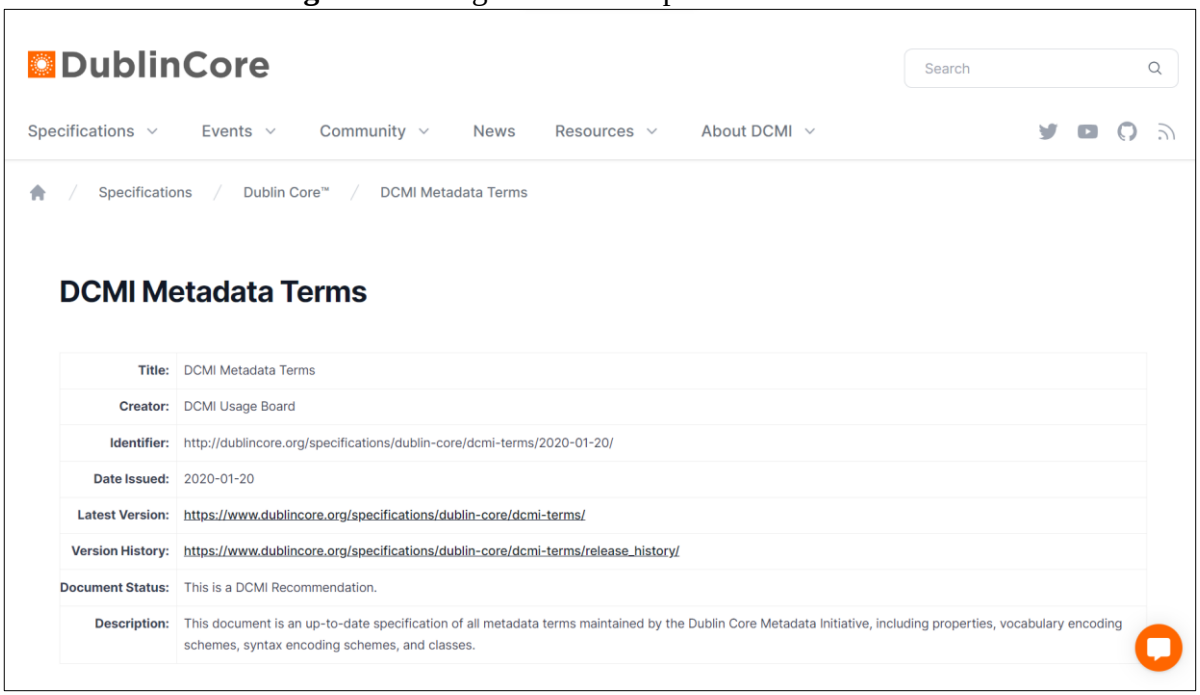
4.3.1 Padrão *Dublin Core*

O DC tem seu início em Chicago, na 2ª *International World Wide Web Conference*, em 1994, num debate sobre semântica e a *Web* diante da dificuldade da descoberta de recursos. Tal fato, fez a OCLC e o *National Center for Supercomputing Applications* (NCSA)[389] a realizarem o OCLC/NCSA *Metadata Workshop* na cidade norte-americana de *Dublin*, *Ohio*, em 1995, em que se discutiu como um conjunto semântico básico seria útil para a busca e a recuperação de recursos baseados na *Web*. O resultado foi chamado de "metadados *Dublin Core*" com base na localização do *workshop*[390] (DUBLIN CORE METADATA INITIATIVE, c2020a). Consoante Harper (2010) e Sayão (2010) o conjunto de elementos DC é pequeno e simples, de modo que são compreensíveis semanticamente; ademais, o DC é representado por diversas sintaxes, como codificado em HTML ou em XML e estruturado em RDF, propiciando o intercâmbio e o reuso.

[389] Disponível em: http://www.ncsa.illinois.edu/. Acesso em: 10 jun. 2020.

[390] O DC original de treze (depois quinze) elementos foi publicado pela primeira vez no relatório do evento em 1995 (http://www.dlib.org/dlib/July95/07weibel.html). Foi formalizado no RFC 5791 (1998) da *Internet Engineering Task Force* (IETF), na ANSI/NISO Z39.85-2001 e na ISO 15836-2003, do qual as versões mais recentes dessas normas são RFC 5791 (2010), Z39-85-2012, ISO 15836-1:2017 e ISO 15836-2:2019. Surgida nos anos 90 e sendo um projeto da *Association for Information Science and Technology* (ASIS&T), a DCMI é uma organização internacional (o qual um dos membros institucionais é a universidade brasileira UNESP, por exemplo) que desde 2002 exerce o papel de agência de padrões "de fato", mantendo sua própria documentação atualizada dos termos de metadados DCMI. O DCMI *Usage Board* atualmente atua como agência de manutenção da ISO 15836. Disponível em: https://www.dublincore.org/specifications/dublin-core/. Acesso em: 30 maio 2023.

Figura 66 – Página inicial do padrão *Dublin Core*.



Fonte: *Dublin Core Metadata Initiative* (2020b).

Hoje na versão 1.1, o DC é um vocabulário de dois níveis: simples e qualificado. Assim, o DC simples abrange quinze propriedades ou elementos essenciais (o *core*) e o DC qualificado contém elementos adicionais; além de qualificadores que especificam o significado do elemento (refinamento de elemento) ou identificam esquemas na interpretação do seu valor (esquema de codificação) (DUBLIN CORE METADATA INITIATIVE, 2000, 2012, 2020b). Para o escopo da preservação digital, Formenton *et al*. (2017) destaca alguns elementos DC qualificado como, por exemplo, Formato (*format*), Identificador (*identifier*), Direitos (*rights*), Detentor de Direitos (*rightsHolder*) e Proveniência (*provenance*), que embora prestados ao acesso do que exatamente para preservação documentam informações previstas nos metadados de preservação PREMIS.

Independentemente das críticas a estrutura e ao conjunto muito simplista e genérico dos elementos DC (sobretudo, defronte aos demais formatos, como o MARC), o DC é um norteador da interoperabilidade semântica e do consenso entre diversas comunidades no mundo, inclusive executa um papel de liderança na criação de metadados descritivos de arquivamento da *Web* (DOOLEY *et al*., 2017; DOORN; TJALSMA, 2007; HARPER, 2010). Como identificado pelo grupo de trabalho WAM da OCLC *Research*, mediante Dooley e Bowers (c2018), Samouelian e Dooley (c2018) e Venlet *et al*. (c2018), o esquema de metadados descritivos do DC na versão

1.1 é amplamente utilizado na descrição de *websites* arquivados pelos usuários do *Archive-It*[391], um serviço de arquivamento da *Web* por assinatura da iniciativa internacional *Internet Archive*.

4.3.2 Padrão MODS

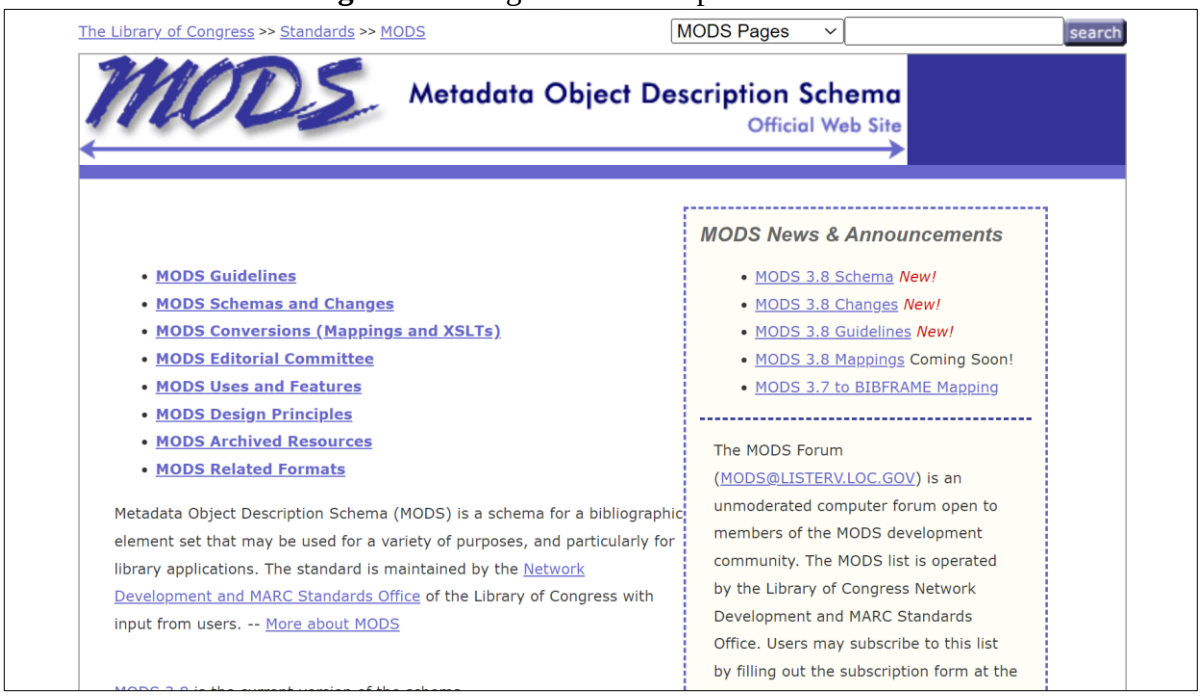
Projetado pela Biblioteca do Congresso dos Estados Unidos em 2002, o esquema MODS pode ser adotado em particular para aplicações de bibliotecas[392]. Expresso em XML, este padrão de metadados descritivos inclui um subconjunto de campos MARC 21 e usa etiquetas baseadas em palavras e não numéricas, permitindo uma fácil compreensão (LIBRARY OF CONGRESS, 2016, 2018a). Como vantagens do MODS, mediante Guenther (2003) e Mccallum (2004), nota-se que o MODS é mais simples que o MARC completo e enseja uma descrição mais rica frente ao DC qualificado; ademais, no MODS há o reagrupamento de certos elementos MARC e, em alguns casos, o que está em vários elementos MARC é reunido num único elemento MODS[393].

[391] Lançado em 2006, o *Archive-It* é a principal solução de arquivamento da *Web* para uma variedade de instituições de patrimônio cultural, sendo utilizado na criação, armazenamento e acesso a coleções de conteúdos da *Web*. Além da funcionalidade central de capturar e preservar conteúdos *Web*, a aplicação *Archive-It* possibilita que os usuários adicionem, importem e exportem metadados descritivos bem como permite a navegação pública e a pesquisa de texto completo via *website* oficial. Disponível em: https://archive-it.org/blog/learn-more/. Acesso em 30 maio 2023.

[392] O MODS foi elaborado e é mantido pelo *Network Development and* MARC *Standards Office* da Biblioteca do Congresso americano, com o apoio dos usuários e *experts*. No *website* oficial do padrão há descrições de projetos que utilizam o MODS, como nos repositórios do *Academic Commons* (https://academiccommons.columbia.edu/) e da *Digital Library Collections* (https://dlc.library.columbia.edu/) pelas bibliotecas da Universidade de *Columbia* nos Estados Unidos. Disponível em: http://www.loc.gov/standards/mods/registry.php. Acesso em: 30 maio 2023.

[393] Também no *website* oficial do padrão temos amostras de como os elementos MODS são usados num registro completo de um *site* arquivado (http://www.loc.gov/standards/mods/userguide/examples.html#archived_website) ou até de um artigo de jornal (http://www.loc.gov/standards/mods/userguide/examples.html#journal_article), por exemplo. Disponível em: http://www.loc.gov/standards/mods/userguide/examples.html. Acesso em: 30 maio 2023.

Figura 67 – Página inicial do padrão MODS.



Fonte: *Library of Congress* (2023d).

Atualmente na versão 3.8, o esquema MODS possui um conjunto de vinte elementos de metadados descritivos de nível superior, onde concede informações bibliográficas que integram demais esquemas XML, tais como o METS e o PREMIS. Sob o enfoque da preservação digital, Formenton *et al*. (2017) observam três elementos MODS: Informação de Origem (<*titleInfo*>), Item Relacionado (<*relatedItem*>) e Condição de Acesso (<*accessCondition*>). Para os autores estes elementos documentam informações úteis que auxiliam os metadados de preservação, seja na comprovação da autenticidade, integridade e proveniência dos objetos como na identificação dos direitos do recurso eletrônico que intervêm na preservação, acesso e uso dos seus conteúdos.

Embora os elementos MODS herdam a semântica dos elementos MARC, a conversão de um registro MARC original para MODS e depois o retorno para MARC resulta em perda de dados ou de especificidade na marcação. Logo, o MARC XML[394] deve ser usado antes para uma troca sem perdas (GUENTHER, 2003; LIBRARY OF CONGRESS, 2016). Sobre exemplos de uso do MODS no arquivamento da *Web*, Guenther e Myrick (2007) indicam o Arquivo da *Web* da Biblioteca do Congresso americano, criado originalmente no projeto "*Mapping the Internet Electronic Resources Virtual Archive*" (MINERVA) em parceira com o

[394] Igualmente desenvolvido e mantido pelo *Network Development and* MARC *Standards Office* da Biblioteca do Congresso americano, o MARC XML pode ser usado para a descrição original do recurso em XML, representação do registro MARC completo em XML, metadados em XML reunidos com o recurso, ou esquema de extensão para o METS. Disponível em: http://www.loc.gov/standards/marcxml/marcxml-overview.html. Acesso em: 30 maio 2023.

Internet Archive, o qual é composto por coleções de *websites* arquivados[395] que são catalogados aplicando o MODS.

4.3.3 Padrão EAD

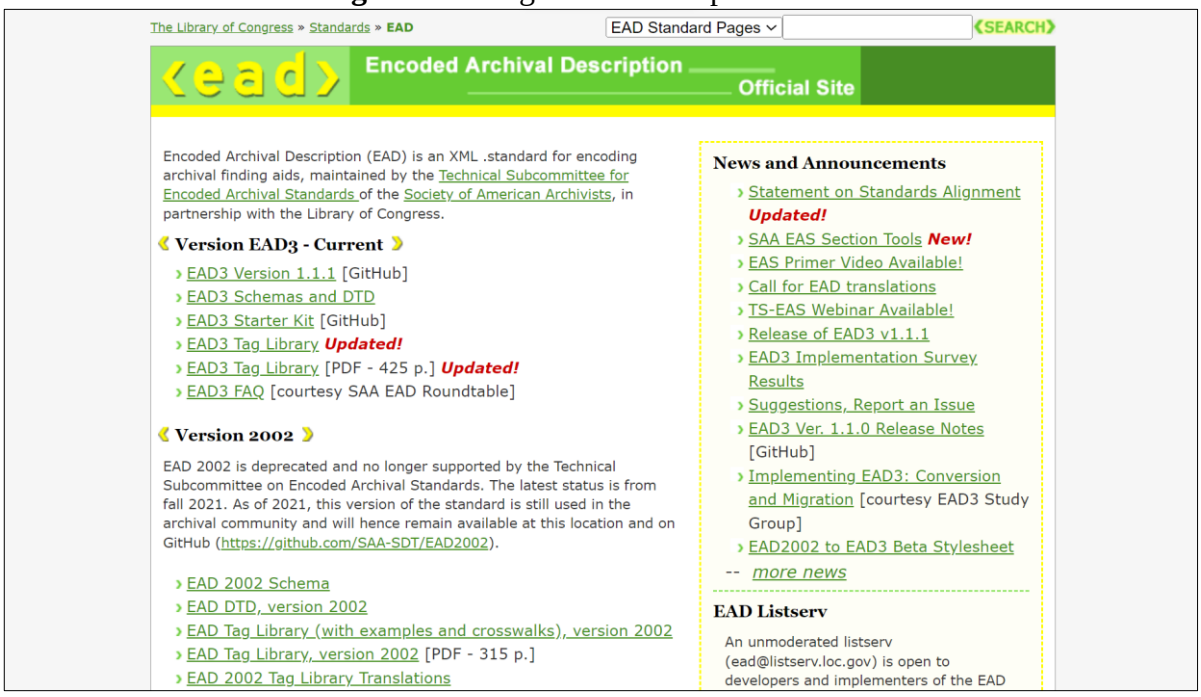
O esquema EAD originou-se em um projeto da biblioteca da Universidade da Califórnia, *Berkeley*, em 1993. Dirigido por Daniel Pitti[396], o projeto *Berkeley* visou desenvolver um padrão de codificação não-proprietário para instrumentos de pesquisa[397] legíveis por computador, como inventários, índices, registros, guias e documentos criados por arquivos, bibliotecas, museus e repositórios para apoiar o uso de suas coleções (LIBRARY OF CONGRESS, 2013). Conforme Allison-Bunnell (2016) e Pala (2017) a versão EAD3 se centra na simplificação do padrão e no aumento de clareza e consistência semântica ante as versões EAD 1.0 e EAD 2002, promovendo a interoperabilidade e a melhora da funcionalidade em ambientes internacionais e multilíngues.

[395] Dos exemplos de coleções com *websites* arquivados na Biblioteca do Congresso americano, estão: *websites* da biblioteca de 2016 até hoje; 11 de setembro de 2001; eleições americanas de 2000 até hoje; literatura brasileira de cordel de 2011 até hoje; eleições presidenciais no Brasil de 2010 até hoje; transição papal de 2005 e 2013; e jogos olímpicos de inverno de 2002. Disponível em: https://www.loc.gov/websites/collections. Acesso em: 30 maio 2023.

[396] PITTI, Daniel V. Encoded Archival Description: the development of an encoding standard for archival finding aids. **American Archivist**, [*S. l.*], v. 60, n. 3, p. 268-283, 1997. Disponível em: https://americanarchivist.org/doi/pdf/10.17723/aarc.60.3.f5102tt644q123lx. Acesso em: 30 maio 2023.

[397] Instrumento de pesquisa (*finding aid*) refere-se ao "[...] meio que permite a identificação, localização ou consulta a documentos ou a informações neles contidas." (ARQUIVO NACIONAL, 2005, p. 108).

Figura 68 – Página inicial do padrão EAD.



Fonte: *Library of Congress* (2023b).

Hoje na versão 1.1.1 do EAD3, este padrão XML[398] tem um conjunto de cento e sessenta e cinco elementos descritivos e oitenta e cinco atributos, que fornece informações bibliográficas que se alinham a outros esquemas XML, tais como o EAC-CPF[399] (SOCIETY OF AMERICAN ARCHIVISTS, 2019). No propósito da preservação digital, Formenton *et al*. (2017) observam certos elementos EAD 2002 mantidos na versão 1.1.1 do EAD3, como a Descrição Arquivística (<*archdesc*>). De acordo com os autores os padrões DC, MODS e EAD mesmo que sejam mais aplicáveis à descoberta, busca, recuperação ou localização de recursos ao invés da preservação, são esquemas úteis para o registro de metadados descritivos de amparo ao PREMIS e o METS.

Ainda que a falta de recursos e de conhecimento/*expertise* disponível numa instituição influencie à sua adoção, nos últimos vinte anos, como levantado por Eidson e Zamon (2019), o EAD mantêm-se relevante pelo grande número de arquivos que o adotaram e continuam a usá-lo atualmente para publicar seus instrumentos de pesquisa *online*. Dos exemplos de uso do EAD

[398] No início o EAD foi criado em SGML, com a participação da SAA e de uma equipe de especialistas. Depois, o EAD foi reestruturado para uma maior compatibilidade com o formato XML e os preceitos da ISAD(G). Assim, as versões 1.0 e 2002 do EAD em XML foram lançadas, nesta ordem, em 1998 e 2002. Hoje na versão EAD3, de 2019, o padrão é mantido pelo *Technical Subcommittee for Encoded Archival Standards* (TS-EAS) da SAA e a Biblioteca do Congresso americano. Disponível em: http://www.loc.gov/ead/eaddev.html. Acesso em: 30 maio 2023.

[399] Desenvolvido e mantido pelo TS-EAS da SAA com a biblioteca estadual alemã de Berlin (*Staatsbibliothek zu Berlin*), o EAC-CPF é um esquema XML para a ISAAR (CPF) usado em estreita associação com o padrão EAD, mas não está limitado a ele. Disponível em: https://eac.staatsbibliothek-berlin.de/about/. Acesso em: 2 mar. 2021.

no arquivamento da *Web*, temos o Arquivo *Online* da Califórnia[400], que fornece acesso público e gratuito a descrições detalhadas de coleções de fontes primárias mantidas por instituições em todo o estado da Califórnia, como o arquivo da *Web* da Universidade da Califórnia em *Irvine*[401] (https://oac.cdlib.org/findaid/ark:/13030/c8q81jn9/), através de instrumentos de pesquisa EAD.

4.3.4 Padrão VRA *Core*

Desenvolvido pela VRA[402] em 1996, o VRA *Core* é um esquema para a descrição de obras culturais visuais – incluindo pinturas, desenhos, esculturas, arquitetura, fotografias etc. – , como de imagens que as documentam. É usado como um formato independente e como um esquema de extensão do METS para objetos que contêm recursos de patrimônio cultural[403] (VISUAL RESOURCES ASSOCIATION, 2014). Segundo Lima, Santos e Santarém Segundo (2016) e Lubas, Jackson e Schneider (2013) este padrão tem versões, sendo a VRA 1.0 (1996) baseada no CDWA, a VRA 2.0 (1996) que indicou a busca de padrões CCO e a VRA 3.0 (2000) que semelha ao DC em simplicidade, número de elementos e qualificadores.

---

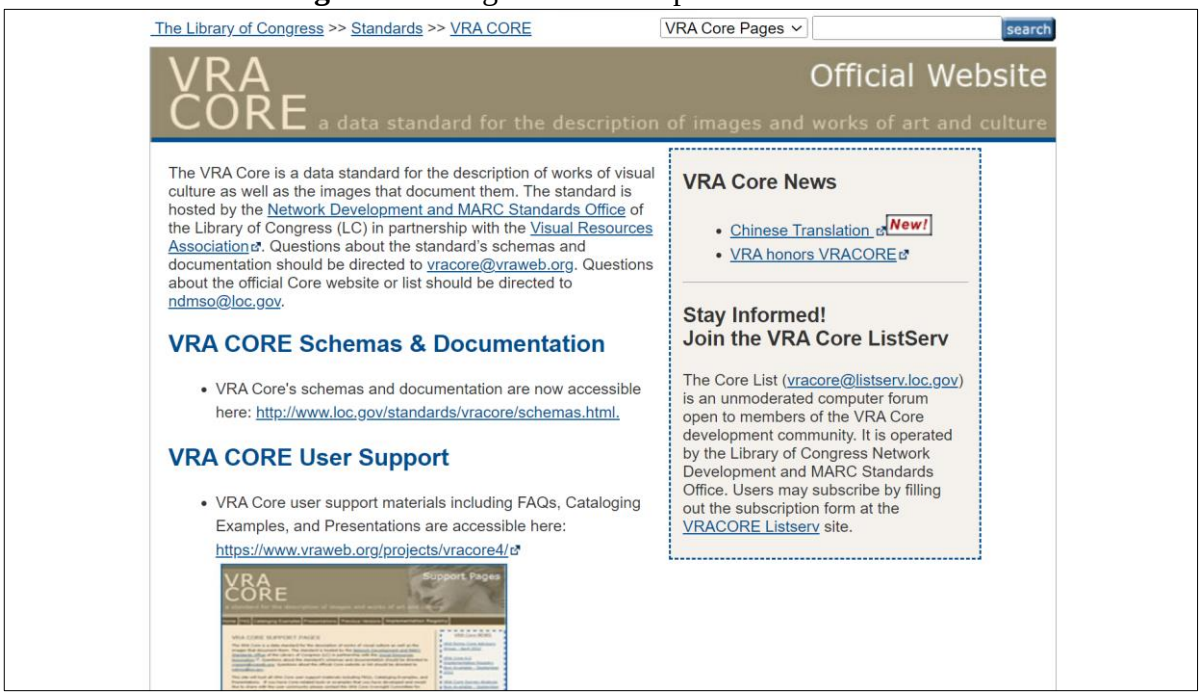
[400] Lançado em 2002, como efeito direto da criação do padrão EAD, o *Online Archive of California* (OAC) permite acesso a guias *online* com descrições detalhadas que facilitam a descoberta de coleções e a localização de objetos individuais. Esses guias são vitais para a compreensão do conteúdo de uma coleção e para determinar se é provável que ela atenda suas necessidades de pesquisa. Disponível em: https://oac.cdlib.org/about/. Acesso em: 30 maio 2023.

[401] O arquivo da *Web* da Universidade da Califórnia no campus de *Irvine* contém *sites* capturados desde 2015 que preservam a presença na *Web* da universidade, a fim de proteger sua memória institucional, prover acesso contínuo a registros de valor histórico e administrativo duradouro e documentar suas atividades acadêmicas etc. Os *sites* são acessados via *Internet Archive*. Disponível em: https://archive-it.org/collections/5613/. Acesso em: 30 maio 2023.

[402] Formalizada em 1982, a VRA é uma organização multidisciplinar dedicada a promover a pesquisa e a educação na área da gerência de imagem e mídia nos meios educacional, cultural e comercial. Esta associação internacional busca prover liderança na área de recursos visuais, desenvolver padrões (o CCO, por exemplo) e ceder ferramentas e oportunidades educacionais à comunidade. Disponível em: http://vraweb.org/about/. Acesso em: 30 maio 2023.

[403] O VRA *Core* foi elaborado e é mantido pelo VRA *Core Oversight Committee*, sendo hospedado pelo *Network Development and* MARC *Standards Office* da Biblioteca do Congresso americano junto com a VRA. No *site* oficial do padrão há descrições de coleções que usam o VRA *Core*, como nas *History of Art Visual Resources Collections* (https://quod.lib.umich.edu/h/hart?page=index) pela Universidade de *Michigan* e nas coleções da Universidade de *Cornell* nos Estados Unidos. Disponível em: http://core.vraweb.org/vracore_registry.html. Acesso em: 30 maio 2023.

Figura 69 – Página inicial do padrão VRA *Core*.



Fonte: *Library of Congress* (2023f).

Atualmente na versão 4.0, lançada em 2007, o esquema VRA *Core* em XML suporta a interoperabilidade e troca de registros. O VRA *Core* 4.0 dispõe dezenove elementos descritivos e nove atributos globais[404], onde o elemento *wrapper* de nível superior – Obra (*work*), Coleção (*collection*) ou Imagem (*image*) – inclui nele os demais dezoito elementos em registros individuais[405] (VISUAL RESOURCES ASSOCIATION, 2007, 2014). Para fins de preservação digital, notamos elementos, como Localização (*location*), Direitos (*rights*) e Fonte (*source*), que podem apoiar o PREMIS e o METS na identificação e definição da fidedignidade, autenticidade integridade, proveniência e contexto de obras culturais originais como de suas representações.

A despeito de sua especificidade e da imposição de certas restrições à criação de *links* para registros não VRA *Core* acrescido ao fato de ser menos comum diante de outros formatos,

[404] No VRA *Core* 4.0 'elementos' são elementos de metadados, tidos equivalentes aos campos numa base de dados; os 'subelementos' também são elementos, relacionados hierarquicamente a elementos; e os 'atributos' qualificam ou relacionam os metadados em diferentes elementos ou subelementos entre si. É construído em torno de três tipos de registro, o qual a Obra é um evento ou objeto único de produção cultural, a Imagem é a representação visual do evento ou objeto, em parte ou no todo; e a Coleção possibilita a catalogação de grupos de materiais, como conjuntos de obras ou de imagens. Cada Obra e Imagem possui sua própria descrição. Esses registros estão relacionados via o elemento Relação (*relation*). Disponível em: http://core.vraweb.org/vracore_faq.html. Acesso em: 30 maio 2023.

[405] Também no *site* oficial do padrão temos registros completos de exemplo para mostrar como a estrutura do VRA *Core* pode ser aplicada a várias combinações de Obra, Imagem ou Coleção em diversas categorias distintas como, por exemplo, pinturas históricas (http://core.vraweb.org/examples/html/example047_full.html) ou obras múltiplas (http://core.vraweb.org/examples/html/example030_full.html); além de artes decorativas e performáticas, filmes e fotografias, dentre outras. Disponível em: http://core.vraweb.org/vracore_examples.html. Acesso em: 30 maio 2023.

como colocado por Eíto-Brun (2015) e Senander III (2013), é possível criar *links* no esquema para instrumentos de pesquisa e o processo de conversão de registros VRA *Core* para registros MARC é bastante simples, direto e eficiente. Quanto aos exemplos de uso do VRA *Core* 4.0 no arquivamento da *Web*, indiretamente temos o arquivo da *Web* da biblioteca da Universidade de *Cornell*[406] que contém *sites* de coleções catalogadas com o VRA Core, como *Mysteries at Eleusi Images of Inscriptions*[407] e *Billie Jean Isbell Andean Collection: Images from the Andes*[408].

4.3.5 Padrão PREMIS

O PREMIS remete o nome de um grupo de trabalho patrocinado pela OCLC e *Research Libraries Group* (RLG) nos Estados Unidos de 2003 a 2005. Esse grupo criou um relatório final em 2005 chamado *Data Dictionary for Preservation Metadata*[409], que define um conjunto básico de unidades semânticas[410], implementável e de larga aplicação, para apoiar a preservação digital em repositórios[411] (CAPLAN, 2017; PREMIS EDITORIAL COMMITTEE, 2015). Conforme Guenther, Dappert e Peyrard (c2016), Lavoie e Gartner (c2013) e Sayão (2010) o dicionário de dados PREMIS não tem como alvo certas classes de metadados, como metadados descritivos e metadados técnicos de formato específico, acordando-se a outros padrões

[406] O arquivo da *Web* da Universidade de *Cornell* coleta e preserva desde 2011 registros de valor histórico, jurídico, fiscal e/ou administrativo para a universidade. Abrange registros e publicações oficiais; documentos audiovisuais; etc. que documentam a história da instituição. Esta coleção *Web* inclui *websites* externos e de domínio *cornell.edu*, acessados via *Internet Archive*. Disponível em: https://archive-it.org/collections/2566. Acesso em: 30 maio 2023.
[407] Disponível em: https://wayback.archive-it.org/2566/*/http://eleusis.library.cornell.edu/. Acesso em: 30 maio 2023.
[408] Disponível em: https://wayback.archive-it.org/2566/*/http://isbellandes.library.cornell.edu/. Acesso em: 30 maio 2023.
[409] PREMIS WORKING GROUP. **Data dictionary for preservation metadata**: final report of the PREMIS working group. Dublin, Ohio: Online Computer Library Center (OCLC); Mountain View, California: Research Libraries Group (RLG), May c2005. Disponível em: https://www.loc.gov/standards/premis/v1/premis-dd_1.0_2005_May.pdf. Acesso em: 30 maio 2023.
[410] Unidade semântica alude a "[...] um pedaço de informação ou conhecimento."; por sua vez, "[...] elemento de metadados é uma forma definida de representar tal informação num registro de metadados, esquema ou banco de dados." (CAPLAN, 2017, p. 5, tradução nossa). Para Dappert e Enders (2010) o primeiro termo são as propriedades que descrevem os objetos digitais e seus contextos, eventos no ciclo de vida, agentes envolvidos na preservação e direitos, e o segundo termo define como implantar o primeiro numa especificação de implementação de metadados.
[411] No *website* oficial do padrão há descrições de projetos que utilizam o PREMIS, como no Repositório de Objetos Digitais Autênticos – RODA (https://demo.roda-community.org/) da *Keep Solutions* em Portugal, no *Système de Préservation et d'Archivage Réparti* – SPAR (https://www.bnf.fr/fr/spar-systeme-de-preservation-et-darchivage-reparti) da BnF e no sistema *Archivematica* (https://www.archivematica.org/pt-br/) da *Artefactual* no Canada. Disponível em: https://www.loc.gov/standards/premis/registry/. Acesso em: 30 maio 2023.

(METS[412], MADS[413] e Z39.87/MIX[414], por exemplo) para cobrir funcionalidades adicionais; ademais, embora muito influenciado pelo OAIS, o PREMIS provê informações que vão além do escopo do repositório.

Figura 70 – Página inicial do padrão PREMIS.

Fonte: *Library of Congress* (2023e).

Hoje na versão 3.0[415], emitida em 2015, o dicionário de dados PREMIS cede orientações para organização e pensamento sobre metadados de preservação. É estruturado sob um modelo

[412] Também no *website* oficial do padrão temos registros de amostra em XML de uso do PREMIS 3.0 com o METS (https://www.loc.gov/standards/premis/v3/sample-records/PREMIS%203%20in%20METS_example%201.xml) ou até sozinho (https://www.loc.gov/standards/premis/v3/sample-records/PREMIS%203%20example%201.xml), por exemplo. Disponível em: https://www.loc.gov/standards/premis/examples.html. Acesso em: 30 maio 2023.

[413] Mantido pelo MODS/MADS *Editorial Committee* com o *Network Development and* MARC *Standards Office* da Biblioteca do Congresso americano e a ajuda dos usuários, o MADS é compatível com o MARC 21 para dados de autoridade e serve como um complemento do MODS para fornecer metadados sobre as entidades autoritativas usadas nas descrições deste padrão. Disponível em: http://www.loc.gov/standards/mads/. Acesso em: 30 maio 2023.

[414] Desenvolvido pelo *Network Development and* MARC *Standards Office* da Biblioteca do Congresso americano em parceria com o NISO *Technical Metadata for Digital Still Images Standards Committee* e *experts* interessados para gerenciar coleções de imagens digitais, os metadados técnicos MIX podem ser usados de forma independente ou como um esquema de extensão com o METS para expressar atributos de imagens fixas digitais, como formato e tamanho do arquivo, resolução etc. Disponível em: http://www.loc.gov/standards/mix/. Acesso em: 30 maio 2023.

[415] O PREMIS *Editorial Committee* coordena as revisões e a implementação do padrão, que consiste no dicionário de dados, um esquema XML e na documentação de suporte hospedados pelo *Network Development and* MARC *Standards Office* da Biblioteca do Congresso americano. Após o lançamento da versão 1.0 do dicionário de dados em maio de 2005, a PREMIS *Maintenance Activity* sucedeu o grupo de trabalho PREMIS e passou a ser incumbida por manter, dar suporte e coordenar as revisões futuras no dicionário de dados PREMIS. Sendo assim, as versões 1.1, 2.0, 2.1, 2.2, 2.3 e, a última, 3.0 do dicionário de dados PREMIS foram lançadas, respectivamente, em 2005, 2008, 2011, 2012, 2014 e 2015. Disponível em: https://www.loc.gov/standards/premis/. Acesso em: 30 maio 2023.

de dados – e implementações, tal como o esquema XML padrão associado[416] – que define quais coisas precisam ser descritas (as entidades) e quais propriedades ou informações necessitam ser conhecidas pelo sistema de preservação para serem ditas sobre elas (as unidades semânticas), relacionando entre si quatro tipos diferentes de entidades: Objetos (*Objects*), Eventos (*Events*), Agentes (*Agents*) e Direitos (*Rights*) (GUENTHER; DAPPERT; PEYRARD, c2016; PREMIS EDITORIAL COMMITTEE, 2015). Para Formenton *et al*. (2017) devido ao PREMIS aplicar o modelo de informação OAIS e os requisitos de preservação de objetos digitais (autenticidade, proveniência etc.), todas as suas entidades/unidades semânticas são vitais à preservação digital.

Embora a falta de treinamento/*expertise* e de integração com o sistema existente possam trazer barreiras à sua adoção em instituições de patrimônio cultural (ALEMNEH; HASTINGS, 2010), o dicionário de dados PREMIS fornece uma estrutura notável para descrever e preservar ambientes computacionais (*hardware*, *software* etc.) que suportam a renderização ou execução dos objetos digitais e a sua utilização em longo prazo. Como exemplo, o PREMIS é adotado na descrição de ambientes de renderização para conteúdos *Web* da BnF, que hospeda o arquivo da *Web* francesa (DAPPERT *et al*., 2013). Outro exemplo de uso do PREMIS no arquivamento da *Web*, incluem Bailey e LaCalle (2015) e Rowell e Krewer (2016) que apresentam a visão do *Internet Archive* sobre como os metadados de preservação PREMIS interage com o formato WARC.

4.3.6 Padrão METS

Criado pela *Digital Library Federation* (DLF)[417], o METS teve como precursor o projeto *Making of America* II (MOA2)[418], de 1997, que forneceu um formato de documento XML para codificar metadados descritivos, administrativos e estruturais para obras textuais e baseadas em imagens. Expresso em XML, o METS possibilita codificar os metadados

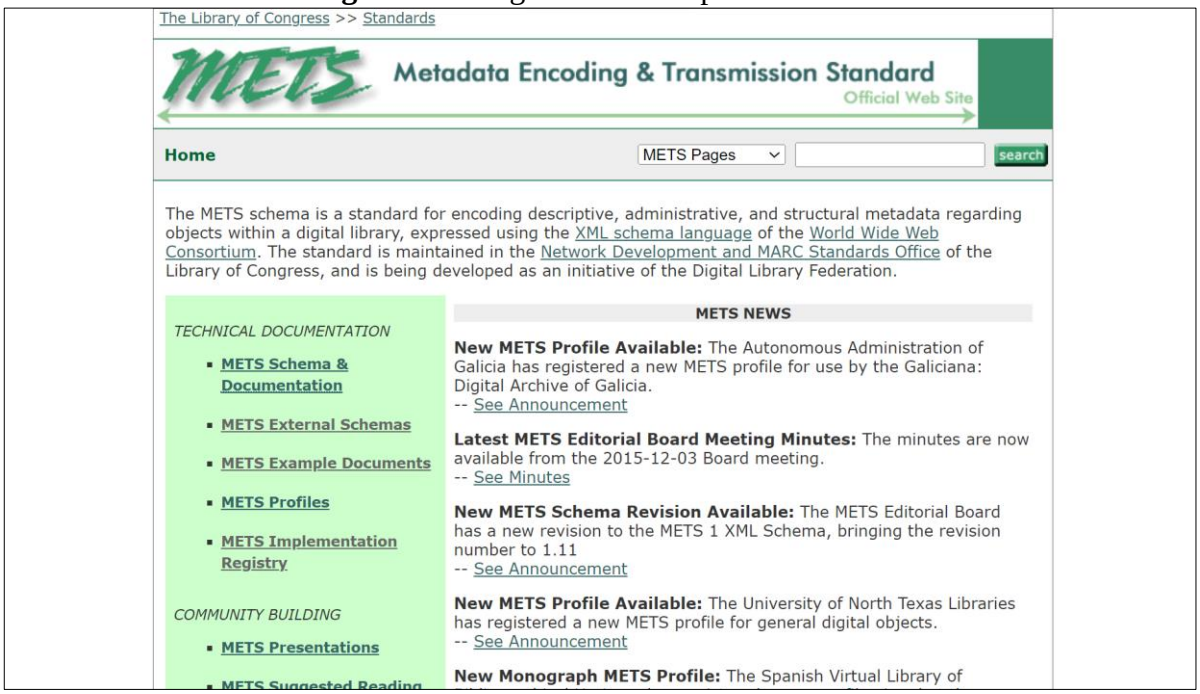
[416] Patrocinada pela Biblioteca do Congresso americano, a PREMIS *Maintenance Activity* cede um esquema XML que, para Caplan (2017), condiz diretamente ao dicionário de dados, permitindo a descrição de Objetos, Eventos, Agentes e Direitos como o uso do PREMIS representado em XML para a troca de metadados entre sistemas de preservação. Disponível em: https://www.loc.gov/standards/premis/v3/premis-v3-0.xsd. Acesso em: 30 maio 2023.

[417] Formada em 1995, a DLF é uma comunidade servida ao *Council on Library and Information Resources* (CLIR) que reúne instituições e profissionais dedicados a pesquisa, aprendizado, justiça social e bem público via *design* criativo e uso inteligente de tecnologias de bibliotecas digitais, cujo parte dos membros iniciaram os esforços do MOA2 DTD e do formato sucessor METS. Disponível em: https://www.diglib.org/about/. Acesso em: 30 maio 2023.

[418] HURLEY, Bernard J. *et al*. **The Making of America II Testbed Project**: a digital library service model. Washington, DC: Digital Library Federation (DLF): Council on Library and Information Resources (CLIR), Dec. 1999. 36 p. Disponível em: https://clir.wordpress.clir.org/wp-content/uploads/sites/6/pub87.pdf. Acesso em: 30 maio 2023.

necessários na gestão de objetos de bibliotecas digitais em um repositório e na troca destes objetos entre repositórios (ou entre repositórios e seus usuários)[419] (LIBRARY OF CONGRESS, 2017). Segundo Cantara (2005), Mcdonough (2006) e Sayão (2010) o METS é um mecanismo flexível para organizar todos os metadados associados ao objeto digital, exprimir as ligações complexas entre múltiplas classes de metadados e, adicionalmente, associar um objeto com comportamentos ou serviços, oferecendo suporte à interoperabilidade, escalabilidade[420] e preservação digital de longo prazo.

Figura 71 – Página inicial do padrão METS.



Fonte: *Library of Congress* (2023c).

Atualmente na versão 1.12.1[421], de 2019, o esquema METS dirige-se a codificar objetos complexos em bibliotecas digitais. Para compartilhar documentos XML consoante ao METS e

[419] No *website* oficial do padrão encontram-se descrições de projetos que adotam o METS, como no repositório de preservação digital Merritt (https://merritt.cdlib.org/) mantido pelo UC3 nos Estados Unidos; na Biblioteca Nacional Digital (http://bndigital.bnportugal.gov.pt/) da Biblioteca Nacional de Portugal; e no sistema de gerência de coleções digitais CONTENTdm (https://www.oclc.org/en/contentdm.html) da OCLC. Disponível em: https://www.loc.gov/standards/mets/mets-registry.html. Acesso em: 30 maio 2023.

[420] A escalabilidade "[...] é um atributo desejável de uma rede, sistema ou processo. O conceito denota a capacidade de um sistema de acomodar um número cada vez maior de elementos ou objetos, processar volumes crescentes de trabalho normalmente e/ou ser suscetível ao alargamento." (BONDI, 2000, p. 195, tradução nossa).

[421] O METS *Editorial Board* mantém o controle editorial do padrão, que integra o esquema XML, o esquema XML do perfil METS e a documentação oficial do METS mantidos pelo *Network Development and* MARC *Standards Office* da Biblioteca do Congresso americano. Após o lançamento da versão 1.0 em fevereiro de 2001, o esquema METS recebeu o Registro NISO em 2004 (com renovação em 2006) além de um conjunto de revisões. Portanto, as versões 1.1 e 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 1.9.1, 1.10, 1.10.1 e 1.11, 1.12 e, a última, 1.12.1 do esquema METS foram lançadas, respectivamente, nos anos de 2002, 2003, 2004, 2005, 2006, 2007, 2009, 2010, 2012, 2013, 2015, 2018 e 2019. Disponível em: https://www.loc.gov/standards/mets/mets-home.html. Acesso em: 30 maio 2023.

estabelecer práticas comuns, o padrão define os componentes de um perfil METS e o esquema XML para codificá-lo. Estes perfis descrevem em detalhes uma classe de documentos METS[422] para criar e processar documentos METS de acordo com um perfil específico[423], sendo que um documento METS constitui sete seções principais: Cabeçalho METS (<*metsHdr*>), Metadados descritivos (<*dmdSec*>), Metadados administrativos (<*amdSec*>), Arquivo (<*fileSec*>), Mapa estrutural (<*structMap*>), Links estruturais (<*structLink*>*)* e Comportamento (<*behaviorSec*>) (DIGITAL LIBRARY FEDERATION, 2010). Na preservação digital, Formenton *et al*. (2017) salientam que um documento METS pode atuar na execução dos pacotes de informação OAIS.

Apesar da imperfeita correlação entre METS e PREMIS (em especial, para informação sobre Agente) e de haver alguma sobreposição no uso das duas estruturas juntas que pode causar variações na representação dos dados (CAPLAN, 2017), os perfis METS registrados[424] mitigam em certa medida, como citado por Lavoie e Gartner (c2013), os problemas de interoperabilidade relativos à flexibilidade no METS. Por exemplo, o perfil METS para capturas de *sites* do projeto ECHO *DEPository*[425] na Universidade de *Illinois* em *Urbana-Champaign* (HABING, 2006) visa a transferência e preservação digital do conteúdo de captura da *Web* entre repositórios[426]. Outro exemplo dentro do arquivamento da *Web*, incluem Truman (2016) e Veikkolainen e Lager (2016) que expõem o arquivo *Web* finlandês[427] mantido pela Biblioteca Nacional da

---

[422] Documento METS refere-se "[...] ao documento XML elaborado em conformidade com o esquema METS."; em contraste, objeto METS trata-se de "[...] todo o artefato digital representado pelo documento METS, incluindo qualquer conteúdo ou metadados referenciados externamente necessários para constituir um objeto completo." (DIGITAL LIBRARY FEDERATION, 2010, p. 25, tradução nossa).

[423] Também no *site* oficial do padrão existem exemplos de documentos METS com MODS, MIX ou PREMIS em XML para registro bibliográfico (https://memory.loc.gov/diglib/ihas/loc.afc.afc9999005.1153/mets.xml), imagem com vídeo (https://digitalassets.lib.berkeley.edu/jpnprints/ucm/mets/caawucm_7_1_00027801.xml) ou coleções de manuscritos (http://fedora.library.northwestern.edu/fedora/get/inu:inu-afrmap-4333843/source), além de áudio, vídeo, mapa etc. Disponível em: https://www.loc.gov/standards/mets/mets-examples.html. Acesso em: 9 jul. 2020.

[424] *Registered Profiles*. Disponível em: https://www.loc.gov/standards/mets/mets-registered-profiles.html. Acesso em: 30 maio 2023.

[425] O ECHO DEPository, acrônimo de *Exploring Collaborations to Harness Objects in a Digital Environment for Preservation*, foi um projeto de sete anos (2004-2010) financiado pelo *National Digital Information Infrastructure and Preservation Program* (NDIIPP) da Biblioteca do Congresso americano. Liderado pela Universidade de *Illinois* em *Urbana-Champaign*, o projeto explorou maneiras de compartilhamento e preservação de informações digitais, como gravações de vídeo, publicações do governo na *Web*, *sites* etc., no qual os parceiros colaboraram para criar ferramentas, práticas, avaliações e pesquisas que auxiliarão na seleção e preservação de recursos eletrônicos em repositórios. Disponível em: http://www.digitalpreservation.gov/partners/echodep.html. Acesso em: 30 maio 2023.

[426] Este perfil deve estar em acordo com as regras do perfil pai ECHO *Dep Generic* METS *Profile for Preservation and Digital Repository Interoperability* (https://www.loc.gov/standards/mets/profiles/00000015.html) que possui como foco principal permitir a interoperabilidade do repositório e a preservação digital do conteúdo do repositório.

[427] O arquivo da *Web* finlandês tem uma amostra diversa e representativa de conteúdo *online* desde 2006 (incluindo páginas HTML, vídeos *Youtube*, *Twitter*, arquivos jpg, mp4, doc etc.), que pode acessado e usado nas instalações de bibliotecas gratuitas de depósito legal, como Biblioteca Nacional da Finlândia e certas bibliotecas universitárias, baseado nas leis de materiais culturais e de direitos autorais de seu país. Este arquivo é acrescido anualmente por coletas automáticas de domínios finlandeses (*site*s de domínio *.fi* e *ax*.) e esforços de coleta focados na Finlândia, jornais, notícias e temas. Disponível em: https://verkkoarkisto.kansalliskirjasto.fi/va/info. Acesso em: 1 mar. 2021.

Finlândia, onde o conteúdo compreende arquivos em formato WARC nos pacotes de informação METS.

4.4 Análise dos padrões de metadados à luz da preservação digital no arquivamento da *Web*

Os padrões e esquemas de metadados DC, MODS, EAD, VRA *Core*, PREMIS e METS possuem características em comum e algumas singularidades. As ponderações realizadas aqui basearam-se nos princípios da preservação digital de longo prazo, nas definições do modelo de informação OAIS, nas informações expressas pelos metadados de preservação e nos metadados descritivos e administrativos para o arquivamento da *Web*, sobretudo, com os elementos WAN de Dooley e Bowers (c2018), os elementos de Kim e Lee (2007) e os metadados de Banos *et al*. (c2013) e de iniciativas na área (incluindo aqueles exibidos em registros de coleções de arquivos da *Web*, como das bibliotecas da Universidade de *Cornell* e do Congresso americano), os quais foram descritos no trabalho. De forma não exaustiva, o Quadro 2 sintetiza os aspectos básicos dos padrões aludidos e os elementos de metadados (ou as unidades semânticas para o caso do PREMIS) considerados para esta pesquisa como importantes na preservação de conteúdos *Web*.

Quadro 2 – Padrões e elementos de metadados de apoio à preservação digital no arquivamento da *Web*.

(continua)

Padrão	Características	Elementos de metadados úteis para a preservação digital no arquivamento da *Web*	
DC qualificado (versão 1.1)	- Esquema XML ou em outras sintaxes tido como flexível, extensível, simples e interoperável; e - Aplicável à descoberta de recursos na *Web* e para o arquivamento da *Web* devido ao seu uso geral pelos usuários do *Archive-It*.	▪ Título (*title*) ▪ Criador (*creator*) ▪ Assunto (*subject*) ▪ Descrição (*description*) ▪ Colaborador (*contributor*) ▪ Data (*date*) ▪ Tipo (*type*) ▪ Formato (*format*)	▪ Identificador (*identifier*) ▪ Fonte (*source*) ▪ Língua (*language*) ▪ Relação (*relation*) ▪ Cobertura (*coverage*) ▪ Direitos (*rights*) ▪ Detentor de Direitos (*rightsHolder*) ▪ Proveniência (*provenance*)
MODS (versão 3.7)	- Esquema XML derivado do MARC 21 considerado mais rico que o DC e mais simples que o MARC completo; - Pode ser usado junto com o MADS ou, ainda, como um esquema de extensão para o METS; e - Aplicável aos objetos de bibliotecas digitais e para *sites* arquivados, tais como os das coleções do arquivo da *Web* da Biblioteca do Congresso americano.	▪ Informação de Título (<*titleInfo*>) ▪ Nome (<*name*>) ▪ Tipo de Recurso (<*typeOfResource*>) ▪ Gênero (<*genre*>) ▪ Informação de Origem (<*originInfo*>) ▪ Língua (<*language*>) ▪ Descrição Física (<*physicalDescription*>) ▪ Resumo (<*abstract*>) ▪ Índice (<*tableOfContents*>) ▪ Nota (<*note*>) ▪ Assunto (<*subject*>)	▪ Item Relacionado (<*relatedItem*>) ▪ Identificador (<*identifier*>) ▪ Localização (<*location*>) ▪ Condição de Acesso (<*accessCondition*>) ▪ Parte (<*part*>) ▪ Extensão (<*extension*>) ▪ Informação de Registro (<*recordInfo*>)
EAD3 (versão 1.1.1)	- Esquema XML e DTD minucioso que é compatível com a norma de descrição arquivística ISAD(G); - Pode ser usado junto com o EAC-CPF e inclui tanto a indicação de elementos correspondentes no MARC, no MODS, na ISAD(G) e na HTML como mapeamentos (*crosswalks*) para o MARC 21, o MODS e a ISAD(G); e - Aplicável à codificação de instrumentos de pesquisa de arquivos como, por exemplo, os instrumentos de pesquisa do Arquivo *Online* da California que cedem descrições detalhadas das coleções do arquivo da *Web* da Universidade da Califórnia em *Irvine* nos Estados Unidos.	▪ Título da Unidade (<*unittitle*>) ▪ Origem (<*origination*>) ▪ Nome Pessoal (<*persname*>) ▪ Nome da Organização (<*corpname*>) ▪ Nome de Família (<*famname*>) ▪ Cabeçalhos de Acesso Controlado (<*controlaccess*>) ▪ Assunto (<*subject*>) ▪ Resumo (<*abstract*>) ▪ Acréscimos (<*accruals*>) ▪ Informação de Aquisição (<*acqinfo*>) ▪ Biografia ou História (<*bioghist*>) ▪ Escopo e Conteúdo (<*scopecontent*>) ▪ Histórico de Custódia (<*custodhist*>) ▪ Nota Descritiva de Identificação (<*didnote*> ▪ Outros Dados Descritivos (<*odd*>)	▪ Gênero/Característica Física (<*genreform*>) ▪ Descrição Física (<*physdesc*>) ▪ Objeto de Arquivo Digital (<*dao*>) ▪ Identificação da Unidade (<*unitid*>) ▪ Língua do Material (<*langmaterial*>) ▪ Língua (<*language*>) ▪ Físico Estruturado (<*physdescstructured*>) ▪ Material Relacionado (<*relatedmaterial*>) ▪ Condições de Acesso (<*accessrestrict*>) ▪ Condições de Uso (<*userestrict*>) ▪ Localização Física (<*physloc*>) ▪ Informação de Processo (<*processinfo*>) ▪ Repositório (<*repository*>) ▪ Agência de Manutenção (<*maintenanceagency*>) ▪ Histórico de Manutenção (<*maintenancehistory*>)

Quadro 2 – Padrões e elementos de metadados de apoio à preservação digital no arquivamento da *Web*.

(continuação)

Padrão	Características	Elementos de metadados úteis para a preservação digital no arquivamento da *Web*	
		▪ Data da Unidade (<*unitdate*>)	
VRA *Core* (versão 4.0)	- Esquema XML simples que pode ser usado junto com o CCO e inclui a indicação de elementos equivalentes no CDWA, no DC e no CCO; e - Aplicável à descrição de obras culturais originais e suas reproduções, como certas coleções da biblioteca da Universidade de *Cornell* que têm um arquivo da *Web* com os *sites* destas.	▪ Obra, Coleção ou Imagem (<*work*>, <*collection*>, <*image*>) ▪ Agente (<*agent*>) ▪ Contexto Cultural (<*culturalContext*>) ▪ Data (<*date*>) ▪ Descrição (<*description*>) ▪ Inscrição (<*inscription*>) ▪ Localização (<*location*>) ▪ Material (<*material*>) ▪ Medidas (<*measurements*>)	▪ Relação (<*relation*>) ▪ Direitos (<*rights*>) ▪ Fonte (<*source*>) ▪ Estado/Edição (<*stateEdition*>) ▪ Período/Estilo (<*stylePeriod*>) ▪ Assunto (<*subject*>) ▪ Técnica (<*technique*>) ▪ Referência Textual (<*textref*>) ▪ Título (<*title*>) ▪ Tipo de Obra (<*worktype*>)
PREMIS (versão 3.0)	- Esquema XML que enfoca o repositório de preservação e sua gestão; - Pode unir-se a outros padrões, como MODS, DC, EAD, METS etc., para cobrir metadados fora do seu escopo e funções adicionais; e - Aplicável ao apoio da preservação de objetos digitais, tal como na descrição de ambientes de renderização para conteúdos da *Web*.	▪ Identificador do objeto (*objectIdentifier*) ▪ Categoria de objeto (*objectCategory*) ▪ Nível de preservação (*preservationLevel*) ▪ Propriedades significativas (*significantProperties*) ▪ Características do objeto (*objectCharacteristics*) ▪ Nome original (*originalName*) ▪ Armazenamento (*storage*) ▪ Informação de assinatura (*signatureInformation*) ▪ Função do ambiente (*environmentFunction*) ▪ Designação do ambiente (*environmentDesignation*) ▪ Registro do ambiente (*environmentRegistry*) ▪ Extensão de ambiente (*environmentExtension*) ▪ Relacionamento (*Relationship*) ▪ Identificador de evento de vinculação (*linkingEventIdentifier*) ▪ Identificador de declaração de direitos de vinculação (*linkingRightsStatementIdentifier*)	▪ Informação detalhada do evento (*eventDetailInformation*) ▪ Informação do resultado do evento (*eventOutcomeInformation*) ▪ Identificador do agente de vinculação (*linkingAgentIdentifier*) ▪ Identificador de objeto de vinculação (*linkingObjectIdentifier*) ▪ Identificador do agente (*agentIdentifier*) ▪ Nome do agente (*agentName*) ▪ Tipo de agente (*agentType*) ▪ Versão do agente (*agentVersion*) ▪ Nota de agente (*agentNote*) ▪ Extensão de agente (*agentExtension*) ▪ Identificador do evento de vinculação (*linkingEventIdentifier*) ▪ Identificador de declaração de direitos de vinculação (*linkingRightsStatementIdentifier*)

Quadro 2 – Padrões e elementos de metadados de apoio à preservação digital no arquivamento da *Web*.

(conclusão)

Padrão	Características	Elementos de metadados úteis para a preservação digital no arquivamento da *Web*	
		▪ Identificador do evento (*eventIdentifier*) ▪ Tipo de evento (*eventType*) ▪ Data e hora do evento (*eventDateTime*)	▪ Identificador de ambiente de vinculação (*linkingEnvironmentIdentifier*) ▪ Declaração de direitos (*rightsStatement*) ▪ Extensão de direitos (*rightsExtension*)
METS (versão 1.12.1)	- Esquema XML flexível que organiza e vincula várias formas de metadados aos objetos digitais num sistema; - Pode estruturar os pacotes PSI, PAI ou PDI do modelo de informação OAIS e incluir outros padrões na seção de Metadados Descritivos, como DC, EAD etc., e ter o PREMIS na seção de Metadados Administrativos; e - Aplicável à transferência e preservação digital do conteúdo de captura da *Web* (*websites*) entre repositórios através do perfil METS do projeto ECHO *DEPository*.	▪ Agente (<*agent*>) ▪ Identificador Alternativo (<*altRecordID*>) ▪ Referência de Metadados (<*mdRef*>) ▪ Invólucro (*wrapper*) de Metadados (<*mdWrap*>) ▪ Metadados Técnicos (<*techMD*>) ▪ Metadados de Direitos de Propriedade Intelectual (<*rightsMD*>) ▪ Metadados de Fonte (<*sourceMD*>) ▪ Metadados de Proveniência Digital (<*digiprovMD*>) ▪ Grupo de Arquivo (<*fileGrp*>) ▪ Arquivo (<*file*>)	▪ Localização de Arquivo (<*FLocat*>) ▪ Conteúdo de Arquivo (<*FContent*>) ▪ Fluxo de Bytes (*byte stream*) de Componente (<*stream*>) ▪ Arquivo de Transformação (<*transformFile*>) ▪ Divisão (<*div*>) ▪ Indicador (pointer) de Arquivo (<*fptr*>) ▪ Indicador (pointer) METS (<*mptr*>) ▪ Link do Mapa Estrutural (<*smLink*>) ▪ Comportamento (<*behavior*>) ▪ Definição de Interface (<*interfaceDef*>) ▪ Mecanismo (<*mechanism*>)

Fonte: Elaborado pelo autor.

Primeiramente, é importante ressaltar que todos os padrões de metadados analisados no Quadro 2 são expressos na sintaxe XML e, em certo ponto, são flexíveis e/ou extensíveis. Como padrão aberto e legível para computadores e humanos, o formato XML atende as necessidades de preservação digital e permite a descrição e o intercâmbio de diversos tipos de dados na *Web* e em outros ambientes; ademais, facilita a integração e o uso combinado de diferentes esquemas baseados nesta mesma linguagem de marcação, como os esquemas externos para uso conjunto com o METS[428] que incluem o EAC-CPF, MARC, MIX etc. Aliás, há a DCMI, no caso do DC, e o *Network Development and* MARC *Standards Office* da Biblioteca do Congresso americano, para os outros padrões analisados (salvo o EAD e o VRA *Core* que são mantidos também, nessa

[428] *External schemas for use with METS*. Disponível em: https://www.loc.gov/standards/mets/mets-extenders.html. Acesso em: 30 maio 2023.

ordem, pelo TS-EAS da SAA e pelo VRA *Core Oversight Committee*), em que uniformizam a descrição e a representação das informações mediante esquemas para a codificação de valores.

O uso conjunto de vários padrões de metadados, impelido por indicações de metadados externos, elementos equivalentes/correspondentes e mapeamentos (*crosswalks*) ou pela adoção comum de sintaxes, normas e vocabulários que retratam a flexibilidade e a extensibilidade dos esquemas e aumentam a interoperabilidade de dados, se faz aceitável visto a alta complexidade dos tipos de recursos a serem descritos e as diversas etapas dos processos de preservação digital de longo prazo e de arquivamento da *Web*. Baseando-se na ambiguidade do escopo dos metadados de preservação, inferimos ainda que todas as classes de metadados – descritivos, linguagens de marcação etc. – são vitais ao alcance da preservação de conteúdos *Web*. Por hora, é verossímil que não temos como fixar qual é o único padrão que garanta plenamente a preservação digital, mas os padrões existentes podem se completarem para documentar as informações exigidas na gerência da preservação e do acesso utilizável de objetos digitais complexos, como os *websites*.

Entre os metadados assíduos nos padrões analisados estão os identificadores, que podem ser contidos na unidade *objectIdentifier* PREMIS e nos elementos Identificador e Relação DC; Item Relacionado, Identificador e Localização MODS; Material Relacionado, Identificação da Unidade e Objeto de Arquivo Digital EAD; Referência Textual VRA; Localização de Arquivo, Referência de Metadados, Definição de Interface, Mecanismo e Indicador METS. Julgando as relações de um único *site* que está sendo descrito com qualquer coleção a qual pertença ou com demais recursos, os identificadores provêm a identificação exclusiva e distintiva do recurso que os metadados se referem como a sua localização eletrônica. Logo, o registro de um URL do *site* arquivado (de acesso, captura etc.) e do URL para o recurso relacionado refletem os metadados definidos na Informação de Descrição de Preservação (sobretudo, referência e contexto) OAIS e os princípios da preservação digital de manter o contexto e de identificar e localizar os objetos.

De fato, o DC, o MODS, o EAD e o VRA *Core* são mais cabíveis à descrição de recursos digitais para fins de sua descoberta, recuperação, apresentação e interoperabilidade. Mesmo que os escopos destes padrões de metadados encontrem-se inerentemente voltados à etapa de acesso ao invés de exatamente para preservação por longo prazo, alguns dos seus elementos descritivos são úteis em apoiar os metadados de preservação PREMIS. Assim, as informações cedidas por eles permeiam aspectos de representação e de preservação, como características e dependências técnicas, cadeia de custódia ou propriedade, alterações feitas, procedência, relações estruturais físicas e lógicas, direitos etc., os quais são pertinentes na gerência de objetos digitais arquivados e que, em algum grau, traduzem parte dos contornos dos metadados de preservação OAIS e os

princípios da preservação digital de garantir a fidedignidade, a autenticidade e a integridade dos objetos e de manter o contexto, a proveniência e a recuperação dos mesmos ao longo do tempo.

Sob a luz da preservação digital no arquivamento da *Web* e pautando-se nos exemplos de descrições de conteúdos *Web* arquivados de Dappert *et al*. (2013), *Digital Library Federation* (2010), Dooley e Bowers (c2018), Habing (2006) e *Library of Congress* (2018a), distribuímos os elementos de metadados indicados no Quadro 2 dos padrões DC, MODS, EAD e VRA *Core* segundo as informações que eles poderão registrar para *websites* e coleções de *sites* arquivados:

- Título, Criador, Assunto, Colaborador e Língua DC; Informação de Título, Localização, Nome, Língua, Índice, Assunto e Parte MODS; Título da Unidade, Origem, Objeto de Arquivo Digital, Nome Pessoal, Nome da Organização, Nome de Família, Cabeçalhos de Acesso Controlado, Assunto, Língua do Material, Língua e Repositório EAD; e Período/Estilo, Agente, Localização, Assunto e Título VRA; que exprimem o nome atribuído ao recurso descrito, o tópico, o idioma, a pessoa/organização responsável por criar o seu conteúdo intelectual ou fazer contribuições a ele e a instituição/repositório que detém o recurso. Por exemplo, o nome e o idioma do *site* ou coleção arquivada, a entidade que criou seu conteúdo ou fez contribuições secundárias e a instituição responsável pela sua seleção, curadoria ou gestão, além dos assuntos temáticos, nomes de lugares geográficos e nomes de entidades usados para o tópico principal que descreve o conteúdo arquivado ou *site*.
- Descrição, Data e Cobertura DC; Informação de Origem, Resumo e Nota MODS; Data da Unidade, Nota Descritiva de Identificação, Outros Dados Descritivos, Informação de Processo, Resumo, Acréscimos, Biografia ou História, Escopo e Conteúdo, Histórico de Custódia e Informação de Aquisição EAD; e Contexto Cultural, Localização, Descrição, Data e Período/Estilo VRA; que apontam um relato do conteúdo, escopo e contexto do recurso descrito, um período de tempo ligado a um evento no ciclo de vida do recurso, e a cadeia de custódia, a procedência e o tópico ou âmbito espaço-temporal do recurso. Por exemplo, a indicação de datas de direitos autorais ou de quando o *site* foi iniciado/se fez inativo, foi/começou a ser arquivado e o URL foi capturado (com a sua frequência), além de proveniência (se um *site* faz parte de uma coleção temática mais ampla etc.) e de um decreto legal ou outra razão para selecionar o *site* ou conteúdo para arquivamento.
- Fonte e Proveniência DC; Informação de Origem MODS; Data da Unidade EAD; e Fonte VRA; que expressam uma referência à fonte das informações registradas sobre o recurso descrito como a um outro recurso de que ele é derivado, as mudanças na custódia

e propriedade do recurso desde a sua criação; e a origem do recurso, incluindo local de origem/publicação, publicador/originador e datas associadas como, por exemplo, a data e o local em que o *site* arquivado foi criado, publicado ou emitido e a data de sua captura.

- Relação DC; Item Relacionado MODS; Material Relacionado EAD; e Relação VRA; que informam uma referência para outro recurso relacionado ao recurso que está sendo descrito. Por exemplo, as relações todo/parte entre um único *site* arquivado e qualquer coleção de *sites* arquivados a qual pertença (com a inclusão do seu título), entre um *site* arquivado e uma coleção de arquivos analógicos ou quaisquer outros materiais digitais, como as páginas da *Web* constituintes do *site* e as imagens e vídeos que integram o *site*.
- Direitos e Detentor de Direitos DC; Condição de Acesso MODS; Condições de Acesso e Condições de Uso EAD; e Direitos VRA; que registram os direitos do recurso descrito, a pessoa/organização que tem ou administra tais direitos, as restrições (ou a sua falta) e as condições que afetam o acesso, a renderização e uso do recurso. A título de exemplo, a indicação para uso no local e de um período que o conteúdo arquivado ou *site* é restrito, se o acesso ao conteúdo está aberto e os titulares de direitos permitem reuso após acesso.
- Tipo e Formato DC; Tipo de Recurso, Gênero e Descrição Física MODS; Cabeçalhos de Acesso Controlado, Gênero/Característica Física, Físico Estruturado e Descrição Física EAD; e Obra, Coleção ou Imagem, Material, Medidas, Tipo de Obra e Técnica VRA; que expõem a natureza do recurso descrito e o seu formato, dimensões, técnica e estilo. Por exemplo, a indicação se o conteúdo arquivado é *website*, arquivos *Web*, mídia social etc. e que se trata de uma coleção com um número específico de *sites* arquivados.
- Extensão e Informação de Registro MODS; Localização Física, Agência de Manutenção e Histórico de Manutenção EAD; e Inscrição e Estado/Edição VRA; que documentam informações sobre o recurso com o uso de mais de um esquema, a localização física do recurso e a instituição/serviço responsável por sua criação, manutenção e divulgação, a identificação da edição do recurso e o seu histórico de criação, revisões, atualizações e outras alterações, além de informações para a gerência e a interpretação do registro de metadados como, por exemplo, a origem ou proveniência do registro do *site* ou coleção arquivada (gerado por máquina ou não etc.) e o seu idioma, data em que foi criado pela primeira vez, organização que o criou ou alterou sua versão original e regras usadas para o conteúdo da descrição (ou seja, vocabulários controlados, normas de catalogação etc.).

Apesar de estar fora dos seus propósitos, os metadados técnicos MIX, *Text*MD, MPEG-7, *Audio*MD e *Video*MD e os dados de autoridade MADS e EAC-CPF podem ainda ser usados

com o PREMIS junto ao METS[429] no amparo ao registro das circunstâncias de criação (data de criação, nome do dispositivo de criação etc.), do histórico de mudanças feitas (documentadas, autorizadas etc.), das características e dependências técnicas (tamanho, *hardware* etc.) e demais aspectos de materiais audiovisuais, de texto e em mais formatos integrados nos *sites* arquivados, bem como ao registro de dados sobre agentes com funções na criação e contribuição, na seleção, curadoria ou gestão, nos direitos, na renderização e nas ações que afetam esses materiais. Sendo assim, estes padrões apoiam a interoperabilidade, a gerência e a preservação de objetos digitais complexos, como *sites* que incluem vários formatos e tipos de conteúdo, devendo ponderá-los para a definição e a validação da procedência, autenticidade e integridade dos seus conteúdos.

Por sua vez, o PREMIS retrata o uso prático dos conceitos de metadados de preservação delineados no modelo de informação OAIS e, em seguimento, reflete os requisitos e princípios da preservação digital, o que faz com que todas as suas unidades semânticas sejam importantes na preservação a longo prazo de *sites* arquivados. Por isso, inspirando-se em Guenther, Dappert e Peyrard (c2016) que ilustram relações entre as unidades semânticas do dicionário de dados PREMIS e as categorias de informação OAIS, salientamos as unidades *significantProperties*, *environmentFunction*/*Designation*/*Registry*/*Extension* e *relationship* (informação de contexto e proveniência, informações estruturais e outras representações OAIS) que podem detalhar, por exemplo, que apenas o conteúdo precisa ser mantido para uma página *Web* contendo animações que não foram tidas vitais; o ambiente que suporta a renderização e a execução de um *site*; e as relações envolvendo ambiente técnico e relações estruturais entre partes integrantes de um *site*.

Finalmente, num repositório OAIS, o METS serve como um esquema central na gestão de *websites* arquivados e na transferência destes objetos entre sistemas (ou entre sistemas e seus usuários), incluindo o DC, MODS, EAD, VRA *Core*, MARC XML etc. na seção de Metadados Descritivos como o PREMIS, MIX, *Text*MD, *Audio*MD/*Video*MD etc. na seção de Metadados Administrativos do documento METS. Na segunda seção, as unidades PREMIS (por exemplo, *eventIdentifier*/*Type*/*DateTime* ou *eventDetailInformation*/*OutcomeInformation*) registram no elemento Metadados de Proveniência Digital (<*digiprovMD*>) quaisquer ações relacionadas à preservação realizadas nos vários arquivos que compõem um *site* ou quais modificações foram feitas a um objeto digital (*website*) e/ou suas partes constituintes durante seu ciclo de vida que, segundo *Digital Library Federation* (2010), podem ser usadas para julgar como esses processos alteraram ou corromperam a capacidade do objeto de representar com precisão o item original.

[429] Disponível em: https://www.loc.gov/standards/premis/premis-mets.html. Acesso em: 30 maio 2023.

Assim, os metadados descritivos e administrativos (<*mdRef*> e <*mdWrap*>; <*techMD*>, <*rightsMD*>, <*sourceMD*> e <*digiprovMD*>) podem ser externos ao documento METS, sendo que estes últimos registram relações de original/derivado entre arquivos, como arquivos foram criados e armazenados etc. Úteis aos requisitos da preservação digital, as seções de Cabeçalho METS e Arquivo (<*agent*> e <*altRecordID*>; <*transformFile*>, <*fileGrp*>, <*file*>, <*FLocat*>, <*FContent*> e <*stream*>) inclui metadados sobre o documento METS em si e lista (por formato etc.) arquivos que formam o conteúdo de *sites*. Outras seções são Comportamento (<*behavior*>, <*interfaceDef*> e <*mechanism*>) para renderizar ou exibir o *site*; e Mapa Estrutural e Ligações Estruturais (<*div*>, <*fptr*> e <*mptr*>; <*smLink*>) que ordenam os *hiperlinks* entre arquivos que compõem os objetos ou entre outros objetos, como uma página *Web* com imagem hiper ligada a outra página, registrando a estrutura de hipertexto dos *sites* arquivados separada dos arquivos HTML do próprio *site* e que pode ser mostrada aos usuários para sua compreensão e navegação do conteúdo (DIGITAL LIBRARY FEDERATION, 2010; LIBRARY OF CONGRESS, 2017).

Na prática, os principais problemas da preservação digital derivam das particularidades dos objetos que pretende manter o acesso, a recuperação e o uso no decorrer do tempo. Um dos exemplos de objetos digitais complexos são os *sites* que contém tanto uma ampla gama de *links* de hipertexto para permitir a navegação de uma página da *Web* para outra como vários arquivos e formatos com alta dependência de tecnologias para o seu acesso, interpretação, renderização – ou execução etc. – e uso que se tornam obsoletas com o tempo; aliás, estão sujeitos à dinâmica e efemeridade da *Web* onde os seus conteúdos são criados e publicados e, por isto, são perdidos ou sofrem rapidamente alterações em sua forma original. Logo, essas facetas obrigam a refletir as questões de autenticidade, integridade e contexto dos *websites* arquivados e, ainda, a elucidar as distinções entre *sites* ao vivo e suas versões fixas arquivadas como a utilidade de abordagens mistas de descrição para um único *site* ou uma coleção arquivada devido a sua heterogeneidade.

Como um dos aspectos sobre a garantia do processo de preservação digital, a adoção de metadados para a preservação por longo prazo auxilia nas tomadas de decisões e no controle de requisitos legais, de versões, da continuidade de acesso, uso e interpretação e de outras questões atreladas ao arquivamento de objetos informacionais em sistemas. Os esquemas de metadados podem prover a interoperabilidade dos objetos digitais entre repositórios ou serviços, incentivar o uso comum de vocabulários, tesauros e listas controladas (padrões de valor de dados) – como, LCSH, *Internet* MIME *types*, ISO 639, ISO 8601 e RFC 4646[430]. –, ou normas, regras e códigos de catalogação bibliográfica e arquivística (padrões de conteúdo de dados) – tais como, RDA,

[430] Disponível em: https://www.ietf.org/rfc/rfc4646.txt. Acesso em: 30 maio 2023.

AACR2, DACS e ISBD –; e permitir a descrição conjunta ou a inclusão de metadados de outros esquemas XML com indicações para metadados externos, como no elemento Extensão MODS.

A pesquisa feita identificou, sistematizou e analisou padrões e esquemas de metadados para o arquivamento da *Web*, debatidos na Ciência da Informação e áreas afins. Os metadados descritivos e técnicos DC, MODS, EAD, VRA *Core*, MIX etc. e os dados de autoridade MADS e EAC-CPF têm uma aplicação mais voltada a apoiar o PREMIS e o METS, seja em permitir a identificação e localização como oferecer dados técnicos, de renderização, integridade e fixidez, direitos e agentes com funções nas ações que afetam *sites* arquivados. Incorporando metadados descritivos, estruturais e administrativos (e de preservação, como PREMIS), o METS é útil em simplificar a ordenação e a gerência das partes constituintes dos *sites* e seus metadados, vincular de forma hierárquica os distintos arquivos (textos, imagens, áudios etc.) que compõem os *sites* e, em adição, gerir tais objetos complexos atuando como um PSI, PAI e PDI num sistema OAIS.

Diferentes tipos de metadados são importantes no arquivamento da *Web*, mas este trabalho focou os metadados descritivos e administrativos (sobretudo, de preservação). Certos elementos de metadados ou unidades semânticas dos padrões identificados puderam ser sinalizados nesta pesquisa como sendo úteis à preservação de *websites* em sistemas de arquivamento digital. Por exemplo, no DC, que mostrou ser um expoente para o arquivamento da *Web* por suas semelhanças com os elementos WAN de Dooley e Bowers (c2018) e uso nos elementos de Kim e Lee (2007), no *Internet Archive*, no Arquivo.pt e outras iniciativas notáveis da área, os elementos indicados no Quadro 2 incluem algumas das informações definidas nas unidades do dicionário de dados PREMIS, tais como os direitos de propriedade intelectual e seus titulares, a identificação única e persistente, as relações todo/parte e de derivação e as dependências técnicas do objeto digital.

Os resultados do trabalho cedem um respaldo teórico, técnico e estruturado de padrões e esquemas de metadados que podem ser usados em arquivos da *Web* concebidos para atender a preservação e fornecer o acesso duradouro de conteúdos *Web* arquivados. Tanto os elementos de metadados como as unidades semânticas apontadas na pesquisa para preservação digital no arquivamento da *Web* colaborarão sobre a escolha dos padrões de metadados de acordo com as necessidades das organizações públicas, privadas, sem fins lucrativos, de pesquisa e patrimônio cultural que estão interessadas e/ou envolvidas em iniciativas nacionais e internacionais na área ou, ainda, a percepção das informações a serem previstas e exigidas para assegurar a descrição, a preservação e a gestão consistente dos *sites* arquivados num sistema, que foram selecionados e coletados a partir de um domínio eletrônico, evento, local ou tópico (ciência e tecnologia etc.).

Evidencia-se que a garantia de preservação digital no arquivamento da *Web* só será factível com a adoção efetiva de padrões de metadados em suporte à administração do arquivamento e da manutenção do acesso permanente e utilizável dos conteúdos *Web* no tempo. Estas estruturas de descrição definirão a identidade e a persistência, a coerência e a compreensibilidade, o acesso e a representação, as funcionalidades, a autenticidade, integridade e confiabilidade, o contexto e a proveniência de *websites* selecionados, coletados e armazenados em sistemas de informação para intuitos de preservação, além de determinar a descoberta, a recuperação, a apresentação, a navegação e a arquivabilidade de *websites* como a interoperabilidade semântica entre sistemas.

5 ARQUIVAMENTO DA *WEB*: definições, motivações e processo de seleção

A julgar pela hipótese de que existe uma carência de estudos nacionais de preservação digital e arquivamento da *Web* no campo CTS de um modo geral constatado pela busca deste assunto junto as principais revistas associadas à área, e em particular no campo da Ciência da Informação e as suas áreas afins (Arquivologia, Biblioteconomia e Museologia) verificada por levantamento bibliográfico e revisão da literatura produzida neste tema, como já referido, os quais investiguem, analisem e sistematizem em profundidade os critérios de seleção aplicáveis à preservação digital e o arquivamento de conteúdos publicados na *Web*, é que se reconheceu a necessidade de identificar e descrever quais critérios de seleção de conteúdos *Web* poderiam ser considerados pelas organizações – sobretudo, as instituições de memória e as universidades – , que estão desenvolvendo os seus arquivos da *Web*, para que estas pudessem contemplar a preservação digital e o arquivamento da *Web*. Objetiva-se ainda, mais especificamente verificar como critérios de seleção de conteúdos da *Web* no âmbito da preservação digital e do arquivamento da *Web* têm sido discutidos pela Ciência da Informação e áreas afins, apontando critérios de seleção que poderiam atender às demandas de estruturação de arquivos da *Web* institucionais de forma mais adequada para a preservação a longo prazo de informações digitais.

Para este fim, é realizado uma pesquisa exploratória, que utiliza como método de coleta de dados a pesquisa bibliográfica e documental a partir de uma revisão de literatura narrativa referente ao arquivamento da *Web* no escopo da preservação digital (CORDEIRO *et al*., 2007; GIL, 2010; SEVERINO, 2016). Deste modo, através da análise do conteúdo (BARDIN, 2016; CAVALCANTE; CALIXTO; PINHEIRO, 2014) da revisão de produções científicas, tais como artigos de periódicos, anais de eventos, dissertações e livros, buscados no *Google Scholar* e na SciELO e nas bases *Scopus* e *ScienceDirect* (*Elsevier*), *Web of Science* (*Clarivate Analytics*) e *Emerald Insight* (*Emerald Publishing*) disponíveis via Portal de Periódicos CAPES e, também, de *sites*, políticas e diretrizes de iniciativas de arquivos da *Web* no mundo, foram reconhecidas e sistematizadas algumas definições para os conceitos de arquivamento da *Web* e de arquivo da *Web*; as razões para se arquivar *sites* apoiado em uma série de casos de uso; e uma descrição do processo de seleção, destacando a definição de uma política de seleção para conteúdo *Web*, a seleção e as suas fases, a documentação do processo de seleção e dos critérios de seleção, a manutenção e o contexto da política, os principais métodos de seleção, e a definição de critérios de seleção e de limites do recurso da *Web* selecionado, baseado nas experiências identificadas.

Portanto, este capítulo se propõe a apresentar os resultados e as análises dos conteúdos coletados, e o resultado deste mapeamento e análise previu contribuir em possíveis delimitações

do conjunto de diretrizes e políticas a serem adotadas por instituições de patrimônio cultural (bibliotecas, arquivos e museus) como pelas universidades públicas nacionais interessadas e/ou envolvidas com a seleção, a coleta/captura, o armazenamento, a recuperação, a preservação e o acesso permanente de informações produzidas, publicadas e/ou difundidas no ambiente da *Web*.

5.1 O conceito de arquivamento da *Web* e de arquivo da *Web*

O arquivamento da *Web* (*Web archiving*) tem as suas origens no domínio da preservação digital (DAY, c2006) e, desde meados dos anos 90, o termo arquivo da *Web* (*Web archive*) "[...] tem sido usado para descrever qualquer coleção da *Web online* e, consequentemente, [...]" o termo arquivamento da *Web* "[...] tem sido utilizado para descrever o ato de coletar e preservar a *Web online* e torná-la disponível.", conforme Brügger (c2018, p. 77, tradução nossa). Brügger (2011, p. 25, tradução nossa) conceitua o arquivamento da *Web* como "[...] qualquer forma de preservação deliberada e proposital de material da *Web*." Também trata do "[...] arquivamento de *website*(s) e/ou página(s) da *Web* disponíveis publicamente." (SHIOZAKI; EISENSCHITZ, 2009, p. 91, tradução nossa), abrangendo "[...] atividades de seleção, captura, armazenamento, preservação e gerência do acesso a instantâneos de recursos[431] da *Internet* ao longo do tempo." com o objetivo principal de "[...] preservar um registro da *Web* perpetuamente, o mais próximo possível de sua forma original, para vários propósitos acadêmicos, profissionais e privados.", segundo *International Organization for Standardization* (2013, não paginado, tradução nossa).

Em complementação a estas definições gerais, alguns estudos igualmente interpretam o arquivamento da *Web* sob a perspectiva de ser um processo. Niu (2012a, não paginado, tradução nossa) define arquivamento da *Web* como sendo "[...] o processo de coleta de dados que foram registrados na *World Wide Web*, armazenando-os, garantindo que os dados sejam preservados em um arquivo e disponibilizando os dados coletados para pesquisas futuras." No *International Internet Preservation Consortium* (c2022c, não paginado, tradução nossa) o arquivamento da *Web* consiste no "[...] processo de coleta de porções da *World Wide Web* para garantir que a informação seja preservada em um arquivo para futuros pesquisadores, historiadores e o público." De acordo com Rockembah (2018, p. 9) o arquivamento da *Web* é "[...] um processo que compreende coletar, armazenar e disponibilizar a informação retrospectiva da *World Wide*

[431] Recurso (*resource*) é "qualquer documento do arquivo representado por um URL." (LIBRARY OF CONGRESS, [2022?c], não paginado, tradução nossa) e, especificamente no contexto da *Web*, o termo refere-se a "[...] qualquer arquivo que faça parte de um *site*.", conforme *The National Archives* ([2022?a], não paginado, tradução nossa).

Web para futuros pesquisadores." Por sua vez, para Bailey e LaCalle (2015, p. 6, tradução nossa) o arquivamento da *Web* constitui "[...] o processo de coletar partes do conteúdo da *Web*, preservar as coleções e, em seguida, fornecer o acesso aos arquivos - para uso e reutilização."

Tais conceituações manifestam que o arquivamento da *Web* necessita de "[...] sistemas informáticos criados com o objectivo de preservar e manter acessível a informação publicada na *Web* após esta deixar de estar disponível em-linha [...]" (GOMES, 2010, p. 1). Conforme Costa, Gomes e Silva (2017, p. 192, tradução nossa) os primeiros arquivos da *Web* surgiram apenas em 1996 e "[...] são uma nova forma de instituições de patrimônio cultural incumbidas a preservar [...]" informações publicadas na *Web* e "[...] conteúdos nascidos em formatos não digitais que foram depois digitalizados e publicados *online* [...]", como também "[...] são um tipo especial de bibliotecas digitais." que tem "[...] a responsabilidade de preservar a informação para as gerações futuras." Da mesma maneira, um arquivo da *Web* remete a "todo o conjunto de recursos rastreados da *Web*[432] ao longo do tempo, compreendendo uma ou mais coleções" (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2013, não paginado, tradução nossa), ou ainda "[...] uma coleção de URLs[433] arquivadas reunidas por tema, evento, área temática, ou endereço *Web*.", em acordo com Bailey e LaCalle (2015, p. 6, tradução nossa).

Para mais, arquivo da *Web* "[...] é um registro dos recursos *Web*.", que "[...] pode incluir HTML e imagens, *scripts*, folhas de estilo, assim como vídeo, áudio e outros elementos que compõem as páginas *Web* e aplicações da *Web*, tudo em um único arquivo." (RHIZOME.ORG, [2022], não paginado, tradução nossa). Também os arquivos da *Web* em geral são constituídos "[...] de múltiplas cópias, com registro de data/hora, compiladas da mesma página da *Web* ou páginas coletadas em momentos diferentes.", sendo que preferencialmente "[...] capturam e preservam não apenas o texto, imagens e conteúdo informativo, mas também a funcionalidade, aparência e sensação da *Web*.", de acordo com *Society of American Archivists* (c2022k, não

---

[432] Rastreador (*crawler*, ou rastreador da *Web* – *Web crawler* –, robô – *robot* – e aranha – *spider* –) remete a "[...] um programa de computador que navega automaticamente pelas páginas *Web*, recuperando o conteúdo e seguindo recursivamente os *links* nelas contidos, de acordo com parâmetros pré-definidos." sendo "[...] normalmente usados pelos mecanismos de busca para construir seus índices de busca e para o arquivamento da *Web*." (BROWN, c2006, p. XIII, tradução nossa) ou, conforme *Bibliothèque Nationale Du Luxembourg* ([c2022], não paginado, tradução nossa), refere-se ao *software* que "[...] escaneia cada elemento de um *site*, seguindo cada *link* e rastreando cada componente em cada página." e que, ao serem utilizados para a indexação na *Web* pelos mecanismos ou motores de pesquisa, possibilitam "[...] resultados de busca mais rápidos e eficientes através de rastreamentos frequentes."

[433] URL refere-se a um tipo de URI, ou melhor, "[...] protocolo para identificar recursos em rede [...]", que "[...] identifica tanto o recurso quanto a sua localização." atuando "[...] como um endereço para recursos em rede, como conteúdo *Web*." (BROWN, c2006, p. XIII, tradução nossa). Já em *International Organization for Standardization* (2013) e em *Library of Congress* ([2022?c]) consisti num subconjunto do URI que especifica a localização de um recurso da *Web* e o protocolo para recuperá-lo. Também remete "o endereço de um recurso (como um documento ou *site*) na *Internet*." (BIBLIOTHÈQUE NATIONALE DU LUXEMBOURG, [c2022], não paginado, tradução nossa).

paginado, tradução nossa). Portanto, em concordância com Brügger (c2018, p. 78, tradução nossa) podemos entender que o termo arquivo da *Web* é utilizado para indicar os resultados do processo de arquivamento da *Web*, isto é, da "[...] ampla atividade de coletar e preservar a *Web online* e torná-la disponível, seja qual for a parte da *Web* em questão (pública ou privada) [...]".

Não obstante, para uma melhor compreensão destes dois conceitos se faz necessário a definição de três termos: rastreamento (*crawl*), captura (*capture*) e coleta (*harvest*). Tais termos podem ser utilizados de maneira indistinta significando "[...] o processo de *download* de todos [...]" os "[...] arquivos indispensáveis para reproduzir completamente um *website*, preservando, em última análise, a forma original do conteúdo recuperado." (LIBRARY OF CONGRESS, [2022?c], não paginado, tradução nossa). Também em *The National Archives* ([2022?a], não paginado, tradução nossa) os três termos supramencionados (incluindo os termos instantâneo – *snapshot* – e arquivo – *archive* –) compõem "[...] o processo de copiar informações digitais da *Web* ao vivo para o arquivo da *Web* utilizando um rastreador." e, de acordo com *International Organization for Standardization* (2013, não paginado, tradução nossa), os termos rastreamento e coleta dizem respeito ao "processo de navegação e cópia de recursos usando um rastreador".

Em contrapartida, cada um desses três termos pode ser conceituado individualmente. O termo rastreamento remete ao "ato de navegar na *Web* automaticamente e metodicamente para indexar ou baixar conteúdo e outros dados da *Web*." (PENNOCK, c2013, p. 35, tradução nossa). Já o termo captura para *International Organization for Standardization* (2013, não paginado, tradução nossa) trata-se da "cópia de um recurso rastreado em um determinado momento", E, por fim, o termo coleta "descreve o processo de rastreamento e *download* de partes da *Internet*, muitas vezes usado como sinônimo de rastreamento da *Web* no contexto de arquivamento da *Web*.", como *Bibliothèque Nationale Du Luxembourg* ([c2022], não paginado, tradução nossa).

Por fim, se faz necessário ainda definir os conceitos de preservação digital e de curadoria digital e diferenciá-los do conceito de arquivamento da *Web*. Sendo um termo "guarda-chuva" conceitual, a preservação digital constitui "[...] o conjunto coordenado e contínuo de processos e atividades que garantem o armazenamento a longo prazo e sem erros de informação digital, com meios para recuperação e interpretação, durante todo o período em que a informação for necessária." (STATE LIBRARY OF NEW SOUTH WALES, 2022, não paginado, tradução nossa). Logo, visto que o conceito de preservação digital também "[...] aplica-se ao conteúdo que nasce digital, bem como ao conteúdo que é convertido para o formato digital." como Dollar e Ashley (2012, p. 317, tradução nossa), o arquivamento da *Web* dentro da preservação digital envolve então os processos e atividades estratégicas que lidam com a preservação de conteúdos digitais da *Web* pública (e/ou privada), ramificando-se com a preservação de mídias sociais etc.

Já a curadoria digital é uma área de investigação e de prática interdisciplinar "[...] que trata do gerenciamento do objeto digital durante todo o seu ciclo de vida.", compondo "[...] um processo mais completo [...]" o qual envolve o "[...] planejamento, avaliação e reavaliação das ações em prol da curadoria do objeto digital e que engloba a preservação digital como parte do seu ciclo." (SIEBRA *et al*, 2013, p. 1). Aliás, a curadoria digital "[...] une as tecnologias e boas práticas do arquivamento e da preservação digital e dos repositórios digitais confiáveis com a gestão dos dados científicos [...]", garantindo a sustentabilidade dos dados digitais para o futuro, de acordo com Sayão e Sales (2012, p. 189). Tendo a preservação digital apenas como uma de suas etapas, o processo de curadoria e preservação dos dados/objetos digitais inclui diferentes estágios, onde destacamos dois: a avaliação do dado e a seleção do que será objeto de curadoria e de preservação por longo prazo; e o arquivamento ou, melhor, a transferência do dado para um arquivo, repositório etc. (HIGGINS, 2011; SAYÃO; SALES, 2012; SIEBRA *et al*., 2013).

Assim, a curadoria digital se faz um conceito mais amplo que a preservação digital e o arquivamento da *Web*, sendo que para este último a relação com o conceito de curadoria digital pode-se ainda conceber pela existência de tomada de decisões de seleção por profissionais da informação (bibliotecários, arquivistas etc.) e especialistas-curadores no desenvolvimento das coleções *Web*, os quais trazem limitações para o arquivamento da *Web* e aos arquivos da *Web*.

5.2 Por que arquivar *websites*?

Como indicado por Pennock (c2013), apesar de se configurar numa das principais razões para arquivar *sites* (sobretudo, na comunidade de patrimônio cultural), meramente preservar o conteúdo da *Web* porque, caso contrário, o mesmo será perdido é um argumento fraco quando comparado a obrigação legal que algumas instituições possuem de capturar e arquivar conteúdo *Web*. Por exemplo, a BnF se fundamenta juridicamente no depósito legal da *Web* francesa, através da Lei nº 2006-961, de 2006, conhecida como Lei DADVSI[434], garantindo que *sites* do domínio francês sejam coletados e preservados, consultados em salas de leitura e reproduzidos segundo o Código de Propriedade Intelectual (do francês *Code de la Propriété Intellectuelle*)[435] (BIBLIOTHÈQUE NATIONALE DE FRANCE, c2022b). Os Regulamentos de Bibliotecas de Depósito Legal (obras não impressas) (do inglês *Legal Deposit Libraries - Non-Print Works* -

[434] Lei relativa aos direitos de autor e direitos conexos na sociedade da informação (do francês *loi relative au droit d'auteur et aux droits voisins dans la société de l'information*). Disponível em: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000266350. Acesso em: 30 maio 2023.

[435] Disponível em: https://www.legifrance.gouv.fr/codes/id/LEGITEXT000006069414/. Acesso em: 30 maio 2023.

Regulations)[436], de 2013, também habilitam as Bibliotecas de Depósito Legal do Reino Unido[437] que compõem o UKWA a coletar qualquer *site* baseado neste país, preservá-los para as gerações futuras, e disponibilizá-los aos usuários em suas instalações (UK WEB ARCHIVE, [2022?]).

Além dos requisitos legislativos para se coletar conteúdos da *Web* sob depósito legal os tornando parte do patrimônio de nações, outras iniciativas de arquivos da *Web* – como o *Coca-Cola Web Archive* em colaboração com o serviço da *Hanzo Archives* – arquivam *sites* tanto para apoio processual onde parte/totalidade do *site* pode ser solicitada em tribunal como para conformidade normativa para aplicações de gestão de registros ou, ainda, devido a um interesse social que as levem a documentar a evolução e o conteúdo da *Internet* no todo e disponibilizá-lo para usuários, como no caso do *Internet Archive* (PENNOCK, c2013). No *New Zealand Web Archive* da Biblioteca Nacional da Nova Zelândia, por exemplo, o arquivo da *Web* é usado para obter um registro visual de como *sites* da Nova Zelândia e do Pacífico mudaram no tempo, e a biblioteca nacional coleta por depósito legal, arquiva e preserva para pesquisa as publicações neozelandesas (livros, *sites*, *blogs* etc.), embasando-se na *National Library of New Zealand (Te Puna Mātauranga o Aotearoa) Act* 2003 e na responsabilidade social de preservar a história social e cultural do país, de acordo com *National Library of New Zealand* ([2022a?], [2022b?]).

As *Stanford Libraries* ([c2022?b]) ou SUL em seu projeto de arquivamento da *Web* listam alguns motivos para seus esforços na área via uma vasta gama de casos de uso local, entre eles:

- Preservação do legado institucional na *Web* – os artefatos impressos que a muito tempo contam a história da universidade deram lugar ao endereço da *Web* "*www.stanford.edu*" como a representação mais consolidada da evolução da universidade. O arquivamento da *Web* é um modo prático de capturar e preservar os registros históricos e legalmente valiosos da instituição, como a "*Stanford University Website Collection*"[438] no serviço *Archive-It* que inclui *sites* dos seus departamentos, laboratórios, institutos, eventos etc.
- Contribuição para aprendizagem – visto que o conteúdo da *Web* se tornou um material primário para investigação e ensino, o arquivamento da *Web*, por exemplo, possibilita a captura das páginas ou instantâneos (*snapshots*) dos *sites* de empresas que são objetos

[436] Disponível em: https://www.legislation.gov.uk/ukdsi/2013/9780111533703/contents. Acesso em: 30 maio 2023.
[437] Existem seis Bibliotecas de Depósito Legal do Reino Unido (do inglês UK *Legal Deposit Libraries*) que fazem parte do sistema de depósito legal do país, a saber: *British Library*; *National Library of Scotland*; *Bodleian Libraries, University of Oxford*; *Cambridge University Library*; *National Library of Wales*; e *Library of Trinity College Dublin, University of Dublin*. Disponível em: https://www.legaldeposit.org.uk/. Acesso em: 30 maio 2023.
[438] Disponível em: https://archive-it.org/collections/5591. Acesso em: 30 maio 2023.

de estudo da *Graduate School of Business*[439] da universidade, e as coleções no *Archive-It* com potencial de valor acadêmico são recursos exclusivos, como a "*Digital Games*"[440] que oferece um contexto complementar para a coleção de Stephen M. Cabrinety[441] que contém *software*, jogos, literatura sobre a indústria de jogos de microcomputação etc.

- Materiais complementares para coleções especiais físicas – o Departamento de Coleções Especiais e Arquivos Universitários seleciona e preserva materiais de valor histórico duradouro para apoiar as necessidades de pesquisa de alunos e docentes da universidade. O arquivamento da *Web* permite adicionar materiais complementares e de outra forma ausentes nestas coleções, como os *sites* arquivados das coleções "Patrick Suppes"[442] e "Philip G. Zimbardo[443] no *Archive-It* que documentam as trajetórias destes professores.
- Gerência de dados governamentais – o governo eletrônico (*electronic government*)[444] oferece novas opções para a divulgação das informações do governo e novos desafios à sua preservação. O arquivamento da *Web* expande o escopo da informação documental que a universidade pode coletar e organizar para as suas comunidades; e as coleções têm respostas aos pedidos da *Freedom of Information Act* (FOIA)[445] e as presenças *Web* de governos da região da Baía de São Francisco na Califórnia, tais como as coleções de *sites* "*Freedom of Information* (FOIA)"[446] e "*Bay Area Govsernments*"[447] no *Archive-It*.
- Salvaguarda de resultados acadêmicos – os projetos de alunos e docentes resultam cada vez mais na criação de *sites* como subprodutos auxiliares ou mesmo centrais. Exemplos incluem as coleções de conteúdos da *Web* "Carolyn Bertozzi"[448], "Carl Djerassi"[449], "*Stanford University Student Organizations Website Collection*"[450] e "*Center for*

[439] Disponível em: https://www.gsb.stanford.edu/faculty-research/case-studies. Acesso em: 30 maio 2023.
[440] Disponível em: https://www.archive-it.org/collections/1023. Acesso em: 30 maio 2023.
[441] Disponível em: http://www.oac.cdlib.org/findaid/ark:/13030/kt529018f2/. Acesso em: 30 maio 2023.
[442] Disponível em: https://archive-it.org/collections/5605. Acesso em: 30 maio 2023.
[443] Disponível em: https://archive-it.org/collections/5604. Acesso em: 30 maio 2023.
[444] Governo eletrônico (ou *e-gov*) é definido como o uso pelo Governo dos Estados Unidos "[...] de aplicações da *Internet* baseadas na *Web* e outras tecnologias da informação, combinadas com processos que implementam estas tecnologias [...]" para "[...] aumentar o acesso e a entrega de informações e serviços do Governo ao público, outras agências e outras entidades governamentais; ou" para "[...] trazer melhorias nas operações do Governo que podem incluir eficácia, eficiência, qualidade de serviço [...]" (UNITED STATES, 2002, não paginado, tradução nossa).
[445] Disponível em: https://www.foia.gov/foia-statute.html. Acesso em: 30 maio 2023.
[446] Disponível em: https://www.archive-it.org/collections/924. Acesso em: 30 maio 2023.
[447] Disponível em: https://www.archive-it.org/collections/903. Acesso em: 30 maio 2023.
[448] Disponível em: https://archive-it.org/collections/6434. Acesso em: 30 maio 2023.
[449] Disponível em: https://archive-it.org/collections/5590. Acesso em: 30 maio 2023.
[450] Disponível em: https://archive-it.org/collections/5593. Acesso em: 30 maio 2023.

Relationship Abuse Awareness"[451] no *Archive-It*, que têm *sites* de professores e dos seus grupos de pesquisas, de órgãos estudantis da universidade e de centros de treinamento.

- Conformidade e gerenciamento de registros – o recredenciamento, as questões jurídicas e as demais ações de conformidade podem requerer o acesso a versões de informações partilhadas nos *sites* da universidade. O arquivamento da *Web* oferece um mecanismo forense na gestão de registros, cumprimento e redução de riscos de litígio para preservar as políticas e a documentação baseadas na *Web* à medida que mudam com o tempo. Por exemplo, a coleção "*Stanford University* COVID-19 *Response*"[452] no *Archive-It* inclui *sites* arquivados que documentam a resposta da instituição à pandemia de coronavírus.

Em adição a estes fatores, Shiozaki e Eisenschitz (2009) realizaram uma pesquisa por questionário com bibliotecas nacionais integrantes do IIPC a fim de esclarecer como elas tentam justificar as suas atividades de arquivamento da *Web*. Os resultados da investigação indicaram que as bibliotecas nacionais exploradas, enquanto organizações do setor público, preveem que:

a) Os benefícios trazidos por suas iniciativas – em especial, a preservação abrangente do patrimônio cultural pelas leis de depósito legal para material da *Web*, a disponibilização ao público dos materiais coletados, evidências de que certos conteúdos já existiam na *Web*, a continuidade do acesso a materiais citados, e a preservação seletiva de recursos valiosos – são superiores aos custos totais, o que é aceitável, conforme os autores, na medida em que a tecnologia permite coletar e reunir de forma remota *sites* publicamente disponíveis a um custo baixo, apesar de que seja difícil mensurar na prática essa relação;
b) Os custos impostos as bibliotecas – como desenvolvimento de sistema (armazenamento, manutenção etc.), custos de equipe ou salários, e operações diárias (seleção, processos de controle de qualidade e preservação a longo prazo, busca de permissões etc.) – são superiores aos custos que recaem sobre as partes interessadas (*stakeholders*, incluindo titulares/detentores de direitos – *rights holders* –, proprietários dos *sites* – *site owners* –, e usuários) como, por exemplo a carga de rede (*network load*) imposta aos proprietários de *sites* e a restrição dos direitos dos titulares de direitos, e segundo os autores neste último grupo os custos são considerados limitados e, quando existem, são aceitáveis; e
c) Todas as bibliotecas fazem esforços para mitigar os riscos legais ligados a questões de propriedade intelectual e de privacidade no processo de arquivar a *Web*, ou seu resultado final, de várias formas (como exemplo, a seleção de material que não inclui conteúdo sensível, a restrição de acesso a dados pessoais, ou o uso da legislação, de contratos com

[451] Disponível em: https://archive-it.org/collections/6063. Acesso em: 30 maio 2023.
[452] Disponível em: https://archive-it.org/collections/13658. Acesso em: 30 maio 2023.

> detentores de direitos, e de políticas de autoexclusão – *opt-out* – ou abordagem na qual antes do início do arquivamento se envia mensagens para os *sites* selecionados onde se anuncia que se rastreará, arquivará e dará acesso a estes, no intuito de se evitar processos judiciais), embora para os autores cada solução legal tenha os seus prós e contras em termos de custos de negociação, escopo de acesso, tamanho e escopo do arquivo da *Web*.

Enquanto biblioteca nacional de fato dos Estados Unidos, a Biblioteca do Congresso como um dos membros fundadores do IIPC[453], por exemplo, embora hoje não seja legalmente obrigada a arquivar *sites*, ela arquiva conteúdo *online* nascido digital em risco por meio do seu programa de arquivamento da *Web* desde 2000, em um esforço para prover acesso e preservar esses objetos efêmeros como a instituição tem feito com materiais impressos (LIBRARY OF CONGRESS, [2022?c]). Para a biblioteca as funções tradicionais para "[...] adquirir, catalogar, preservar e servir materiais de coleção de importância histórica para o Congresso e para o povo americano estendem-se a materiais digitais, incluindo *sites*." que exercem "[...] um papel cada vez mais importante na vida intelectual, comercial e criativa dos Estados Unidos.", conforme *Library of Congress* (2022a, p. 1, tradução nossa). No seu arquivo da *Web*, a biblioteca preserva e cede acesso para pesquisa a *sites* arquivados e, antes do arquivamento, notifica o proprietário do *site* (exceto *sites* do governo americano ou aqueles que utilizam *Creative Commons*[454]) que gostaria de incluir o seu conteúdo no arquivo (LIBRARY OF CONGRESS, 2022a, [2022?e]).

A Biblioteca do Congresso dos Estados Unidos (LIBRARY OF CONGRESS, 2022a) se esforça para construir coleções que registrem a criatividade americana e reflita a diversidade e complexidade do país, com prioridade na aquisição de material de perspectivas e vozes sub-representadas para assegurar a variabilidade de autoria, identidades culturais e outros fatores histórico-culturais. A título de exemplos das suas coleções *Web*, temos o "*Women's and Gender Studies Web Archive*"[455], o "LGBTQ+ *Politics and Political Candidates Web Archive*"[456] e o "*LGBTQ+ Studies Web Archive*"[457], que incluem conteúdo *online* sobre movimentos culturais, sociais e políticos pela igualdade de gênero, órgãos políticos e jurídicos LGBTQ+ nos Estados Unidos, e a história, o saber e a cultura LGBTQ+ americana e no mundo, complementando os

---

[453] Disponível em: https://netpreserve.org/about-us/members/library-congress/. Acesso em: 30 maio 2023.
[454] *Creative Commons* é "[...] um tipo de licença, baseada em direitos autorais, que fornece uma forma padronizada para os criadores concederem a outras pessoas o direito de compartilhar e usar seu trabalho" (SOCIETY OF AMERICAN ARCHIVISTS, c2022d, não paginado, tradução nossa).
[455] Disponível em: https://www.loc.gov/collections/womens-and-gender-studies-web-archive/about-this-collection/. Acesso em: 30 maio 2023.
[456] Disponível em: https://www.loc.gov/collections/lgbtq-politics-and-political-candidates-web-archive/about-this-collection/. Acesso em: 30 maio 2023.
[457] Disponível em: https://www.loc.gov/collections/lgbtq-studies-web-archive/about-this-collection/. Acesso em: 30 maio 2023.

acervos físicos da biblioteca. Junto ao desenvolvimento das suas coleções[458], a biblioteca ainda vem cooperando ativamente com outras organizações para documentar fatos que se manifestam na *Web* ao redor do mundo, como a coleção "*Ukraine Conflict*"[459] construída com a equipe do *Archive-It*, especialistas da Universidade de *Stanford* etc. que documenta o conflito na Ucrânia.

Os casos de uso associados aos arquivos da *Web* podem similarmente fornecer demais motivos para arquivar páginas *Web* que documentam eventos recentes, dados governamentais, reações públicas, notícias históricas, instituições culturais e informações de fonte para pesquisas acadêmico-científicas quanto a uma ampla diversidade de tópicos, com conteúdos produzidos em várias nações e em diferentes idiomas e plataformas (*sites*, *blogs*, mídias sociais etc.), como:

- Eventos espontâneos – constituindo uma classe de conteúdo da *Web* em risco, os eventos espontâneos (isto é, catástrofes, acidentes, revoluções, tópicos sociais populares etc.) podem ocupar brevemente os holofotes do público e depois sumir de vista (STANFORD LIBRARIES, [c2022?b]). Por exemplo, temos a coleção *"#RickyRenuncia web collection (Puerto Rico 2019)*"[460], com reportagens, conteúdo de mídia social etc. sobre eventos que levaram à renúncia do ex-governador de Porto Rico Ricardo Roselló; e "*Ukraine Conflict*"[461] no *Archive-It* do *Internet Archive Global Events*[462], que inclui *blogs*, mídias sociais etc. acerca do conflito na Ucrânia desde 2014. Como mais notícias mudam para *feeds* informais de mídia social de rápida atualização, volumes de dados sobre eventos atuais podem ser perdidos e, assim, os arquivos preservam tal conteúdo como partes do registro histórico, conforme *International Internet Preservation Consortium* (c2022b).
- Preservação de citações e referências na *Web* – os arquivos da *Web* podem servir para a citação de versões de um conteúdo *Web* usadas como referências em obras acadêmicas, aumentando a longevidade da citação e o seu valor para futuros leitores (PENNOCK, c2013). Os arquivos da *Web* fornecem *links* para versões específicas e estáveis do *site*,

[458] Disponível em: https://www.loc.gov/programs/web-archiving/web-archives/. Acesso em: 30 maio 2023.
[459] Disponível em: https://www.archive-it.org/collections/4399. Acesso em: 30 maio 2023.
[460] Disponível em: https://archive-it.org/collections/12491. Acesso em: 30 maio 2023.
[461] Disponível em: https://archive-it.org/collections/4399. Acesso em: 30 maio 2023.
[462] O programa de arquivamento de eventos espontâneos iniciou em 2007 quando o serviço *Archive-It*, do *Internet Archive*, se uniu a pesquisadores da *Virginia Polytechnic Institute and State University* (ou *Virginia Tech*) nos Estados Unidos para arquivar o conteúdo da *Web* relacionado ao assassinato em massa de 16 de abril de 2007 pelo aluno-atirador Cho Seung-hui em seu campus https://www.weremember.vt.edu/) e, desde então, as Coleções de Eventos Espontâneos (em inglês *Spontaneous Event collections*, https://archive-it.org/home/IAGlobalEvents) são criadas pela equipe do *Archive-It* junto com outras organizações, especialistas etc. com o objetivo de estabelecer um corpus de conteúdo *Web* sobre um evento particular, capturar conteúdo de risco em tempos de crise, e também fornecer acesso aberto as coleções para pesquisa e navegação (por exemplo, o "*#blacklivesmatter Web Archive*" https://archive-it.org/collections/4783, a coleção "*Hurricane Harvey 2017*" https://archive-it.org/collections/9323, e a "*2013 Boston Marathon Bombing*" https://archive-it.org/collections/3649) (INTERNET ARCHIVE, [2022a]).

através de identificadores persistentes[463] formais atribuídos a cada recurso ou por uma estrutura de URL consistente e estável para acessar recursos, como é o caso do arquivo da *Web* da Biblioteca do Congresso americano em que os recursos recebem um ID de citação exclusivo, que redireciona para o local do *site* arquivado e, além disto, assegura que os *sites* citados serão localizados mesmo que a estrutura de URL padrão do arquivo seja alterada, de acordo com *International Internet Preservation Consortium* (c2022b).

- Comunicação científica – diferentes coleções de *sites* arquivados de iniciativas contêm conteúdo *Web* que auxiliam na comunicação e na divulgação de assuntos da ciência. No arquivo da *Web* da Biblioteca do Congresso americano, por exemplo, existe a coleção "*Science Blogs Web Archive*"[464] que, julgando os *blogs* de ciência como periódicos ou diários *online* que enriquecem o acervo analógico de revistas científicas da biblioteca, provê recursos para acadêmicos e outros que pesquisam sobre redação, pesquisa, ensino e comunicação científica nos Estados Unidos; e o "*Coronavirus Web Archive*"[465], o qual registra o impacto e a resposta à pandemia de coronavírus em comunidades nos Estados Unidos e no mundo (incluindo grupos marginalizados afetados pela pandemia e que são criadores de conteúdos *Web*, como as comunidades afro, asiáticas e latinas americanas).
- Guerra Russo-Ucraniana – com a invasão russa na Ucrânia em 2022, várias iniciativas têm identificado, coletado, gravado e arquivado *sites* ucranianos antes de serem perdidos durante a guerra a fim de preservar a memória cultural digital do país. Como exemplo, o *Saving Ukraine Cultural Heritage Online* (SUCHO)[466], uma equipe internacional de mais de mil voluntários (incluindo bibliotecários, arquivistas etc.), trabalha para criar arquivos *Web* de *sites* de instituições culturais ucranianas em risco de perda, como o *site* do governo ucraniano sobre o arquivo oficial de *Kharkiv*[467], a partir do envio de URLs ao *Wayback Machine* do *Internet Archive*, do uso do *Conifer* para gerar gravações da

[463] Identificador persistente (*persistent identifier*) trata-se de "[...] uma referência duradoura a um recurso digital." compondo-se de um identificador único para "[...] garantir a proveniência de um recurso digital (que é o que propõe ser) [...]" e de um serviço duradouro "[...] que localiza o recurso ao longo do tempo mesmo quando sua localização muda.", assegurando que "[...] o identificador resolva para a localização atual correta.", visando "[...], assim, resolver o problema da persistência de acesso ao recurso citado, em particular na literatura acadêmica." como, por exemplo, o DOI, o PURL, o URN, o *Handle System* e outros esquemas (*schemes*) de identificadores persistentes (DIGITAL PRESERVATION COALITION, c2015, não paginado, tradução nossa). Para Conselho Nacional de Arquivos (2020, p. 34) refere-se ao "identificador de longa duração de um recurso na *Internet* que se mantém válido mesmo que a tecnologia de acesso ou a localização física do recurso identificado se modifique no tempo."
[464] Disponível em: https://www.loc.gov/collections/science-blogs-web-archive/about-this-collection/. Acesso em: 30 maio 2023.
[465] Disponível em: https://www.loc.gov/collections/coronavirus-web-archive/about-this-collection/. Acesso em: 30 maio 2023.
[466] Disponível em: https://www.sucho.org/. Acesso em: 30 maio 2023.
[467] Disponível em: https://www.sucho.org/archives. Acesso em: 30 maio 2023.

experiência de navegação nos *sites* além de outras tecnologias para rastrear, arquivar e auxiliar na reconstrução dos *sites* (ADAMS; FERNANDEZ, 2022; SERRANO, 2022).

- Ataques terroristas na França – após os atentados de Paris em 2015 e, depois, de Nice em 2016, muitas instituições lançaram coleções especificamente centradas no massacre que atingiu o jornal satírico francês Charlie Hebdo, como "Charlie Hebdo"[468] da equipe do *Archive-It*, que contém mídias sociais, notícias e *sites* institucionais relacionados ao ataque em Paris. A BnF e o *Institut national de l'audiovisuel* (INA)[469] criaram "coleções de emergência" para capturar uma amostra das reações *online* oficiais e populares sobre os ataques através dos rastros deixados na *Internet* e no *Twitter*, incluindo homenagens, apoio, opiniões críticas e hostis etc., que para Schafer *et al*. (2019) oferece um "quadro" de uso potencial para fonte de pesquisa quanto a resposta social aos ataques, juntamente com demais materiais, como artigos de imprensa, fotografias, entrevistas, dentre outros.
- Comunidade LGBT – algumas instituições arquivam *sites* referentes a minorias sexuais e de gênero, em especial, a população LGBT+, a fim de preservar as memórias culturais, sociais e políticas destes grupos no mundo que existem na *Web*. Por exemplo, através do uso do *Archive-It*, a Universidade da Califórnia, em *Berkeley* nos Estados Unidos, criou as coleções "*Southeast Asia* LGBT *Web Archive*"[470] e "*Archiving the* LGBT *Web: Eastern Europe and Eurasia*"[471]; e o UKWA tem a coleção "LGBTQ+ *Lives Online*"[472], o qual oferece um recurso único para pesquisas sobre o assunto e enriquece as coleções impressas das bibliotecas parceiras desta iniciativa. Cocciolo (2016) e Pendse (2014) também exploram os desafios e a utilidade da criação de arquivos *Web* de comunidades LGBT específicas para documentar e preservar para pesquisa acadêmica os movimentos pelos seus direitos, os efeitos da epidemia de AIDS, a luta por aumento de aceitação etc.
- Jornalismo digital – as coleções da *Web* de notícias históricas criadas pelas instituições de arquivamento servem como fonte para o estudo baseado em jornalismo *online*. Por exemplo, existem a "ABC *News – Australian Internet sites*"[473] no *Preserving and Accessing Networked DOcumentary Resources of Australia* (PANDORA *Archive*), com notícias da *Australian Broadcasting Corporation* (ABC) *News*[474] na Austrália; e a

[468] Disponível em: https://archive-it.org/collections/5190. Acesso em: 30 maio 2023.
[469] Disponível em: https://www.ina.fr/institut-national-audiovisuel/collections-audiovisuelles/le-web-media. Acesso em: 30 maio 2023.
[470] Disponível em: https://archive-it.org/collections/6459. Acesso em: 30 maio 2023.
[471] Disponível em: https://archive-it.org/collections/6165. Acesso em: 30 maio 2023.
[472] Disponível em: https://www.webarchive.org.uk/en/ukwa/collection/1151. Acesso em: 30 maio 2023.
[473] Disponível em: http://pandora.nla.gov.au/col/16241. Acesso em: 30 maio 2023.
[474] Disponível em: https://www.abc.net.au/news/. Acesso em: 30 maio 2023.

"*Hurricane Katrina blogs Web collection*"[475] no *Archive-It* da Universidade do Mississippi nos Estados Unidos, o qual documenta uma amostra representativa de *blogs* e jornalismo produzidos por vítimas do furacão Katrina. O futuro da pesquisa acadêmica em jornalismo digital requer o acesso a longo prazo do conteúdo noticioso da *Web*, como *feeds* de *sites* de mídia social, sendo vital arquivar de forma escalável e sistemática esses objetos ou *sites* de notícias dinâmicos e complexos, consoante Broussard e Boss (2018).

- Historiografia da *Web* – os arquivos da *Web* criados por instituições de arquivamento se configuram como fontes historiográficas, onde as informações culturais etc. preservadas nestes repositórios têm possibilidades de uso por historiadores e outros pesquisadores para estudos históricos da *Web* e da *Internet* (RODRIGUES; ROCKEMBACH, 2021). Como exemplo, as coleções "*Personal stories of Australians in war*"[476] e "*Historic gold mining sites*"[477] no PANDORA *Archive* podem ser usadas por historiadores da *Web* para fornecer dados e documentar os estudos, porém é preciso considerar os desafios do uso deste tipo de fonte de pesquisa herdada do passado, sobretudo, as características do *site* arquivado e as limitações no processo de arquivamento, como a ausência de elementos (imagens, *hiperlinks* etc.) ou de páginas *Web* específicas por razões técnicas quando um *site* é arquivado podendo este não corresponder ao *site online*, segundo Brügger (2012).
- Comunidade indígena – diversas bibliotecas e universidades desenvolvem coleções da *Web*, através do *Archive-It*, voltadas nas comunidades indígenas. Como exemplo, temos a "*Hawaii - Hawaiians*"[478] da Universidade do Havaí, que inclui *sites* com informações sobre havaianos indígenas, questões nativas e de soberania; e a "*Policing, Racism, and Indigenous People in Thunder Bay*"[479] da Universidade de *Lakehead* no Canadá, que contém notícias e respostas a eventos em *Thunder Bay* quanto ao racismo anti-indígena e questões de policiamento. No *New Zealand Web Archive* (NATIONAL LIBRARY OF NEW ZEALAND, [2022b?]), *sites* arquivados por e para os povos indígenas *Māori* estão representados em sua coleção o que, segundo Ka'ai-Mahuta (2019), contribui na preservação e disponibilização às gerações futuras das informações digitais partilhadas pelos povos indígenas garantindo a continuidade cultural e transmissão dos seus saberes.
- Comunidade negra – determinadas iniciativas se dedicam a criar coleções centradas nas comunidades e cultura negra pelo mundo. Por exemplo, o UKWA tem a coleção "*Black*

[475] Disponível em: https://archive-it.org/collections/7625. Acesso em: 30 maio 2023.
[476] Disponível em: http://pandora.nla.gov.au/col/12925. Acesso em: 30 maio 2023.
[477] Disponível em: http://pandora.nla.gov.au/col/13023. Acesso em: 30 maio 2023.
[478] Disponível em: https://archive-it.org/collections/1279. Acesso em: 30 maio 2023.
[479] Disponível em: https://archive-it.org/collections/9394. Acesso em: 30 maio 2023.

and Asian Britain"[480], que inclui *sites* referentes a cultura e a história da presença negra e asiática no Reino Unido; e o *Middlebury College* através do *Archive-It* criou a coleção "*Community Responses to Anti-Black Racism and Police Violence*"[481], que inclui *sites* com respostas e reações de indivíduos e organizações ao assassinato de George Floyd em 2020[482] e a luta por justiça social para os negros nos Estados Unidos. Em Rollason-Cass e Reed (2015) os autores também examinam a criação de uma coleção da *Web* em torno do movimento *#blacklivesmatter*[483], isto é, o "*#blacklivesmatter Web Archive*"[484] no *Archive-It*, que provê um recurso valioso para pesquisadores, ativistas e historiadores sobre o movimento contra os maus-tratos de afro-americanos nas mãos das autoridades.

- Mudanças climáticas e desastres naturais – muitas coleções *Web* documentam os efeitos e as respostas aos eventos climáticos ocorridos no mundo. Por exemplo, temos a "*Indian Ocean Tsunami December 2004*"[485] no UKWA, com *sites* de órgãos de ajuda, *sites* para registrar experiências pessoais etc. quanto ao desastre do *tsunami* de 2004 na Ásia; e a "*Japan Earthquake*"[486] no *Archive-It* da *Virginia Tech: Crisis, Tragedy, and Recovery Network*, com *blogs*, *sites* de notícias etc. que retratam os eventos ao redor do *tsunami* e terremoto de 2011 no Japão e a reconstrução pós-desastre. Outros exemplos, incluem Rockembach e Serrano (2021) que demostram a relevância da preservação do conteúdo *Web* sobre mudanças climáticas, mostrando o que foi e o que terá de ser preservado no futuro; e Radinsky e Horvitz (2013) que mineraram a *Web* para prever eventos futuros, achando uma relação entre secas e tempestades em Angola que estimula surtos de cólera.

Assim, segundo *Stanford Libraries* ([c2022?b], não paginado, tradução nossa) "assegurar a capacidade contínua de acesso ao conteúdo da *Web* [...]" "[...] é imperativo para

---

[480] Disponível em: https://www.webarchive.org.uk/en/ukwa/collection/1107. Acesso em: 30 maio 2023.
[481] Disponível em: https://archive-it.org/collections/14467. Acesso em: 30 maio 2023.
[482] Em 25 de maio de 2020, o afro-americano George Perry Floyd Jr. foi assassinado em *Minneapolis*, estado norte-americano de *Minnesota*, depois que o policial branco Derek Chauvin o estrangulou se ajoelhando no seu pescoço durante uma abordagem. Após a morte de Floyd, uma onda de protestos e debates antirracistas e contra a violência policial foi gerada nos Estados Unidos e por todo o mundo e, em 2021, houve o início do julgamento de Chauvin que culminou em sua condenação. Disponível em: https://noticias.uol.com.br/reportagens-especiais/george-floyd-como-negro-morto-pela-policia-inspira-hoje-luta-antirracista/#page4. Acesso em: 30 maio 2023.
[483] *Black Lives Matter* (em português Vidas Negras Importam) é um movimento social, iniciado nos Estados Unidos e difundido ao redor do mundo, que campanha contra a brutalidade policial, a discriminação racial etc. dirigida às comunidades negras. Este movimento iniciou em 2013, quando houve o uso da *hashtag #BlackLivesMatter* nas mídias sociais após a absolvição de George Zimmerman na morte do afro-americano Trayvon Martin em *Sanford*, Flórida e, a partir de 2014, ganhou mais força após as mortes dos afro-americanos Michael Brown pelo policial branco Darren Wilson em *Ferguson*, *Missouri*, e Eric Garner por policiais de Nova Iorque, gerando uma grande onda de protestos e manifestações. Disponível em: https://blacklivesmatter.com/herstory/. Acesso em: 30 maio 2023.
[484] Disponível em: https://archive-it.org/collections/4783. Acesso em: 30 maio 2023.
[485] Disponível em: https://www.webarchive.org.uk/en/ukwa/collection/2435. Acesso em: 30 maio 2023.
[486] Disponível em: https://archive-it.org/collections/2438. Acesso em: 30 maio 2023.

objetivos tão diversos como investigação, ensino, construção de coleção de biblioteca, legado institucional, conformidade legal e gestão da informação governamental." Em Brügger e Finnemann (2013) os autores trazem definições para arquivo da *Web*, caracterizando-o como: I) um arquivo em tempo real (*real-time*), pois a *Web online* é alterada ou excluída rapidamente, o que demanda uma ação inadiável quanto a seleção e a coleta do que se pretende preservar, enquanto ainda estiver *online*; II) uma versão digital "renascida", única, deficiente e editada inexistente antes do ato de arquivamento e não uma cópia de mesma dimensão do que já esteve *online* opondo-se ao material digitalizado ou nascido digital que se baseia, dado que a instituição que procura arquivar a *Web online* define o que arquivar e omitir, os *softwares* e métodos de arquivamento, e a disponibilização; III) multitemporal e multiespacial, porque tem várias versões do mesmo *site*, cada uma de um período de captura diferente; e num momento, certas partes são arquivadas e, depois, outras são, fazendo a extensão do *site* arquivado ser não idêntica ao longo do tempo.

De fato, os arquivos da *Web* constituem "[...] um instantâneo, ou representação, do que estava *online* e acessível ao rastreador no momento do rastreamento e não uma cópia funcional completa de um *site*.", e não são "[...] um '*backup*'[487] de um *site* a partir do qual o *site* original pode ser restaurado posteriormente." (THE NATIONAL ARCHIVES, [2022?c], não paginado, tradução nossa). Por exemplo, na Biblioteca do Congresso os esforços de arquivamento da *Web* se definem em criar uma cópia de arquivo – um instantâneo – o qual reflita o máximo possível a aparência e o funcionamento do *site* na ocasião em que foi coletado, mas reconhecendo que os rastreadores da *Web* têm limitações técnicas e podem não ser capazes de preservar todas as partes de um *site*, de acordo com *Library of Congress* ([2022?c], 2022a); e no UKGWA, um *site* arquivado significa a coleta de uma série de capturas/instantâneos de um *site* durante a sua vida útil, que é tida de "linha do tempo" (*timeline*) por contar a história do *site*, aliás, a iniciativa é um arquivo vivo (*living archive*)[488] em crescimento onde o arquivamento do *site* inicia-se com a primeira captura e continua até o *site* ser fechado (THE NATIONAL ARCHIVES, [2022?a]).

[487] *Backup*, para o contexto da *Web*, constitui uma "[...] uma cópia exata de algumas ou todas as páginas e recursos de um *site* (e em alguns casos, funcionalidade) nos mesmos formatos que ao vivo, o qual pode ser utilizada para restaurar completamente um *site* em caso de problemas ou exclusão." (THE NATIONAL ARCHIVES, [2022?a], não paginado, tradução nossa).

[488] Arquivo vivo (*living archive*) consiste na tentativa de "[...] registrar a memória popular contemporânea em um momento em que essa própria memória está em um momento de fluxo e contestação." (ANTOUN, 2012, não paginado, tradução nossa), e para Rollason-Cass e Reed (2015, p. 243, tradução nossa) remete "capturar conteúdo sobre um evento à medida que ele ocorre, utilizando conteúdo contribuído por membros das comunidades que participam ativamente ou são afetados por eventos e permitindo que a trajetória da coleção evolua, se preciso [...]".

Sendo um repositório importante da nossa história recente e da nossa herança cultural (ALNOAMANY; WEIGLE; NELSON, 2016), arquivos da *Web* fornecem um contexto melhor do que as capturas de tela (*screenshots*) isoladas e oferecem janelas vívidas num momento no tempo, bem como podem documentar a mudança ao longo do tempo e recriar a experiência que um usuário teria se visitasse o *site* ao vivo (*live site*) no dia em que foi arquivado, consoante Bailey e Lacalle (2015) e Rhizome.org [2022]. Estas coleções da *Web* temáticas, por domínios eletrônicos[489], ou baseadas em regiões e eventos, que exigem ferramentas especiais para o seu uso em que coleções inteiras podem ser processadas como dados ou pesquisadores podem visualizar *sites* arquivados página *Web* por página (INTERNATIONAL INTERNET PRESERVATION CONSORTIUM, c2022a), são uma fonte útil de informações únicas e historicamente valiosas para pesquisa devido aos tópicos e à qualidade da coleta e, também, por propiciarem explicar as histórias do passado e a conjecturar eventos futuros através da extração, modelagem e análise da evolução dos dados, em conformidade com Cadavid (2017) e Costa, Gomes e Silva (2017).

International Internet Preservation Consortium (c2022b) e Reynolds (2013) elucidam casos de uso para os arquivos da *Web* (e o arquivamento da *Web*), dentre os quais destacamos:

- Análise de *links* (*link analysis*) – como a coleta de grandes conjuntos de *sites* abrange também a captura de *links* e conexões entre eles, essas redes de *sites* e dados vinculados podem ser extraídos para observar relações entre pessoas, ideias, organizações etc. ao longo do tempo. Assim como em *sites* na *Web* ao vivo, tal análise poderá ser usada com dados de arquivos da *Web* para observar mudanças no tempo ou em períodos do passado.
- Atividade de extensão e educação (*outreach and education*) – já que a *Web* se tornou parte dos serviços de instituições educacionais e de patrimônio cultural, os arquivos da *Web* têm sido utilizados em exposições físicas de museus e em exposições *online*, além disso existem esforços para envolver alunos na criação de coleções de arquivos da *Web* com o propósito de envolvê-los com a história e realçar a relevância de se coletar *sites*.
- Prestação de contas (*accountability*) – como o rastreamento de *sites* ao longo do tempo permite analisar alterações no conteúdo *Web*, esse acesso é útil para garantir a prestação de contas para o conteúdo que não existe mais onde, por exemplo, as empresas podem

[489] Domínio (*domain*, ou domínio de destino – *target domain* – e *site* de destino – *target site* –) consisti no "[...] *site* que é alvo de rastreamento." (THE NATIONAL ARCHIVES, [2022?a], não paginado, tradução nossa), ou em consonância com *Bibliothèque Nationale Du Luxembourg* ([c2022], não paginado, tradução nossa), faz referência a "[...] uma subdivisão da *Internet* designada em um endereço com uma abreviação exclusiva (como *.lu* ou *.com*)."

arquivar seu conteúdo *Web* em defesa contra ações judiciais e os arquivos públicos da *Web* podem mostrar mudanças nas políticas ou práticas de governos, organizações etc.

- Vinculação persistente (*persistent linking*) – enquanto conteúdos *Web* podem mudar ou sumir sem aviso prévio, os arquivos da *Web* oferecem aos usuários *links* para acesso a versões específicas e estáveis do conteúdo em questão via, por exemplo, identificadores persistentes, permitindo que os usuários consultem esse conteúdo e acessem conhecendo exatamente qual versão do *site* está sendo utilizada como uma citação ou referência etc.
- Acesso a conteúdo excluído ou modificado (*access to deleted or modified content*) – os arquivos da *Web* disponibilizam *sites* que já foram excluídos ou alterados de modo que os usuários podem visualizar facilmente conteúdos inacessíveis na *Web* ao vivo, como são os casos das ferramentas *Wayback Machine* e *Memento Time Travel* que permitem aos usuários capturar, acessar e visualizar versões anteriores de *sites* e páginas da *Web*.
- Análise de tendências tecnológicas (*analysis of technology trends*) – *JavaScript*[490], HTML, RDF e mais formatos de arquivo, linguagens de programação e de marcação capturados em coleções de arquivos *Web* servem de linha temporal do desenvolvimento de tecnologias *Web*, e a análise em páginas coletadas pode mostrar mudanças no uso de formatos da *Web* ao longo do tempo, indicando tendências em marcação e formatação.

Niu (2012b) baseada em parte nos casos de uso do IIPC definiu igualmente quais usos e funcionalidades que se esperam que os arquivos da *Web* suportem para as necessidades dos seus usuários, que podem auxiliar a informar o *design* da funcionalidade de futuros arquivos da *Web* a serem construídos como a avaliar ou autoavaliar os arquivos da *Web* existentes, a saber:

- Parâmetros de pesquisa (*search parameters*) – o arquivo da *Web* suporta pesquisa por URL, palavra-chave e booleana, além da pesquisa baseada em domínio que oferece aos usuários a opção de pesquisar conteúdo em um único *site* arquivado específico; restrição dos resultados das pesquisas por data, seja a um período de tempo fixado por dias, meses ou anos; busca restrita por tipo de mídia (HTML, PDF, DOC etc.); interface de pesquisa integrada, em que todas as coleções de conteúdos da *Web* são acessíveis via um portal de pesquisa; distinção entre pesquisa simples e pesquisa avançada com usabilidade para os usuários (por exemplo, sem ferramentas de pesquisa avançada ocultas e *links* de ajuda invisíveis); e pesquisa segura para filtrar conteúdos adultos, como *sites* pornográficos.

[490] *JavaScript* refere-se a "[...] uma linguagem de programação de computador comumente usada para criar efeitos interativos dentro de navegadores *Web*." (THE NATIONAL ARCHIVES, [2022?a], não paginado, tradução nossa) ou, de acordo com Brown (c2006, p. XII, tradução nossa), remete a "[...] uma linguagem de *script* orientada a objetos, comumente usada para adicionar funcionalidade às páginas *Web*."

- Resultados de pesquisa (*search results*) – o arquivo da *Web* apresenta todas as versões arquivadas de um URL e a data e hora em que cada versão foi capturada, e fornece um resumo de conteúdo para cada URL retornada; classifica os resultados pela relevância para a consulta e os agrupa por domínio, mostrando as relações hierárquicas entre *sites* e as páginas da *Web* que pertencem a esses *sites*; fornece identificadores persistentes[491] para páginas *Web* arquivadas, certifica a precisão/confiabilidade[492], a autenticidade[493] e a integridade[494] do conteúdo arquivado, e indica que a versão que está sendo visualizada é um *site* arquivado e como o mesmo pode ser acessado (isto é, *online* ou presencial).
- Navegação (*browsing*) – as coleções no arquivo são organizadas em uma hierarquia, e os usuários podem navegar na hierarquia da coleção, restringir por domínio e período, ou navegar até o resultado desejado; os *sites* podem ser organizados em classes baseado no tema/gênero e por departamentos de governos, e a coleção pode ser organizada em categorias e subcategorias, primeiro por domínios de nações (.br, .pt etc.), depois a *Web*

[491] A estratégia de identificadores persistentes para a preservação digital tem a capacidade de ser criticamente vital para ajudar a estabelecer a autenticidade de um recurso e fornecer o acesso a ele ainda que a sua localização mude, assim como superar os problemas da natureza impermanente das URLs e possibilitar a interoperabilidade entre as coleções; porém, possui certas limitações, como de inexistir um sistema único de identificação permanente aceito por todos (embora os DOIs estejam amplamente implementados) ou de haver uma dependência da manutenção contínua do sistema como custos para se utilizar um serviço (DIGITAL PRESERVATION COALITION, c2015).
[492] Confiabilidade (*reliability*) alude a "credibilidade de um documento arquivístico enquanto uma afirmação do fato." e que "existe quando um documento arquivístico pode sustentar o fato ao qual se refere, e é estabelecida pelo exame da completeza, da forma do documento e do grau de controle exercido no processo de sua produção." (CONSELHO NACIONAL DE ARQUIVOS, 2020, p. 18) ou, conforme Fundação Biblioteca Nacional (2020, p. 18), é o "[...] atributo de um documento arquivístico referente à manutenção de sua fidedignidade e autenticidade."
[493] Autenticidade (*authenticity*) significa "uma característica mecânica de qualquer objeto digital que reflete o grau de confiabilidade no objeto, na medida em que os metadados de suporte que acompanham o objeto deixam claro que o objeto possuído é o que pretende ser." (NATIONAL DIGITAL STEWARDSHIP ALLIANCE, 2013, não paginado, tradução nossa), ou em conformidade com Conselho Nacional de Arquivos (2020, p. 12) diz respeito a "credibilidade de um documento enquanto documento, isto é, a qualidade de um documento ser o que diz ser e que está livre de adulteração ou qualquer outro tipo de corrupção." Para provar que um registro digital – isto é, a evidência de um evento, como uma transação ou decisão – é confiável (*trustworthy*), ele deve demostrar não apenas autenticidade, assegurando que não foi modificado ou alterado, mas também confiabilidade, garantindo que ele representa fielmente o evento que se destina a documentar (DIGITAL PRESERVATION COALITION, [2018e]).
[494] Integridade (*integrity*) é o "estado dos documentos que se encontram completos e que não sofreram nenhum tipo de corrupção ou alteração não autorizada nem documentada." (CONSELHO NACIONAL DE ARQUIVOS, 2020, p. 35) e, para *National Digital Stewardship Alliance* (2013, não paginado, tradução nossa), a verificação de fixidez ou fixidade (*fixity check*) constitui "um mecanismo para verificar se um objeto digital não foi alterado de maneira não documentada.", onde as somas de verificação (*checksums*), os resumos de mensagem (*message digests*) e as assinaturas digitais (*digital signatures*) são alguns exemplos de ferramentas para executar verificações de fixidez (*fixity checks*) e as informações de fixidez (*fixity information*) – isto é, aquelas criadas por tais verificações – "[...] fornecem evidências da integridade e autenticidade dos objetos digitais e são essenciais para permitir a confiança."

de uma nação por outros domínios de primeiro nível ou de nível superior[495] (.gov, .edu etc.), cada subcategoria dividida por gênero (*blogs*, jornais etc.), depois por mídia (PDF, vídeo etc.), entre outras.

- Funcionalidades associadas às políticas para lidar com as solicitações dos usuários: os usuários podem solicitar ou não que determinadas URLs, páginas da *Web* ou conteúdos da *Web* sejam arquivados/adicionados a um arquivo da *Web* ou removidos do mesmo, e o acesso público *online* a um *site* que já está no arquivo da *Web* pode ser bloqueado, devido a motivos comerciais, ou porque o seu conteúdo é sensível ou, também, por não ter permissão dos proprietários de direitos autorais (*copyright owners*) para mostrá-los.
- Serviços personalizados – o arquivo da *Web* fornece uma interface de usuário de arquivo *Web* personalizada (criar conta pessoal), permitindo que os usuários salvem pesquisas realizadas para referência e possível nova pesquisa; e notifica os usuários por *e-mail* sobre atualizações no arquivo, incluindo a publicação de relatórios mensais sobre novos títulos adicionados, ou quando coletas descobrem que uma página da *Web* foi atualizada ou alterada, especificando as alterações feitas num URL durante um período de tempo.
- Mineração de dados (*data mining*)[496] – o arquivo da *Web* pode apresentar gráficos que ilustram como certos *sites* arquivados se associam a determinados eventos num período de tempo, fornecer informações de *link* para uma página da *Web* arquivada (*links* de entrada, de saída e internos), permitir que usuários extraiam um subconjunto do arquivo da *Web* baseado em critérios, como idioma, formato de arquivo e metadados, e exportem o subconjunto extraído para processamento em outro lugar ou processem e analisem os

[495] Domínio de primeiro nível, ou de nível superior e/ou de topo (*top level domain*), representa o mais alto nível de domínios no sistema de nomes de domínio (em inglês *Domain Name System* – DNS) "[...] incluindo domínios de primeiro nível com código de país (por exemplo, *.fr, .de*), que se baseiam nos códigos de território de dois caracteres da abreviatura ISO 3166 do país, e domínios genéricos de primeiro nível (por exemplo, *.com, .net, .org, .paris*.)" (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2013, não paginado, tradução nossa).

[496] *Data mining* refere-se ao "[...] processo de identificação de padrões previamente desconhecidos pela análise de relações em grandes quantidades de dados reunidos a partir de diferentes aplicações." (SOCIETY OF AMERICAN ARCHIVISTS, c2022f, não paginado, tradução nossa), ou em conformidade com a *International Organization for Standardization* (2014, p. 2, tradução nossa) corresponde também ao "processo computacional que extrai padrões através da análise de dados quantitativos de diferentes perspectivas e dimensões, categorizando-os, e resumindo potenciais relacionamentos e impactos". A título de exemplo, ao passo que o corpus em grande escala de *sites* capturados oferece a possibilidade de análise de padrões e tendências modernas textuais (ou mineração de texto), projetos de pesquisa que estudam a frequência de uso de termos ou a análise de sentimentos podem utilizar coleções de arquivos da *Web* para extrair, visualizar e analisar a linguagem empregada em *sites* rastreados (*crawled websites*), oportunizando descobrir relações, tais como a frequência de coocorrência entre os termos, e as emoções adotadas ao discutir tópicos específicos, em acordo com *International Internet Preservation Consortium* (c2022b).

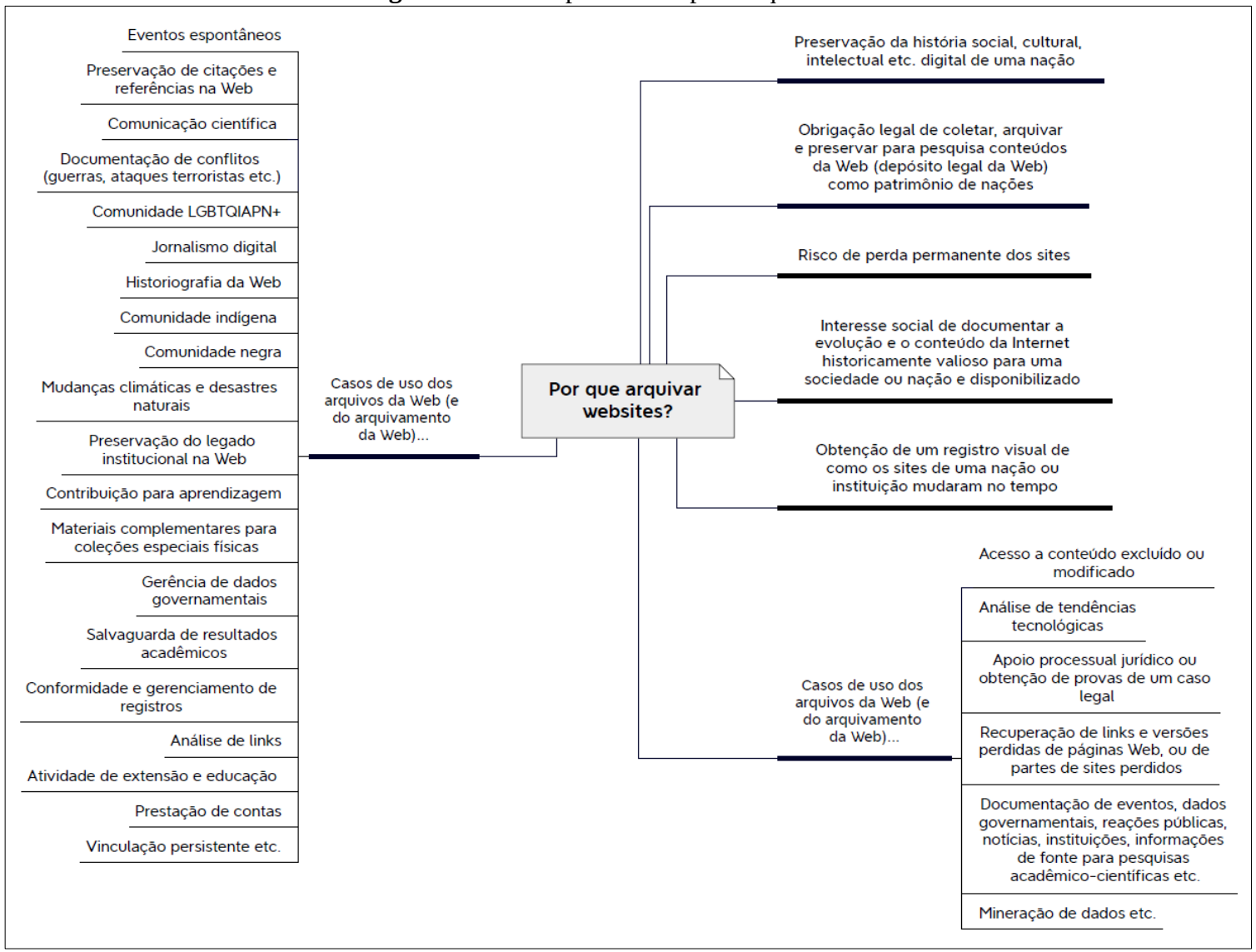
dados no próprio arquivo da *Web*, além de preservar arquivos de *log* (*log files*) do *site* que contém informações quanto a sistemas operacionais, servidores *Web*[497], versões etc.

- Recuperação de pelo menos partes de *sites* perdidos – os usuários e os proprietários de *sites* podem usar arquivos da *Web* para reconstruir uma certa versão de um *site* perdido, mantendo-se a estrutura do *site* original.

Em resumo, existem várias razões para se arquivar *sites* ou, ainda, para a realização do arquivamento da *Web* e para a criação de arquivos da *Web*, que são sintetizados na Figura 72.

[497] Servidor *Web* (*Web server*) constitui "um programa de computador que recebe solicitações HTTP de clientes (geralmente navegadores *Web*), e 'fornece' o conteúdo *Web* solicitado para eles.", porém "o termo também pode ser aplicado ao computador no qual o *software* do servidor *Web* está sendo executado." (BROWN, c2006, p. XIV, tradução nossa).

Figura 72 – Principais razões para arquivar *sites*.



Fonte: Elaborado pelo autor.

À vista disso, em acordo com Pennock (c2013), talvez a questão então não seja tanto "por que arquivar *sites*?", mas sim "por que não arquivar *sites*?". Arquivar a *Web* (*archiving the Web*) para Oury e Poll (2013, p. 133, tradução nossa) denota "[...] selecionar e capturar recursos da *Internet*, armazená-los em arquivos da *Web*, preservá-los e gerir o acesso sustentável aos arquivos.". Sendo uma forma de neutralizar as características frágeis da *Web* e garantir o acesso a longo prazo das informações na *World Wide Web* (BRÜGGER, 2005; MASANÈS, c2006b; SHIOZAKI; EISENSCHITZ, 2009), o processo arquivamento da *Web* é similar ao tradicional arquivamento de documentos em papel ou pergaminho, do qual o acesso aos conjuntos de *sites* arquivados poderá ser disponibilizado à vários perfis de usuários, incluindo jornalistas, gestores de *sites*, historiadores, cientistas, juristas, governos, empresas e público geral, para inúmeros casos de uso, como obter provas de um caso legal, recuperar *links* e versões perdidas de páginas *Web*, estudar conteúdos digitais etc., segundo Gomes (2010) e *The National Archives* (c2011).

Com primeiros registros de ações em 1996, bibliotecas e arquivos nacionais, empresas, consórcios de organizações etc. em todo o mundo estão envolvidos na aquisição e preservação de partes ou de toda a *Web*, de acordo com Costa, Gomes e Silva (2017) e Ferreira, Martins e Rockembach (2018). Muitas destas iniciativas de arquivamento da *Web* são membros do IIPC o que, para *International Internet Preservation Consortium* (c2022a), Niu (2012b) e Shiozaki e Eisenschitz (2009), mostra que estão empenhadas em garantir que seus conteúdos *Web* sejam preservados e disponibilizados, bem como estão entre os arquivos da *Web* mais estabelecidos do mundo e detêm capacidade financeira ou técnica e a intenção de cooperação internacional. Essa atividade é um pouco nova para as instituições patrimoniais (em especial, no papel usual das bibliotecas), o que explica porque a padronização ainda é quase inexistente (DI PRETORO; GEERAERT, 2019; OURY; POLL, 2013), e os arquivos da *Web* sendo uma nova feição destas instituições, com diferentes escopos e frequências de captura etc., normalmente crescem a um tamanho de dados maior do que as bibliotecas digitais, conforme Costa, Gomes e Silva (2017).

Como observado antes, são muitas as razões para se capturar a *Web*, que vão desde os argumentos comuns de que a *Web* faz parte das nossas vidas, a questão da natureza transitória das páginas *Web* ou que materiais únicos estão disponíveis só na *Web*, como ainda o papel das bibliotecas nacionais na coleta, preservação e acesso de material em todos os formatos (TUCK, 2008). Diante desse papel tradicional das bibliotecas, se faz claro que o equivalente digital às publicações em papel deve ser tratado com a mesma prioridade de preservação por relevarmos a informação e menos o formato, obrigando-nos dar passos corretos para salvaguardar as novas formas digitais do nosso patrimônio cultural e intelectual, consoante Campos (2002). Assim, a

partir do final dos anos 90, as bibliotecas nacionais têm selecionado, coletado e armazenado *sites* e publicações culturalmente valiosas de seu domínio nacional como uma função análoga ao típico depósito legal de materiais impressos, baseando-se em leis e regulamentos para coleta da *Web*[498] ou, ainda, em contratos (OURY; POLL, 2013; SHIOZAKI; EISENSCHITZ, 2009).

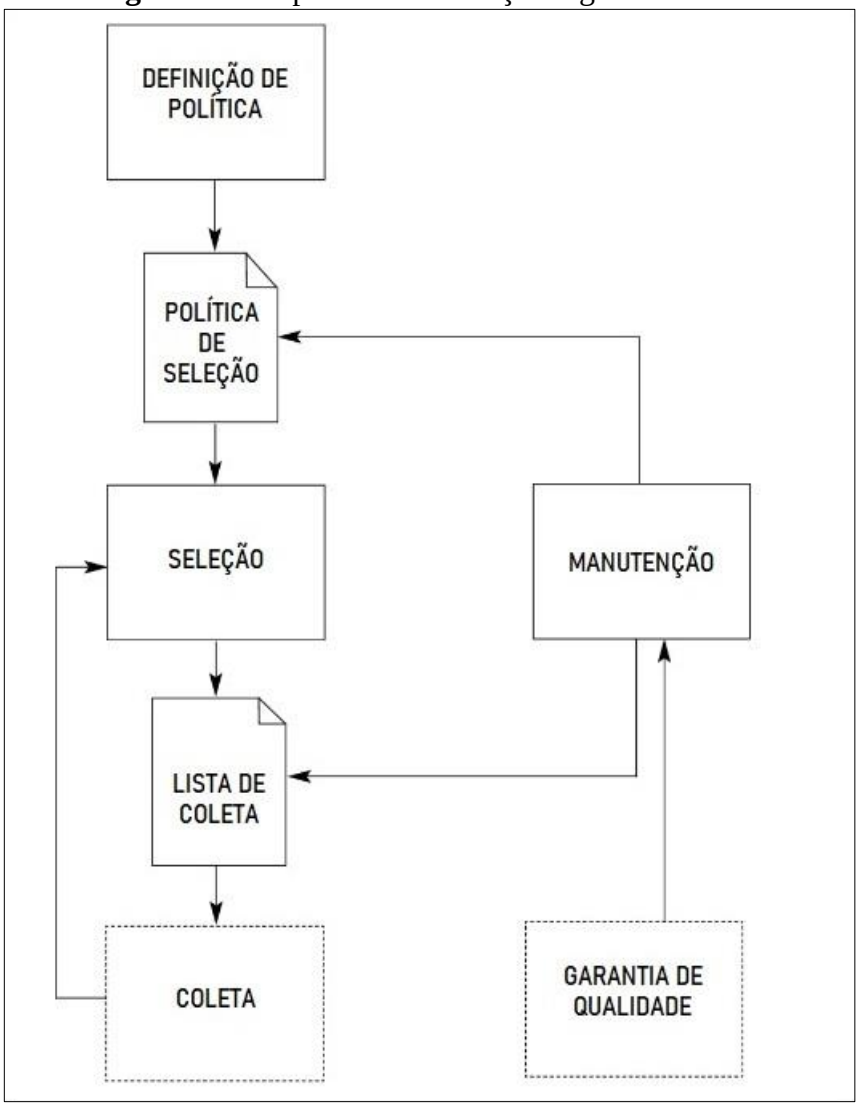
Além do mais, Melo (2020, p. 35) nos pontua que a natureza dinâmica e efêmera do ambiente *Web* "[...] exige que o processo de preservação seja pensado de forma sistêmica desde o princípio, incluindo metodologia de coleta dos dados, estabelecimento de políticas para seleção do conteúdo, técnicas e métodos de armazenamento, preservação digital e acesso." Este conjunto de atividades de preservação da *Web*, ou arquivamento da *Web*, possui, basicamente, a intenção de "[...] preservar a forma original do conteúdo coletado sem modificação.", e para se alcançar tal objetivo é preciso que existam ferramentas, padrões, políticas e melhores práticas que assegurarão o gerenciamento dos arquivos da *Web* ao longo do tempo, em concordância com o *International Internet Preservation Consortium* (c2022a, não paginado, tradução nossa).

5.3 Processo de seleção

Como indicado por Brown (c2006), o processo de seleção (*selection process*) é dividido em diversas fases distintas, segundo ilustrado na Figura 73.

[498] Coleta da *Web* (*Web harvesting*) é um termo usado "[...] para descrever a seleção, cópia e arquivamento de *sites* encontrados na *Internet*." (NATIONAL LIBRARY OF NEW ZEALAND, [2022b?], não paginado, tradução nossa).

Figura 73 – O processo de seleção segundo Brown.



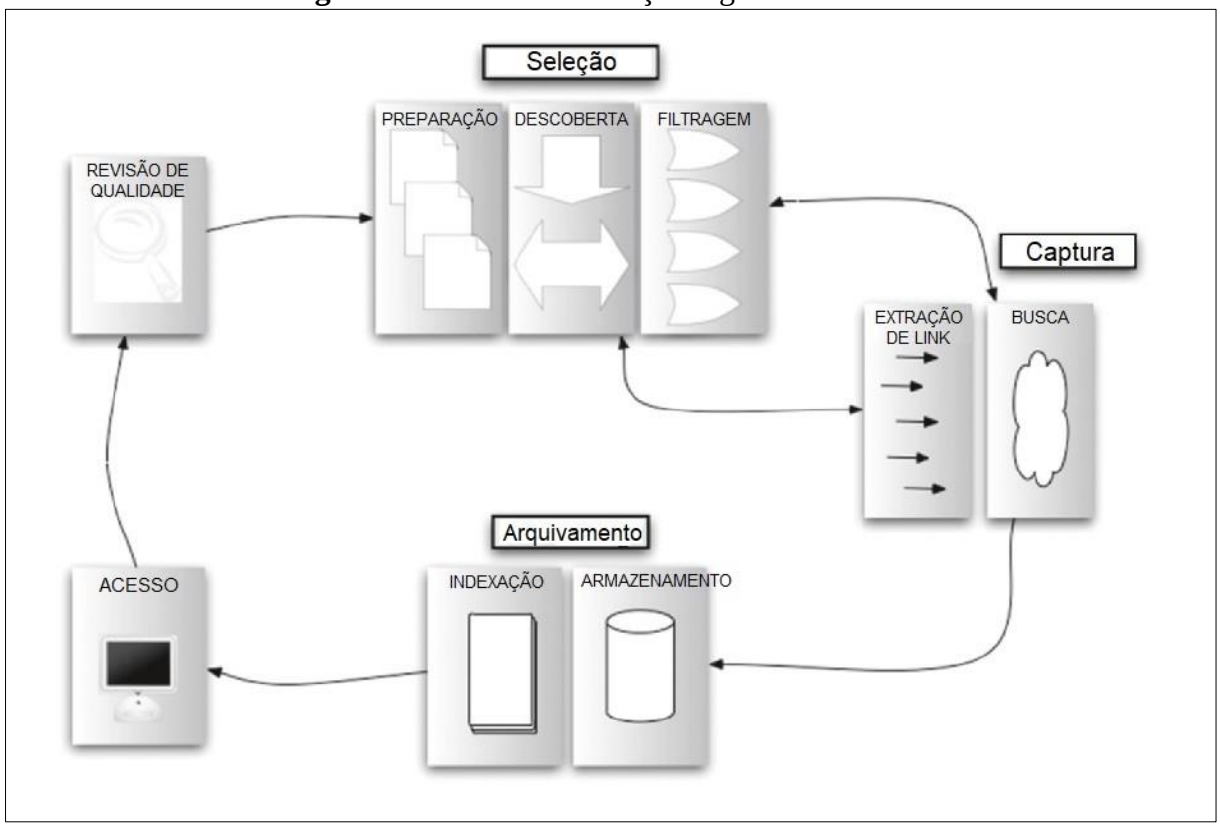
Fonte: Adaptado de Brown (c2006).

Para o autor o processo de seleção inicia-se com o desenvolvimento de uma política de seleção (*selection policy*) e a sua promulgação através de uma lista de recursos da *Web* a serem coletados, sendo que a política de seleção e a lista de coleta (*collection list*) precisam ser mantidas continuamente, e o *feedback* da coleta (*collection*) e a garantia de qualidade (*quality assurance*)[499] dos recursos *Web* devem ser utilizados para refinar o processo.

Esta divisão pode ser completada com o ciclo de seleção (*selection cycle*) apontado em Masanès (c2006a), conforme exposto na Figura 74.

[499] Garantia de qualidade constitui "[...] uma verificação de precisão e completude.", que "usa uma combinação de ferramentas e técnicas automáticas e manuais." (THE NATIONAL ARCHIVES, [2022?a], não paginado, tradução nossa).

Figura 74 – O ciclo de seleção segundo Masanès.



Fonte: Adaptado de Masanès (c2006a).

Segundo o autor a etapa de seleção (*selection*), com as suas três fases (preparação, descoberta e filtragem), ocorre no início de todo o ciclo e deve ser repetida regularmente; demais, ela precede as etapas de captura (*capture*) (que fornece insumo – *input* – e orientação), de arquivamento (*archiving*) e de revisão de qualidade (*quality review*), e vêm após a etapa de arquivamento e acesso (*access*) de rastreamentos prévios, se houver, atendendo os problemas e as mudanças necessárias que a etapa de revisão de qualidade suscitou.

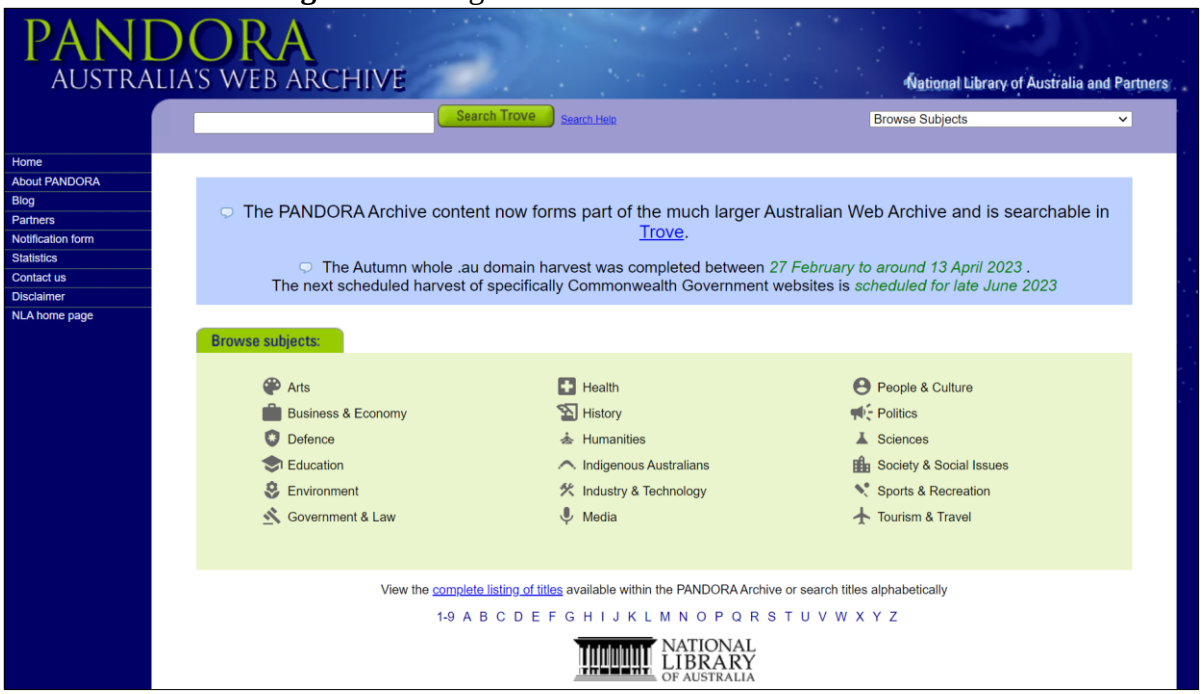
5.4 Definição de uma política de seleção

Em qualquer programa de arquivamento da *Web* uma política de seleção bem definida é uma condição fundamental (BROWN, c2006), e a seleção é uma questão-chave e o primeiro passo para o arquivamento da *Web*, de acordo com Kim e Lee (2007), Masanès (c2006a) e Niu (2012a). Contudo, simplesmente aplicar as abordagens desenvolvidas para seleção de material impresso não é adequado, haja vista que os materiais da *Web* e o próprio arquivamento da *Web* exigem novos métodos e práticas, como observado em Brügger (2011, 2012, 2012, c2018) e Brügger e Finnemann (2013). Uma política de seleção de conteúdo da *Web* permite definir e elucidar, segundo Khan e Rahman (2019) e Biblarz *et al*. (2001), quais os conteúdos do *site* que

devem ser capturados com base nas prioridades, no propósito e no escopo dos conteúdos *Web* já definidos. As escolhas realizadas neste domínio determinam o tipo, a extensão e a qualidade da coleção resultante da instituição de arquivamento, além do mais, sendo a característica de cada instituição ou arquivo da *Web*, a natureza da política de seleção variará evidentemente em função dos requerimentos organizacionais individuais (BROWN, c2006; MASANÈS, c2006a).

Como exemplo, no PANDORA *Archive*, o Arquivo da *Web* da Austrália, cada uma das agências parceiras, incluindo a Biblioteca Nacional da Austrália que gere e mantém o arquivo, selecionam conteúdos *Web* de acordo com as suas políticas de desenvolvimento de coleção ou diretrizes de seleção[500], que fornecem orientação acerca da jurisdição ou áreas temáticas pelas quais são responsáveis (NATIONAL LIBRARY OF AUSTRALIA, 2020). Assim, enquanto a Biblioteca Nacional da Austrália é incumbida de arquivar *sites* e documentos de relevância e significância a nível nacional, as outras agências parceiras visam selecionar os conteúdos os quais reflitam as suas regiões ou áreas de especialização, como o *Australian War Memorial* que seleciona *sites* relativos à história militar australiana, e a *State Library of New South Wales* que arquiva publicações *online* do governo do estado australiano de Nova Gales do Sul além de uma amostra representativa de *sites* não governamentais, conforme *Australian War Memorial* ([2020?]), *National Library of Australia* (2018) e *State Library of New South Wales* (2013).

Figura 75 – Página inicial do PANDORA *Archive*.

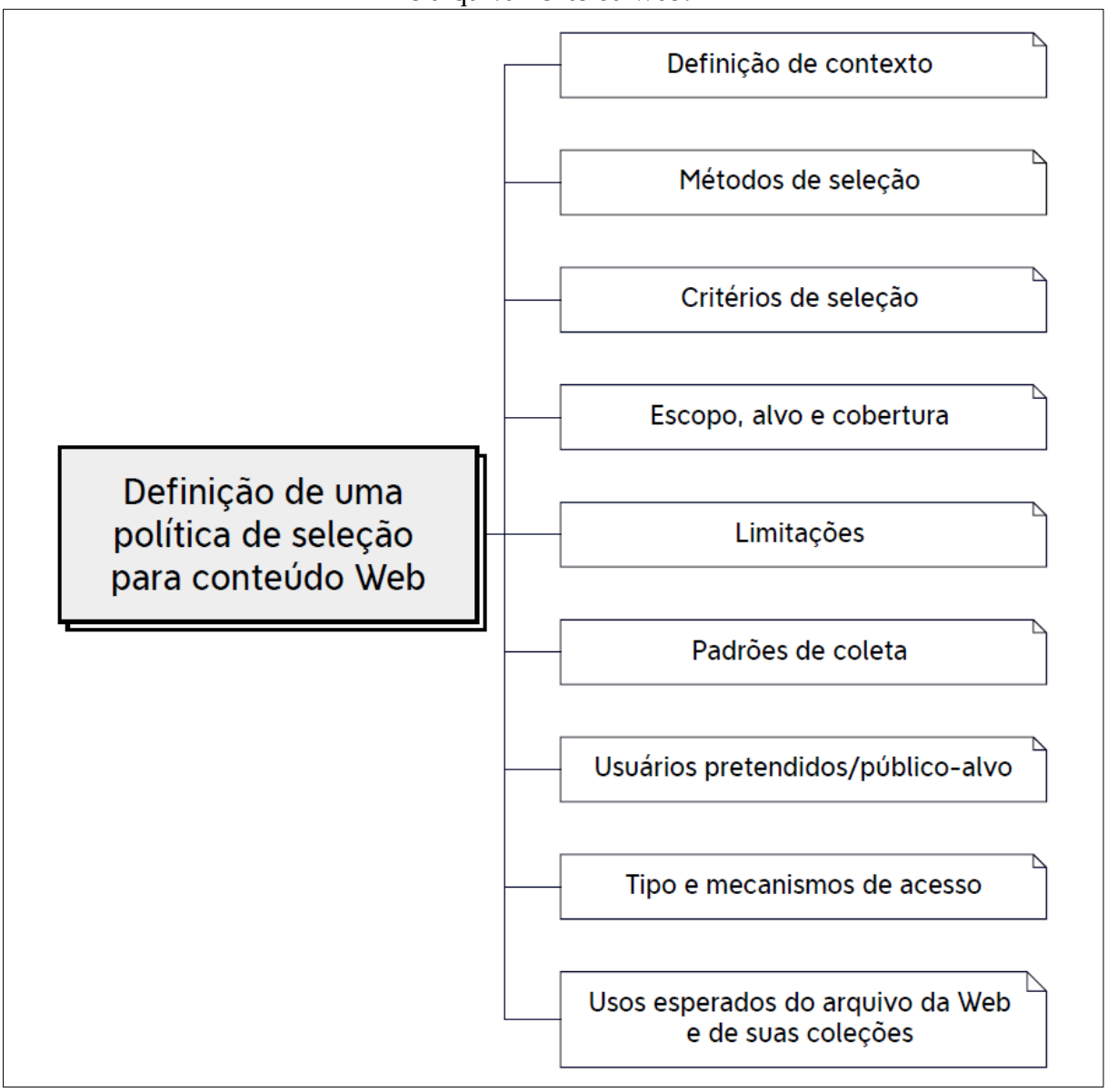


Fonte: *National Library of Australia* ([2023?b]).

[500] Disponível em: https://pandora.nla.gov.au/guidelines.html. Acesso em: 3 jun. 2023.

Apesar da existência de variações nas políticas individuais de seleção organizacional, através de Bragg *et al*. (2013), Brown (c2006), Kran e Rahman (2019) e Masanès (c2006a) a formulação de uma política de seleção de conteúdos *Web* para o arquivamento da *Web* exigirá geralmente alguns procedimentos, conforme exposto na Figura 76.

Figura 76 – Procedimentos para definição de uma política de seleção de conteúdos *Web* para o arquivamento da *Web*.



Fonte: Elaborado pelo autor.

Baseando-se nos autores supracitados os procedimentos para a criação de uma política de seleção para o arquivamento da *Web* em um arquivo da *Web*, indicados na Figura 76, podem ser descritos e exemplificados da seguinte forma:

- Definição de contexto – uma política de seleção de arquivamento da *Web* tem que ser formulada no contexto de uma política de seleção existente em toda a organização, ou

de políticas de seleção internas análogas para outros tipos de recursos, e de quaisquer políticas de seleção externas ou organizacionais de alto nível cabíveis. Por exemplo, no *Library of Congress Web Archive*, a Biblioteca do Congresso americano seleciona *sites* apoiado nas diretrizes das "*Collection Policy Statements*"[501] e "*Supplemental Guidelines for Web Archiving*", segundo *Library of Congress* ([2022?c], 2022a); e no *Australian Web Archive* (AWA)[502], a seleção é feita com base na "*Collection development policy*" da Biblioteca Nacional da Austrália, consoante *National Library of Australia* ([2022a]).

- Métodos de seleção – várias abordagens possíveis podem ser adotadas para a seleção e, portanto, um método apropriado deve ser identificado na política. A título de exemplo, a Biblioteca do Congresso americano, a partir de *Library of Congress* ([2022?c], 2022a), segue em geral uma abordagem baseada em coleções temáticas, tópicas ou de assuntos, onde pela coleta da *Web*, que ocorre em torno de assuntos e eventos, adquiri *sites* selecionados; já a Biblioteca Nacional da Austrália aborda a coleta da *Web* por capturas de páginas *Web* ou instantâneos (*snapshots*) de todo o domínio *Web* (*Web domain*) australiano (*au.*), coleções temáticas de cobertura focada em eventos, questões e setores selecionados, e arquivamento agendado e oportuno de *sites* selecionados com grande valor ou que mudam rapidamente, como *National Library of Australia* ([2022b]).
- Critérios de seleção – a política tem que conter critérios de seleção muito bem definidos e capazes de permitir a tomada de decisões de seleção específicas. Como exemplo, a Biblioteca do Congresso americano considera alguns fatores ao fazer as determinações de coleta, como utilidade em servir as necessidades informacionais atuais ou futuras dos pesquisadores da biblioteca e do Congresso norte-americano, singularidade e qualidade da informação fornecida, conteúdo acadêmico, em risco de perda de conteúdo (em razão da natureza efêmera de determinados *sites*), relação com outros recursos nas coleções da biblioteca, ou se o arquivamento da *Web* é o método apropriado para documentar o tópico ou evento, caso existam outros recursos, conforme *Library of Congress* (2022a).

[501] Disponível em: https://www.loc.gov/acq/devpol/cps.html. Acesso em: 3 jun. 2023.
[502] Disponível em: https://www.nla.gov.au/collections/building-our-collections/australian-web-archive. Acesso em: 3 jun. 2023.

- Escopo (*scope*)[503], alvo (*target*)[504] e cobertura (*coverage*) – após a escolha dos *sites* a serem arquivados, a política deve definir se o arquivamento será de todo o *site* ou de partes dos mesmos; e na descrição do contexto dos objetivos da prática de construção de coleções em uma política de desenvolvimento de coleções, o alvo da coleção, ou seja, o conteúdo a ser arquivado, deve ser descrito, podendo este ser especificado através da definição de critérios de inclusão e exclusão, tais como qualidade, gênero e editores. Por exemplo, a Biblioteca Nacional da Austrália indica em sua política que na coleta de *sites* publicados no domínio *.au*, ela captura *sites* num nível alto para assegurar abrangência, coletando a maioria, mas não todo o conteúdo de cada *site*; já o Arquivo.pt sinaliza que não coleta toda a *Web* pública portuguesa, impondo restrições de número de conteúdos por *site*, de número de *links* que o rastreador percorre desde um endereço inicial até chegar a um conteúdo etc., como *National Library of Australia* ([2022b]) e Arquivo.pt (2021a).
- Limitações – devido ao arquivamento da *Web* ser direto e constantemente prejudicado por dificuldades técnicas para a captura de conteúdos, como a *Web* oculta (*hidden Web*), mídia de *streaming*, conteúdo altamente interativo etc., a política de seleção deve incluir, sempre que possível, estes limites tecnológicos para o que pode ser de fato arquivado. Por exemplo, o *Library of Congress Web Archive*, o UKWA e o UKGWA, sinalizam estas limitações, onde nem todos os *sites* são arquivados por completo, estão ausentes, ou não são exibidos corretamente, havendo lacunas em suas coleções, segundo *Library of Congress* ([2022?e]), *The National Archives* ([2022?c]) e UK *Web Archive* [2022?].
- Padrões de coleta (*gathering patterns*) – a política deve descrever, se possível, os padrões de campanha de arquivamento (*archiving campaign*) para garantir que o resultado final seja coerente com o objetivo geral do desenvolvimento da coleção,

[503] Escopo são "os parâmetros ou restrições impostas em um rastreamento para garantir que ele colete apenas o que é desejado" (PENNOCK, c2013, p. 35, tradução nossa). Este termo também pode se referir tanto ao escopo de um rastreamento como ao escopo do arquivo da *Web*. No primeiro caso consisti no "conjunto de parâmetros que define a extensão de um rastreamento, por exemplo, o número máximo de itens ou a profundidade máxima do *path* que o rastreador deve seguir", podendo ser "[...] tão amplo quanto um domínio de nível superior (por exemplo, *.de*) ou tão estreito quanto um único arquivo."; e no segundo caso remete a "extensão de um arquivo ou coleção da *Web*, conforme determinado pelo mandato legal institucional ou pela política de coleta", segundo *International Organization for Standardization* (2013, não paginado, tradução nossa).

[504] Alvo representa o "conjunto significativo de recursos a serem coletados conforme definido por uma ou mais [...]" sementes (*seeds*, ou URL de destino – *targeted* URL –), isto é, o " URL correspondente à localização de um particular recurso a ser rastreado, utilizado como ponto de partida por um rastreador da *Web*", bem como pelas configurações de rastreamento (*crawl settings*, ou parâmetros de rastreamento – *crawl parameters* –) que tratam da "definição de quais recursos devem ser coletados e a frequência e profundidade necessária para cada conjunto de sementes", conforme *International Organization for Standardization* (2013, não paginado, tradução nossa).

incluindo, ao menos, o(s) gatilho(s) (*trigger*) (agenda/calendário de arquivamento, acontecimentos etc.), a(s) duração(ões) da campanha, e as relações a serem feitas entre campanhas. Exemplo de uma descrição simples: iniciar instantâneos de domínios da *Web* nacionais de três em três meses, com uma duração de campanha de sessenta dias e, usando como pontos de entrada (*entry point*), listar todos os domínios encontrados na campanha anterior. Também chamado de semente (*seed*), que para Pennock (c2013, p. 35, tradução nossa) é "o ponto de partida para um rastreamento, muitas vezes um URL", o ponto de entrada refere-se ao primeiro nó (*node*) "[...] de onde o caminho para outros documentos será encontrado em um processo de rastreamento." como, por exemplo, a página inicial (*homepage*) de um certo *site* (MASANÈS, c2006a, p. 79, tradução nossa).

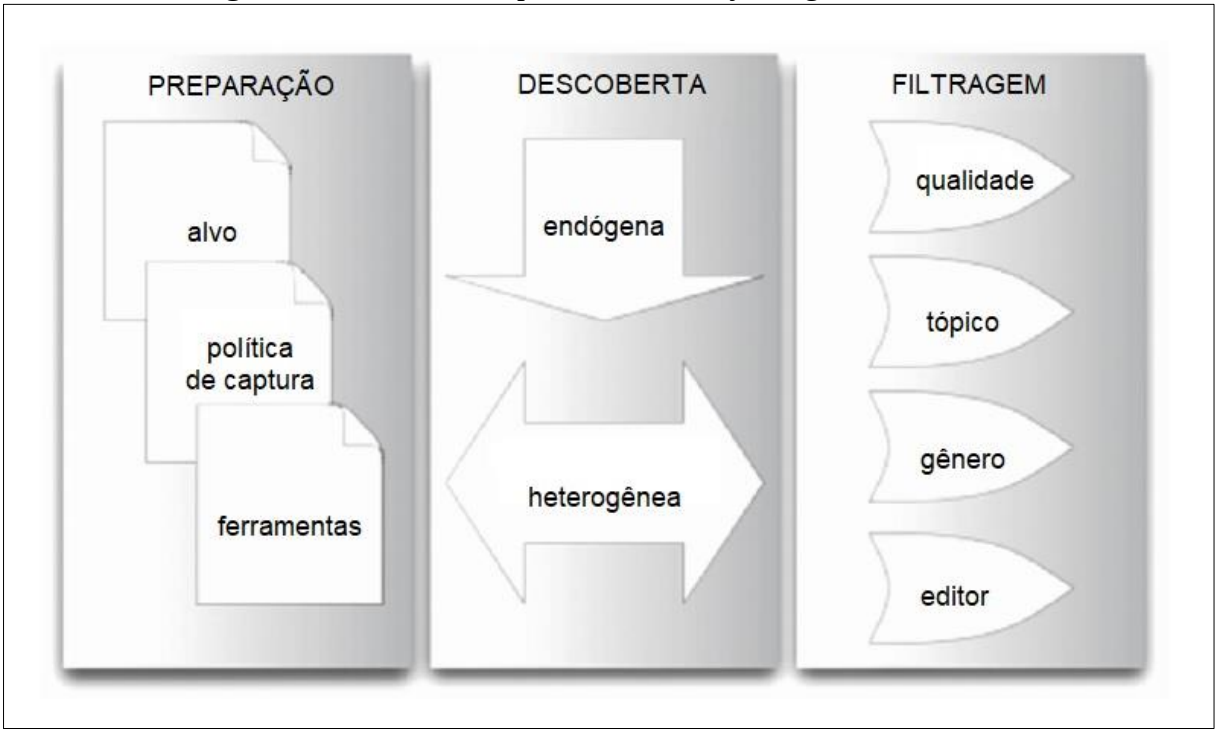
- Outros procedimentos – a definição de uma política deve incluir também os usuários pretendidos ou público-alvo, o tipo e os mecanismos de acesso, e os usos esperados do arquivo da *Web* e de suas coleções.

5.5 Seleção

Depois de definida a política de seleção, que pode ter o processo de seleção (*selection process*) e a abordagem de seleção (*selection approach*) (KRAN; RAHMAN, 2019), ela deve ser promulgada (*enacted*), como já dito antes, resultando na seleção de recursos da *Web* para coleta. Segundo Brown (c2006) isto será articulado mediante uma lista de coleta que é o produto final do processo de seleção e fornecerá a base para realização da coleta efetiva dos recursos *Web*, tendo que abranger uma clara definição dos limites (*boundary definition*) de cada recurso da *Web* na lista devido à natureza interconectada da *Web* (e seus recursos), e uma especificação do tempo e frequência (*timing and frequency*) apropriados de coleta para cada recurso da *Web*.

Também, como observado anteriormente, o processo de seleção pode ser dividido em três fases principais: preparação (*preparation*), descoberta (*discovery*) e filtragem (*filtering* ou *filtration*), segundo mostrado na Figura 77.

Figura 77 – As fases do processo de seleção segundo Masanès.

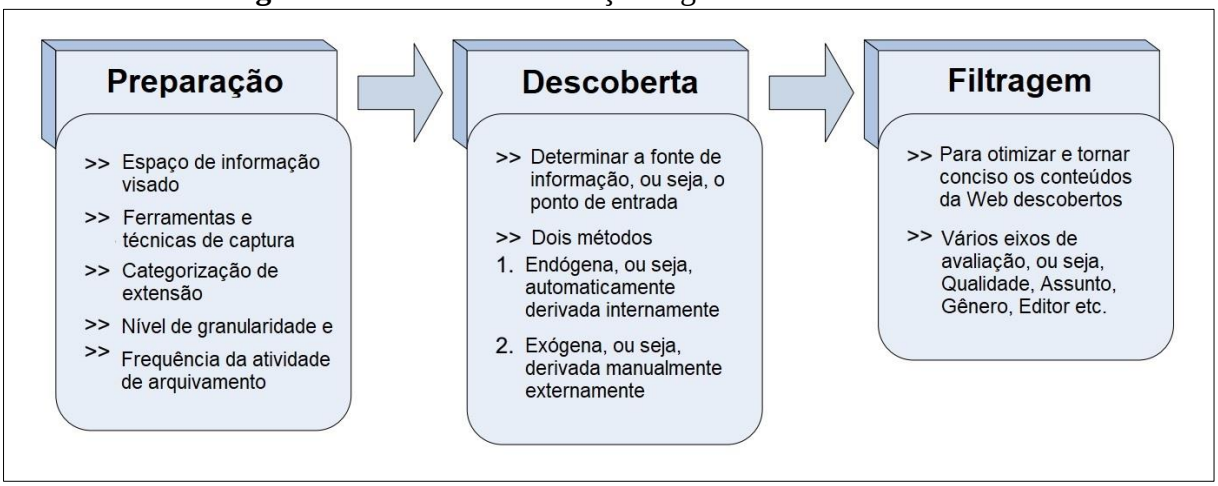


Fonte: Adaptado de Masanès (c2006a).

Em Masanès (c2006a) tais fases podem ocorrer em uma ordem sequencial ou, até certo ponto, serem misturadas, consistindo na: (I) preparação com seu produto principal – *output* – (a definição do alvo – *target* –, a política de captura – *capture policy* –, e a lista de ferramentas – *tools* – a usar); (II) descoberta endógena (*endogeneous*) e heterogênea (*heterogeneous*); e (III) filtragem de acordo com qualidade (*quality*), tópico (*topic*), gênero (*genre*) e editor (*publisher*).

Essas fases se assemelham com as subtarefas do processo de seleção sugerido por Khan e Rahman (2019) os quais, em combinação, proporcionam, até determinado grau, uma seleção qualitativa (*qualitative selection*) dos conteúdos da *Web*, conforme reproduzido na Figura 78.

Figura 78 – Processo de seleção segundo Khan e Rahman.



Fonte: Adaptado de Khan e Rahman (2019).

Para os autores, dentro de uma abordagem sistemática de preservação da *Web*, enquanto em etapas prévias os recursos *Web* são identificados sendo priorizados em função dos requisitos e consultas esperadas da comunidade destinada do arquivo da *Web* pretendido e a técnica de captura viável é identificada a partir da frequência de captura, os conteúdos *Web* agora precisam ser preparados e filtrados para seleção e uma abordagem de seleção (por exemplo, não seletiva, seletiva, de depósito etc.) exequível necessita ser selecionada baseando-se nos conteúdos *Web*.

5.5.1 Preparação

A fase de preparação objetiva "[...] definir o alvo da coleta, a política de captura e as ferramentas para implementá-la." (MASANÈS, c2006, p. 82, tradução nossa) e, de acordo com Khan e Rahman (2019, p. 79, tradução nossa), também "[...] determinar o espaço de informação alvo, a técnica de captura, as ferramentas de captura, a categorização de extensão, o nível de granularidade e a frequência da atividade de arquivamento." Sendo a chave para o sucesso de todo o processo não devendo ser subestimada em tempo e recursos exigidos para realizá-la com êxito, a preparação, tanto para coleções temáticas quanto para coleções centradas em domínios (escopo do arquivo da *Web*), requer o apoio de especialistas em domínio (*domain experts*, isto é, bibliotecários, arquivistas, acadêmicos, pesquisadores e quaisquer referências autênticas – documentos, artigos de pesquisa etc. –) que devem determinar precisamente qual é o espaço de informação alvo, como ele pode ser caracterizado em extensão e granularidade, bem como qual a frequência de captura que será aplicada, conforme Khan e Rahman (2019) e Masanès (c2006).

Também, conforme Khan e Rahman (2019) e Masanès (c2006), a fase de preparação necessita da definição e uso de algumas ferramentas que permitirão descobrir as informações pretendidas na fase subsequente de descoberta, que podem ser divididas em quatro categorias:

- *Hubs* – os *hubs* podem ser diretórios globais ou tópicos, coleção de *sites* ou uma única página *Web* etc. com *links* essenciais para um particular assunto ou tópico. Os *hubs* são mantidos por humanos e fornecem uma fonte valiosa para identificação, sendo que a sua confiabilidade, atualidade (*freshness*) e cobertura tem de ser avaliada periodicamente; ademais, o contato com a pessoa incumbida de um *hub* permitirá entender como os seus insumos (*inputs*) podem ser utilizados e a monitoração dos *hubs* durante a campanha de captura (por exemplo, se eles oferecem RSS) garante a permanência da sua relevância.
- Motores de busca (*search engines*) – essas ferramentas podem facilitar a descoberta do material da *Web* relevante, desde que possam ser definidos termos de consulta precisos relacionados a um assunto ou tópico. O uso de mecanismos de pesquisa especializados pode melhorar muito a pertinência e a atualização dos resultados, além do mais, na fase de preparação é útil definir uma lista de consultas e de mecanismos de pesquisa a serem usados e uma periodicidade de consulta e/ou um mecanismo para obter atualizações (por exemplo, os *feeds* RSS baseados em consultas ou agentes que filtram novos resultados).
- Rastreadores (*crawlers*) – estas ferramentas podem ser utilizadas para extrair de maneira sistemática (e bem definida) conteúdos da *Web*, incluindo textos, imagens, áudio, vídeo e/ou *links*, além do leiaute geral (ambiente) de uma página da *Web* ou de um *site* inteiro.
- Fontes externas (*external sources*) – as fontes externas podem ser fontes não *Web*, como materiais impressos e listas de discussão, que são monitoradas pela equipe de seleção. Estas ferramentas devem ser usadas para fornecer novos recursos e diferentes direções para a coleta pois, por exemplo, dependendo da autoridade das fontes externas, a citação de um item por esta fonte pode, por si mesmo, consistir em motivo para a sua seleção.

Assim, segundo Masanès (c2006), ao final da fase de preparação, temos como resultado: a descrição do alvo da coleção; a política de captura, incluindo o nível de aplicação, a frequência e a extensão da captura de um determinado *site* ou página da *Web*; e a lista de ferramentas que serão usadas para a descoberta e captura junto com uma descrição de como serão empregadas.

5.5.2 Descoberta

A fase de descoberta propõe-se a "[...] determinar a lista de pontos de entrada que serão usados para a captura, bem como a frequência e o escopo dessa captura." (MASANÈS, c2006,

p. 85, tradução nossa) ou, segundo Khan e Rahman (2019, p. 79, tradução nossa), objetiva "[...] determinar a fonte de informação a ser armazenada no arquivo." Esta determinação, de acordo com os autores, pode ser alcançada de duas maneiras: I) uma lista de pontos de entrada criada manualmente é usada para definir a lista de pontos de entrada (isto é, *links*/URLs) para rastrear a coleção manualmente e atualizar a lista ao longo do rastreamento da *Web*; ou II) há uma lista de pontos de entrada criada automaticamente para definir a lista de pontos de entrada, através da extração automática de *links* e da obtenção de uma lista atualizada toda vez no rastreamento.

Existe um limite bem claro entre a descoberta e o rastreamento da *Web* em si para coleta realizada manualmente (mesmo que a lista de pontos de entrada possa ser atualizada com base nos *links* descobertos no rastreamento) e, por outro lado, nas coletas realizadas automaticamente essa distinção é ofuscada em razão de que a maior parte da descoberta ocorre durante o próprio rastreamento da *Web* por extração de *links* (MASANÈS, c2006)[505]. A partir de Khan e Rahman (2019) e Masanès (c2006) observamos três métodos possíveis de descoberta, descritos a seguir:

- Endógena – abordagem usada na seleção ou coleta automática que aproveita a estrutura de *links* da *Web* para atravessar e encontrar material novo e útil. Este método depende da extração de *links* por meio de rastreadores da *Web* e é realizada a partir da exploração da lista de pontos de entrada e do ambiente de *links* de *sites* e páginas da *Web* rastreadas.
- Exógena (*exogenous*) – abordagem utilizada na seleção ou coleta manual que resulta da exploração do conjunto de ferramentas definidas e usadas na fase de preparação. Este método depende da exploração de uma lista de pontos de entrada para *hubs*, mecanismos de pesquisa e documentos não-*Web* (fontes externas).
- Heterogênea – abordagem semelhante a descoberta exógena que usa fontes (isto é, *hubs*, mecanismo de busca e fontes não-*Web*) não ligados a uma comunidade específica. Este método depende plenamente do tipo, da qualidade e da usabilidade destas fontes onde, por exemplo, a utilidade de *hubs* depende maiormente da qualidade e da atualidade da fonte, e as fontes não-*Web* requerem monitoramento específico, adaptado a cada caso.

Além disto, de acordo com Masanès (c2006) é necessário atribuir aos pontos de entrada descobertos tanto uma frequência (por exemplo, "só uma vez", diariamente ou várias vezes ao dia, semanalmente, mensalmente ou a cada x meses) como um escopo de captura (por exemplo, limitar a captura a uma página *Web* ou *sites* específicos, quer dizer, nível do *site* e da página da

[505] Sobre as diferenças entre a seleção manual e automática, veja a seção Métodos de seleção.

Web – definido como domínio, subdomínio[506] ou localização do diretório –, onde as unidades ou medidas podem estar na lista de pontos de entrada ou nas descobertas a partir delas). Para o autor isto pode ser feito individualmente ou com base no agrupamento de pontos de entrada, em nível de coleta ou de campanha de captura com a definição de um ou vários perfis de captura, sendo assim preciso moldar o processo de descoberta em consonância com a política de seleção.

5.5.3 Filtragem

A fase de filtragem objetiva principalmente "[...] reduzir o espaço aberto pela fase de descoberta aos limites definidos pela política de seleção." (MASANÈS, c2006, p. 87, tradução nossa) e, segundo Khan e Rahman (2019, p. 80, tradução nossa), também "[...] otimizar e tornar conciso os conteúdos da *Web* descobertos (espaço de descoberta)." Podendo ser combinada de forma prática ou lógica com a fase de descoberta, a filtragem se faz distintamente relevante para coletar conteúdos *Web* mais específicos, removendo os indesejados ou duplicados; ademais, a filtragem pode ser realizada manualmente ou automaticamente, onde em geral para preservação adota-se uma filtragem automática como método e, por sua vez, a filtragem manual é exigida quando os critérios usados para seleção ou a *Web* não podem ser interpretados por ferramentas automáticas ou robôs, sendo o caso de ser necessário uma caracterização de alto nível, avaliação subjetiva ou conhecimento externo, de acordo com Khan e Rahman (2019) e Masanès (c2006).

Com base em Khan e Rahman (2019) e Masanès (c2006), identificamos vários eixos (*axes*) ou critérios de avaliação que podem ser aplicados, individualmente ou em conjunto, para a política de seleção manual, a saber:

- Qualidade (*quality*) – consisti numa avaliação da autoridade (*authority*) e credibilidade (*credibility*) para recursos secundários (por exemplo, *sites* de análise, comentários etc. sobre uma campanha eleitoral presidencial) e da relevância (*relevance*) e autenticidade (*authenticity*) para recursos da *Web* primários (por exemplo, *sites* de partidos políticos).
- Assunto (*subject*) – o assunto pode ser delimitado ao redor das disciplinas tradicionais por área do conhecimento (Biologia, Artes etc.), ou um evento, indivíduo e organização específica, ou qualquer objeto em geral que será previsto a partir de várias perspectivas (por exemplo, eleições).

[506] Subdomínio (*Sub-domain*) é "[...] frequentemente usado para *microsites* ou *subsites*.", ou melhor, *sites* menores hospedados "[...] dentro ou como parte de um *site* maior.", constituindo "um diretório que prefixa o nome de domínio principal, por exemplo: *https://subdomain.domain.gov.uk/*." (THE NATIONAL ARCHIVES, [2022?a], não paginado, tradução nossa).

- Gênero (genre) – o gênero *Web* é um *site* institucional, *blogs*, páginas *Web* pessoais, fóruns, *wikis*[507] etc. Este critério pode ser tanto o principal critério de seleção para estudos de gênero quanto um critério adicional.
- Editor (*publisher*) – o editor ou o proprietário do *site* pode ser utilizado como base para a seleção. Este critério requer a análise detalhada do *site* uma vez que, na maioria das vezes, não há garantias de que as alegações de identidade sejam legítimas na *Internet*.

5.6 Documentação

Independente dos critérios adotados para a seleção manual ou automática, é imperativa a documentação (*documentation*) cuidadosa do processo de seleção, e os próprios critérios de seleção devem ser arquivados visto que eles podem evoluir com o tempo (MASANÈS, c2006a; STIRLING, CHEVALLIER; ILLIEN, 2012). Antes de tudo, as páginas *Web* renderizadas são específicas do usuário, onde ao serem exibidas podem divergirem dependendo do navegador (*browser*) usado para acessar o conteúdo, da localização geográfica do usuário etc., conforme Vishwasrao (2017). Também, consoante Dougherty *et al*. (2010), cada objeto arquivado parece ser uma representação aproximada do que estava na *Web* ao vivo (*live Web*), não se podendo verificar a sua veracidade com a *Web* ao vivo; além do mais, ao passo que a tecnologia da *Web* avança, a noção de *Web* ao vivo se faz cada vez menos estática em virtude dos objetos da *Web* (*Web objects*) serem fornecidos de maneira distinta para pessoas diferentes, e as impressões (*impressions*) ou instantâneos arquivados de artefatos da *Web* são frequentemente incompletas.

Esta incompletude (*incompleteness*) dos arquivos da *Web* é discutida, em especial, por Brügger e Finnemann (2013). Os autores explicam que, comparado com materiais digitalizados, essa incompletude diz respeito a não que coisas estejam faltando, mas sim que elas podem estar ausentes de maneiras que tornam muito difícil determinar se algo está faltando além de como o quê e onde, sendo que isto se deve a duas razões: I) em oposto de muitas coleções de dados digitalizados, o arquivo da *Web* não tem à sua disposição um original estável para comparar, a *Web* ao vivo desapareceu para sempre; e II) em um nível detalhado, a incompletude raramente é documentada, onde coisas ou conteúdos podem estar faltando sem explicações, deste modo, os arquivos da *Web* e os estudiosos que os usam, carecem de métodos confiáveis para definir se, ou até que ponto, o material de um arquivo encontra-se completo. Ao avaliar para estudo uma coleção temática de *sites* arquivados, é vital que a sua criação tenha sido cuidadosamente

[507] *Wiki* consisti em "[...] uma forma de *site* colaborativo, em que os usuários podem facilmente adicionar e atualizar conteúdo *online*." (BROWN, c2006, p. XIV, tradução nossa).

documentada, permitindo conhecer os critérios e processos para identificação do material *Web*, o período de tempo em que isso foi feito etc. (BRÜGGER, 2012; FOOT; SCHNEIDER, c2006).

Isto posto, o arquivamento da *Web* só pode alcançar uma amostragem de instanciação de conteúdo (*sampling of instantiation of content*) (MASANÈS, c2006c) em que, com o passar do tempo, o contexto original da amostra (*sample*) se perde e não restará nenhuma pista para que os pesquisadores compreendam o que o arquivo representa, em conformidade com Masanès (c2006a). Em outras palavras, este processo pode arquivar só certos momentos (*instantiations*) das páginas da *Web* – enquanto objetos dinâmicos que se apresentam em detalhes na *Web* ao vivo –, com um grau de variação possível, contrastando com as publicações que são consistentes entre todos os usuários individuais que as visualizam juntamente (DOUGHERTY *et al*., 2010; VISHWASRAO, 2017). Para contornar isso, Masanès (c2006a) menciona que é imprescindível documentar cada aspecto do processo de seleção, a ser realizado por meio das suas três fases (preparação, descoberta e filtragem), para fins de proporcionar elementos de análise no futuro.

Para Masanès (c2006a) deve-se documentar, se possível, no nível de item, as fases de descoberta e filtragem, dado que isto irá informar porque um conteúdo está ou não na coleção. O autor menciona, por exemplo, que manter uma lista de URIs que foram descobertos e filtrados será útil para entendermos no futuro como a coleção foi construída e, assim, o que ela representa comparada com a *Web* ao vivo; ou também documentar como a descoberta endógena foi feita será útil para poder refazer o caminho que foi seguido e mapear aqueles que não foram. Quanto a fase de preparação, Masanès (c2006a) destaca os seguintes aspectos a serem documentados:

- O alvo da coleção;
- A política e infraestrutura de captura (isto é, a capacidade técnica, o *software* utilizado, a prioridade, a receptividade – *politeness* – com os servidores etc.); e
- As ferramentas adotadas, incluindo nome, regularidade, contexto de uso, pessoal (*staff*) etc.

De outra forma, a catalogação e o registro (*registration*) de materiais são exigidos para encontrar itens específicos nas coleções digitais utilizando-se de identificadores (*identifiers*); e no caso dos documentos digitais, este registro é chamado de metadados, segundo Hallgrimsson (c2006) e Khan e Rahman (2019). Tido como "documentação descritiva e técnica" por Khan e Rahman (2019), os metadados são "dados que descrevem o contexto, o conteúdo e a estrutura dos registros e sua gerência ao longo do tempo" (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2001, p. 3, tradução nossa). Por sua vez, para o Arquivo da *Web* Suíça (WEBARCHIV SCHWEIZ, 2016, p. 5/6, tradução nossa), tratam-se "[...] de informações sobre dados que permitem que os dados desejados sejam acessados, trocados e gerenciados com a

maior eficiência possível.", e no domínio eletrônico, eles abrangem informações bibliográficas, técnicas e administrativas (formato e tamanho de arquivo, data de transferência, dentre outros).

Em *International Organization for Standardization* (2013), Khan e Rahman (2019), *National Digital Stewardship Alliance* (2013) e Riley (c2017), ainda que não seja consenso, os metadados podem ser divididos em sete categorias conceituais e funcionais, detalhadas a seguir:

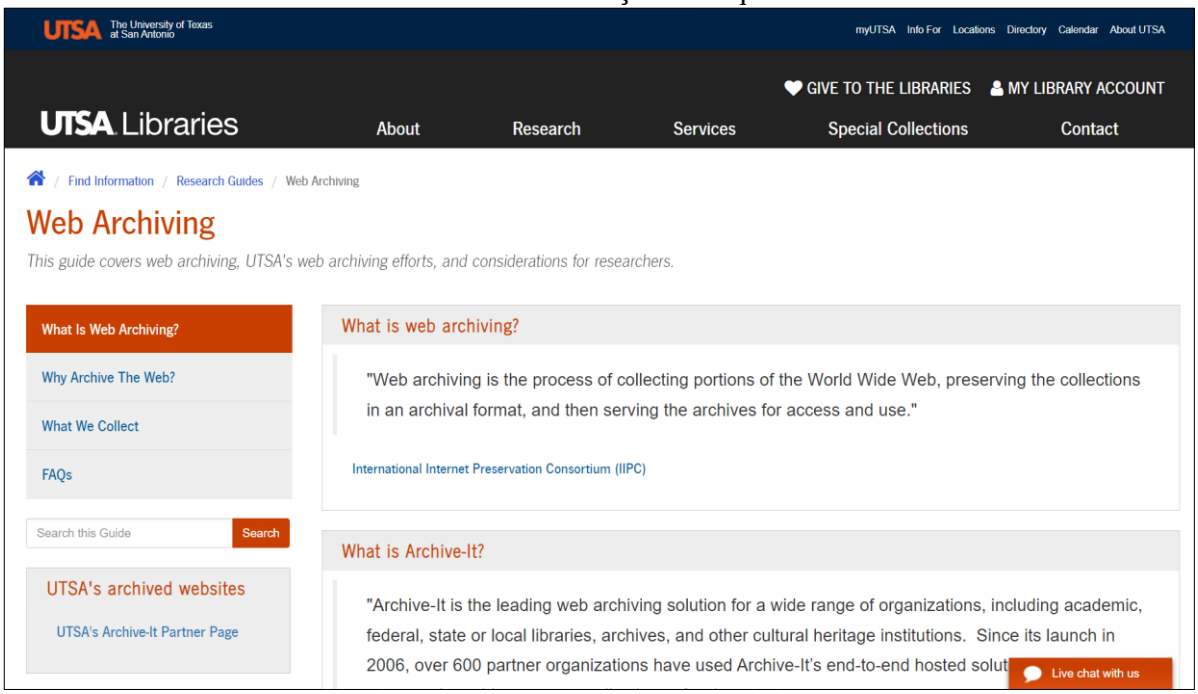
- Metadados descritivos (*descriptive metadata*) – descrevem um recurso/objeto digital e seu conteúdo intelectual para fins de descoberta e identificação. Por incluir elementos, como título, autor(es), resumo e palavras-chave.
- Metadados administrativos (*administrative metadata*) – transmitem informações para o gerenciamento interno e apropriado dos objetos em um repositório, tais como quando e onde um arquivo foi criado, quem pode acessar o arquivo e o seu tipo, dividindo-se em:
  - Metadados técnicos (*technical metadata*) – descrevem o estado técnico e o processo usado para criar ou modificar um objeto digital como, por exemplo, o seu formato de arquivo, além do *software* original e as suas configurações utilizadas no arquivo;
  - Metadados de preservação (*preservation metadata*) – informações contextuais para realizar, documentar e avaliar os processos que apoiam a retenção e a acessibilidade de longo prazo do conteúdo digital. Também definem a autenticidade do conteúdo digital, e registram a cadeia de custódia e a proveniência de um objeto digital; e
  - Metadados de gestão de direitos (*rights management metadata*) – indicam direitos de propriedade intelectual, restrições de usuário e acordos de licença que podem limitar a utilização final do conteúdo digital (incluindo arquivos de metadados).
- Metadados estruturais (*structural metadata*) – descrevem os tipos físicos e lógicos, as versões, relações ou demais características dos arquivos de conteúdo que compreendem um objeto complexo. Por exemplo, como as seções são ordenadas para formar capítulos.
- Linguagens de marcação (*markup languages*) – metadados e sinalizadores (*flags*) para outros recursos estruturais ou semânticos dentro do conteúdo, tais como o formato XML que, para Brown (c2006, p. XIV, tradução nossa), é uma linguagem "[...] de propósito geral, projetada para criar linguagens de marcação de propósito especial para descrever muitos tipos diferentes de dados [...]", destinando-se "[...] a facilitar o intercâmbio de dados, particularmente através da *Internet*."

Diante disso, os metadados de preservação conceitualmente englobam a divisão clássica dos metadados em classes descritivas, administrativas e estruturais, tornando-se claro que estes metadados precisam apoiar uma variedade de funções, dentre elas descoberta e acesso, registro de contexto e proveniência de objetos, documentação de ações e políticas de repositório (DAY,

c2006). Conforme Khan e Rahman (2019) os metadados exercem um papel fundamental na preservação a longo prazo de objetos digitais e são relevantes para identificar os metadados que podem auxiliar na recuperação de um objeto particular do arquivo após a preservação. Por isso, o autor aponta que a preservação e arquivamento digital exigem padrões de metadados a serem identificados e adotados em projetos de arquivamento da *Web* e preservação de conteúdos da *Web*, como os esquemas de metadados descritivos DC, MODS, VRA Core, de metadados de preservação PREMIS, e de codificação de metadados descritivos, administrativos e estruturais METS, com a finalidade de se localizar e assegurar o seu acesso para objetos digitais diversos.

A título de exemplo, na política de arquivos da *Web* das bibliotecas da Universidade do Texas em *San Antonio* nos Estados Unidos (UNIVERSITY OF TEXAS AT SAN ANTONIO LIBRARIES, 2016, 2022), a equipe de arquivamento da *Web* de coleções especiais não apenas identifica um tópico, assunto e/ou tema para uma coleção (se estiver criando uma nova coleção), seleciona recursos *Web* úteis e específicos para rastrear, e define a frequência de alterações ou atualizações do conteúdo e estabelece a frequência de rastreamento de maneira apropriada com isto (semanal, mensal e anual), mas também administra a mecânica de rastreamento e adiciona descrições via metadados. Esta equipe cria metadados adotando o DACS e uma adaptação dos quinze elementos do DC e, em semanas ou meses após a captura completa, os metadados do site são atualizados e melhorados para garantir a precisão das informações descritivas e facilitar a busca e recuperação do conteúdo, como *University of Texas at San Antonio Libraries* (2020).

Figura 79 – Página inicial das bibliotecas da Universidade do Texas em *San Antonio* nos Estados Unidos sobre seus esforços de arquivamento da *Web*.



Fonte: *University of Texas at San Antonio Libraries* (2023b).

Outro exemplo está no relatório "*Archival preservation of Smithsonian Web resources: strategies, principles, and best practices*" dos SIA nos Estados Unidos (SMITHSONIAN INSTITUTION ARCHIVES; DOLLAR CONSULTING, 2001), o qual contém um modelo de metadados de preservação, com requisitos baseados em princípios e exigências de manutenção de registros, melhores práticas de arquivamento e o DC, a ser usado para garantir o rastreamento e a preservação de *sites* da instituição; além de orientações sobre a documentação ou metadados dos *sites* a serem capturados na transferência para um repositório, pautando-se em uma série de perguntas sobre o *design*, uso e manutenção dos *sites*. Tal captura e manutenção de informações técnicas (metadados), segundo *Smithsonian Institution Archives* e *Dollar Consulting* (2001), ajudam a identificar o contexto tecnológico dos *sites* e incitam maior segurança do usuário em sua confiabilidade, sobretudo, se funcionalidades de um *site* não puderem mais ser suportadas.

É importante salientar que a catalogação pode se configurar num dos principais desafios financeiros e técnicos ao construir arquivos da *Web*, e a aplicação de metadados poderá ser um dos maiores desafios na criação de coleções de materiais publicados na *Web*, requerendo um conjunto de habilidades especializadas (MURRAY; HSIEH, 2007). Em Dooley *et al*. (2017), Vlassenroot *et al*. (2019) e Khan e Rahman (2019) a falta de diretrizes de metadados descritivos referentes ao arquivamento da *Web* da mesma maneira é problemática em iniciativas onde um dos seus propósitos é vincular diferentes arquivos da *Web*; deste modo, é preciso aumentar a padronização do gerenciamento de metadados junto com a criação e a seleção de metadados

pertinentes para garantir a capacidade de descoberta dos arquivos da *Web* e torná-los acessíveis aos seus usuários. Para isso, a adoção de ferramentas semiautomáticas de geração de metadados se faz determinante para o futuro levando-se em conta a complexidade da operação e os custos de originação manual dos metadados (KHAN; RAHMAN, 2019; MIRANDA; GOMES, 2009).

Sobre a abordagem de geração de metadados e a riqueza dos metadados gerados, Niu (2012a) cita que isto dependerá da escala (*scale*) do arquivo da *Web* e dos recursos disponíveis na instituição de arquivamento. Segundo a autora os arquivos da *Web* muito grandes respaldam-se normalmente na geração automática de metadados em que informações de metadados, tais como o registro de data e hora (*time stamp*) quando o recurso da *Web* foi coletado, o tamanho em *bytes*, o código de *status* (por exemplo, 404 para não encontrado), o URI, o tipo MIME (por exemplo, texto/HTML) etc., podem ser criadas ou capturadas por rastreadores ou, até, extraídas de *meta tags* das páginas HTML; por sua vez, os arquivos da *Web* de pequena escala podem criar metadados manualmente via marcação (*tagging*), comentário ou classificação (*rating*) de usuários como, por exemplo, metadados administrativos oriundos das notas criadas pela equipe de arquivamento em um processo de captura e revisão, ou metadados descritivos provenientes da contribuição de especialistas de uma área que é o tema central da coleção *Web* em questão.

Portanto, na criação de arquivos da *Web* é decisivo um planejamento estratégico eficaz que, por conseguinte, exige um processo de preservação compreensível e planejado, o qual deve resultar em um arquivo da *Web* bem organizado que englobe o conteúdo a ser preservado e as informações contextuais indispensáveis para a interpretação do conteúdo (KHAN; RAHMAN, 2019). Aliás, conforme Brügger (2005), se é importante para a análise posterior de um *site* que todos os seus elementos de expressão sejam de fato arquivados (inclusive como apareceram e foram posicionados), ou se ao arquivar a estrutura de uma página se torna ciente de alterações nela que por alguma razão são significativas (por exemplo, no caso de alguém desejar analisar o histórico do *design* de *sites* e necessitar ter certeza da construção da página), então pode ser útil combinar o emprego de *software* para arquivar *sites* inteiros com documentação estática ou dinâmica, seja sob a forma de imagens estáticas ou de uma gravação de tela (*screen recording*).

Com isto, Brügger (2005) indica que os elementos que mais tarde possam vir a faltar na estrutura arquivada serão de fato documentados. Para o autor em *sites* menores pode-se optar por documentar tudo, e em *sites* maiores certas áreas devem ser selecionadas baseando-se, por exemplo, nas áreas que são essenciais para a análise posterior, ou no caso de que certas páginas da *Web* serem especialmente relevantes, seja porque compõem o "cruzamento" (*crossroads*) de navegação, interligando muitas páginas (página inicial, as primeiras páginas de áreas menores etc.), ou porque têm alta frequência de atualização, entre outros. Acerca da decisão de se usar

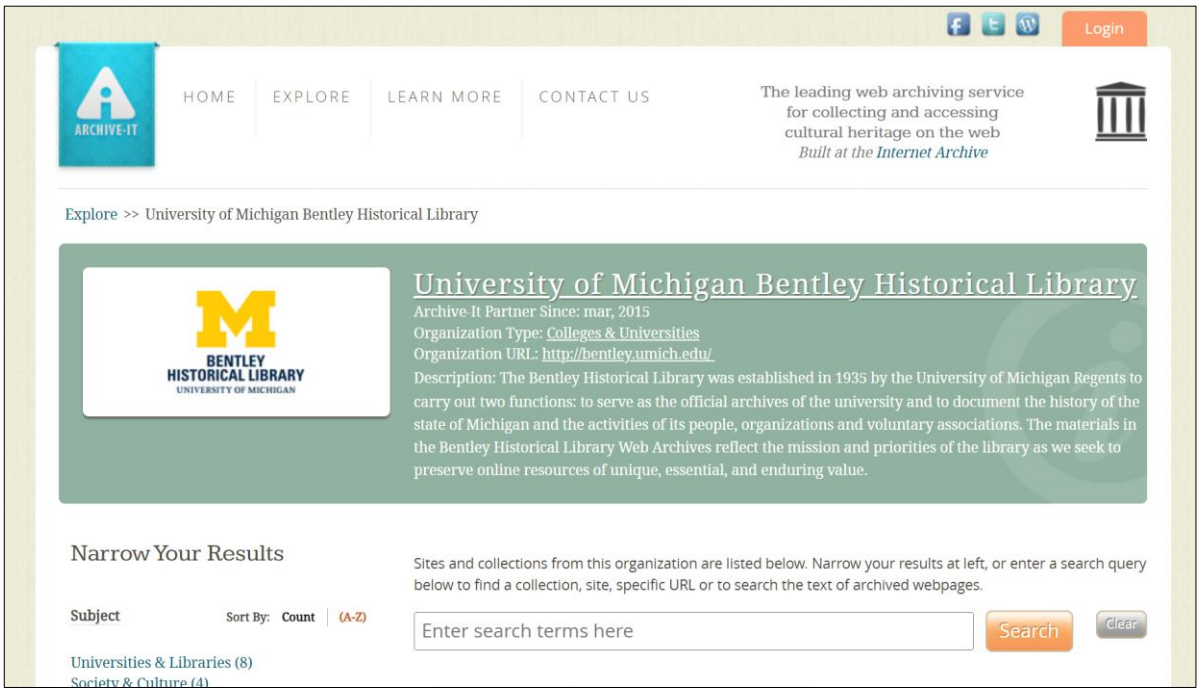
gravações ou imagens estáticas, Brügger (2005) ainda aponta que o primeiro método é rápido uma vez que somente é preciso se mover para o material *Web* a ser documentado (após isso ele é arquivado), porém a gravação pode depois ser difícil de navegar se for longa e resulta em poucos arquivos grandes; já as imagens estáticas exigem muitos processos de arquivamento únicos, são mais fáceis de navegar posteriormente, mas resultam em muitos arquivos pequenos.

5.7 Manutenção

De acordo com Brown (c2006) um programa de arquivamento da *Web* geralmente irá operar um ciclo contínuo de seleção e coleta, devendo a política de seleção (e a lista de coleta) ser mantida e atualizada regularmente para garantir a sua constante relevância e adequação ao fim que se destina, podendo isto ser aperfeiçoado pelo *feedback* dos resultados do processo de controle de qualidade da pré-coleta (*pre-collection quality control process*). Também, segundo o autor, a política de seleção não deve permanecer estática, tendo de ser atualizada para refletir as mudanças de fatores internos e externos, tais como as novas prioridades organizacionais e os evoluções na *Web*; similarmente, a lista de coleta, seja ou não parte da política de seleção, será claramente dinâmica, pois à medida que os recursos *Web* são coletados, novos recursos podem ser identificados que precisam ser considerados para a seleção ou, ainda, podem ser revelados pontos fortes e limitações em tecnologias de coleta exigindo revisões da abordagem de seleção.

A título de exemplo, na política de desenvolvimento de coleção de arquivos da *Web* da Biblioteca Histórica *Bentley* da Universidade de *Michigan* nos Estados Unidos (BENTLEY HISTORICAL LIBRARY, 2016), os arquivos da *Web* da universidade são descritos como um trabalho em andamento devido à natureza dinâmica e efêmera dos recursos *online* e, como tal, as coleções requerem manutenção/conservação regulares e reavaliação do conteúdo, assim, os arquivistas: identificam, avaliam e selecionam novos *sites* com base nos critérios e prioridades definidas; realizam a garantia de qualidade em todas as capturas para determinar o sucesso das mesmas e verificar a exatidão (*accuracy*) dos URLs alvo; e atualizam e melhoram os metadados do *site* para assegurar a exatidão da informação descritiva e facilitar a recuperação de conteúdo.

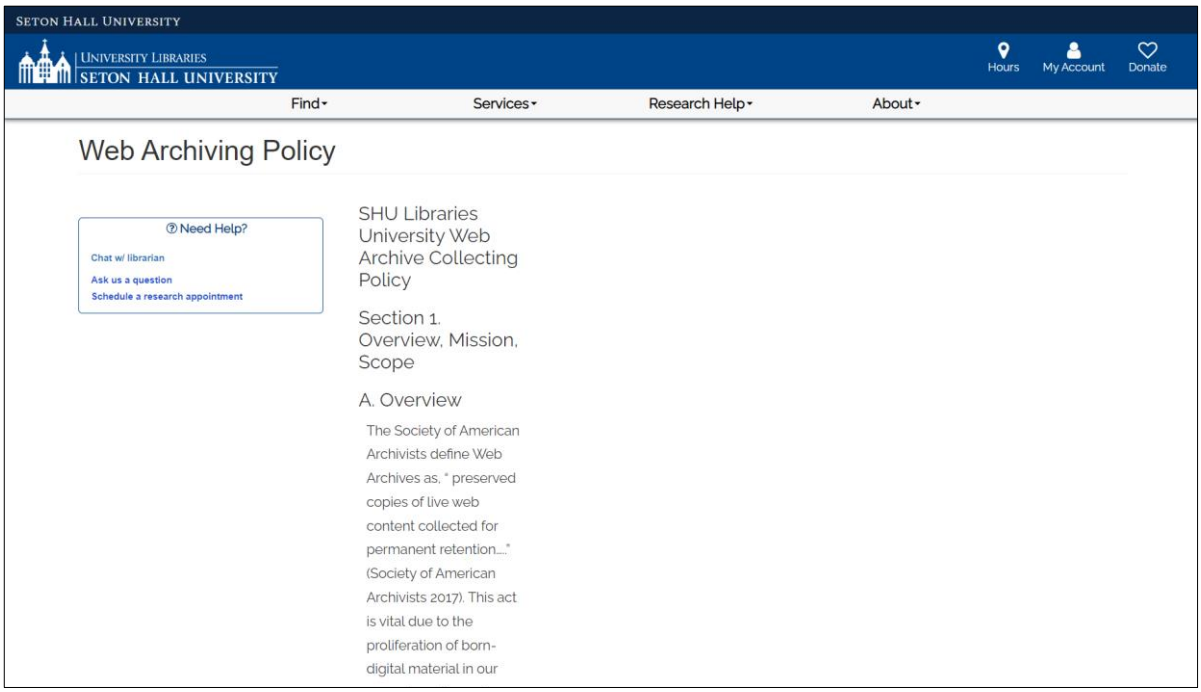
Figura 80 – Página inicial da coleção de arquivos da *Web* da Biblioteca Histórica *Bentley* no *Archive-It*.



Fonte: *Bentley Historical Library* ([2023?]).

Outro exemplo está na política de coleção do arquivo da *Web* das bibliotecas da *Seton Hall University* (SHU) nos Estados Unidos (SETON HALL UNIVERSITY LIBRARIES, [2022?]), que contém páginas da *Web* de valor histórico para esta universidade, onde as coletas de *sites* existentes são revisadas trimestralmente a fim de verificar se houveram mudanças significativas nos *sites* para iniciar a re-coleta; a lista atual de *sites* coletados é revisada periodicamente, sendo que possíveis acréscimos a essa lista específica são discutidas; e os metadados descritivos de rastreamentos novos e existentes são igualmente mantidos atualizados com novas informações.

Figura 81 – Página inicial da política de arquivamento da *Web* das bibliotecas da Universidade *Seton Hall* nos Estados Unidos.



Fonte: *Seton Hall University Libraries* ([2022?]).

À vista disto, em acordo com Brown (c2006), é fundamental que a política de seleção inclua procedimentos para manter a atualidade da política, converter as decisões de seleção em atualizações da lista de coleta, e aprovar e implementar alterações; além do mais, a regularidade com que essa manutenção (*maintenance*) deve ser realizada dependerá do método de seleção empregado e da frequência da coleta pois, para o autor, em um programa de arquivamento da *Web* que esteja envolvido numa coleta de alta frequência terá provavelmente de reexaminar as suas políticas com mais regularidade do que aquele que meramente executa coletas esporádicas.

5.8 Contexto de seleção

Sobre o contexto em uma política de seleção de arquivamento da *Web*, Brown (c2006) indica o seguinte:

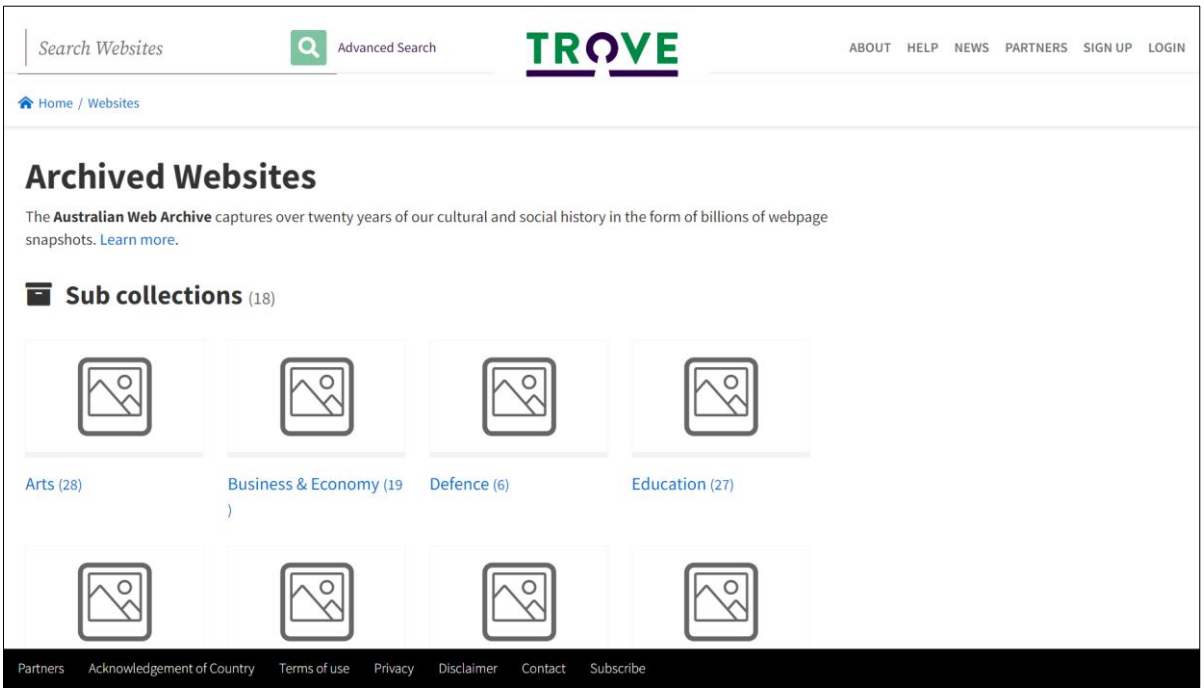
- Ao passo que o arquivamento da *Web* surgiu, nas últimas décadas, como um assunto de grande preocupação para uma série de organizações em nível internacional, é importante que as políticas individuais de seleção organizacional sejam vistas num contexto global.
- É raro que qualquer organização tenha a capacidade ou o anseio de realizar uma coleção de fato abrangente da *Web*, e a natureza interconectada da *Web* traz implicações para o contexto e a integridade de cada recurso da *Web* coletado. Como exemplo, os recursos armazenados num arquivo da *Web* podem depender de outros recursos *Web* mantidos

por um arquivo externo para prover informações contextuais/conteúdo, ou seja, o *link* externo de um *site* arquivado (que antes seria inativo) poderá ser um *link* para uma cópia arquivada do *site*, coletada por outra instituição ao mesmo tempo. Tais questões podem ser enfrentadas pela construção de políticas de seleção colaborativas e complementares.

- Enquanto as decisões de seleção em qualquer organização com um mandato curatorial pautam-se nas demandas observadas de um determinado grupo de usuários, as políticas de seleção são na prática motivadas pela demanda local e não por uma consciência de bem global, porém existem iniciativas de políticas de seleção colaborativa como, por exemplo, no UKWA que atendem às exigências de instituições-membros individuais e às de outras instituições, os quais expressam um otimismo da possibilidade de encontrar políticas de seleção em um contexto bem mais amplo do que os limites das organizações.
- A política pode ser criada no contexto de uma política de seleção já existente em toda a organização ou políticas de seleção similares para outros tipos de recursos. Por exemplo, uma instituição patrimonial – biblioteca, arquivo etc. – pode ter uma política de seleção existente para publicações periódicas com a qual a nova política precisará se associar.

Exemplificando, na política de desenvolvimento de coleções da Biblioteca Nacional da Austrália (NATIONAL LIBRARY OF AUSTRALIA, [2022b]) é especificado a política de coleta de *sites* (ou intenção de coleta – *collecting intent* –) junto com a política de coleta de todos os outros tipos de materiais que a instituição procura coletar (livros, música, fotos etc.), ademais, o arquivo da *Web* australiano (NATIONAL LIBRARY OF AUSTRALIA, [2022a]), que consiste das coleções do PANDORA *Archive* (com coletas seletivas por tema ou evento), do *Australian Government Web Archive* (com coleção de *sites* do governo da *Commonwealth*) e da coleção *Australian Domain Harvest* (com conteúdos coletados nos domínios australianos *.au*), resulta de um esforço colaborativo com outras instituições culturais parceiras (bibliotecas estaduais, territoriais etc.) com o objetivo de aumentar a coleção eficiente de materiais da *Web*.

Figura 82 – Página inicial das coleções do arquivo da *Web* australiano através do serviço de descoberta *Trove*.



Fonte: *National Library of Australia* ([2023?a]).

Assim, a percepção do contexto mais geral em que a política de seleção de conteúdo da *Web* irá funcionar é um fator primordial para a elaboração da política em si (BROWN, c2006).

5.9 Métodos de seleção

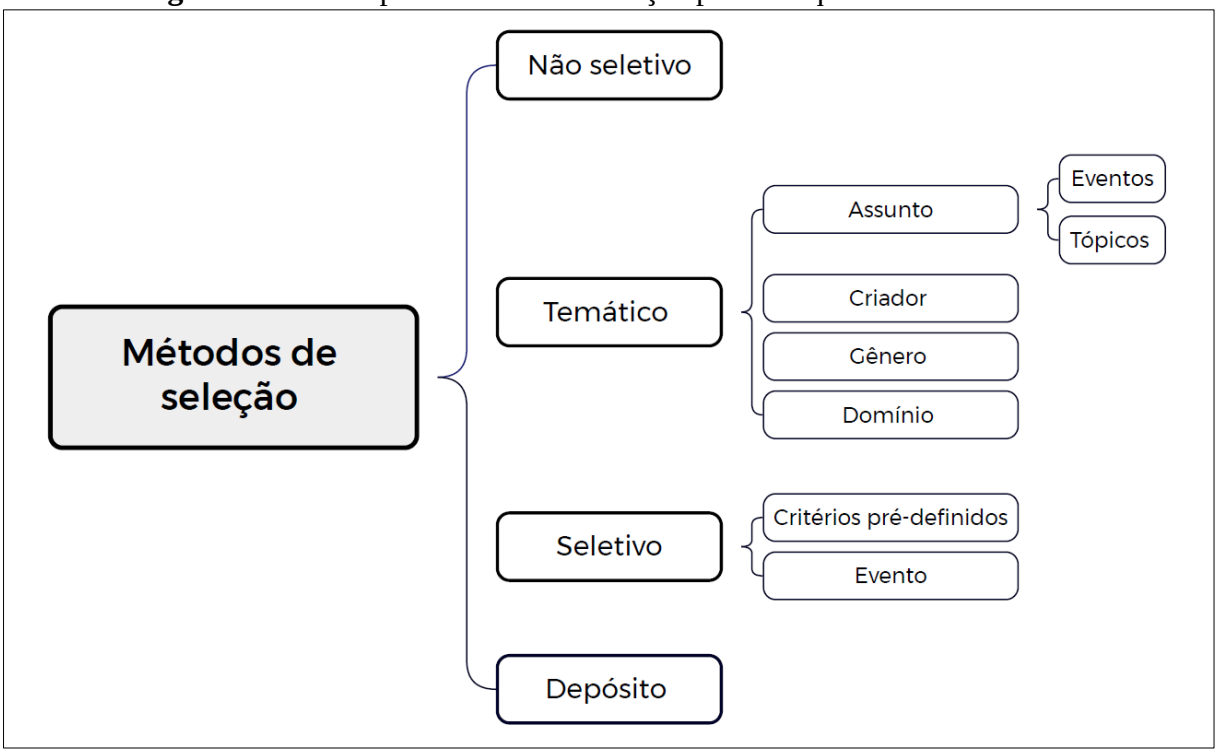
Há várias abordagens diferentes de seleção que podem ser categorizadas em função do seu escopo, e cada método de seleção implica em algum grau de conflito de decisão (*trade-off*) entre a amplitude (*breadth*) e a profundidade (*depth*) da coleta, do qual as escolhas podem ser condicionadas por questões de recursos, incluindo as tecnologias de coleta disponíveis para a instituição coletora/de arquivamento (BROWN, c2006). Esta abordagem de seleção (*selection approach*) poderá ser automática ou manual. Se por um lado a seleção manual é rara e muito trabalhosa por requerer ferramentas automáticas para encontrar o conteúdo *Web* e, em seguida, a revisão manual dessa coleção para identificar o subconjunto que deve ser capturado, por outro a seleção automática é amplamente adotada em políticas de projetos de coleta, preservação e arquivamento da *Web*, em consonância com Day (2003b), Khan e Rahman (2019) e Melo (2020).

Sobre a oposição entre a seleção manual e a coleta automática em massa (*bulk automatic harvesting*), Masanès (c2006a) pontua que a primeira é falsamente apenas manual e a segunda é erroneamente tida como abrangente, pois o arquivamento da *Web* sempre supõe alguma forma

de seletividade ainda que seja executado em larga escala e por ferramentas automáticas, sendo que essa seletividade e o determinismo das ferramentas (que tem impacto suficiente na coleção final resultante) ocorrem na descoberta e na captura do material. Para o autor a abrangência em oposição à seletividade é um mito, porque: I) o tamanho e a versatilidade da *Web* não permitem descobrir e capturar todos os instantâneos possíveis de conteúdo para todos os usuários; II) há uma seletividade padrão dos rastreadores de grande escala relativo à extensão, profundidade e tempo de rastreamento de *sites* (que dependem da capacidade de extrair *links*, pontos de entrada etc.); e III) a seleção manual de conteúdos dificilmente ocorre sem exigir o uso de ferramentas automáticas de descoberta, como motores de busca, e mesmo que a descoberta seja inteiramente manual, a captura é feita geralmente com o uso de ferramentas baseadas na extração de *links*; aliás, essas ferramentas tem pelo menos um viés de captura incluído, como definição de escopo, exclusões implícitas ou explícitas de conteúdo por tipo de formato, priorização de captura etc.

Em Brown (c2006), Kran e Rahman (2019) e Murray e Hsieh (2007) identificamos um conjunto de quatro métodos principais para selecionar conteúdos *Web* para o arquivamento da *Web*, conforme exposto na Figura 83.

Figura 83 – Principais métodos de seleção para o arquivamento da *Web*.



Fonte: Elaborado pelo autor.

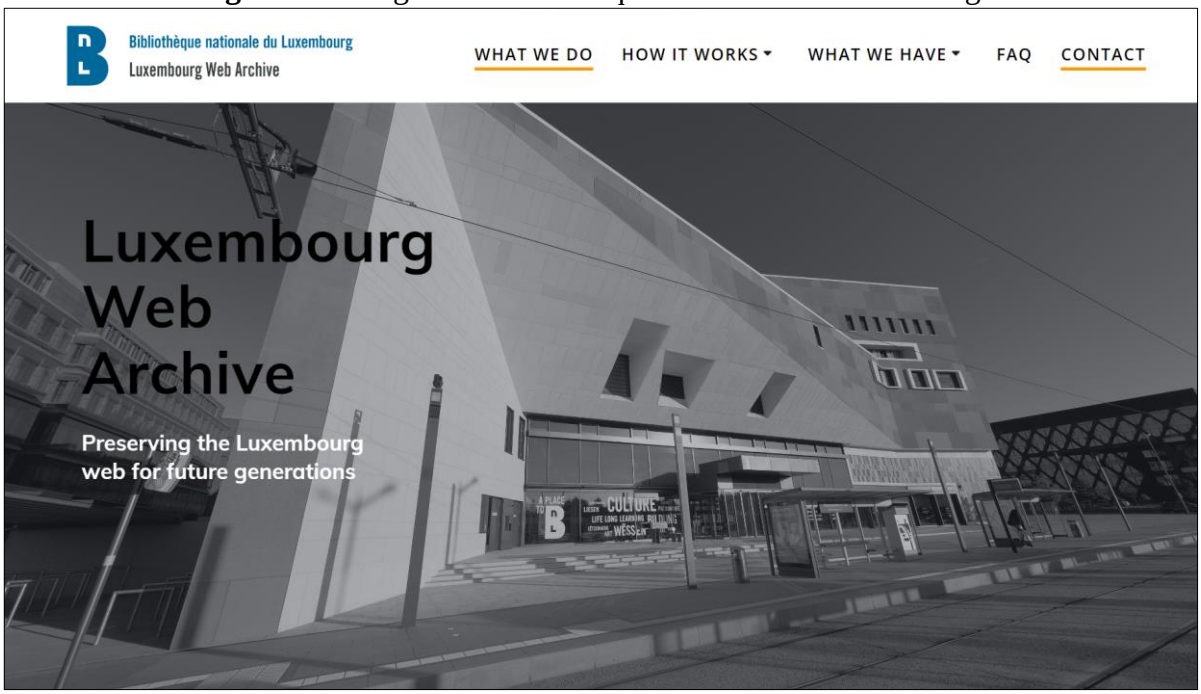
Respaldando-se nos autores supramencionados cada um dos métodos para a seleção de conteúdos num arquivo da *Web*, indicados na Figura 83, podem ser descritos da seguinte forma:

- Não seletivo (*unselective*) – abordagem que consiste em coletar tudo o que for possível, onde todo o *site* e seus domínios e subdomínios associados são baixados para o arquivo da *Web*. Além de ser referida como seleção de domínio (*domain selection*) e seleção em massa (*bulk selection*), também é chamada de seleção ou coleta automática (*automatic harvesting or selection*) em que um rastreador faz a coleta. É uma abordagem de coleta tecnicamente mais barata e rápida e produz um quadro amplo da *Web* como um todo, porém gera enormes volumes de dados não triados, duplicados e potencialmente inúteis consumindo muitos recursos, e pode ter implicações legais se feita em um contexto onde a permissão para coletar é exigida. Por exemplo, o *Internet Archive* usa essa abordagem se baseando em argumentos, como: a natureza interconectada da *Web*, o qual implica que o contexto completo de algum *site* acaba por abranger toda a *Web* e, assim, só pode ser preservado se tudo for coletado; o fato da seleção ser um processo caro e demorado e que, por isso, é mais simples e mais eficiente evitar decisões específicas de seleção; e que todo processo de seleção é subjetivo e prejulga a relevância que será dada a recursos da *Web* específicos pelas gerações futuras e, sendo assim, tal abordagem é inteiramente objetiva e não restringe as possíveis exigências dos futuros pesquisadores. No entanto, a iniciativa não é atualmente capaz de coletar todo o conteúdo da *Web* mundial, como a *deep Web*[508], e o *Internet Archive* admite que essa abordagem deve ser completada por métodos de coleta mais focados e seletivos, permitindo maior profundidade e qualidade.
- Temático (*thematic*) – abordagem alternativa que envolve definir critérios de seleção temática e que é considerada semisseletiva na medida em que cada recurso *Web* dentro do tema poderia potencialmente ser coletado. Trata-se de um método de coleta vantajoso no sentido que ele pode reduzir o escopo da coleta a um tamanho mais administrável do que a abordagem não seletiva, e também é muito provável que, ao coletar todos os recursos *Web* tematicamente relacionados, parte do contexto mais significativo de cada recurso da *Web* será preservado. Exemplificando, no arquivo da *Web* de Luxemburgo da *Bibliothèque Nationale du Luxembourg* ([c2022]) é realizado alguns tipos de coleta da *Web*, incluindo rastreamentos direcionados (*targeted crawls*) que buscam coletar o máximo de informações possíveis sobre um tópico ou evento específico dentro de um número seletivo de endereços durante um período limitado, mas esta abordagem é mais

[508] *Deep Web* remete a "parte da *Web* que não pode ser rastreada e indexada por motores de busca, particularmente constituída de recursos que são gerados dinamicamente ou protegidos por senha" (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2013, não paginado, tradução nossa).

demorada na definição da lista de sementes (*seed list*[509], ou melhor, a lista de todas as sementes – *seed*, ou endereço de URL usado como ponto de partida para rastreamentos da *Web* – que foram usadas para construir uma coleção) e na definição da profundidade e frequência das coletas.

Figura 84 – Página inicial do arquivo da *Web* de Luxemburgo.



Fonte: *Bibliothèque Nationale du Luxembourg* ([c2023?]).

Pode-se prever uma variedade de abordagens temáticas, como:

- Assunto (*subject*) – a seleção se baseia em cima do assunto, como eventos e tópicos. Por exemplo, nos métodos do arquivo da *Web* de Luxemburgo (BIBLIOTHÈQUE NATIONALE DU LUXEMBOURG, [c2022]), além de rastreamentos de domínio (*domain crawls*) e coleta de eventos, temos também coleções temáticas que cobrem um tópico/área de interesse específica com maior prioridade para o arquivo (e isto pode estar ligado à relevância do tópico ou ao ritmo da mudança de informações); e na política de desenvolvimento de coleções da Biblioteca Nacional da Austrália (NATIONAL LIBRARY OF AUSTRALIA, [2022b]), é indicado que a biblioteca aborda a coleta da *Web* por meio de determinadas atividades, incluindo coleções temáticas com uma cobertura centrada em eventos, questões prioritárias e setores

[509] *Seed* (ou URL de semente – *seed* URL –) corresponde a "[...] um URL que o rastreador é instruído a capturar." (THE NATIONAL ARCHIVES, [2022?a], não paginado, tradução nossa), ou, para *Library of Congress* ([2022?c], não paginado, tradução nossa), diz respeito ao "[...] ponto de entrada ou de partida do rastreador e o ponto de acesso dentro do arquivo.", sendo "[...] tipicamente o URL selecionado para arquivamento [...]" onde "o rastreador segue os *links* das páginas do URL de semente para as páginas subsequentes."

selecionados a fim de capturar assuntos de importância nacional, como campanhas eleitorais federais e desastres (por exemplo, incêndios florestais de 2019-2020 etc.).

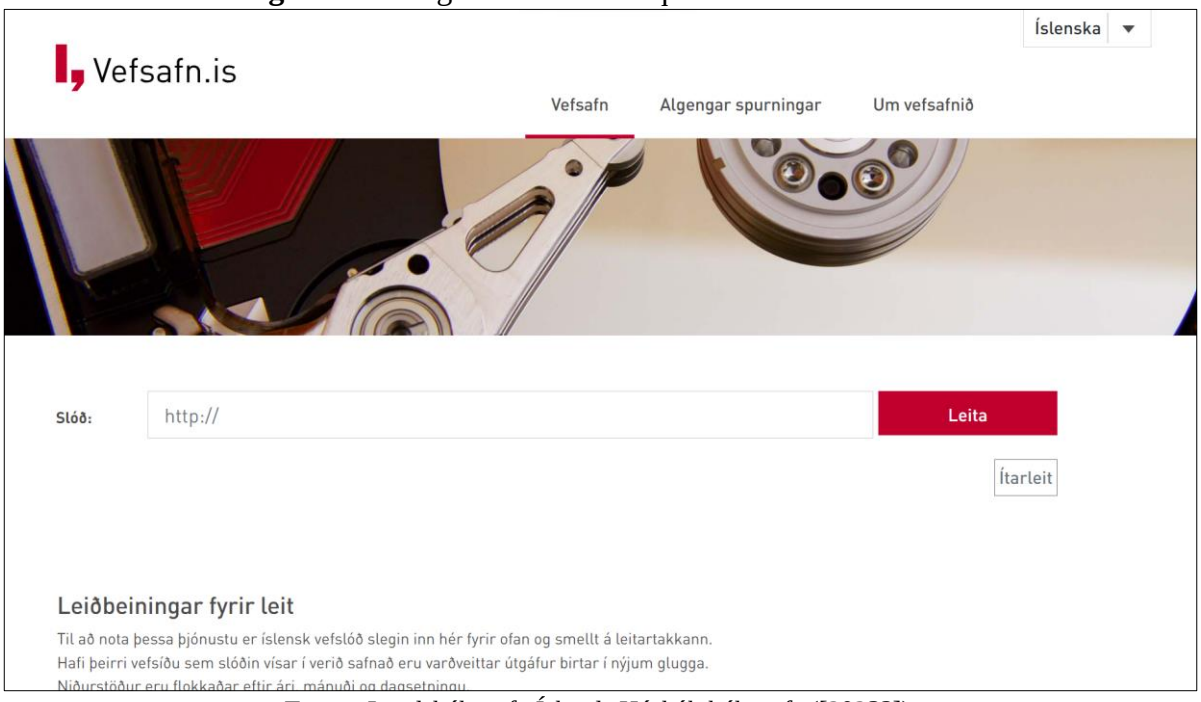
- Criador (*creator*) – a seleção é definida a partir do indivíduo e/ou organização responsável por criar ou gerir um recurso *Web* particular como, por exemplo, uma editora ou agência governamental. Exemplificando, no UKGWA dos *The National Archives* ([2022?a], [2022b]) preserva-se os *sites* e conteúdos públicos de contas de mídia social (por exemplo, *Twitter*, *Flickr* e *Youtube*) do governo central do Reino Unido, sendo que são coletados registros dos departamentos e órgãos do governo, entre eles: material publicado pelos departamentos de Estado, seja em sua própria família de *sites* ou no *gov.uk*; agências; órgãos públicos não departamentais; *sites* do *National Health System* (NHS) com foco nacional; consultas públicas; e outros.
- Gênero (*genre*) – o escopo da seleção se baseia em gêneros específicos de recursos *Web*, como publicações, *blogs*, registros governamentais etc. Exemplificando, no arquivo da *Web* Húngaro da Biblioteca Nacional da Hungria (em húngaro *Országos Széchényi Könyvtár* – OSZK *Webarchívum*) há coletas por localização geográfica, tópico e/ou gênero, do qual nesta última se inclui a seleção de periódicos eletrônicos (revistas *online*, jornais, boletins, portais de notícias etc.), páginas do *Facebook*, canais de *podcast*, entre outros (ORSZÁGOS SZÉCHÉNYI KÖNYVTÁR, c2022).

Figura 85 – Página inicial do arquivo da *Web* Húngaro.

Fonte: *Országos Széchényi Könyvtár* (c2023).

- Domínio (*domain*) – o escopo da seleção é determinado em função de domínios *Web* específicos, tais como *.uk* e .br, do qual o termo domínio tem um sentido que vai além do puramente técnico onde, por exemplo, o domínio do governo britânico abrange *sites* em uma série de domínios *Web*, como *.gov.uk* e *.org.uk*. Por exemplo, no arquivo da *Web* islandesa da Biblioteca Nacional e Universitária da Islândia (LANDSBÓKASAFN ÍSLANDS HÁSKÓLABÓKASAFN, [2022]), é coletado o conteúdo hospedado sob o domínio *.is* com o código do país da Islândia, além de *sites* selecionados sob domínios internacionais *.com*, *.net* etc. e domínios nacionais, como *sites* registrados em *.dk* com o código do país da Dinamarca; e no arquivo da *Web* de Luxemburgo (BIBLIOTHÈQUE NATIONALE DU LUXEMBOURG, [c2022]), se realiza rastreamentos bianuais e em grande escala de todos os domínios .lu (*bnl.lu*, *data.bnl.lu* etc.), mas estes rastreamentos de domínio são bastante lentos na captura de *sites* que estão mudando rápido ou desapareceram entre duas coletas.

Figura 86 – Página inicial do arquivo da *Web* islandesa.



Fonte: *Landsbókasafn Íslands Háskólabókasafn* ([2023?]).

- Seletivo (*selective*) – abordagem mais restrita que consiste em identificar recursos para coleta, como uma única publicação *Web* ou *site*. Visando preservar recursos ou porções específicas da *Web* com base em um conjunto de critérios, estratégias e/ou parâmetros predefinidos para materiais em coleções e nos acessos e informações fornecidas pelo arquivo da *Web*, a decisão de inclusão sob esta abordagem poderia ser tomada no nível

de *site* (isto é, os *sites* de um domínio selecionado, como arquivar *sites* educacionais do domínio de alto nível *.pk*); de página *Web* (ou seja, as páginas de um *site* selecionado, como arquivar as páginas iniciais – *homepages* – de todos os *sites* educacionais); e de conteúdo da *Web* (isto é, o conteúdo a ser preservado, como arquivar todas as imagens das *homepages* dos *sites* educacionais). É uma abordagem útil para a gestão de direitos de propriedade intelectual (por exemplo, nos casos em que são necessárias permissões explícitas para a coleta) e a compreensão mais detalhada das propriedades e qualidades dos recursos individuais coletados beneficiando outros processos (por exemplo, garantia de qualidade, catalogação, preservação e entrega); e quando o número de *sites* a serem arquivados for muito grande, ou se o arquivamento tiver como alvo toda a *Web* e quiser reduzir o escopo identificando os recursos em que há mais interesse, ou para iniciar um projeto piloto de preservação identificando o que é possível e pode ser gerido. Porém, a coleta de recursos isoladamente acentua o problema da preservação do contexto, o que exige atenção à definição de limites para assegurar a coleta de recursos significativos e completos. Ademais, quanto maior for o grau de seletividade usado, mais subjetiva será a coleta resultante, coibindo as exigências ainda desconhecidas dos futuros usuários, e esta abordagem faz suposições implícitas/explícitas quanto ao material da *Web* que não será selecionado para preservação e que, como tal, poderá ser perdido para a posteridade. Exemplificando, no arquivo da *Web* dinamarquesa (*Netarkivet*), a Biblioteca Nacional da Dinamarca (em dinamarquês *Det Kongelige Bibliotek*) coleta só material disponível na *Internet* e se utiliza de diferentes estratégias de coleta, entre elas a coleta seletiva de vários tipos de *sites* como, por exemplo, todos os meios de comunicação dinamarqueses, partidos políticos, associações, ministérios e agências, vídeos do *Youtube* e perfis selecionados de mídia social, além da coleta de dois ou três eventos anualmente como, por exemplo, a pandemia do coronavírus (DET KONGELIGE BIBLIOTEK, [2022b]).

Figura 87 – Página inicial do arquivo da *Web* dinamarquesa.

Fonte: *Det Kongelige Bibliotek* ([2023?]).

Assim, a abordagem seletiva pode ser baseada num critério pré-estabelecido ou evento:

- Critério (*criteria*) – envolve a seleção de recursos *Web* com base em vários critérios pré-definidos. Dependendo do objetivo geral de preservação, um critério simples ou complexo de seleção de conteúdo pode ser definido como, por exemplo, todos os recursos pertencentes a uma organização, todos os recursos de um gênero (*blogs* de ciência etc.), recursos *Web* que tratam de uma comunidade específica em uma instituição (funcionários, ou discentes, professores etc.), todas as publicações de uma organização individual ou grupo de organizações, todas os recursos que podem beneficiar usuários externos (historiadores, pesquisadores, ex-alunos etc.), e outros.
- Evento (*event*) – consiste na seleção de *sites* ou recursos da *Web* baseados em vários eventos temporais, sendo que os *sites* baseados em eventos têm duas características: atualizações muito frequentes; e o conteúdo do *site* é perdido após um curto período (algumas semanas ou meses) como, por exemplo, o início e o fim de um mandato ou ano acadêmico, a duração de uma atividade (projeto de pesquisa etc.). Aliás, os arquivistas podem se centrar em *sites* que abordam eventos importantes nacionais ou internacionais, tais como desastres, eleições, a Copa do Mundo de Futebol etc. Exemplificando, no arquivo da *Web* Dinamarquesa, as coleções de eventos coletam material da *Internet* sobre eventos que são considerados como tendo um impacto significativo na história da Dinamarca, e um "evento" é definido pelo fato deste

receber muita atenção do público que aciona novos *sites* e, também, é em grande parte processado em *sites* já existentes, podendo ser "previsível" (*predictable*) (por exemplo, eleições parlamentares/municipais etc.) e "imprevisível" (*unpredictable*) (por exemplo, desastres naturais etc.) (DET KONGELIGE BIBLIOTEK, [2022a]).

- Depósito (*deposit*) – abordagem que preserva materiais depositados em um arquivo da *Web* por editores com base em requisitos de depósito exigidos pelo governo ou acordos voluntários (por exemplo, uma editora comercial de periódicos acadêmicos que deposita o seu conteúdo publicado em uma agência de arquivo, como uma biblioteca ou arquivo nacional). Neste método, o pacote de informações (*information package*), contendo uma cópia do *site* com arquivos associados que podem ser acessados através de diferentes *hyperlinks*[510], é submetido pelo administrador ou proprietário do *site* e pode ser aplicável a uma pequena coleção de alguns *sites* ou, caso contrário, o proprietário do *site* pode iniciar o projeto de preservação como, por exemplo, uma empresa que inicia um projeto de preservação de seu *site*. Devido ao papel ativo dos editores, autores, proprietários etc. de conteúdo nesta abordagem específica, ela pode vir a oferecer uma solução potencial para a captura de conteúdo da *deep Web*, o qual não pode ser capturado por rastreadores da *Web*. Exemplificando, no arquivo da *Web* Dinamarquesa, a Biblioteca Nacional da Dinamarca é responsável por coletar e preservar a parcela dinamarquesa da *Internet* como parte da Lei Dinamarquesa de Depósito Legal, sendo que com a Lei atual e vigente nº 636[511], de 2005, o depósito legal foi estendido para incluir outras publicações além da impressa, como mais recentemente material de *Internet*; ademais, a biblioteca orienta as editoras, autores e outros sobre o que e como entregar (*upload*) as publicações e como realizar o depósito legal digital (DET KONGELIGE BIBLIOTEK, [2022b], [2022c]).

Para além desta tipologia de métodos de seleção, Pennock (c2013), ao considerar que as políticas de seleção para os conteúdos da *Web* são normalmente concordantes com políticas mais amplas de coleta organizacional, delimita os escopos das coleções da *Web* baseadas em coleta de domínio e no arquivamento seletivo, destacando os seus pontos fortes e fracos, como:

- Coleções de domínio (*domain collections*) – visam reunir todos os *sites* associados a um país (isto é, coleções de domínios nacionais), podendo incluir *sites* que terminam com o sufixo de domínio nacional (.br etc.), *sites* hospedados naquele país com um sufixo de domínio diferente (.*com* etc.) e *sites* hospedados no exterior cujo conteúdo se centra na

[510] *Hyperlink* (ou *link*) refere-se a "estrutura de relacionamento utilizada para vincular informações na *Internet*" (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2013, não paginado, tradução nossa).
[511] Disponível em: https://www.retsinformation.dk/eli/lta/2005/636. Acesso em: 3 jun. 2023.

nação coletora. São coleções mais abrangentes, porém as limitações atuais da tecnologia de arquivamento fazem com que os *sites* nestas coleções (em especial, os maiores e mais complexos) estejam muitas vezes incompletos (por exemplo, renderizações incorretas e a extensão total do *site* não capturada). Todavia, o escopo total destas coleções permite que as relações com outros *sites* e o conteúdo externo vinculado sejam mais prováveis de serem mantidos se comparado com um *site* arquivado dentro de uma coleção seletiva.

- Coleções seletivas (*selective collections*) – procuram selecionar *sites* individuais para inclusão numa coleção conforme a sua relevância para a política de coleta do órgão ou instituição. Tomado a forma de "coleções especiais" sobre um tema, assunto ou evento específico, são coleções que focam esforços em *sites* tidos particularmente valiosos e, ao passo que essa medida de valor discutível exige que a qualidade dos *sites* arquivados alcance um nível mínimo, é mais provável que os *sites* estejam "completos", ou melhor, que será mais provável que os arquivos estejam presentes e renderizados de maneira correta, ainda que os *links* para *sites* externos sejam mais propensos a serem quebrados.

Também Pennock (c2013) destaca outros problemas sobre os métodos de seleção em coleções *Web*: I) os limites artificiais que as coleções com escopo impõem, pois os *sites* dessas coleções se conectam a outros *sites* que não são capturados como parte de uma coleção, o que pode ser frustrante aos pesquisadores que irão ter *links* quebrados; II) o provável ou involuntário e não reconhecido viés de seleção dos arquivos da *Web* seletivos, já que a seleção de *sites* é um processo manual que reflete os interesses particulares ou o saber da pessoa que escolhe os *sites* para a coleção e que, pela dinamicidade da *Internet*, pode não se manter a par das novas fontes (sobretudo, para coleções de eventos) e, por isso, as coleções seletivas estão sujeitas ao risco de serem tendenciosas sem intenção e seu valor de pesquisa restrito, mas isto pode ser minimizado ao manter informações dos selecionadores e seus interesses; e III) a sobreposição de seleções para instituições de coleta da mesma nacionalidade, porque traz o questionamento de qual será o custo-benefício de várias instituições arquivarem o mesmo *site*, se isso favorece os usuários e qual o seu impacto sobre o proprietário do *site*, porém isto pode ser auxiliado por uma política clara de arquivamento que controle estas questões e trate-as de um modo mutuamente benéfico.

Uma política de seleção que resolva ou amenize os problemas supracitados, seja pela interconectividade, tamanho e dinamicidade da *Web* ou pelos escopos dos diferentes métodos e abordagens de seleção aplicados na construção de coleções da *Web*, necessita da especificação precisa e explícita aos usuários-pesquisadores, proprietários, autores e editores de *sites* etc. dos critérios adotados na seleção de conteúdos para o arquivo da *Web* (além da definição dos limites dos *sites* arquivados como do arquivamento da *Web* desenvolvido), a serem discutidos a seguir.

5.10 Critérios de seleção

Depois de identificado o método de seleção apropriado, Brown (c2006) sinaliza que este deve articular-se como um conjunto de critérios de seleção específicos dos quais precisam ser devidamente detalhados para permitir decisões individuais de seleção que, por sua vez, podem ser expressas numa lista de recursos *Web* a serem coletados. De acordo com Vlassenroot *et al*. (2019) temos uma ampla variação em relação às estratégias e critérios de seleção presentes em arquivos e bibliotecas nacionais. Para o autores a política de seleção de arquivos nacionais a respeito do arquivamento da *Web* difere no fato de que se limita em maior parte aos registros públicos de organizações governamentais, como o UKGWA dos *The National Archives* (c2014, [2022?a]); já no caso das bibliotecas nacionais, Vlassenroot *et al*. (2019) indicam que o escopo da coleção é maior dado que o arquivamento da *Web* é tido como parte da legislação de depósito legal ou como um complemento às coleções mais tradicionais de publicações eletrônicas ou em papel em países sem legislação de depósito legal, tal como é o caso da *Bibliothèque Nationale de France* (c2022b) e das bibliotecas britânicas de depósito legal do UK *Web Archive* ([2022?]).

Considerando as dificuldades para uma única organização, como o *Internet Archive*, de arquivar a *World Wide Web* de forma exaustiva satisfazendo todas as necessidades já que este ambiente está em constante mudança e muitos conteúdos perdem-se antes de serem arquivados (GOMES; FREITAS; SILVA, 2006), vários países estão criando os seus próprios arquivos da *Web* nacionais com estratégias e critérios de seleção específicos. Como os recursos são escassos e nem toda a *Web* pode ser preservada, Costa, Gomes e Silva (2017) sinalizam que a política de seleção da maioria das iniciativas de arquivamento da *Web* centra-se em preservar as partes mais significativas da *Web* a partir de sua própria perspectiva. Entretanto, definir os limites de uma *Web* nacional não é uma tarefa simples e as políticas de seleção são controversas, aliás, os arquivistas da *Web* definem estratégias para preencher os arquivos de acordo com o escopo de suas ações e os recursos disponíveis, como o uso dos limites das *Webs* nacionais como critérios de seleção que influenciam a cobertura de seus arquivos (GOMES; FREITAS; SILVA, 2006).

Na prática, embora seja importante preservar os *links* entre *sites*, Stirling, Chevallier e Illien (2012) afirmam que é preciso "cortar" esses *links* para permitir a preservação do que de outra forma seria um espaço infinito. Para os autores a *Web* nos leva a reconsiderar a questão do que deve e não deve ser preservado, sendo que este tipo de seleção pode ser comparado às técnicas tradicionais de arquivamento onde toda a preservação é baseada na seleção do material a preservar e a destruição do resto. Além do mais, enquanto os pesquisadores e os historiadores

podem vir a aceitar a ideia de não se ter tudo em um arquivo da *Web* na condição de que a razão das escolhas e critérios tomados sejam claramente especificados e justificados, os estudiosos digitais (*digital scholars*) quando utilizam arquivos da *Web* para as suas pesquisas devem levar em consideração como o conteúdo da *Web* arquivado é selecionado e quem é responsável por fazer essa seleção (STIRLING; CHEVALLIER; ILLIEN, 2012; VLASSENROOT *et al*., 2019).

Outro tópico digno de atenção é o papel dos estudiosos digitais, junto com membros do público em geral, na seleção de conteúdo para arquivos da *Web*. Se por um lado em algumas instituições de arquivamento existem especialistas em coleta específica como responsáveis por realizar a seleção, por outro temos casos em que a seleção na instituição é uma responsabilidade compartilhada entre um grande número de indivíduos, cada um dedicando somente um tempo à seleção do conteúdo da *Web*, em acordo com Vlassenroot *et al*. (2019). Exemplificando estes modelos de colaboração das instituições de arquivamento com parceiros externos, destacamos:

- Na criação de coleções *Web* de "emergência" sobre os ataques terroristas na França, a BnF lançou um apelo para receber sugestões da rede de correspondentes de seu próprio depósito legal da *Web*[512] e da rede internacional do IIPC (SCHAFER *et al*., 2019).
- Nas coleções "*Women's and Gender Studies Web Archive*" e "LGBTQ+ *Studies Web Archive*" da Biblioteca do Congresso dos Estados Unidos, os *sites* são selecionados pelo especialista em coleção de estudos de mulheres e gênero e pelo especialista em coleção de estudos LGBTQ+ da biblioteca, em consulta com *experts* nestas duas áreas.
- Na criação da coleção "*Ukraine Conflict*" do *Internet Archive Global Events*, houve a cooperação entre a equipe *Archive-It* e especialistas no assunto da Universidade de *Stanford*, da Biblioteca do Congresso americano, da *Global Investigative News Network* e do *Ukrainian Research Institute*[513] da Universidade de *Harvard* nas áreas de jornalismo investigativo, estudos russos e eurasianos.
- Na coleção "UTSA*'s '#BlackLivesMatter: Critical Perspectives' Coursework*"[514] acerca de um curso multidisciplinar oferecido na Universidade do Texas em *San Antonio* que examinou os contextos socioculturais e históricos do movimento *#BlackLivesMatter*, os arquivistas das coleções especiais das bibliotecas da universidade trabalharam junto ao corpo docente e os alunos do curso para reunir

[512] Disponível em: https://www.bnf.fr/fr/la-bnf-archive-le-web-du-coronavirus#bnf-une-collecte-qui-s-appuie-sur-un-r-seau-de-correspondants-tendu. Acesso em: 3 jun. 2023.
[513] Disponível em: https://huri.harvard.edu/. Acesso em: 3 jun. 2023.
[514] Disponível em: https://archive-it.org/collections/7885. Acesso em: 3 jun. 2023.

materiais de aula ou trabalhos e criaram uma amostra digital a fim de tornar este conteúdo disponível para acesso a longo prazo.

- Na coleção "*2016 Summer Olympics and Paralympics*"[515] do IIPC no *Archive-It*, além da contribuição de instituições membros do IIPC na sugestão de *sites* para inclusão na coleção, o grupo de trabalho CDG criou um formulário público para que indivíduos contribuíssem com a seleção de temas associados aos Jogos Olímpicos e Paraolímpicos de 2016, realizado no Rio de Janeiro (BYRNE, 2016; ROCKEMBACH, 2018).

Particularmente, os critérios de seleção têm de abordar três questões, descritas a seguir (BROWN, c2006; MASANÈS, c2006a):

- Conteúdo (*content*) – estabelecer critérios para determinar a natureza dos recursos *Web* qualificados para a seleção em relação ao seu conteúdo intelectual. O nível de detalhe necessário variará em função do método de seleção escolhido, visto que as abordagens temáticas e não seletivas demandam critérios adicionais mínimos e a abordagem seletiva irá exigir uma maior preparação. Exemplificando, dentro do programa de arquivamento da *Web* das coleções especiais das *University of Texas at San Antonio Libraries* (2022) é usado como critério de seleção de conteúdos *Web*, por exemplo, os *sites* de entidades e comunidades nas áreas temáticas de Educação Bilingue, Cultura de Alimentos (*Food Culture*), Estudos de Gênero, Estudos de Raça/Etnia, Estudos de Sexualidade, Estudos Fronteiriços (*Border Studies*), e de cultura e história de *San Antonio* e do sul do Texas.
- Extensão (*extent*) – estabelecer critérios para determinar a extensão dos recursos da *Web* selecionados, como a indicação de que não serão coletados *links* externos de *sites*. Este aspecto dependerá também do método de seleção escolhido (com o maior detalhe sendo necessário às abordagens seletivas), e terá de ser mais desenvolvido no âmbito da lista de coleta onde os limites técnicos em detalhes devem ser especificados. Aliás, vale frisar que não existe referência a objetos na *Web*, uma vez que URLs fornecem referências a locais e não a objetos o que torna-se um desafio aplicar uma política de seleção em um ambiente estruturado apenas em termos de localização; e o conceito de escopo se define como a extensão da coleção desejada, delimitada por critérios topológicos (o domínio da *Web* italiana, por exemplo), temáticos (os *sites* associados à biologia, por exemplo), baseados no gênero (por exemplo, *blogs*), tempo (um *site* obsoleto desde os últimos dois anos, por exemplo) e outros, em que a cada novo *link* descoberto em um rastreamento deverá ser avaliado para averiguar se ele se enquadra ou não no escopo. Exemplificando,

[515] Disponível em: https://archive-it.org/collections/7235. Acesso em: 3 jun. 2023.

no arquivo da *Web* da Catalunha da *Biblioteca de Catalunya* (c2011) é coletado somente *sites* e partes de *sites* abertos e acessíveis pela *Internet*, porém não é capturado nenhuma página *Web* que exija senha, formulário etc. (por exemplo, áreas reservadas a assinantes de uma publicação), assim como os *links* para imagens e outros elementos de um *site* externo que podem não ser exibidos a menos que tenham sido capturados pelo arquivo.

- Tempo e frequência (*timing and frequency*) – estabelecer princípios para determinar o momento e, se cabível, a frequência da coleta da *Web*. Em razão da natureza dinâmica da *Web* e da natureza frágil e transitória dos recursos *Web*, a coleta deve ocorrer o mais rapidamente possível após o fato em que o conteúdo foi considerado digno de seleção, e muitos recursos *Web* serão coletados em intervalos periódicos, mas há certos recursos que poderão ser coletados numa única ocasião, como um *site* de inquérito público com um tempo de vida fixo. No *Australian War Memorial* (2003), por exemplo, se realiza uma análise técnica de cada título e é tomado uma decisão sobre a frequência de captura desejável com base no padrão de publicação e na relevância da informação. Assim, a determinação do momento e da frequência com que um recurso *Web* deve ser coletado será definido na lista de coleta, porém isto pode ser influenciado por quatro fatores:
  - Ciclo de vida (*lifecycle*) – a natureza do recurso *Web* pode ser definida em relação ao seu ciclo de vida ativo, podendo este ser aberto ou de duração limitada. A título de exemplo, *sites* de organizações podem existir e desenvolver-se por um período indefinido e *sites* baseados em eventos podem ter um prazo de conclusão previsto podendo depois o conteúdo se tornar fixo e o *site* deixar de ser atualizado/mantido. No arquivo da *Web* islandesa da *Landsbókasafn Íslands Háskólabókasafn* ([2022]) todo o conteúdo islandês é coletado três vezes por ano e os domínios selecionados são coletados semanalmente, e são realizadas coletas temporárias de eventos, como eleições, onde cópias de páginas *Web* relativas ao evento são feitas semanalmente.
  - Taxa de alteração de conteúdo (*rate of content change*) – à medida que o conteúdo de um recurso *Web* pode ser dinâmico ou fixo e a frequência com que tal conteúdo muda pode variar enormemente, na definição da frequência de coleta será preciso avaliar a taxa de alteração. Como exemplo, uma página *Web* comum será atualizada regularmente ao passo que um artigo de periódico será provavelmente publicado de uma forma pronta, e *sites* ou páginas *Web* individuais podem permanecer estáticos durante meses ou anos enquanto outros podem mudar bastante várias vezes por dia. Exemplificando, no UK *Web Archive* ([2022?]) a maioria da coleção é coletada como parte do seu grande rastreamento anual de domínios no Reino Unido (*sites*

registrados em *.uk*, *.scot*, *.wales*, *.cymru* etc.), sendo que os *sites* selecionados são arquivados com mais frequência em função da sua relevância para uma coleção, ou da taxa de alteração do *site*, como é o caso dos inúmeros *sites* de notícias que são coletados diariamente em oposto a outros que são visitados com menor frequência.

- Avaliação de risco (*risk assessment*) – enquanto os recursos da *Web* estão sujeitos a fatores de risco de natureza financeira (por exemplo, o término do financiamento para um *site* de projeto), organizacional (quando uma mudança nas prioridades da organização leva à remoção ou não manutenção de recursos *Web*, por exemplo) e tecnológica (por exemplo, quando a obsolescência tecnológica torna um recurso da *Web* inacessível), o monitoramento e a avaliação desses riscos devem informar o processo de seleção. A monitoração poderá ser manual ou com o uso de ferramentas automáticas para monitorar a disponibilidade de um *site* e rastrear a frequência e a duração de qualquer tempo de inatividade (por exemplo, o uso de *software* de servidor *Web* desatualizado ou a ocorrência de períodos de inatividade frequentes), e se o risco for julgado um fator relevante, a política de seleção deve definir os tipos de risco a controlar (e como isso será conseguido) e uma justificação para avaliar o impacto desse risco em recursos da *Web* específicos. Exemplificando, nas políticas do programa de coleta de recursos da *Web* das *Columbia University Libraries* (c2021a) há vários critérios que conduzem o processo de seleção de *sites* para arquivamento, incluindo o risco perceptível de longevidade do *site* (sendo que os rastreamentos são feitos tri/semestralmente para *sites* atualizados ativamente); e na política de desenvolvimento de coleção das *Stanford Libraries* ([c2022?b]) além de se preferir conteúdos da *Web* atuais, existe também a priorização de conteúdos em risco, incluindo aqueles com interesse ou propósito limitado no tempo, sujeitos à censura de governo, divulgados por organizações imaturas e eventos espontâneos.
- Atualidade/significância (*topicality/significance*) – para a definição da frequência de coleta deve-se incluir uma avaliação subjetiva da atualidade e do valor inerente do recurso *Web*. O momento e a frequência com que cada recurso *Web* selecionado deve ser coletado precisará ser revisto regularmente como parte da manutenção da lista de coleta, pois a atualidade percebida de um *site* irá provavelmente mudar com o tempo exigindo um ajuste na frequência e, portanto, quatro opções básicas são possíveis: a coleta repetida (*repeated collection*, isto é, a coleta por uma série de instantâneos estáticos ao longo do tempo de recursos dinâmicos com ciclos de vida abertos em intervalos repetidos – semanais, semestrais ou anuais –); a coleta *ad-*

> *hoc* (isto é, a coleta de recursos que mudam a taxas imprevisíveis em resposta a um evento de gatilho, como alguma forma de monitoramento automático/manual do recurso ou um alerta de uma fonte externa – por exemplo, informantes humanos para identificar novos itens para coleta –); a coleta única (*one-off collection*, isto é, a coleta pontual de recursos com conteúdo fixo – por exemplo, uma publicação *online* etc. –, ou que mudam durante um período e depois se tornam fixo – por exemplo, o *site* para um inquérito público governamental); e a coleta abrangente (*comprehensive collection*, isto é, a captura do ciclo de vida completo de um recurso *Web* dinâmico e aberto, especialmente, no caso de um ambiente de gerenciamento de registros em que transações *online* devem ser preservadas para fins probatórios, sendo necessário a coleta para arquivamento ser integrada ao fluxo de trabalho de gerencia do *site*). Exemplificando, na *Bibliothèque Nationale de France* (c2022a, c2022c) a frequência e a profundidade (total ou parcial de um *site*) das coleções são ajustadas à natureza dos *sites* e ao ritmo das suas atualizações para manter versões sucessivas representativas de sua evolução e, especificamente, nas suas coleções "direcionadas" (ou seja, as coleções atuais como *sites* de referência em acordo com os demais acervos da BnF, as coleções de projetos em cooperação que documentam temas transversais ou grandes eventos, e as coletas de emergência que referem-se a acontecimentos inesperados com forte impacto na sociedade e que são veiculadas espontaneamente nas mídias sociais), há parâmetros de frequência e profundidade variáveis em milhares de *sites* selecionados por bibliotecários, especialistas e/ou pesquisadores. Uma destas coleções trata da coleta de conteúdo sobre Covid-19[516], onde a equipe do depósito legal digital seleciona e arquiva os conteúdos de *sites*, *blogs* e mídias sociais formando um todo coerente, significativo e representativo, e determina a frequência das coletas feitas por meio de robôs, seja várias vezes ao dia para as redes sociais, uma vez ao dia para *sites* de imprensa nacional e regional etc.

De fato, através de Gomes, Freitas e Silva (2006), nenhum critério sozinho resolve a seleção do conteúdo a arquivar num arquivo da *Web* nacional e, assim, combinações tem de ser usadas. Ao analisar estratégias de seleção de conteúdos para um arquivo da *Web* nacional, os autores pontuam que os critérios de preencher automaticamente o arquivo só com documentos em *sites* sob domínio de primeiro nível com código do país (.pt para Portugal, por exemplo) ou em servidores *Web* localizados fisicamente no país, resultam na exclusão de uma alta quantia

[516] Disponível em: https://www.bnf.fr/fr/la-bnf-archive-le-web-du-coronavirus. Acesso em: 3 jun. 2023.

de materiais *Web* hospedados em domínios genéricos (*.com*, *.net* etc.) ou em servidores fora do país. Outro critério importante que Gomes, Freitas e Silva (2006) apontam é o de selecionar os tipos de mídia que se irá armazenar conforme os recursos disponíveis e o escopo do arquivo, pois os custos e a complexidade da preservação de materiais crescem com a variedade de mídias arquivadas e podem se tornar inviáveis. Sobre este último critério, por exemplo, a *Library of Congress* (2022a) declara que os formatos incluídos num *site* selecionado podem ser materiais audiovisuais, imagens etc. ou itens afins necessários para apoiar a pesquisa no assunto coberto.

Iniciativas em bibliotecas nacionais e universitárias, órgãos de pesquisa, instituições financeiras e de memória

Muitas iniciativas de arquivos da *Web* ao redor do mundo (algumas delas membros do IIPC) expõem em geral os seus critérios de seleção e as frequências de capturas de *sites*, como:

- *Archivo de la Web chilena* – como um serviço que objetiva preservar *sites* chilenos para garantir o acesso da informação gerada em caso da sua perda, o arquivo da *Web* chileno da Biblioteca Nacional do Chile fez coletas da *Web*, conforme a legislação de depósito eletrônico na lei chilena nº 19.733 de 2001 (revisada em 2013)[517], que buscam abranger momentos atuais do país e a obra/história da biblioteca, como os *sites* de candidaturas presidenciais de 2013 e do *Servicio Nacional del Patrimonio Cultural* (SNPC) (isto é, *sites* de domínio *.cl*, *gob.cl* etc.); aliás, a biblioteca define coletas uma vez/várias vezes ao dia, por semana ou em certas datas pela quantia de conteúdo carregado nos *sites*, e por restrições na lei chilena nº 17.336 de propriedade intelectual de 1970 (revisada em 2017)[518] os *sites* que não são do SNPC só são consultados nas instalações de bibliotecas regionais e da biblioteca nacional[519] (BIBLIOTECA NACIONAL DE CHILE, [2022]).

[517] Disponível em: https://www.bcn.cl/leychile/navegar?idNorma=186049&buscar=19.733. Acesso em: 3 jun. 2023.
[518] Disponível em: https://www.bcn.cl/leychile/navegar?idNorma=28933. Acesso em: 3 jun. 2023.
[519] Disponível em: http://www.bibliotecanacionaldigital.gob.cl/bnd/612/w3-propertyvalue-761832.html. Acesso em: 3 jun. 2023.

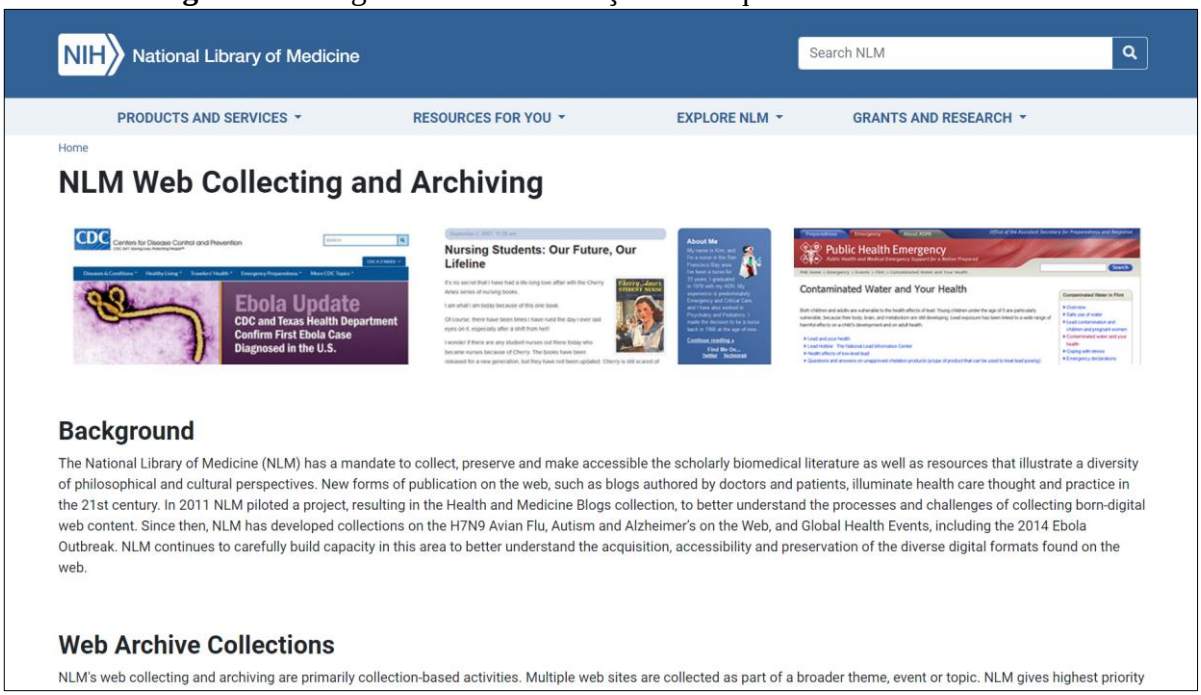
Figura 88 – Página inicial do arquivo da *Web* chilena.

Fonte: *Biblioteca Nacional de Chile* ([2023?]).

- NLM*'s Web archive collections* – criada desde 2011 em apoio ao desenvolvimento da coleção da NLM com foco em eventos e tópicos que documentam conteúdos na *Web*, nas coleções de arquivos da *Web* da NLM o arquivamento é orientado pelas "*Collection Development Guidelines of the* NLM*: Web Content*"[520] onde além de *sites* de domínio da instituição ("*nlm.nih.gov*") são coletados seletivamente *sites* de notícias, conteúdo de mídia social etc. que registrem os desafios e os progressos na pesquisa em biomedicina, documentem a prática e o ensino da medicina, demostrem como os serviços de saúde são organizados, prestados e financiados, relatem a execução de políticas que afetam os serviços de saúde, ilustrem a percepção pública da prática médica e da saúde pública, e ajudem na compreensão e impacto dos eventos globais de saúde; aliás, as coleções de *sites* arquivados estão disponíveis *online* (https://archive-it.org/organizations/350) para uso acadêmico através do *Archive-It* (NATIONAL LIBRARY OF MEDICINE, 2019).

[520] Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK518732/. Acesso em: 3 jun. 2023.

Figura 89 – Página inicial das coleções de arquivos da *Web* da NLM.



Fonte: *National Library of Medicine* (2023).

- *Web Archive Singapore* (WAS) – sendo uma coleção de *sites* de Singapura datada de 2006 que consiste em organizações, notícias, eventos etc. preservados para informação, referência e pesquisa sobre a história e o patrimônio desse país, o arquivo da *Web* de Singapura da Biblioteca Nacional de Singapura arquiva todos os *sites* de domínio *.sg* disponíveis publicamente e que são coletados conforme a *National Library Board Act* 1995 (2020 *revised edition*)[521] além de *sites* não registrados sob o domínio singapuriano e que são de propriedade de singapurianos ou instituições do país e são coletados por solicitação de permissão, e os *sites* da coleção são arquivados no mínimo uma vez por ano ou com mais frequência se tiverem relevância histórica, podendo ser acessados *online* ou, em determinados casos, *in loco* (NATIONAL LIBRARY BOARD, c2022).

[521] Disponível em: https://sso.agc.gov.sg/Act/NLBA1995. Acesso em: 3 jun. 2023.

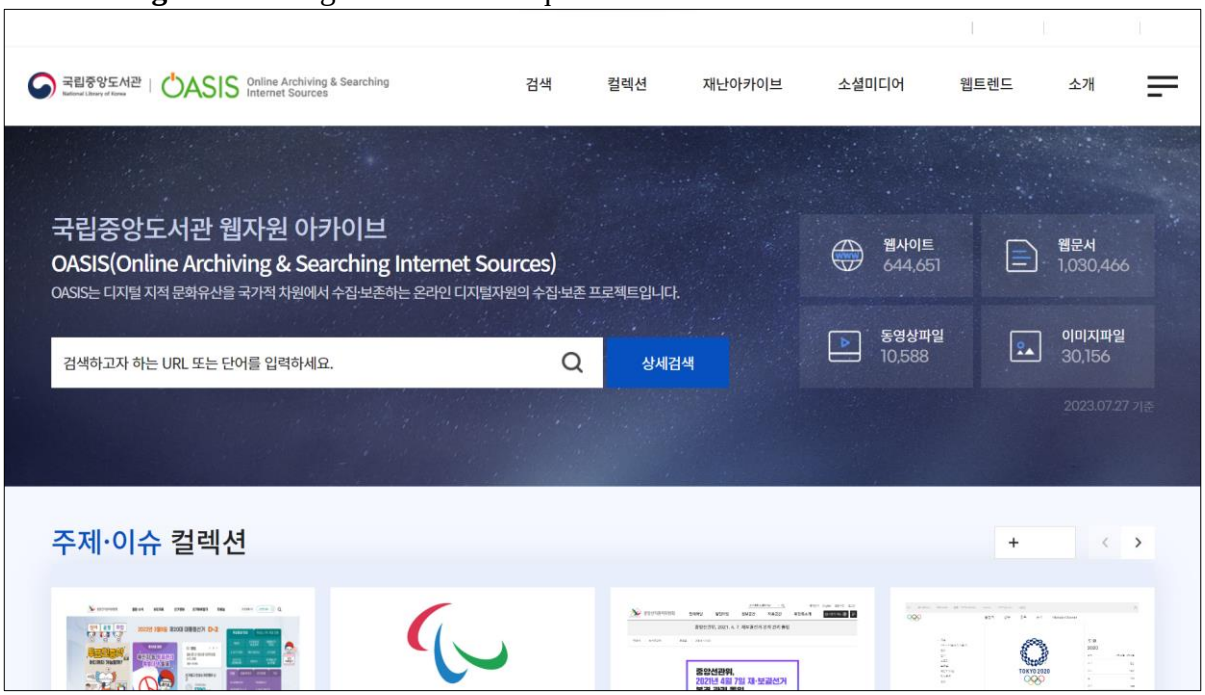
Figura 90 – Página inicial do arquivo da *Web* de Singapura ou WAS.

Fonte: *National Library Board* (c2023).

- *Online Archiving & Searching Internet Sources* (OASIS)*'s National Library of Korea* – compondo um projeto criado em 2004 para preservar o patrimônio cultural intelectual digital em nível nacional e fornecê-lo às gerações vindouras, o OASIS da Biblioteca Nacional da Coreia coleta materiais da *Web* e *sites* no domínio de nível superior *.kr* com o código do país da Coreia além de *sites* importantes de domínios genéricos *.com*, *.org* etc. de acordo com a Lei da Biblioteca[522], sendo que os *sites* de destino são coletados todos os anos, os *sites* principais ou aqueles relacionados a problemas são coletados a cada um dia, semana, mês e três/seis meses; ademais, devido os *sites* coletados serem protegidos pela lei coreana de direitos autorais[523], a biblioteca exigi ao proprietário do *site* (detentor dos direitos autorais) permissão para a divulgação e o uso das informações em suas instalações ou em outros locais (NATIONAL LIBRARY OF KOREA, c2006).

[522] Disponível em: https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%EB%8F%84%EC%84%9C%EA%B4%80%EB%B2%95. Acesso em: 3 jun. 2023.
[523] Disponível em: https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%EC%A0%80%EC%9E%91%EA%B6%8C%EB%B2%95. Acesso em: 3 jun. 2023.

Figura 91 – Página inicial do arquivo da *Web* da Coreia do Sul ou OASIS.



Fonte: *National Library of Korea* (c2006).

- *Web Archiving Project* (WARP) – desde 2002, o WARP da Biblioteca Nacional da Dieta no Japão arquiva *sites* de instituições japonesas, como do governo, da biblioteca nacional, de universidades, empresas etc., além de *sites* de eventos culturais realizados no país (ou seja, *sites* de domínio *.jp*, *go.jp*, *or.jp*, *co.jp* etc., ou registrados sob extensões genéricas *.com*, *.org*, *.net* etc.); demais, os *sites* de instituições nacionais são arquivados uma vez por mês e os *sites* de outras instituições que também estão sujeitas a coleta em acordo com a *National Diet Library Law* de 1948[524] são arquivados quatro vezes por ano, sendo que a maioria das coleções de *sites* arquivados estão disponíveis *in loco* nas instalações da biblioteca, porém uma parte dos *sites* é fornecida *online* na *Internet* com a permissão dos criadores/proprietários de *sites* (NATIONAL DIET LIBRARY, c2013).

[524] Disponível em: https://www.ndl.go.jp/en/aboutus/laws.html. Acesso em: 3 jun. 2023.

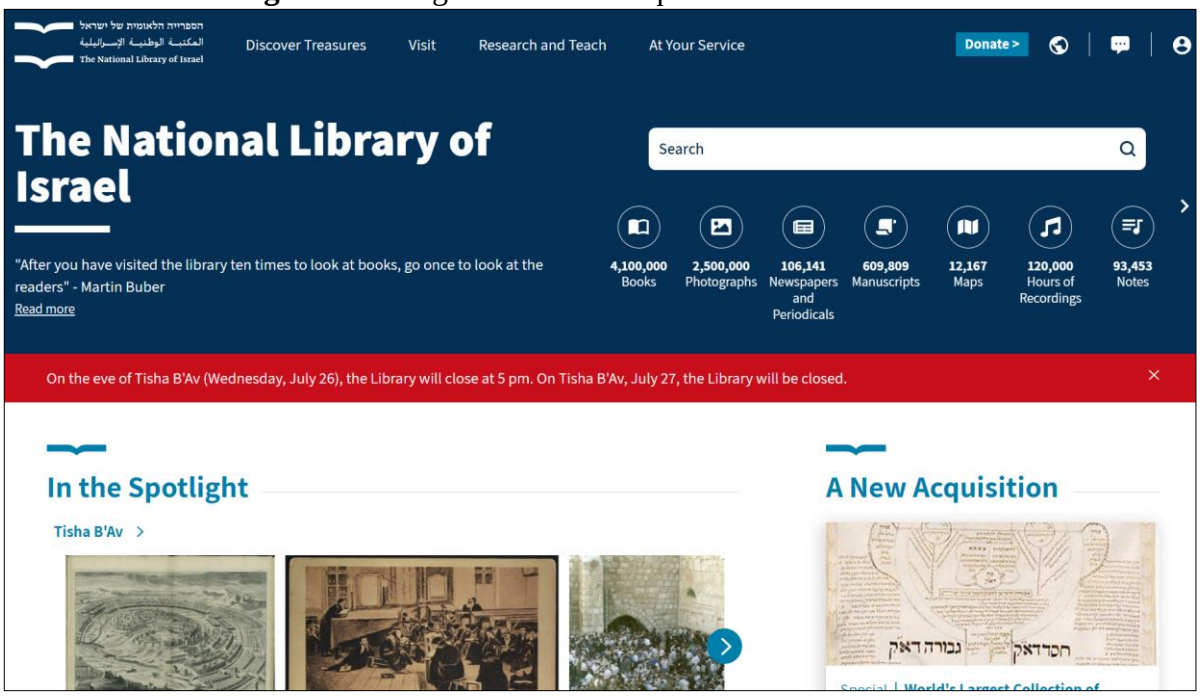
Figura 92 – Página inicial do arquivo da *Web* do Japão ou WARP.

Fonte: *National Diet Library* ([2023?]).

- *Israeli Internet Archive* – iniciado em 2013, o arquivo da *Web* israelense da Biblioteca Nacional de Israel destina-se a documentar e preservar as publicações *online* publicadas nos *sites* públicos de Israel para pesquisadores, estudantes etc. hoje e no futuro e, assim, inclui todos os *sites* cujo URLs terminam em *.il* e alguns *sites* adicionais que estão sob a geolocalização israelense ou são relevantes para o público israelense, sendo estes *sites* digitalizados e copiados cerca de duas vezes por ano e podendo ser consultados apenas nas instalações da biblioteca devido a Lei de Direitos Autorais de Israel (do inglês *Israeli Copyright Act*, 2007)[525]; além do mais, a Lei da Biblioteca Nacional de Israel (do inglês *National Library Law*, 2007)[526] imputa a tarefa da biblioteca o dever de preservar tudo o que for publicado impresso ou eletronicamente (*Internet*) relativo ao discurso público e a cultura israelense e ao Estado de Israel (NATIONAL LIBRARY OF ISRAEL, 2022).

[525] Disponível em: https://www.tau.ac.il/law/members/birnhack/IsraeliCopyrightAct2007.pdf. Acesso em: 3 jun. 2023.
[526] Disponível em: https://web.nli.org.il/sites/NLI/English/library/aboutus/renewal/Documents/National%20Library%20Law.pdf. Acesso em: 3 jun. 2023.

Figura 93 – Página inicial do arquivo da *Web* israelense.



Fonte: *National Library of Israel* (2023).

- *New Zealand Web Archive* – iniciado em 1999, o arquivo da *Web* da Nova Zelândia, como parte da coleção da Biblioteca Alexander Turnbull da Biblioteca Nacional do país, realiza coletas da *Web* seletivas desde 1999 (*selective harvesting*, a seleção de *sites* de alto valor) e de domínio desde 2008 (*domain harvesting*, a coleta anual do máximo de material que for tecnicamente possível com mínima intervenção humana), adquirindo *sites* no domínio de código de país *.nz*, ou *sites* sob as extensões genéricas *.com*, *.net* e *.org* que podem ser hospedados em máquinas localizadas fisicamente no país, ou *sites* selecionados no exterior que são cobertos pela Lei da Biblioteca Nacional da Nova Zelândia de 2003; além disto, os *sites* arquivados sobre o país, os neozelandeses e ao Pacífico para pesquisa e preservação de longo prazo podem ser acessados *online*[527] ou, em caso de restrições de acesso à sua publicação, só *in loco* na biblioteca em salas de leitura (NATIONAL LIBRARY OF NEW ZEALAND, [2022a?], [2022b?], [2022c?]).
- *Patrimoni Digital de Catalunya* (PADICAT): *L'Arxiu Web de Catalunya* – criado desde 2005 pela Biblioteca da Catalunha na Espanha para capturar, preservar e fornecer acesso permanente a toda a produção digital cultural, científica e geral catalã gerada e arquivar a *Web* catalã, o PADICAT ou arquivo da *Web* da Catalunha realiza a captura sistemática de *sites* sob o domínio *.cat*, além de *sites* de instituições que a biblioteca possui acordo

[527] Disponível em: https://natlib-primo.hosted.exlibrisgroup.com/primo-explore/search?vid=NLNZ. Acesso em: 3 jun. 2023.

de colaboração, *sites* tidos úteis a partir da pesquisa por navegação e *sites* recomendados por usuários (ou seja, *sites* de domínio *.es* com o código do país da Espanha, *sites* sob extensões genéricas *.com*, *.org* etc.); demais, as versões antigas de páginas *Web* podem ser acessadas *online* (https://www.padicat.cat/ca/cerca-i-descobreix) e visualizadas por *Wayback Machine*, e os *sites* arquivados são capturados no mínimo duas vezes por ano, havendo aumento de frequência no futuro (BIBLIOTECA DE CATALUNYA, c2011).

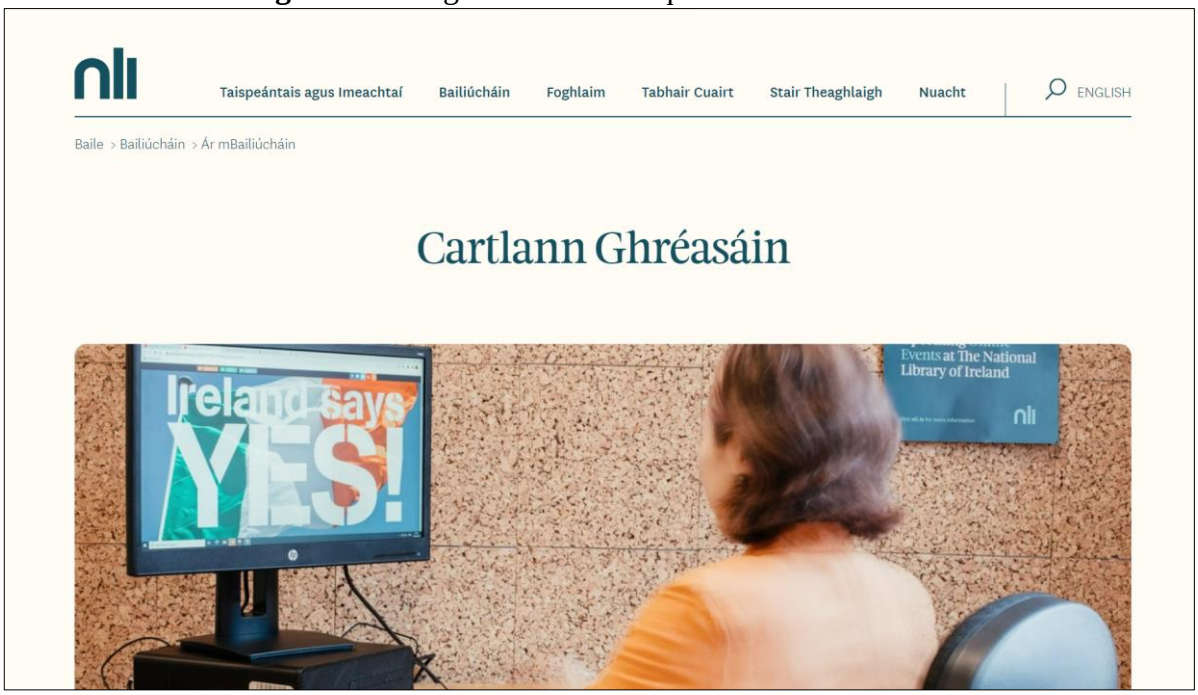
Figura 94 – Página inicial do PADICAT ou arquivo da *Web* da Catalunha.

Fonte: *Biblioteca de Catalunya* (c2011).

- *Irish Web archive (Cartlann Ghréasáin)* – desde 2011, a Biblioteca Nacional da Irlanda (do irlândes *Leabharlann Náisiúnta na Héireann*) faz o arquivamento da *Web* de forma seletiva[528], mas em 2017 ela empreendeu um rastreamento de domínio do espaço *Web* irlandês, realizado pelo *Internet Archive*, dando particular destaque a *sites* do governo irlandês e de órgãos públicos, *sites* em língua irlandesa e *sites* de ensino superior (ou seja, *sites* registrados em *.ie*, *gov.ie*, *nd.edu*, *ucd.ie*, *.com*, *.org*, *.net* etc.); além disto, no arquivo da *Web* irlandês a biblioteca reconhece ser de responsabilidade do proprietário do *site* cumprir a legislação nacional de proteção de dados e direitos autorais e, assim, ela preserva esses materiais de interesse público e os disponibiliza para fins de pesquisa e estudo, e os *sites* arquivados podem ser acessados *online* no *Archive-It* e exibidos por meio da *Wayback Machine* (NATIONAL LIBRARY OF IRELAND, c2022; [201-?]).

[528] Disponível em: https://www.nli.ie/en/udlist/web-archive-collections.aspx. Acesso em: 19 dez. 2022.

Figura 95 – Página inicial do arquivo da *Web* irlandês.

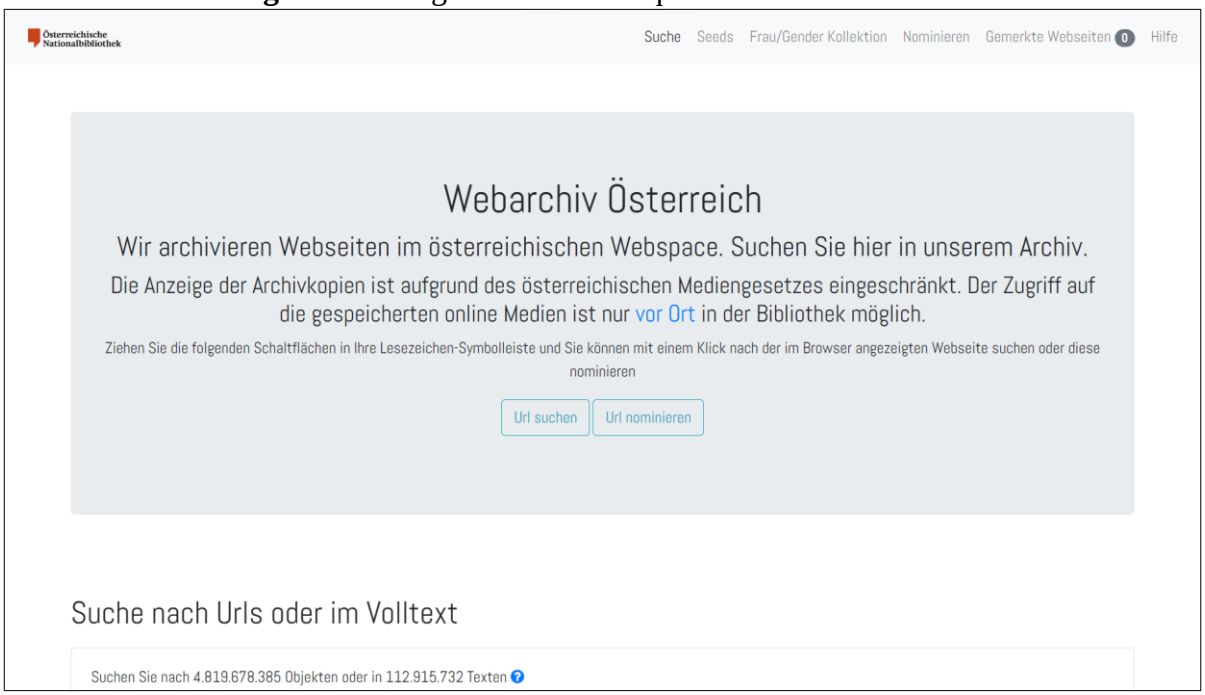


Fonte: *National Library of Ireland* (c2023).

- *Web Archive Austria* (*Webarchiv Österreich*) – como o depósito legal austríaco para mídia *online* entrou em vigor em 2009 com a alteração da Lei de Imprensa da Áustria (do alemão *Mediengesetzes*)[529] permitindo até quatro rastreamentos por ano de *sites* sob o domínio *.at* e de *sites* relativos à Áustria, o arquivo da *Web* da Áustria da Biblioteca Nacional Austríaca procura coletar e arquivar mantendo utilizável para fins científicos a *Internet* austríaca a partir das estratégias de coleta de domínio a cada dois anos (isto é, domínio *.at* inteiro e outros domínios de nível superior relativos à Áustria, como *.com*, *.org*, *.info*, *.eu* etc.), de coleta seletiva (isto é, páginas *Web* que estão sujeitas a mudanças frequentes, tais como meios de comunicação etc.) e de coleta de eventos em intervalos adequados (isto é, conteúdo *online* para ocasiões especiais e grandes eventos, tais como eleições), sendo que a seleção das páginas é realizada por curadores conforme a política de coleta do arquivo (ÖSTERREICHISCHEN NATIONALBIBLIOTHEK, [2022]).

[529] Disponível em: https://www.ris.bka.gv.at/Dokumente/BgblAuth/BGBLA_2009_I_8/BGBLA_2009_I_8.html. Acesso em: 3 jun. 2023.

Figura 96 – Página inicial do arquivo da *Web* da Áustria.



Fonte: *Österreichischen Nationalbibliothek* (c2023).

- *Estonian Web Archive* (*Eesti veebiarhiiv*) – como um banco de dados gerenciado pela Biblioteca Nacional da Estônia (do estoniano *Eesti Rahvusraamatukogu*) para preservar recursos de informação *online* pertencentes ao patrimônio cultural do país, o arquivo da *Web* estoniano coleta *sites* da Estônia e de instituições estatais publicados no domínio *.ee* com o código do país da Estônia ou, ainda, outros domínios de nível superior (os *sites* registrados em *.eu* com o código da União Europeia, por exemplo), além de *sites* importantes para o país de domínios genéricos *.com*, *.org*, *.net* etc. que são arquivados e disponibilizados *online* gratuitamente (https://veebiarhiiv.digar.ee/) com a permissão do detentor dos direitos autorais, através da Lei de Preservação de Cópias (do estoniano *Säilituseksemplari seadus*, 2016)[530] (EESTI RAHVUSRAAMATUKOGU, 2022).

[530] Disponível em: https://www.riigiteataja.ee/akt/107072016001. Acesso em: 3 jun. 2023.

Figura 97 – Página inicial do arquivo da *Web* estoniano.

Fonte: *Eesti Rahvusraamatukogu* ([2023?]).

- *Spletni arhiv Narodne in Univerzitetne Knjižnice* – NUK (NUK *Web Archive*) – desde 2008, o arquivo da *Web* da Biblioteca Nacional e Universitária da Eslovênia ou NUK preserva uma amostra da *Web* eslovena com critérios de captura gerais, incluindo obras de autores eslovenos, em língua eslovena e/ou sobre o país como parte do patrimônio cultural da sua nação, e critérios especiais, como domínio (isto é, publicações *online* – *sites*, jornais, artigos etc. – publicadas no domínio *.si* com o código do pais da Eslovênia e sob os domínios gerais *.eu*, *.com*, *.info* etc.) e formatos de dados (isto é, publicações em XML ou em formatos estendidos e padronizados – HTML, PDF etc. –); assim, *sites* são coletados como parte do patrimônio cultural da sua nação, com base na lei eslovena de cópia obrigatória de publicações (do esloveno *Zakon o obveznem izvodu publikacij*, 2006)[531] e nas regras sobre tipos e seleção de publicações eletrônicas para cópia obrigatória (do esloveno *Pravilnik o vrstah in izboru elektronskih publikacij za obvezni izvod*, 2007)[532] (NARODNA IN UNIVERZITETNA KNJIŽNICA, c2022).

[531] Disponível em: https://www.uradni-list.si/glasilo-uradni-list-rs/vsebina?urlid=200669&stevilka=2977. Acesso em: 3 jun. 2023.
[532] Disponível em: https://www.uradni-list.si/glasilo-uradni-list-rs/vsebina?urlid=200790&stevilka=4422. Acesso em: 3 jun. 2023.

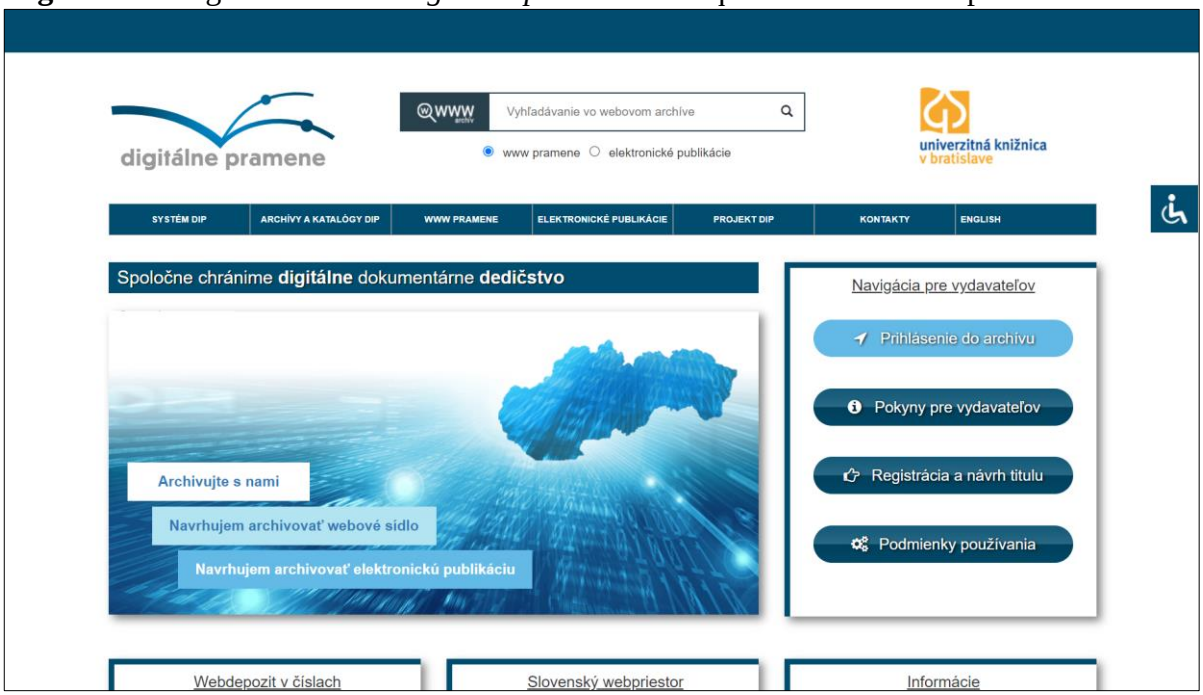
Figura 98 – Página inicial do arquivo da *Web* da Eslovênia ou NUK.

Fonte: *Narodna in Univerzitetna Knjižnica* (c2023).

- *Digital Resources – Web harvesting and E-Born Content Archiving (Digitálne pramene – Webharvesting a archivácia e-Born obsahu)* – a fim de coletar, preservar a longo prazo e dispor fontes originais de informação digital, o sistema de fontes digitais da Biblioteca da Universidade de Bratislava foca em coletas nacionais, temáticas e seletivas de *sites* de valor científico e cultural que são parte do patrimônio da República Eslovaca (isto é, *sites* de domínio *.sk*, *.cz* e *.eu* com os códigos de país da Eslováquia e República Checa e a extensão de domínio geográfico para a União Europeia, ou de *sites* registrados em *.com* etc.), e os *sites* devem atender critérios, como *sites* publicados em seu território, ou na língua eslovaca, ou de autores eslovacos, ou sobre o país e os eslovacos, segundo a política de coleta do sistema[533]; aliás, o acesso aos *sites* é gratuito e local, regido por disposições legais, contratos licença[534] e normas de uso justo das fontes de informação (UNIVERZITNÁ KNIŽNICA V BRATISLAVE, [2022?a], [2022?b], [2022?c], [2022?d]).

[533] Disponível em: https://www.webdepozit.sk/dokumenty/DIP_Politika-zberu_WWW.docx. Acesso em: 3 jun. 2023.
[534] Disponível em: https://www.webdepozit.sk/dokumenty/ZMLUVA_UKB.doc. Acesso em: 3 jun. 2023.

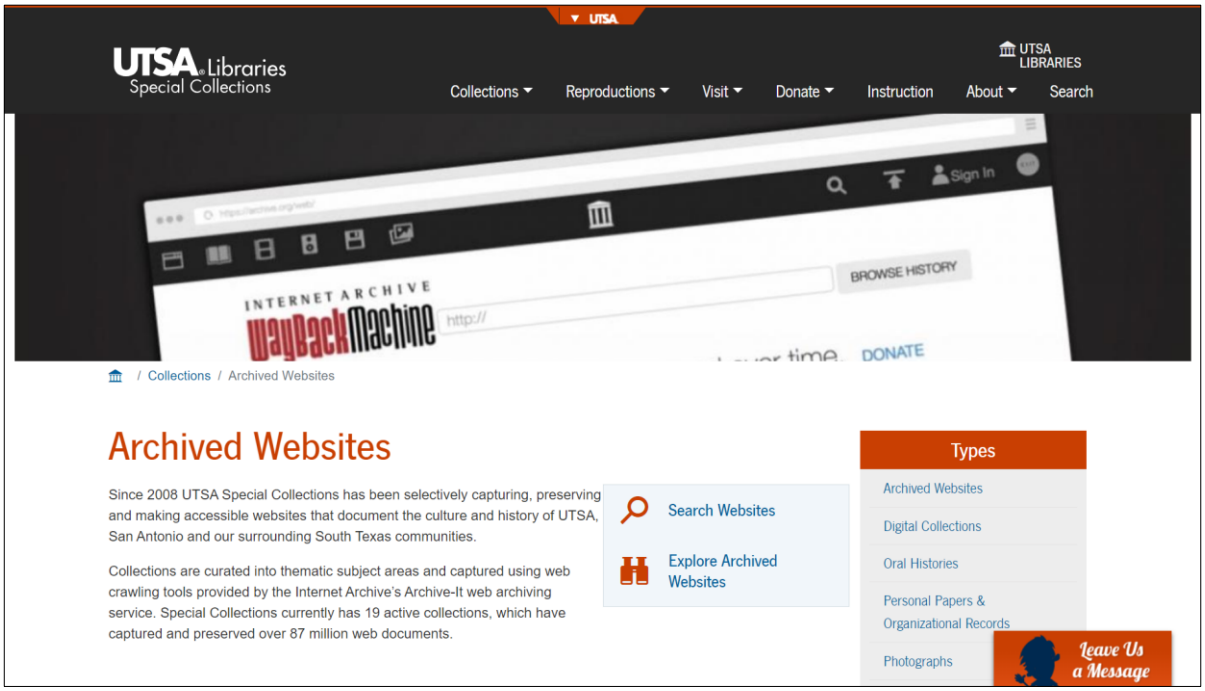
Figura 99 – Página inicial do *Digitálne pramene* ou arquivo da *Web* da República Eslovaca.



Fonte: *Univerzitná Knižnica v Bratislave* ([2023?]).

- *University of Texas at San Antonio* (UTSA) *Libraries* – a fim de capturar, preservar e tornar conteúdos da *Web* acessíveis, para uso de acadêmicos da UTSA e do mundo, que documentam a cultura e a história da UTSA, de *San Antonio* e das comunidades do sul do estado americano do *Texas*, as coleções especiais das bibliotecas da UTSA, pelo seu programa de arquivamento que adota o *Archive-It* desde 2008, utiliza como critérios de inclusão em sua política de seleção os materiais que, por exemplo, devem complementar os arquivos físicos/digitais adicionais mantidos pelas coleções, ou estar associados com a história, administração ou cultura da UTSA, ou apoiar a pesquisa e o ensino da UTSA, sendo que o foco geográfico deve centrar-se em *San Antonio* e no sul do *Texas*; assim, coletam-se *sites* hospedados ou afiliados à UTSA e *sites* de organizações e comunidades em várias áreas e eventos atuais (ou seja, *sites* de domínio *utsa.edu*, ou em *.org*, *.com*, *.net* etc.) (UNIVERSITY OF TEXAS AT SAN ANTONIO LIBRARIES, 2020, 2022).

Figura 100 – Página inicial das coleções de *sites* arquivados das bibliotecas da Universidade do Texas em *San Antonio* nos Estados Unidos.



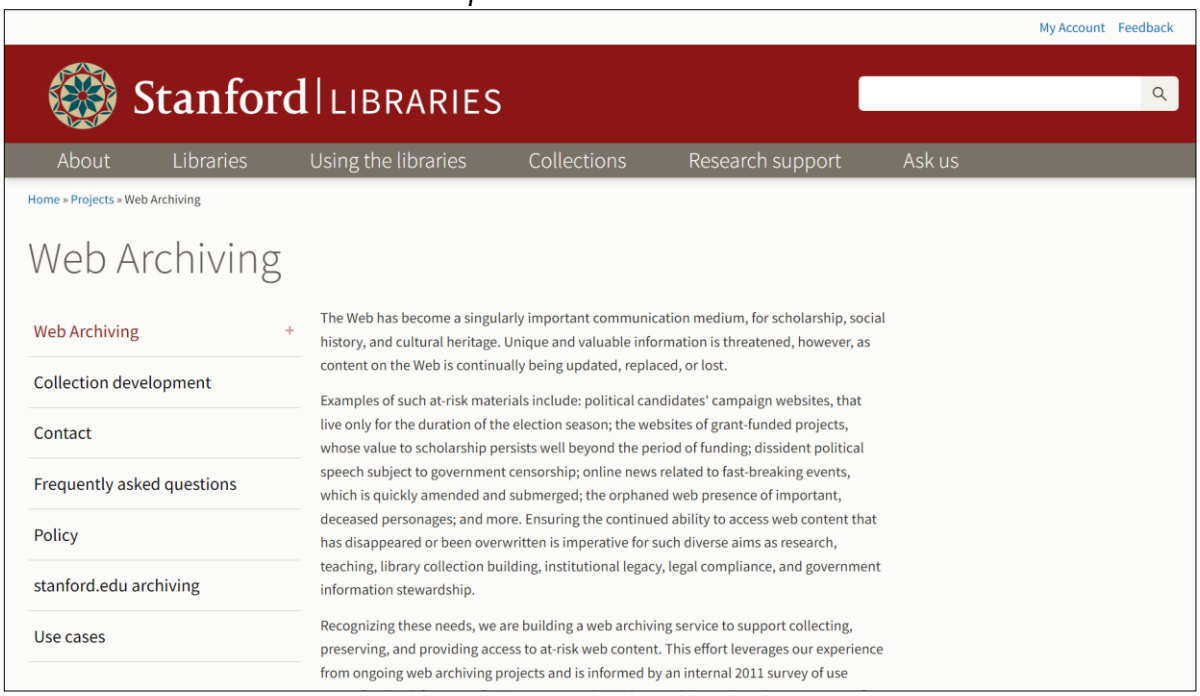
Fonte: *University of Texas at San Antonio Libraries* ([2023?a]).

- *Harvard Library* – envolvidas no arquivamento da *Web* desde 2009 (com primeiras coletas feitas em 2007), as bibliotecas da Universidade de *Harvard* nos Estados Unidos, através das suas coleções de arquivos *Web* no *Archive-It*[535], coletam mensalmente ou anualmente e cedem acesso para pesquisa aos registros culturais e históricos de *Harvard* na *Web* e da presença *online* da universidade na *Web* (incluindo *sites* de departamentos e *sites* de iniciativas de pesquisa, publicações pessoais e profissionais de indivíduos e registros de organizações afiliadas a universidade – professores, alunos, funcionários, centros, clubes, jornais, revistas etc. – e outros), além de *sites* para estudo sobre diversos temas que complementam ou relacionam-se com as coleções impressas mantidas pelas bibliotecas da universidade (ou seja, *sites* de domínio *harvard.edu* ou *sites* registrados sob as extensões genéricas *.com*, *.org*, *.gov*, *.info* etc.) (HARVARD LIBRARY, c2022).
- SUL – envolvidas no arquivamento da *Web* desde 2007, as bibliotecas da Universidade de *Stanford* nos Estados Unidos para fins de preservar registros da universidade arquiva regularmente muitos *sites* de *Stanford* (ou seja, *sites* de domínio *stanford.edu*), e em sua política de desenvolvimento de coleção, prioriza-se certas categorias de conteúdo *Web* mais ameaçadas, incluindo aqueles de interesse/propósito limitado no tempo, sujeitos à censura governamental, disseminados por organizações imaturas e eventos espontâneos,

[535] Disponível em: https://preservation.library.harvard.edu/web-archives-collections. Acesso em: 3 jun. 2023.

bem como conteúdos atuais, resultados de mecanismos de busca e *links* encurtados que são menos comuns em arquivos da *Web* existentes, ou áreas específicas que reflitam as coleções especiais[536], as pesquisas[537], a equipe de especialistas-curadores[538], o grupo de pesquisadores, docentes etc., e a geografia e a história[539] da universidade (isto é, *sites* de domínios genéricos *.com*, *.org*, *.net* e outros) (STANFORD LIBRARIES, [c2022?b]).

Figura 101 – Página inicial sobre o arquivamento da *Web* nas bibliotecas da Universidade de *Stanford* nos Estados Unidos.



Fonte: *Stanford Libraries* ([2023?a]).

- CUL*'s Web resources collection program* – iniciado em 2008, o programa de coleta de recursos da *Web* das bibliotecas da Universidade de *Columbia* nos Estados Unidos tem arquivado *sites* afiliados à instituição e *sites* de organizações cujos arquivos impressos são mantidos na universidade (ou seja, *sites* de domínio *columbia.edu* e *sites* registrados em *.org*, *.com*, *.net* etc.), e em suas políticas existe uma variedade de critérios que orienta o seu processo de seleção de *sites* para arquivamento, como a relevância do assunto para a pesquisa atual e o ensino, o risco notado de longevidade do *site* e a complementaridade dos *sites* com as coleções impressas existentes nas CUL, além de *sites* selecionados em áreas temáticas compatíveis aos pontos fortes existentes na coleção CUL; em adição, a

[536] Disponível em: https://library.stanford.edu/libraries/spc/about. Acesso em: 3 jun. 2023.
[537] Disponível em: https://www.stanford.edu/research/. Acesso em: 3 jun. 2023.
[538] Disponível em: https://library.stanford.edu/people/specialists. Acesso em: 3 jun. 2023.
[539] Disponível em: https://www.stanford.edu/about/history/. Acesso em: 3 jun. 2023.

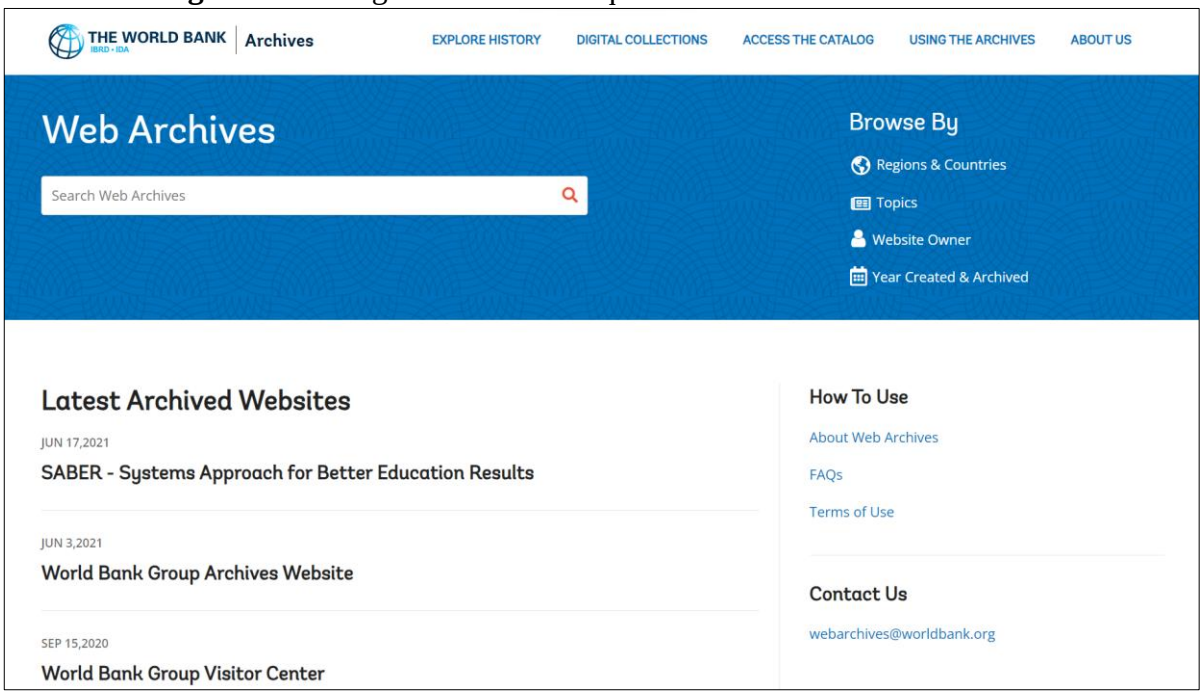
identificação de *sites* específicos para arquivamento e o desenvolvimento de temas de coleção é definido pelo coordenador do programa em conjunto com proprietários dos *sites*, especialistas e pesquisadores (COLUMBIA UNIVERSITY LIBRARIES, c2021a).

- SHU *Web Archive* – no arquivo da *Web* da SHU, os arquivistas selecionam conteúdo baseado na política de preservação das bibliotecas da SHU[540] e nos critérios de inclusão para captura estão *sites* (ou URLs) contendo informações oficiais da SHU hospedados em servidores *Web* da universidade, como *sites* de domínio *shu.edu*, ou hospedados por empresas privadas que possuem só informações da universidade e não contenham uma combinação de informações públicas e privadas, além de *sites* privados que tenham informações significativas sobre a SHU e/ou possam auxiliar na formação da política da universidade, e os *sites* são fornecidos *in loco* para uso educacional respeitando-se os direitos dos proprietários dos *sites*; em adição, a frequência dos rastreamentos dos *sites* (isto é, anuais, semestrais, semanais, únicos etc.) esta sincronizada com frequência das alterações feitas nos *sites* (SETON HALL UNIVERSITY LIBRARIES, [2022?]).
- Arquivo.pt – iniciado em 2008, o arquivo da *Web* portuguesa, como um serviço da FCT do Ministério da Educação e Ciência de Portugal que visa a preservação da informação na *Web* pública portuguesa (isto é, todos os conteúdos hospedados sob o domínio *.pt* e outros fora deste domínio que sejam de interesse para a comunidade portuguesa), impõe algumas restrições na coleta exaustiva da *Web*, como tamanho máximo dos conteúdos baixados da *Web*, número de conteúdos por *site* e número de *links* que o rastreador *Web* percorre desde um endereço inicial até chegar a um conteúdo; adicionalmente, o arquivo faz três/quatro coletas por ano da *Web* com cerca de 90% dos conteúdos coletados ao final de sete dias, sendo que quando ocorrem eventos relevantes (por exemplo, eleições) realiza-se coletas extraordinárias de *sites* selecionados, e os *sites* arquivados podem ser acessados *online* para uso educativo, científico e de pesquisa (ARQUIVO.PT, 2021a).
- *World Bank's Web Archives Program* – lançado em 2007, o programa de arquivos da *Web* do Banco Mundial pretende arquivar não todos os *sites* antigos da instituição e, sim, somente aqueles que foram descontinuados ou atualizados significativamente e que na avaliação do Banco possuem valor histórico e de pesquisa a ser preservado; ademais, a sua coleção de arquivos da *Web* remonta a 1998 e inclui *sites* que foram substituídos por mudanças de *design*, tecnologia etc., *sites* que não são mais atualizados, e capturas

[540] Disponível em: https://library.shu.edu/library/preservation-policy. Acesso em: 3 jun. 2023.

das páginas ou instantâneos de *sites* ao vivo selecionados, os quais podem ser acessados e visualizados *online* (https://www.worldbank.org/en/webarchives) junto a um perfil de informações (metadados) contendo, por exemplo, o nome do proprietário do *site*, o ano de criação do *site* e a data de arquivamento do *site* (WORLD BANK GROUP, c2022).

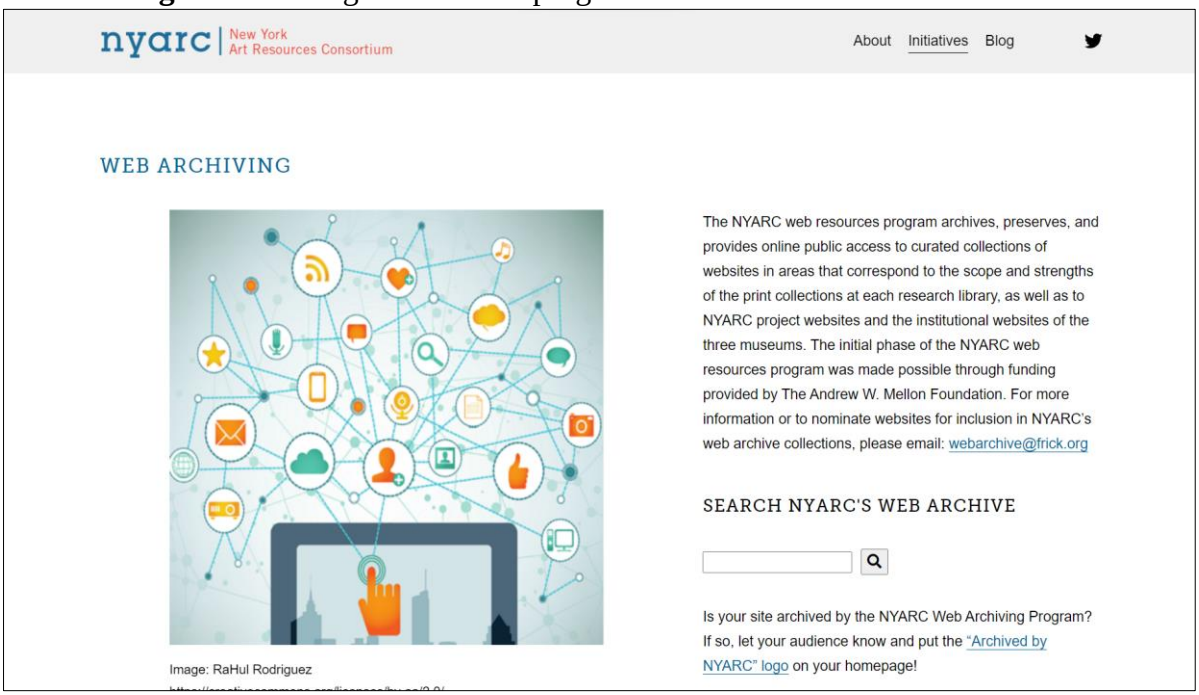
Figura 102 – Página inicial dos arquivos da *Web* do Banco Mundial.



Fonte: *World Bank Group* (c2023).

- *New York Art Resources Consortium* (NYARC) *Web Archiving Program* – como um programa que desde 2012 arquiva, preserva e fornece acesso público *online* pelo serviço *Archive-It* a coleções selecionadas de *sites* em áreas que coincidem com o escopo das coleções impressas nas bibliotecas de pesquisa dos três principais museus de arte de Nova York nos Estados Unidos que integram o NYARC (isto é, *The Brooklyn Museum*, *The Frick Collection* e *The Museum of Modern Art*), o arquivo da *Web* do NYARC se centra, por exemplo, no arquivamento de *sites* de artistas modernos, catálogos de leilões baseados na *Web* e *sites* que atendem às diretrizes de desenvolvimento de coleção do NYARC (ou seja, *sites* de domínio *.org*, *.com* etc.), havendo pedidos de permissão aos proprietários dos *sites*; além disso, os *sites* que mudam com regularidade são arquivados com maior frequência (NEW YORK ART RESOURCES CONSORTIUM, [2022]).

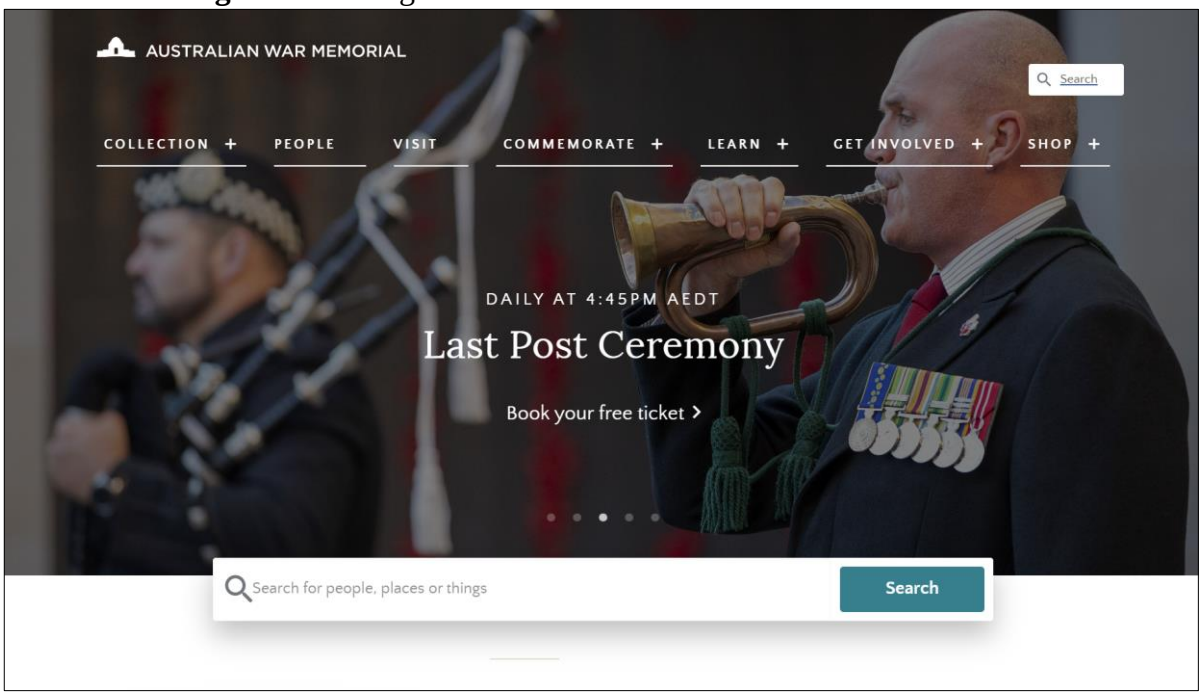
Figura 103 – Página inicial do programa de recursos da *Web* do NYARC.



Fonte: *New York Art Resources Consortium* ([2023?]).

- *Australian War Memorial* – devido a exigência de arquivar e preservar a parte relevante do patrimônio documental da Austrália que está publicamente disponível na *Internet*, o Memorial de Guerra Australiano, como uma agência de coleta especial do PANDORA *Archive*, seleciona, por exemplo, materiais onde a proporção significativa do conteúdo é sobre a história militar australiana ou sobre um assunto de importância social, política, cultural, religiosa, científica ou econômica para o tema, além de *sites* localizados num servidor australiano ou estrangeiro (isto é, *sites* registrados em *.au*, *.com*, *.org* etc.) sobre experiências de australianos em tempos de conflito ou em forças não australianas e/ou de cidadãos não australianos servindo nas forças australianas; em adição, alguns títulos *online* selecionados são arquivados de forma abrangente e outros têm um instantâneo por mês, semestre, ano ou só uma vez (AUSTRALIAN WAR MEMORIAL, [2020?]).

Figura 104 – Página inicial do Memorial de Guerra Australiano.



Fonte: *Australian War Memorial* (c2023).

É, portanto, primordial que todo programa de arquivamento da *Web* seja fundamentado em uma política de seleção consistente e claramente estruturada que possibilite a definição de um conjunto identificável de recursos da *Web* para coleta (com limites e frequências de coleta muito bem delimitados), do qual independentemente da abordagem de seleção empregada (não seletiva, temática etc.) essa terá de considerar tanto as exigências da instituição incumbida pela coleta quanto as características dos materiais a serem coletados (BROWN, c2006). Também a seleção manual demonstra utilidade para uma comunidade ou objetivo particular em que uma avaliação de alto nível de itens é imprescindível e, caso não possa ser feita por robôs, a seleção humana deve ser usada tornando-se preciso organiza-la adequadamente; ademais, mesmo em rastreamentos holísticos, isto é, o arquivamento da *Web* realizado por rastreamentos abertos utilizando a extração de *links* para a descoberta, existe um nível de seletividade e priorização que necessita ser admitido, organizado e justificado, em conformidade com Masanès (c2006a).

Para mais, de acordo com Masanès (2005, c2006c) idealmente qualquer arquivo da *Web* deve ser completo, mas o arquivamento da *Web* é muitas vezes uma questão de escolhas já que um processo de arquivamento perfeito e completo é inalcançável, como observaremos a seguir.

5.11 Definição dos limites

Depois de implementada a política de seleção, ela gerará uma lista de recursos *Web* para serem coletados; esta lista, incluída na própria política (caso seja estática) ou existente como um documento autônomo (se for muito dinâmica), constitui a articulação da política de seleção em nível técnico tendo de ser acionável com o método de coleta pertinente e, como tal, deve definir detalhadamente os limites de cada recurso da *Web* selecionado para que seja possível coletá-lo (BROWN, c2006). Como indaga Masanès (2005), arquivar um *site* significa omitir qualquer documento vinculado fora de seu domínio? Do contrário, até que profundidade os *links* externos devem ser seguidos pelo rastreador *Web*? Para o autor não há uma resposta geral para essas perguntas, apenas respostas específicas baseadas no objetivo principal que orienta o arquivamento; aliás, escolhas tem de ser feitas sobre quais características ou funcionalidades devem ser preservadas, e se o *site* não é, sobretudo, uma coleção de páginas estáticas, é preciso o arquivista da *Web* se concentrar na interação das funcionalidades (não meramente para fins de navegação) e mais de modo geral na experiência que o *site* oferece no contexto arquivístico.

De acordo com Brown (c2006) os recursos da *Web* são definidos em função de um URL que fornece um endereço único para eles na *Web*, e os limites (*boundaries*) de um recurso serão definidos em termos de um URL (e/ou um nome de domínio – *domain name* –), opcionalmente qualificado por um conjunto de parâmetros. Para o autor um URL é composto de três elementos:

- Esquema (*scheme*) – define o formato do URL, normalmente usando um protocolo de comunicação, como HTTP e *File Transfer Protocol* (FTP). Por exemplo, '*http://*'.
- Nome de domínio – diz respeito ao "[...] nome único de um servidor *Web*.", onde "em um URL, o nome de domínio segue imediatamente o identificador do esquema de protocolo de comunicações em rede [...]" (BROWN, c2006, p. XI-XII, tradução nossa) como, por exemplo, '*www.mysite.com*' em '*www.mysite.com/aboutmysite.html*'. Em *International Organization for Standardization* (2013, não paginado, tradução nossa) também é a "cadeia de caracteres de identificação que define um domínio de autonomia administrativa, autoridade ou controle na *Internet*, definido pelas regras e processos [...]" do sistema de nomes de domínio (ou DNS) que é o "sistema de nomeação global hierárquico e distribuído usado para identificar entidades conectadas à *Internet*". Assim, define o *host* (ou hospedeiro, servidor) para o recurso *Web*, sendo que consiste de dois ou mais rótulos (*labels*), separados por pontos. Em '*www.nationalarchives.gov.uk/*', por exemplo, o rótulo mais à direita é o domínio de nível superior, que especifica ou um código de país, como '*uk*' para o Reino Unido (ou um domínio genérico, como '.*com*'

para uma entidade comercial); e o rótulo à esquerda deste define o domínio de segundo nível[541], que descreverá a entidade de hospedagem (por exemplo, '*microsoft.com'*) ou determinará um domínio genérico para qualificar um código de país (por exemplo, '*.gov.uk*'). Um nome de domínio totalmente qualificado também inclui o nome de *host* (*host name*) do servidor *Web*, usando o rótulo mais à esquerda (por exemplo, '*www*').

- *Path* – especifica a localização do recurso *Web* na estrutura de diretório do servidor *Web host* (*host Web server*). Trata do "[...] nome que indica a localização de um arquivo de computador em um sistema de arquivos." e, no âmbito da *Web*, o *path* (caminho) "[...] é a parte de um URL que segue o nome de domínio, e indica a localização de um arquivo no servidor *Web host* [...]" (BROWN, c2006, p. XII, tradução nossa). Por exemplo, '*preservation/webarchive/default.htm*' em '*www.nationalarchives.gov.uk/preservation/webarchive/default.htm*', onde o URL aponta para um arquivo chamado '*default.htm*' localizado dentro do *path* do diretório '*preservation/webarchive*', hospedado no domínio '*nationalarchives.gov.uk*' no servidor *Web host* '*www*'.

Para Brown (c2006) se um único recurso da *Web*, como um documento, for coletado isoladamente, então a lista de coleta precisará apenas indicar o URL desse recurso (por exemplo, '*www.loc.gov/acq/devpol/webarchive.pdf*'), porém se um *site* inteiro, ou subconjunto de um *site*, tiver sido selecionado, isto em geral será definido como um nome de domínio, e talvez um *path* de um *site*, que contém todos os recursos a serem coletados. Em ambos os casos, de acordo com o autor, parâmetros serão necessários para qualificar isto, onde a natureza destes dependerá do método de coleta adotado, como:

- O número de níveis da estrutura do diretório a ser coletado; e
- Se *links* externos serão ou não ser seguidos e, em caso afirmativo, até qual profundidade.

Também segundo Brown (c2006) a definição dos limites de um recurso específico pode exigir alguma análise do recurso ao vivo, e deve ser revista como parte da garantia de qualidade da pré-coleta. Além do mais, o autor pontua que deve haver cuidado para garantir que o domínio seja identificado corretamente, pois *sites* podem tanto utilizar vários nomes de *host* distintos, tais como '*wwwr*' e '*wvvw2*', fornecendo assim acesso a diferentes conteúdos, como também podem incluir subdomínios (por exemplo, o *site* '*www.uktradeinvest.gov.uk*', o qual contém os

[541] Domínio de segundo nível (*second level domain*) consiste em "subdivisões dentro dos domínios de primeiro nível para categorias específicas de organizações ou áreas de interesse (por exemplo, *.gov.uk* para *sites* governamentais, *.asso.fr* para *sites* de associações)", segundo *International Organization For Standardization* (2013, não paginado, tradução nossa).

subdomínios '*www.trade.uktradeinvest.gov.ule*' e '*www.invest.uktradeinvest.gov.ule*'), que podem vir a não ser coletados, a menos que sejam explicitamente identificados e os métodos de coleta estejam adequadamente configurados para assim proceder com as suas devidas capturas.

De outra forma, enquanto a política tradicional de aquisição (*acquisition*) teve que lidar com limitações (*limitations*) financeiras para aquisição, armazenamento etc., o arquivamento da *Web* é igualmente direto e permanentemente prejudicado por dificuldades técnicas para a captura de conteúdos *Web* (MASANÈS, c2006a). Em acordo com Cadavid (2017) e Masanès (c2006a) a *Web* avança mais rápido que as tecnologias para seu arquivamento e este processo está limitado por várias restrições tecnológicas, do qual diferentes tipos de tecnologia desafiam as técnicas atuais de captura, como conteúdo dinâmico e/ou de *streaming* (*streaming content*), *link* profundo (*deep linking*), novos formatos, arquivos multimídia etc., existindo, por essa razão, limites difíceis para o que realmente pode ser arquivado. Isto pode ser associado com a definição proposta por Brugger e Finnemann (2013) sobre os dois tipos gerais de incompletude (*incompleteness*) do arquivo da *Web* em comparação com o que antes estava *online*, a saber:

I. Devido ao fato de que o material não se encontra mais *online*, o usuário de um arquivo da *Web* perderá certas informações sobre a *Web* que em geral estão disponíveis na *Web online* como, por exemplo, resultados de pesquisa ou informações quanto ao estado atual da *Web* (isto é, número de nomes de domínio, usuários etc.) (BRÜGGER, 2012); e
II. À vista de uma combinação indefinida da estratégia de arquivamento escolhida, seleções deliberadas, erros de arquivamento e insuficiências técnicas, num nível mais detalhado, elementos individuais da *Web* e possibilidades de interação podem estar ausentes no arquivo da *Web*, tais como vídeo transmitido (*streamed*), imagens, gráficos e *hiperlinks*.

Neste último tipo de incompletude dos arquivos da *Web*, há ainda limitações inerentes ao rastreamento de *sites* protegidos por senha, e ao reconhecimento do arquivo *robots.txt*[542], que visa restringir o acesso de rastreadores da *Web* a arquivos ou diretórios específicos em um *site* (FELLOWS *et al*., 2008); ou a falhas em rastreamentos de arquivamento para arquivar arquivos *flash* e em linguagem *JavaScrip* (HOCKX-YU, 2011); ou ao não enfoque das tecnologias de arquivamento da *Web* em bibliotecas nacionais para acessar e coletar material da *deep Web*[543], que corresponde à parte não indexada da *Web* (CADAVID, 2017); dentre outras

[542] *Robots.txt* (ou padrão de exclusão de robôs – *robots exclusion standard* –, ou protocolo de exclusão de robôs – *robots exclusion protocol* –) se refere a "[...] um pequeno arquivo em um *site* que informa aos rastreadores da *Web* e outros robôs da *Web* como o *site* deve ser rastreado." (THE NATIONAL ARCHIVES, [2022?a], não paginado, tradução nossa) ou, de acordo com *International Organization for Standardization* (2013, não paginado, tradução nossa), trata-se de um "protocolo usado para impedir que rastreadores da *Web* acessem todo ou parte de um *site*".
[543] Disponível em: https://archive-it.org/collections/11718. Acesso em: 3 jun. 2023.

dificuldades. Especificamente sobre as limitações nas tecnologias de rastreamento, Pennock (c2013) ressalta alguns tipos de conteúdo problemáticos para serem capturados por rastreadores da *Web*, como:

- Banco de dados ou conteúdo dinamicamente conduzido. Por exemplo, páginas *Web* que são geradas através de um banco de dados em resposta a uma solicitação do usuário;
- Arquivos multimídia transmitidos; e os conteúdos protegidos por senha, como já citados antes, onde os rastreadores podem lidar com isto se fornecidos com a senha, porém sem a senha, tais conteúdos são de difícil acesso; e
- Alguns tipos de menus orientados por *JavaScript*. Por exemplo, quando os URLs são gerados por mecanismos dinâmicos.

De acordo com a autora o conteúdo gerado dinamicamente e o conteúdo protegido por senha podem ser enquadrados numa categoria que costuma ser chamada de *deep Web*, do qual é composta de conteúdos que são difíceis para os rastreadores *Web* visualizarem e, assim, terem acesso. Outras questões que interrompem e/ou impedem que um rastreador da *Web* progrida, incluem as regras de proibição do arquivo *robots.txt*, como também já citado antes, que podem vir a não permitir a coleta de muitos conteúdos de um *site* em particular; a complexidade técnica ou configurações do servidor de um *site*; ou os limites operacionais no tamanho do rastreamento em que um escopo de rastreamento excede a quantidade de memória disponível do rastreador *Web* para armazenar *hosts* descobertos ou URLs programados, e as armadilhas de rastreador (*crawler traps*)[544], como é o caso de carrinhos de compras – *shopping carts* – ou de calendários *online* com páginas dinâmicas e sem data final fixada (KOERBIN, c2021; PENNOCK, c2013).

Por sua vez, o que é coletado depende do que o rastreador *Web* ou robô identifica e é realmente capaz de coletar com sucesso (KOERBIN, c2021). Conforme Brügger (2005, 2011, c2018) e Masanès (c2006c), infelizmente, muitas vezes os elementos de um *site* não são todos arquivados quando é usado um *software* para arquivar *sites* inteiros, o que pode ser resultado de: I) condições específicas do programa, tais como quando o *software* de arquivamento não é configurado corretamente, não existe memória suficiente, ou por alguma razão desconhecida, determinados elementos não são arquivados etc.; e II) da dinâmica de atualização (*dynamics of updating*), onde o conteúdo da *Web* pode ter se alterado durante o processo de arquivamento, e

[544] Armadilha de rastreador consisti na "página da *Web* (ou conjunto dela) que fará com que um rastreador caia ou siga infinitamente as referências a outros recursos considerados de pouco ou nenhum valor", sendo que dentro das configurações ou parâmetros de rastreamento pode-se incluir filtros para excluir essas armadilhas, além de "[...] polidez do rastreador (número de solicitações por segundo ou minuto enviadas ao servidor que hospeda o recurso) [...]" e conformidade com o arquivo *robots.txt* (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2013, não paginado, tradução nossa).

não sabemos claramente se, onde e quando isso acontece, sendo imprevisível e irregular; isto constitui numa das principais razões para as deficiências relativas ao tempo, o qual o documento *Web* arquivado é incompleto em comparação com o que já esteve na *Web* ao vivo pois algo é perdido no processo de arquivamento, devido à assincronia entre a atualização e arquivamento.

Neste sentido, conforme Vlassenroot *et al*. (2019), determinadas instituições fazem uso de vários critérios de exclusão (*exclusion criteria*), alguns dos quais se referem à legalidade do conteúdo onde é consenso nas legislações nacionais que pornografia infantil, ódio, discurso xenófobo, racista ou de incitação à violência etc. constituem conteúdos ilegais, como exemplo:

- Na política de métodos de arquivamento da *Web* e diretrizes de coleta das bibliotecas da Universidade do Texas em *San Antonio* nos Estados Unidos (UNIVERSITY OF TEXAS AT SAN ANTONIO LIBRARIES, 2016), listam-se os tipos de conteúdo *Web* que normalmente não são coletados pela instituição, como *sites* protegidos por senha, *sites* criados por estudantes individuais, bases de dados, e *sites* públicos que têm pedidos de exclusão *robots.txt*; e
- Nas diretrizes de seleção de publicações para arquivamento como parte do PANDORA *Archive* do *Australian War Memorial* em *Canberra* na Austrália (AUSTRALIAN WAR MEMORIAL, [2020?]), listam-se os tipos de publicações que não são selecionadas para preservação, tais como *sites* promocionais e publicidade, teses, esboços e trabalhos em andamento, registros organizacionais e *sites* que servem só para organizar informações da *Internet* (por exemplo, diretórios e portais).
- Na antiga política de coleta para arquivamento de *sites* usada no PANDORA *Archive* pela *National Library of Australia* (2018), o qual foi substituída em 2016 por sua nova política de desenvolvimento de coleções da instituição de acordo com *National Library of Australia* ([2022b]), listaram-se categorias de publicações que geralmente não foram coletadas, tais como *sites* de notícias, listas de discussão, salas de bate-papo, quadros de avisos e grupos de notícias, jogos, conjuntos de dados (*datasets*), artigos e trabalhos individuais, e rascunhos e trabalhos em andamento.

Iniciativas em bibliotecas nacionais e universitárias, órgãos de pesquisa, instituições financeiras e de memória

Para mais, além de explicitar o que é, qual a sua importância e como funciona o processo de arquivamento da *Web*, os tipos de conteúdos coletados ou os critérios de seleção dos *sites*, a frequência de capturas, o tamanho do arquivo, os direitos autorais dos *sites* arquivados etc., os

arquivos da *Web* usualmente esclarecem as razões pelas quais alguns *sites* e conteúdos das suas coleções *Web* estão incompletos, indisponíveis ou não são exibidos corretamente, assim como as diretrizes para *sites* preserváveis que possibilitam um arquivamento da *Web* mais bem sucedido, e como lidam com as regras de proibição do arquivo *robots.txt*, entre eles destacamos:

- *Archivo de la Web chilena* – no arquivo da *Web* chilena, a Biblioteca Nacional do Chile sinaliza que os *sites* coletados podem ter *hiperlinks* para outros *sites* ou mídias sociais que não fazem parte do *site* coletado ou da seleção do arquivo e, assim, não aparecerão ou não estarão operacionais; em adição, a biblioteca relata a eficácia do arquivamento da *Web* por meio de uma escala com porcentagens, onde: 100%, corresponde a coleta em que resgata todos os elementos e tecnologias de um *site*; 95%, existem problemas com o *Flash* e com a exibição e/ou *links* para redes sociais; 90%, além dos problemas anteriormente citados, há também problemas para coletar determinadas imagens; 89% ou menos, temos juntos com os problemas acima, dificuldades em identificar elementos *JavaScript*; e 10% ou menos, o *site* é considerado incompatível para ser coletado a partir da tecnologia atualmente disponível (BIBLIOTECA NACIONAL DE CHILE, [2022]).
- NLM*'s Web archive collections* – nas coleções de arquivos da *Web* da NLM, a biblioteca Nacional de Medicina dos Estados Unidos explica os motivos para que alguns *sites* da coleção parecem incompletos. A NLM alega que tenta coletar todos os componentes necessários para renderizar uma página fielmente, contudo, nem sempre isto é possível, visto que algumas partes de um *site* podem ser protegidas por senha ou bloqueadas para rastreadores *Web*; e sobre a situação dos direitos autorais destas coleções, a NLM cita que está criando coleções de arquivos *Web* de acordo com as suas políticas de uso justo (*fair use*), pautadas no *Code of Best Practices in Fair Use for Academic and Research Libraries*[545] da *Association of Research Libraries* (ARL), em que julga uso justo criar coleções de *sites* e outros materiais da *Internet* baseadas em eventos e tópicos e torná-los disponíveis para o uso acadêmico (NATIONAL LIBRARY OF MEDICINE, 2019).
- WAS – no arquivo da *Web* de Singapura da Biblioteca Nacional de Singapura, é dito que a iniciativa procura capturar cada *site* da forma mais abrangente possível como foi exibido na *Internet*, mas indica limitações técnicas para isto, sobretudo, conteúdos que requerem a entrada do usuário ou *plugins* para renderização etc., além disto, em *sites* grandes o rastreador *Web* pode arquivar só uma parte do *site* e não sua totalidade devido

[545] Disponível em: https://www.arl.org/wp-content/uploads/2014/01/code-of-best-practices-fair-use.pdf. Acesso em: 3 jun. 2023.

a restrições de tempo e tamanho de dados, e alguns *links* no *site* arquivado podem não funcionar porque apenas uma página do *site* foi selecionada para arquivamento, dado que só tal página atendeu os critérios de seleção para preservação; em adição, a iniciativa traz diretrizes que ajudam a tornar um *site* amigável ao rastrear, como incluir no *design* do *site* os padrões da *Web*[546] e as diretrizes de acessibilidade[547] do W3C para um melhor arquivamento e exibição do *site* arquivado (NATIONAL LIBRARY BOARD, c2022).

- OASIS*'s National Library of Korea* – no projeto OASIS da Biblioteca Nacional da Coreia, a iniciativa esclarece a razão de alguns recursos da *Web* coletados aparecerem quebrados, bem como o porquê de páginas da *Web* não serem visíveis em seu arquivo da *Web*. O projeto OASIS aponta que *sites* profundos feitos com banco de dados e *sites* dinâmicos escritos em *Flash* e *Java* não podem ser coletados porque o seu rastreador da *Web* não pode extrair as informações de *links* HTML conectados entre os recursos da *Web* a serem coletados, e o OASIS coleta apenas as páginas do *site* de destino da coleta e não conteúdos *Web* relacionados à parte externa do *site*; adicionalmente, a biblioteca elaborou um guia[548] para orientar os desenvolvedores acerca de como construir *sites* e páginas da *Web* para que os rastreadores *Web* ou robôs possam coletar e preservar com segurança e precisão cada *site* de destino (NATIONAL LIBRARY OF KOREA, c2006).
- WARP – no projeto WARP da Biblioteca Nacional da Dieta no Japão, é indicado que existem alguns tipos de arquivos que a iniciativa não pode arquivar devido a limitações técnicas, como arquivos armazenados em um banco de dados, ou definidos para excluir *bots* (isto é, robôs ou rastreadores da *Web*), ou com códigos de caracteres corrompidos, ou que podem ser transmitidos ou reproduzidos à medida que são baixados (*download*), ou cujos *links* são gerados de forma dinâmica com *JavaScript*; ademais, o WARP pontua outros tipos de conteúdos que são difíceis de coletar, incluindo os arquivos de *streaming* e conteúdos dinâmicos ou dados armazenados no banco de dados e exibidos apenas após a realização de uma pesquisa, onde as requisições (consultas) são enviadas ao servidor por ações do usuário (por exemplo, busca, rolar a tela etc.), e um programa do lado do servidor gera resultados e retorna dados (NATIONAL DIET LIBRARY, c2013, 2014).
- *Israeli Internet Archive* – no arquivo da *Internet* Israelense, a Biblioteca Nacional de Israel indica que o arquivo não conterá todos os *sites* Israelenses e, sim, que a varredura

[546] Disponível em: https://www.w3.org/standards/. Acesso em: 3 jun. 2023.
[547] Disponível em: https://www.w3.org/WAI/standards-guidelines/. Acesso em: 3 jun. 2023.
[548] Disponível em: https://www.nl.go.kr/oasis/contents/O6060100.do?page=1&schM=view&id=41&schBcid=pds_oasis&schFld=bd_title. Acesso em: 3 jun. 2023.

(*scan*) automática copiará páginas iniciais (*homepages*) e páginas conectadas às páginas iniciais e assim sucessivamente, em níveis adicionais, mas limitados, de profundidade, significando que a varredura copiará páginas *Web* dos *sites* até uma certa profundidade, e nenhuma outra; além do mais, a biblioteca igualmente sinaliza que as páginas da *Web* que são protegidas por senha e que contêm dados e informações privadas, como detalhes médicos ou financeiros em *sites* de bancos, de compras financeiras, de planos de saúde etc., não serão copiadas para o arquivo (NATIONAL LIBRARY OF ISRAEL, 2022).

- PADICAT: *L'Arxiu Web de Catalunya* – no arquivo da *Web* da Catalunha da Biblioteca da Catalunha, a iniciativa visa preservar os *sites* exatamente como estavam no momento da captura, mas por anomalias no *software* de visualização de arquivos e inconsistências no arquivamento desses *sites* (por exemplo, as exclusões do arquivo *robots.txt*) alguns *sites* podem não ser exibidos corretamente (ou seja, *links* externos, *sites* que exigem senha, formulários etc.); em adição, o PADICAT apresenta recomendações para evitar algumas das dificuldades na coleta dos recursos da *Web* como anomalias na navegação e visualização das versões capturadas dos *sites*, incluindo o arquivo *robots.txt*, onde o PADICAT respeita os *sites* que usam elementos de exclusão, e os *links* ou imagens, *scripts* etc. de outros *sites* externos onde não serão exibidos corretamente uma vez que o *site* tenha sido capturado pelo PADICAT (BIBLIOTECA DE CATALUNYA, c2011).
- *Irish Web archive* (*Cartlann Ghréasáin*) – no arquivo da *Web* Irlandês, a Biblioteca Nacional da Irlanda especifica que existem limitações técnicas no arquivamento da *Web* que indicam os tipos de *sites*/conteúdos que não podem ser arquivados, como caixas de busca (*search boxes*), menus, *sites* que dependem de *JavaScrip/Flash* e *feeds* ou *hashtags* de mídia social ao vivo; ademais, a biblioteca sinaliza que alguns *sites* arquivados estão incompletos ou são exibidos incorretamente por limitações ao que não pode, por agora, ser capturado, como vídeos incorporados (*embedded videos*) no *site* e características dinâmicas (o *Google Maps* etc.), e o fato do não acesso a todos os *links* do *site* arquivado reflete que, às vezes, os *links* em um *site* não são capturados devido a estarem fora de escopo para o arquivamento mostrando, assim, que eles eram externos aos URLs do *site* determinadas para a captura (NATIONAL LIBRARY OF IRELAND, c2022; [201-?]).
- *Luxembourg Web Archive* – no arquivo da *Web* de Luxemburgo da Biblioteca Nacional de Luxemburgo, é explicitado que a iniciativa busca capturar cada *site* exatamente como ele é, com a maior profundidade e detalhe possível, porém existem limites para o que se pode alcançar e muitas razões pelas quais os arquivos da *Web* não são perfeitos, como

National Library of Luxembourg ([2022]). Além das diretrizes seguidas por rastreadores da *Web* que excluem a captura de alguns conteúdos, a iniciativa pontua que há limites técnicos e orçamentários que criam lacunas e capturas incompletas em um arquivo da *Web*, tais como a possibilidade de realizar poucas capturas por ano para a maioria dos *sites*, e que cada versão arquivada de um *site* mostra o conteúdo que estava disponível no momento da captura e quaisquer alterações que ocorram entre as capturas não farão parte do arquivo (BIBLIOTHÈQUE NATIONALE DU LUXEMBOURG, [2021]).

- *Web Archive Switzerland* (*Webarchiv Schweiz*) – no arquivo da *Web* Suíça, a NB ou Biblioteca Nacional da Suíça, explica que áreas protegidas (por exemplo, *intranets* e dados privados protegidos por acesso) não são copiados pelo rastreador *Web*, bem como o arquivo *robots.txt* e *meta tags* são ignorados pelo rastreador com a justificativa de que se eles forem levados em consideração ao copiar, existirá o risco de que o *site* não seja reproduzido completo e corretamente em sua apresentação; diante da dificuldade de se arquivar alguns *sites* devido a alto volume de dados, animações *Flash*, funções de menu etc., a biblioteca orienta que uma característica útil de um *site* amigável ao rastreador da *Web* são os *links* em formato HTML/XHTML que não estão incorporados em *Flash* e *JavaScript*, e opções alternativas de navegação através de uma versão baseada em texto ou mapa do *site* (*sitemap*)[549] (SCHWEIZERISCHE NATIONALBIBLIOTHEK, [2022]).

[549] Mapa do *site* é "[...] uma lista de páginas dentro de um *site*." (THE NATIONAL ARCHIVES, [2022?a], não paginado, tradução nossa).

Figura 105 – Página inicial do arquivo da *Web* Suíça.

Fonte: *Schweizerische Nationalbibliothek* (2023).

- UKWA – no arquivo da *Web* do Reino Unido, a iniciativa aponta que os *sites* coletados visam a refletir o mais completamente possível como o *site* parecia e se comportava na *Internet* num momento no tempo, mas certos elementos nos *sites* arquivados podem não estar presentes seja porque apenas uma página do *site* foi destinada ao arquivo ou, ainda, devido os rastreadores *Web* não conseguirem capturar conteúdo que requer *plugins* para renderização, conteúdo de banco de dados que requer entrada do usuário, componentes interativos baseados em *scripts* de programação etc.; em adição, o UKWA indica dicas para tornar o *site* compatível com rastreadores de modo que possam capturar o máximo de seu conteúdo, como a criação de um mapa do *site* XML[550] para que todo o conteúdo do *site* possa ser rastreado, e o uso do arquivo *robots.txt* para impedir o acesso a áreas do *site* que podem causar problemas se forem rastreadas (UK WEB ARCHIVE, [2022?]).
- *National Library of Scotland* (*Leabharlann Nàiseanta na h-Alba*) – na coleta da *Web* para o UKWA da Biblioteca Nacional da Escócia, a biblioteca cita que visa que a cópia arquivada seja uma representação mais precisa possível do original buscando-se coletar todos os recursos de um *site*, como HTML, imagens, *Cascading Style Sheets* (CSS) e *scripts*, porém conteúdos podem não ser reunidos devido a limitações técnicas,

[550] Mapa do *site* XML (XML *sitemap*) trata-se de "[...] um arquivo que lista as páginas de um *site* em formato XML [...]" "[...] para ajudar os rastreadores a indexá-las." (THE NATIONAL ARCHIVES, [2022?a], não paginado, tradução nossa).

incluindo captura de mídia de *streaming* e conteúdo interativo; além disso, a biblioteca aponta que segue como regra o protocolo de exclusão de robôs, o que faz com que algum conteúdo seja restringido das ações de rastreamento pelo proprietário do *site*, no entanto, em certas ocasiões, ela opta por ignorar o *robots.txt*, como no caso de um conteúdo ser necessário para renderizar uma página ou for tido de valor curatorial e se enquadrar nos limites da legislação inglesa de depósito legal (NATIONAL LIBRARY OF SCOTLAND, c2022).

- OSZK *Webarchívum* – no arquivo da *Web* Húngaro, a Biblioteca Nacional da Hungria indica que nem sempre o rastreador pode lidar com páginas *Web* dinâmicas que contêm muito *JavaScript*, ou que requerem intervenção do usuário, fazendo com que as cópias arquivadas do *site* estejam incompletas onde, por exemplo, as páginas se desfazem, os *links* levam a mensagens de erros, algumas funções interativas não funcionam etc.; aliás, a biblioteca traz orientações tanto para *sites* compatíveis com rastreadores (por exemplo, o fato da parte valiosa do conteúdo não ser muito profunda a partir da página inicial e poder ser acessada via *links* e não só por um formulário de pesquisa) como para *sites* amigáveis para arquivamento de alta qualidade (por exemplo, a existência de metadados detalhados incorporados no cabeçalho das páginas *Web* e em outros documentos – isto é, imagens, arquivos PDF etc. – que facilitam a identificação de metadados automáticos sobre os objetos digitais coletados) (ORSZÁGOS SZÉCHÉNYI KÖNYVTÁR, c2022).
- *Web Archive Austria* (*Webarchiv Österreich*) – no arquivo *Web* Austríaco, a Biblioteca Nacional Austríaca indica que conteúdos ou *sites* externos, *intranets* e dados privados protegidos por acesso não são arquivados; adicionalmente, a biblioteca estabelece que o arquivo *robots.txt* e *meta tags* são ignorados pelo rastreador *Web* com a justificativa de que ela opera o arquivamento da *Web* como parte de seu mandato legal pela Lei de Mídia (em alemão *Mediengesetzes*[551]) para arquivar o espaço *Web* austríaco, que deve ser conferido prioridade (ÖSTERREICHISCHEN NATIONALBIBLIOTHEK, [2022]).
- *Estonian Web Archive* (*Eesti veebiarhiiv*) – no arquivo da *Web* Estoniano, a Biblioteca Nacional da Estônia indica que, apesar dos *sites* arquivados poderem ser navegados num navegador (*browser*) utilizando o *software Wayback* e tornando a experiência do usuário equivalente a navegar na *Internet*, é possível haver deficiências na funcionalidade da versão do arquivo como dificuldades para os *sites* ricos em *scripts* e *streaming* de mídia; demais, a biblioteca traz dicas de como construir um *site* para que ele seja corretamente

[551] Disponível em: https://www.ris.bka.gv.at/Dokumente/BgblAuth/BGBLA_2009_I_8/BGBLA_2009_I_8.html. Acesso em: 3 jun. 2023.

exibido no arquivo, como incluir informações (metadados) para ser possível descrevê-lo precisamente; evitar o uso de seções, pois o rastreador *Web* visitará o *site* e navegará só por *links*; e preferir adicionar à página *Web* do *site* um *link* direto para o arquivo de mídia que pode ser baixado, podendo o rastreador salvá-lo e, mesmo que o *player* do *site* não possa reproduzi-lo, o usuário que navega na página poderá baixar o arquivo de vídeo, áudio ou outra mídia e visualizá-lo (EESTI RAHVUSRAAMATUKOGU, 2022).

- *Spletni arhiv* NUK (NUK *Web Archive*) – o arquivo da *Web* da Biblioteca Nacional e Universitária da Eslovênia indica que visa armazenar cada *site* em sua totalidade, porém sinaliza que isto nem sempre é possível devido a certas partes da *Web* serem impossíveis de capturar e preservar de fato, assim, algumas funcionalidades de seus *sites* arquivados são limitadas ou inexistentes, incluindo conteúdos que podem ser acessados *online* via registro com um nome de usuário e senha, e elementos de *design* ou, talvez, de conteúdo, para sites com *design* específico; além do mais, o arquivo indica recomendações ao criar um *site* para facilitar o processo de sua captura, tais como a criação de um mapa do *site* para assegurar que todo o conteúdo seja abrangido, uma vez que partes do *site* podem não ser identificadas pelo rastreador *Web* (por exemplo, páginas que utilizam *Flash* ou *JavaScript* para navegação) (NARODNA IN UNIVERZITETNA KNJIŽNICA, c2022).
- *Icelandic Web Archive* (*Íslenska vefsafnið*) – no arquivo da *Web* islandesa, a Biblioteca Nacional e Universitária da Islândia pontua que não respeita as regras do *robots.txt*, o qual limitam o acesso ao conteúdo (por exemplo, imagens) que está comprovadamente no escopo da coleção *Web* da biblioteca, e a sua experiência mostra que são poucos os proprietários que usam essas políticas para manter os rastreadores *Web* longe de partes sensíveis de *sites*; ademais, a biblioteca indica algumas diretrizes para que o rastreador acesse e colete mais facilmente os materiais, incluindo se os *sites* atendem aos padrões de acessibilidade (especialmente, os do W3C), a não dependência do *site* das funções do *JavaScript* etc. (LANDSBÓKASAFN ÍSLANDS HÁSKÓLABÓKASAFN, [2022]).
- *Digital Resources – Web harvesting and E-Born Content Archiving* (*Digitálne pramene – Webharvesting a archivácia e-Born obsahu*) – no sistema de informação da Biblioteca da Universidade de Bratislava, é indicado a adoção da ferramenta *online ArchiveReady* (http://archiveready.com/), que avalia se um *site* está pronto para arquivamento e/ou se um *site* será arquivado corretamente por arquivos da *Web* de modo que seja fácil para estes projetos acessá-lo e preservá-lo; sendo assim, esta aplicação *Web* verifica um *site* (isto é, HTML, imagens, CSS, mapas do *site* etc.) depois de inserir o seu URL, a seguir analisa e cria um breve resumo da avaliação da sua arquivabilidade (*archivability*), ou

seja, faz avaliações complexas para calcular a capacidade de arquivamento do *site* por facetas de arquivabilidade: acessibilidade, coesão, metadados e conformidade com os padrões (BANOS, [c2017], UNIVERZITNÁ KNIŽNICA V BRATISLAVE, [2022?b]).

- UTSA *Libraries* – nos arquivos da *Web* das bibliotecas da Universidade do Texas em *San Antonio*, é explicitado que nem todo o conteúdo de um recurso *Web* arquivado pode ser capturado, seja porque proprietários de *sites* podem proibir os rastreadores de acessar as suas páginas e/ou exigir um *login* para acessar o seu *site* ou porque há conteúdos que não podem ser capturados (por exemplo, *feeds* de mídia social que exigem *login* para acessá-los, caixas de comentários – *comment boxes* –); além disso, a iniciativa indica alguns tipos de conteúdo da *Web* que podem não ser capturados adequadamente, como conteúdo baseado em *JavaScript* e vídeos do *YouTube*, e quais conteúdos ela não coleta como é o caso dos *sites* protegidos por senha, bases de dados, calendários (*calendars*) e recursos da *Web* não pertencentes à UTSA com solicitações de exclusão do arquivo *robots.txt* (UNIVERSITY OF TEXAS AT SAN ANTONIO LIBRARIES, 2016, 2022).
- SUL – no arquivo da *Web* das Bibliotecas da Universidade de *Stanford* (ou SWAP) são explicitadas algumas limitações para as capacidades das tecnologias de arquivamento da *Web*, como a difícil captura de *streaming* de multimídia, plataformas de mídia social e conteúdo servido via *JavaScript* assíncrono etc., além das razões de se encontrar *links* quebrados no arquivo da *Web* (que é mais discreto do que a própria *Web*) que refletem o conteúdo não coletado porque não existia quando fora coletá-lo, ou foi muito difícil de capturar com as tecnologias disponíveis ou, ainda, foi deliberadamente excluído do escopo de captura; adicionalmente, a SUL específica diretrizes de arquivamento para tornar mais fácil arquivar o *site* com maior fidelidade, tais como manter os *links* estáveis, adotar formatos/dados duráveis, e conformidade com padrões abertos da *Web* definidos pelo W3C (MELO; ROCKEMBACH, 2020; STANFORD LIBRARIES, [c2022?b]).
- CUL*'s Web resources collection program* – no programa de coleta de recursos *Web* das Bibliotecas da Universidade de *Columbia*, indicam-se determinadas razões para que um *site* arquivado possa parecer incompleto no arquivo, como conteúdos desafiadores ou impossíveis de serem capturados e/ou reproduzidos (por exemplo, menus de navegação guiados por *JavaScript* etc.), e a impossibilidade de capturar arquivos que não estão conectados e têm de ser recuperados de um banco de dados via uma consulta do usuário (por exemplo, uma base de dados de publicações que requer a execução de uma busca para acessar publicações); aliás, a CUL apresenta diretrizes para *sites* preserváveis que ajudam a garantir o sucesso do arquivamento, como a inclusão de um mapa do *site* que

fornece *links* para todo o conteúdo de um *site* específico assegurando que os rastreadores da *Web* encontrarão os conteúdos (COLUMBIA UNIVERSITY LIBRARIES, c2021a).

- Arquivo.pt – no arquivo da *Web* portuguesa, a FCT do Ministério da Educação e Ciência de Portugal informa que os conteúdos e funções constantes no *site* do Arquivo.pt são disponibilizados no estado em que se encontram, e a Unidade de Computação Científica ou, melhor, a Fundação para a Computação Científica Nacional (FCCN) da FCT, não é responsável pela completa exatidão, completude e atualização dos conteúdos, salvo se tiver agido de forma fraudulenta ou com negligência grosseira (ARQUIVO.PT, 2022a). No Arquivo.pt (2022b) são indicadas recomendações essenciais para criar páginas *Web* que podem ser arquivadas e acessadas com eficiência no tempo, como um *link* para cada conteúdo e homepage compatível com rastreadores (no caso da organização do *site*), e *links* em HTML, textos publicados em formatos textuais e tipo de mídia e codificação do conjunto de caracteres identificados corretamente (no caso do conteúdo da página).
- *World Bank's Web Archives Program* – no programa de arquivos da *Web* do Banco Mundial é explicado o motivo de determinados arquivos PDFs não estarem disponíveis e a razão de certas imagens não serem exibidas no arquivo da *Web*. O programa justifica que, ao arquivar um *site*, o rastreador *Web* pode ter encontrado um problema técnico ao tentar copiar um arquivo PDF, e o arquivo pode ter sido corrompido ou não estar mais disponível; além do mais, especificamente no caso de imagens, o programa menciona que problemas de direitos autorais ou algum outro impedimento não técnico pode ter lhe ainda impedido de exibir essas imagens nos *sites* (WORLD BANK GROUP, c2022).
- NYARC *Web Archiving Program* – no programa de arquivamento da *Web* do NYARC, a iniciativa sinaliza que procura preservar a aparência e a funcionalidade de cada *site* como ele apareceu num determinado momento, porém alguns *sites* arquivados não são completamente capturados porque apenas uma página do *site* foi destinada à inclusão na coleção de arquivos, bem como complicações técnicas podem limitar a capacidade dos rastreadores *Web* de capturar mídia rica (*rich media*), conteúdo de banco de dados e outros componentes interativos e, com isto, certos elementos de um *site* arquivado podem não estar presentes; demais, o NYARC explicita algumas medidas para garantir que o conteúdo de criadores de *sites* seja mais facilmente preservado para o futuro, como incluir um mapa do *site*, fornecer *links* em formatos HTML/XHTML e evitar a adoção de formatos proprietários (NEW YORK ART RESOURCES CONSORTIUM, [2022]).

Sendo assim, todos esses limites têm de ser incluídos numa política de seleção sempre que possível, já que eles afetarão a qualidade do arquivo resultante (MASANÈS, c2006a). Há

que considerar, como Koerbin (c2021), que o mundo *online* atual dos cidadãos não se restringe nas divisas territoriais das nações e, sim, varia onde quer que seja pela *Web* sem fronteiras e, é claro, nas plataformas de mídia social, em que conceitos e limites de publicação e comunicação são confusos ou inexistentes e os formatos (junto às plataformas de publicação internacional) podem ser os mais desafiadores em termos técnicos e jurídicos. Logo, para o autor, existe uma grande quantidade de conteúdo que não está publicado só em domínios nacionais (por exemplo, *sites* australianos registrados em *.au*), mas também em domínios internacionais (por exemplo, *sites* de organizações, notícias baseadas na Austrália e outros, sob extensões genéricas *.com* etc.); bem como a missão de depósito legal de *sites* tem fronteiras jurisdicionais e uma coleção representando uma nação e seu povo é essencialmente condicionada pelo alcance desta missão.

No entanto, Koerbin (c2021) afirma que, ao julgar os resultados práticos do objetivo de coleta abrangente, devemos reconhecer que não se trata de coletar todo e cada recurso, em quaisquer formas e períodos, e sim empregar métodos, tecnologias e garantias disponíveis para coletar em prazo e escala ideais de modo a oferecer uma representação inteligível do conjunto.

6 CONSIDERAÇÕES FINAIS

Na prática, o arquivamento da *Web* se configura como um processo e o arquivo da *Web* constitui o seu produto. Assim, o processo de arquivamento da *Web* envolve atividades que se centram em realizar a preservação digital de informações registradas na *Web*, sejam elas nato-digitais ou digitalizadas e publicadas *online*, armazenando-as em um sistema de repositório (ou arquivo da *Web*) a fim de assegurar o seu acesso utilizável a longo prazo para uma comunidade de usuários, como organizações, profissionais, governos, pesquisadores, acadêmicos e público em geral. Os registros coletados da *Web*, ou melhor, os *sites* arquivados (e as informações neles contidas) podem ser vistos, lidos e navegados no arquivo da *Web* como se estivessem na *Web* ao vivo mas, em acordo com *The National Archives* (c2011, [2022?c]), eles são preservados como instantâneos (ou representações) das informações em determinados momentos ao longo da sua vida útil que, junto das imperfeições técnicas ocorridas no arquivamento que afetam a sua exibição e funcionamento/comportamento correto, reforçam muito o argumento de que as coleções dos arquivos da *Web* não devam ser vistas e usadas como *backups* completos de *sites*.

Por sua vez, lidando comumente com partes da *Web* pública e *online*, os arquivos da *Web* são um novo tipo de instituição de memória que consiste de uma ou diversas coleções de *sites*/páginas *Web* arquivadas compostas de HTML, textos, imagens, áudios, *scripts*, *JavaScript* e folhas de estilo ou CSS (que dão a eles dinamicidade, aparência e comportamento) além de outros elementos. Estas coleções de endereços de URLs rastreados, criadas individualmente por uma instituição ou colaborativamente entre diversas organizações de arquivamento, podem ser organizadas de várias formas como, a título de exemplo, por coleções especiais (Covid-19 etc.) e categorias (patrimônio e cultura, personalidades, educação etc.) no WAS da *National Library Board* (c2022); por áreas temáticas (sociedade, saúde, humanidades, tecnologia, pesquisa etc.) e eventos (eleições etc.) no arquivo da *Web* eslovena da *Narodna in Univerzitetna Knjižnica* (c2022); ou por coleções temáticas (tópico/tipo de material, como música etc.), domínios *Web* (*sites* registrados em *.es*, *.gal* etc.), eventos em destaque (ataque terrorista etc.) e *sites* em risco de desaparecimento, no arquivo da *Web* espanhola da *Biblioteca Nacional de España* ([2023]).

Sobre os atuais motivos para se arquivar *sites*, como vimos no trabalho, há primeiro um conjunto de razões mais gerais e, em algum grau, "românticas" para que instituições (sobretudo, bibliotecas e arquivos nacionais) desenvolvam arquivos da *Web* e, em segundo, outros fatores que de fato as levam a engajarem-se na área. No primeiro grupo temos a ideia da urgência de se preservar os conteúdos *Web* pois eles estão sendo perdidos para sempre ou de se salvaguardar o patrimônio histórico-cultural digital em risco produzido pelas nações pautando-se na missão

social de arquivar esses conteúdos enquanto uma extensão das tarefas clássicas das instituições completando coleções físicas existentes e, também, os inúmeros casos de uso dos arquivos da *Web* (por exemplo, como fonte de dados permanentes para pesquisa e ensino, ou para análise de eventos e períodos do passado e, talvez, antever fatos futuros etc.). Já no segundo grupo existe a questão da legislação que pode obrigar e, juntamente, conferir o direito e a proteção das instituições de capturar e arquivar sob depósito legal materiais da *Web* ou, por outro lado, impor que conteúdos publicados no passado (registros de organizações, governos etc.) sejam arquivados para conformidade normativa e apoio como provas em defesa de processos judiciais.

Conforme foco deste trabalho, o arquivamento da *Web* inicia-se com o estabelecimento de uma política de seleção (e avaliação dos conteúdos *Web* a serem preservados) para, assim, iniciar a captura, armazenagem e disponibilização dos *sites* arquivados em sua forma original (aparência, navegabilidade etc.) no momento em que foram coletados (LOHNDORF, [2023]; ROCKEMBACH, 2019). A criação de uma política de seleção de arquivamento da *Web*, como notamos no trabalho, passa pelos processos de definição de contexto (a integração da política com todas as políticas de seleção internas, externas e/ou organizacionais), a identificação de limitações (a explicitação de limites para a captura dos conteúdos que afetam a sua exibição completa/correta e a qualidade da coleção *Web* resultante) e a descrição dos padrões de coleta (o gatilho – isto é, domínios *Web* na agenda de arquivamento, eventos etc. – e sua frequência de coleta, a duração da campanha de arquivamento e as relações entre as campanhas). Outros procedimentos, não menos importantes, delineiam essencialmente os critérios para a seleção de conteúdos acordados na política, entre eles a determinação dos métodos de seleção, do escopo, cobertura e alvo (extensão), do tipo e mecanismos de acesso, dos usos previstos da coleção *Web* e seu público-alvo e, evidentemente, dos próprios critérios para tomada de decisões de seleção.

Sobre estes últimos procedimentos, primeiramente, os escopos dos principais métodos de seleção usados na criação de coleções *Web* temáticas, seletivas, de domínio etc. nos revelam diferentes critérios, sobretudo, nas abordagens temáticas e seletivas (BROWN, c2006; KRAN; RAHMAN, 2019; MURRAY; HSIEH, 2007). No método temático, a seleção pode basear-se em um assunto de relevância nacional ou internacional, tais como tópicos/áreas de interesse e eventos; ou em um indivíduo e/ou organização (por exemplo, universidades) que criou ou gere o recurso *Web*; ou em gêneros de recursos *Web*, como *blogs*; ou em domínios *Web*, incluindo *sites* de domínio nacional (.br, gov.br, .edu.br etc.), regional (por exemplo, *.cat* no PADICAT) e internacional (*.com*, *.net* etc.). De outro modo, no método seletivo a decisão de inclusão pode ser no nível de *site* (os *sites* de um domínio selecionado), de página *Web* (as páginas de um *site* selecionado) e de conteúdo da *Web* (o conteúdo a ser preservado do *site*); além disto, a seleção

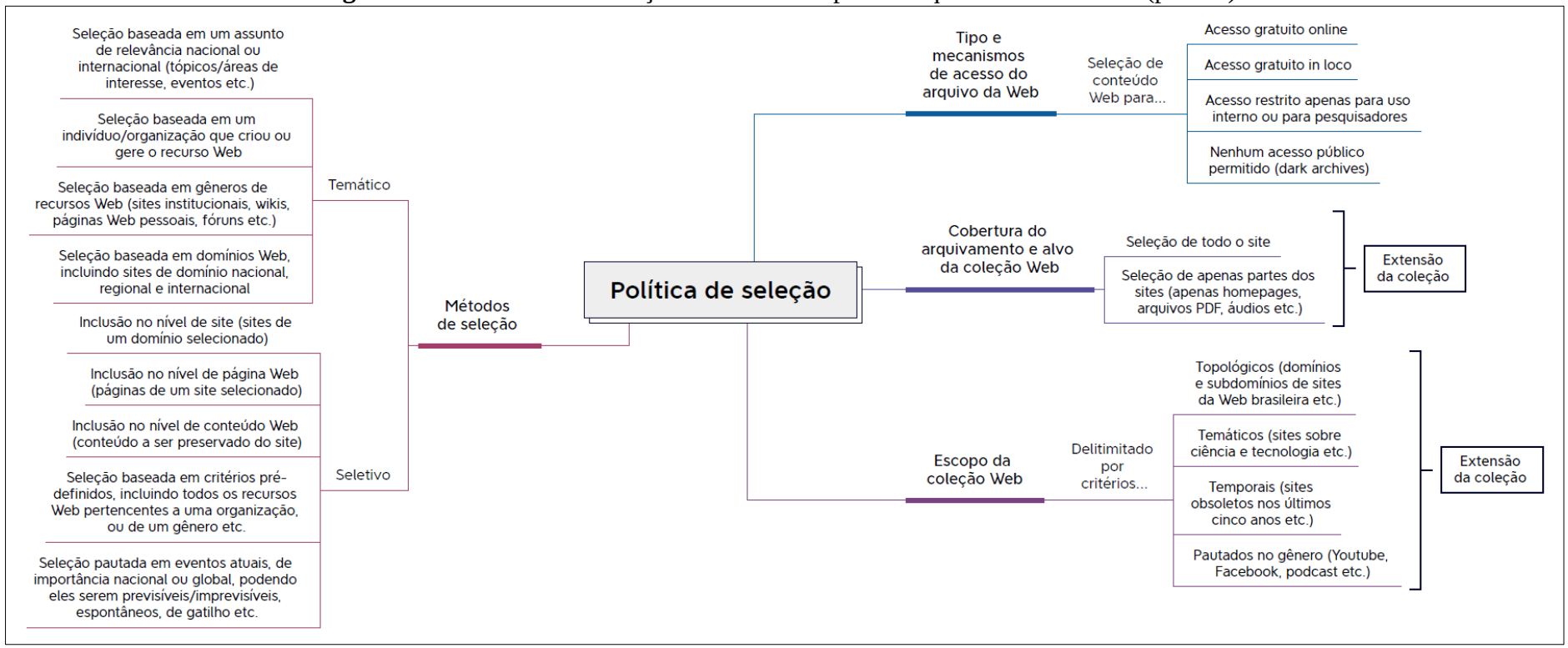
pauta-se tanto em critérios pré-definidos, incluindo todos os recursos *Web* de uma organização (ou grupo de organizações) ou de um gênero, recursos *Web* sobre uma comunidade específica numa organização (docentes, alunos etc.) e todos os recursos *Web* que podem vir a beneficiar usuários externos (jornalistas, cientistas etc.), como também em eventos atuais, de importância nacional ou global, podendo eles serem previsíveis/imprevisíveis, espontâneos, de gatilho etc.

Em segundo, o escopo da coleção pode ser delimitado por critérios de seleção temáticos, topológicos, pautados no gênero etc.; e a cobertura do arquivamento e o alvo da coleção trazem consigo tanto os critérios de se selecionar todo o *site* ou só partes deles (vídeos, arquivos PDFs, *hyperlinks* etc.) onde, de certa forma, reflete os limites sinalizados nos arquivos da *Web*, como ainda os critérios de inclusão e exclusão de conteúdo para a coleção segundo qualidade, gênero, editores e outros, se relacionando com os eixos de avaliação da fase de filtragem no processo de seleção sugerido por Khan e Rahman (2019) e por Masanès (c2006a). Também na definição do tipo e dos meios de acesso, os arquivos da *Web* em geral são acessados gratuitamente *online* e, por vezes, só *in loco* (por exemplo, no WAS e no *New Zealand Web Archive*), ou podem ser arquivos ocultos (*dark archives*) com nenhum acesso público permitido (por exemplo, no *Coca Cola Web Archive*), ou terem acesso restrito só para uso interno ou para pesquisadores, sendo que em caso de concessão do acesso ao público deve-se respeitar direitos autorais, contratos com titulares de direitos e permissões dos proprietários de *sites*, segundo Shiozaki e Eisenschitz (2009) e Vlassenroot *et al*. (2019); e na definição dos usos esperados do arquivo da *Web* e dos usuários pretendidos, os arquivos dirigem-se basicamente à pesquisa, podendo essa ser para uso científico, profissional, apoio legal etc., e os usuários podem ser da instituição de arquivamento (funcionários, alunos etc.) e/ou usuários externos (ex-alunos e ex-funcionários, público geral etc.).

Finalmente, além dos critérios de seleção discutidos na literatura científica e expostos no trabalho, nas iniciativas de arquivos da *Web* levantadas identificam-se diferentes frequências de captura dos *sites*. Os *sites* selecionados nos arquivos da *Web* podem ser coletados tanto uma vez ou várias vezes, seja ao dia, por semana e ao ano, como também em certas datas de maneira extraordinária em resultado de eventos importantes (por exemplo, eleições), e a frequência dos rastreamentos dos *sites* é dependente da sua prioridade definida na coleta, ou da sua relevância histórica, ou da regularidade das alterações de seu conteúdo. Aliás, obedecendo leis nacionais e políticas institucionais de desenvolvimento de coleções, e empregando coletas de domínios nacionais ou institucionais, seletivas etc. dos *sites*, seja como parte do patrimônio documental de países ou registros de organizações (universidades, bibliotecas nacionais, governos etc.), as iniciativas analisadas adotam um conjunto específico e, em algum grau, não padronizado, de

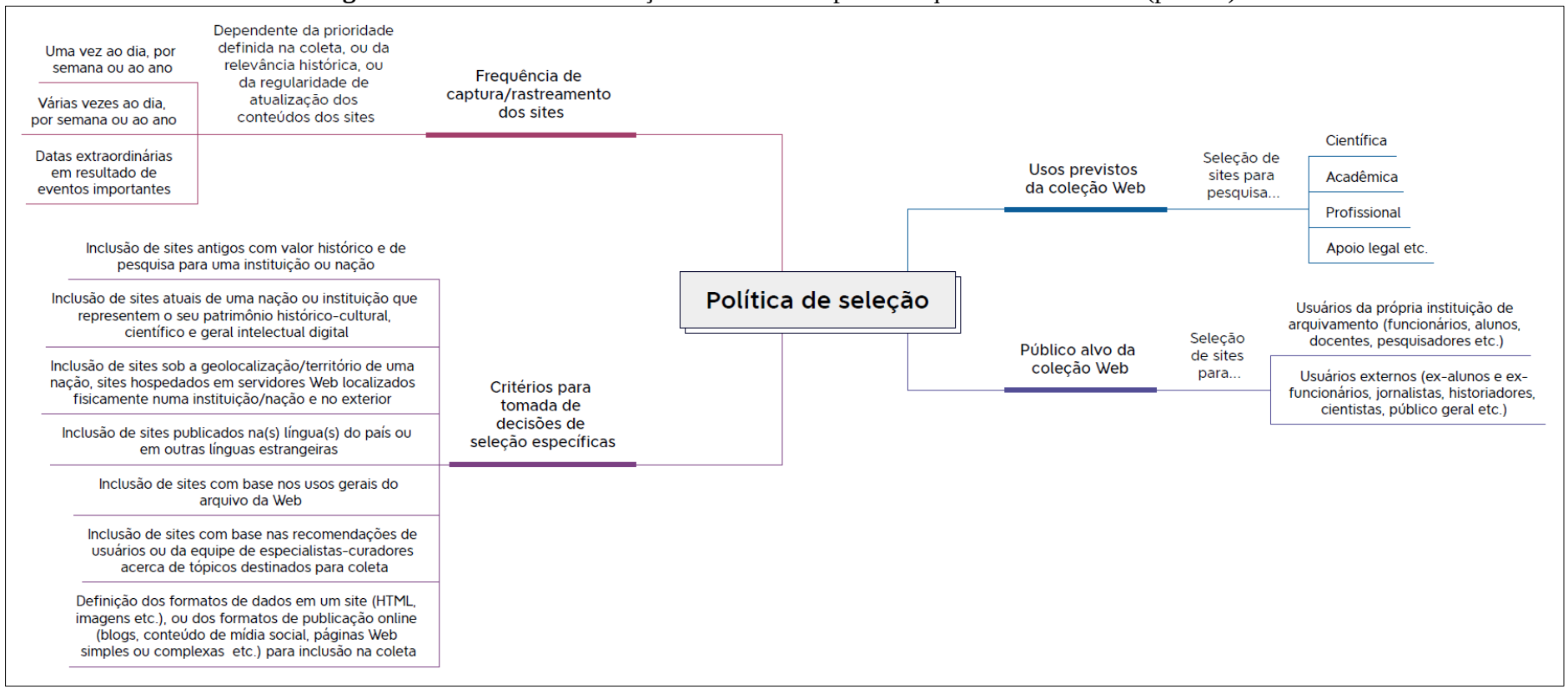
critérios para tomada de decisões de seleção de materiais da *Web* pública, que são sintetizados nas Figuras 106 e 107.

Figura 106 – Critérios de seleção de conteúdos para o arquivamento da *Web* (parte 1).



Fonte: Elaborado pelo autor.

Figura 107 – Critérios de seleção de conteúdos para o arquivamento da *Web* (parte 2).



Fonte: Elaborado pelo autor.

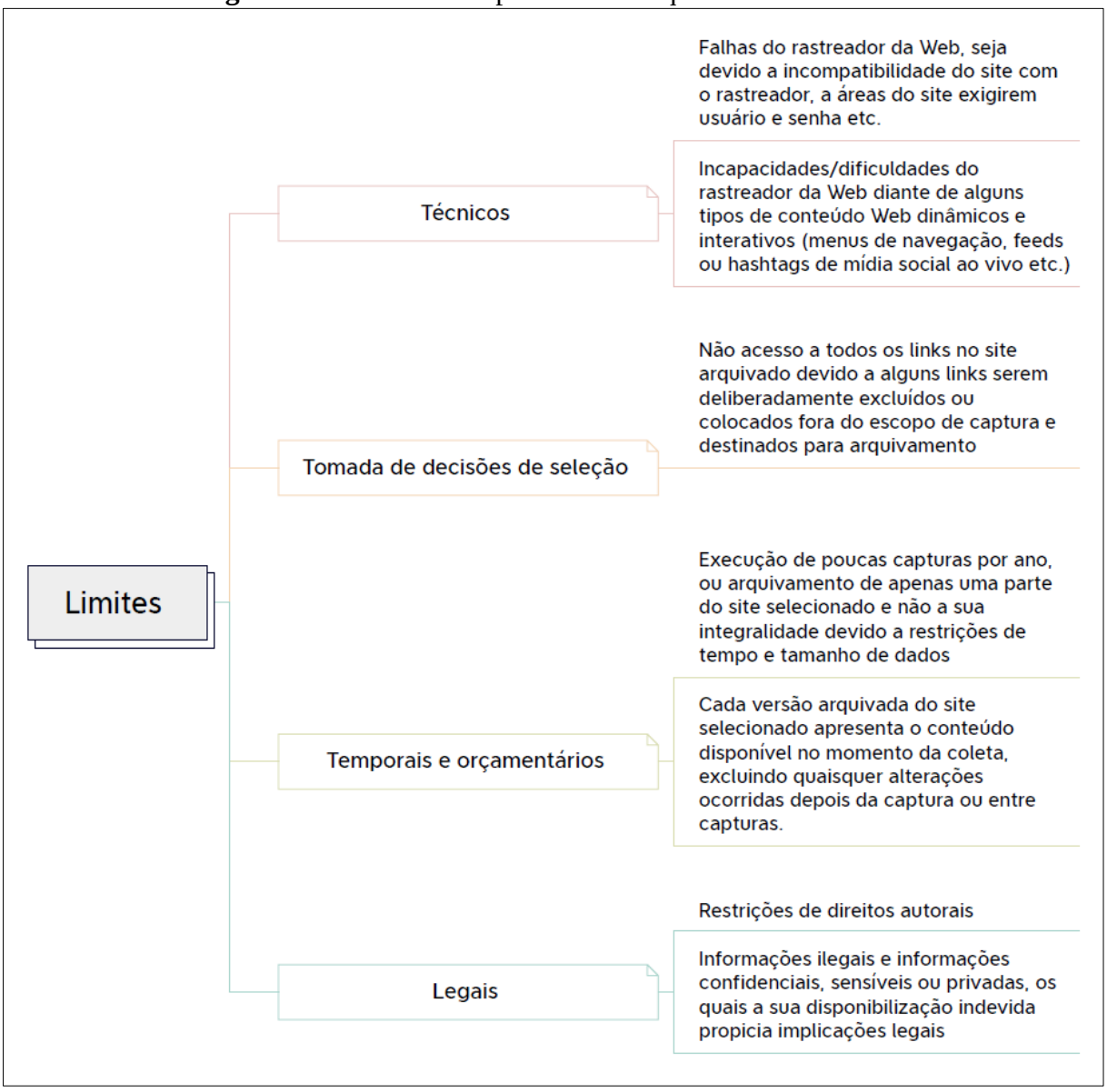
Apontados nas Figuras 106 e 107, os critérios de seleção de conteúdos *Web* identificados a partir das estratégias, políticas de seleção e iniciativas de arquivos da *Web* discutidas no trabalho, os quais são aplicáveis na estruturação de diretrizes e políticas institucionais de preservação digital e arquivamento da *Web*, podem ser descritos da seguinte forma:

- Critérios para *sites* "antigos" e atuais – inclusão de *sites* antigos com valor histórico e de pesquisa para uma instituição ou nação, em especial, *sites* que foram descontinuados ou atualizados consideravelmente, substituídos por mudanças de *design*, de tecnologia etc., ou *sites* que não são mais atualizados (e que correm riscos de perda permanente); e a inclusão de *sites* atuais de uma nação ou instituição que represente significativamente o seu patrimônio histórico-cultural, científico e geral intelectual digital, produzido e/ou publicado na *Web*, sendo que esta coleção de *sites* públicos destinados para coleta pode consistir em registros acerca de eventos, educação, personalidades, esportes, governo e política, locais, natureza e meio ambiente, comunicações, artes, patrimônio e cultura etc.
- Critérios geográficos e linguísticos – a inclusão de *sites* sob a geolocalização/território de uma nação (por exemplo, *sites* no domínio de nível superior .br com o código do país do Brasil – incluindo org.br, edu.br, gov.br etc. –, de propriedade ou não de brasileiros ou de órgãos do país); ou *sites* hospedados em servidores *Web* localizados fisicamente numa instituição ou nação e no exterior (por exemplo, *sites* de domínio "ufscar.br" da UFSCar, como *sites* de departamentos etc., e *sites* importantes para a universidade ou ao Brasil de domínios genéricos *.org*, *.com* etc.); ou *sites* publicados na(s) língua(s) do país ou em outras línguas estrangeiras (por exemplo, *sites* em português brasileiro e em inglês sobre o Brasil e os brasileiros, a UFSCar, de autores brasileiros natos ou não etc.).
- Critérios relativos ao assunto – a inclusão de *sites* cuja a proporção significativa do seu conteúdo é sobre a área temática destinada para coleta no arquivo da *Web*, como ciência no Brasil; ou *sites* sobre um ou vários assuntos de relevância social, cultural, religiosa, política, científica, econômica, educacional etc. para a área temática de interesse, como cortes de verbas, *fake news* e desinformação, e mulheres na ciência. Por exemplo, no arquivo da *Web* de Luxemburgo da *Bibliothèque Nationale du Luxembourg* ([c2022]) emprega-se alguns critérios de seleção, entre eles: topicalidade (*topicality*), que consiste no questionamento se o *site* contém informações sobre o tópico da coleção; e relevância (*relevance*), que trata do questionamento do quão importante é o conteúdo do *site* para o tópico e qual a probabilidade de haver informações mais interessantes num momento posterior, e este critério é útil para definir a profundidade da coleta já que, às vezes, não seja pertinente o *site* inteiro para o tópico e, sim, só algumas de suas páginas/conteúdos.

- Critérios pautados nos usos do arquivo da *Web* e em seus usuários ou especialistas – a inclusão de *sites* a partir dos intuitos gerais do arquivo, como *sites* que complementam ou relacionam-se com os arquivos impressos (e talvez coleções digitais) mantidos pela instituição e *sites* que apoiam a pesquisa e ensino desta; ou com base nas recomendações de usuários ou, ainda, da equipe de especialistas-curadores sobre tópicos destinados para a coleta no arquivo da *Web*. Exemplificando, nas diretrizes de coleta da *Web* da *Library of Congress* (2022a) são delimitados alguns fatores ao recomendar conteúdo arquivável da *Web* para aquisição, entre eles: utilidade em atender às necessidades informacionais atuais ou futuras do Congresso americano e dos pesquisadores da biblioteca; conteúdo acadêmico; e relação com os demais recursos informacionais nas coleções da biblioteca.
- Critérios relativos ao formato – a definição de quais formatos de dados incluídos em um *site* farão parte ou não da coleta, como HTML, PDFs, arquivos *Office*, imagens, áudio etc.; ou quais formatos de publicação *online* serão ou não aceitos na coleta, como *blogs*, publicações em XML, conteúdo de mídias sociais, páginas *Web* simples e estáticas ou complexas dinâmicas e interativas. Como exemplos, nas diretrizes de coleta da *Web* da *Library of Congress* (2022a) é especificado que tenta-se reunir os objetos multiformatos de um *site* e, por limitações técnicas nas ferramentas de arquivamento da *Web*, exclui-se alguns conteúdos, tais como *podcasts* e *YouTube*; e no arquivo da *Web* da Áustria da *Österreichischen Nationalbibliothek* ([2022]) coleta-se vários recursos, como materiais não publicados em formato impresso e que são cientificamente relevantes e citados em formato digital (ou seja, artigos, *preprints*, relatórios etc. em *sites* de cientistas e projetos de pesquisa), com foco em texto e imagem, e exclui-se da coleta recursos cujo conteúdo é principalmente jogo, *software*, listas de discussão, publicidade, grupos de notícias etc.

Ademais, embora as iniciativas de arquivos da *Web* estudadas neste trabalho busquem, antes de mais nada, capturar os *sites* precisamente como eles são originalmente em aparência, funcionalidade, comportamento etc. (representação) na *Internet*, com a maior profundidade e detalhamento (metadados) completo possível e com foco em preservar porções significativas da *Web* baseado em seus próprios princípios e julgamentos, existem impedimentos para o que na prática elas podem coletar e arquivar e outros motivos para que uma parte das coleções *Web* construídas por elas sejam incompletas ou exibidas incorretamente, de acordo com a Figura 108.

Figura 108 – Limites do processo de arquivamento da *Web*.



Fonte: Elaborado pelo autor.

Os limites do processo de arquivamento da *Web*, apontados na Figura 108, podem ser descritos da seguinte maneira:

- Limites técnicos – abrange as falhas do rastreador da *Web* em coletar conteúdos em um *site*, seja devido a incompatibilidade do *site* com o rastreador, a partes/áreas do *site* que exigem usuário e senha (por exemplo, intranets e informações privadas protegidas por um *login* de acesso etc.) ou a esse conteúdo ter acesso bloqueado para rastreadores (por exemplo, o arquivo *robots.txt*); além das incapacidades/dificuldades destas tecnologias de arquivamento da *Web* disponíveis perante a alguns tipos de conteúdo *Web* dinâmicos e interativos que são desafiadores ou impossíveis de serem capturados e/ou reproduzidos adequadamente no momento atual, tais como arquivos armazenados em banco de dados

e que requerem intervenção do usuário, *streaming* de multimídia, menus de navegação e outros conteúdos guiados/servidos por *JavaScript*, componentes interativos pautados em *scripts* de programação, caixas de busca e *feeds* ou *hashtags* de mídia social ao vivo.

- Limites atrelados a tomada de decisões de seleção – como o não acesso a todos os *links* no *site* arquivado, seja porque apenas uma página do *site* (por exemplo, *homepages*) foi destinada para arquivamento da *Web* ou porque alguns *links* de outros *sites*/conteúdos *Web*, externos aos URLs do *site* definidas para captura, foram deliberadamente excluídos ou colocados fora do escopo de captura para o arquivamento do arquivo da *Web*.
- Limites temporais e orçamentários – como a possibilidade de executar algumas capturas por ano (frequência) para a maior parte dos *sites* selecionados, ou de arquivar só uma parte do *site* selecionado e não a sua integralidade por restrições de tempo e tamanho de dados, ou o fato de cada versão arquivada de um *site* selecionado apresentar o conteúdo que estava disponível no momento em que a captura foi executada pelo rastreador *Web*, excluindo quaisquer alterações ocorridas depois da captura ou, mesmo, entre capturas.
- Limites legais – inclui as restrições de direitos autorais, e as informações ilegais (como exemplo, conteúdos com pornografia infantil, discurso de ódio, xenófobo, racista ou de incitação à violência) e informações confidenciais, sensíveis ou privadas protegidas por senha e sobre indivíduos e organizações (por exemplo, registros de governos, detalhes médicos ou financeiros em *sites* de bancos e de compras financeiras), dos quais a sua disponibilização indevida a pesquisadores e/ou ao público propiciará implicações legais.

Diante disto, conforme vimos no trabalho, muitas iniciativas oferecem orientações para criação de *sites* preserváveis que auxiliam a torná-los compatíveis/amigáveis ao arquivamento de alta qualidade, facilitando o acesso, a coleta e o rastreamento correto dos seus conteúdos por rastreadores *Web*, como: o WAS, o OASIS, o Arquivo.pt, o Arquivo da *Web* Suíça, o UKWA, o Arquivo da *Web* estoniano, o Arquivo da *Web* da Biblioteca Nacional e Universitária da Eslovênia, o Arquivo da *Web* Húngaro, o Arquivo da *Web* islandesa, o SWAP, o programa de arquivamento da *Web* do NYARC e o programa de coleta de recursos *Web* das Bibliotecas da Universidade de Columbia. A atenção nestas recomendações (junto da adoção de tecnologias e padrões internacionais da *Web*) pode, por exemplo, como Melo e Rockembach (2020), conduzir a um alto grau de capacidade de arquivamento dos *sites* pelos arquivos da *Web* ou, melhor, de arquivabilidade de *sites* (*website archivability*) para preservação digital, que se trata da "[...] extensão em que um *site* atende às condições para a transferência segura de seu conteúdo para um arquivo da *Web* visando a sua preservação." (BANOS *et al.*, c2013, p. 11, tradução nossa).

Há de se levar em consideração que a identificação de critérios de seleção adotados em iniciativas de preservação digital e arquivamento da *Web* ou, melhor, nos arquivos da *Web*, é um tema hoje delicado. Não existe um debate profundo destes critérios nas iniciativas que foram levantadas no trabalho e, se existem, está implicitamente em suas políticas onde muitas vezes não é uma prioridade publicadas e/ou disponibilizadas ao público e pesquisadores, tornando-se difícil encontrar critérios de seleção bem definidos. Aliás, embora não seja o foco do trabalho, é importante salientar que temos diferenças, seja ela conceitual, prática etc., entre as palavras "preservação" e "arquivamento" da *Web*, o que traz a discussão de que talvez uma ou mais iniciativas de arquivos da *Web* analisadas podem desenvolver o arquivamento da *Web*, mas não cheguem ao nível de realizar com efeito a preservação digital dos conteúdos *Web*. A título de exemplo, no arquivo da *Web* da Áustria (ÖSTERREICHISCHEN NATIONALBIBLIOTHEK, [2022]) não encontramos qualquer menção clara que a iniciativa procura realizar a preservação digital de longo prazo dos *sites* coletados e arquivados referentes ao espaço da *Web* austríaca.

Além disso, a palavra "arquivo da *Web*" podendo significar um sistema de repositório, ou um tipo de instituição de memória, ou uma organização que coleta e preserva partes da *Web*, ou um tipo de biblioteca digital, ou as coleções *Web* arquivadas etc., denota que este conceito não está muito bem definido na área recente do arquivamento da *Web*. Especificamente para o primeiro significado, por exemplo, a definição do arquivo da *Web* como uma espécie de sistema de repositório talvez não seja o mais ideal, pois nem todo repositório possui a função única e exclusiva de preservação digital. Visto que muitas das iniciativas levantadas no trabalho fazem parte do IIPC, esta organização compõe um modelo relevante de consórcio global de bibliotecas nacionais e outras instituições internacionais consolidadas que se ajudam entre si para preservar o conteúdo da *Internet*, contudo, é notório a alta presença de membros de países ricos da Europa e da América do Norte. Apesar dos esforços na área nos últimos anos, a realidade de consórcio para arquivamento da *Web* ainda está distante do caso brasileiro, porém a biblioteca nacional do Brasil seria pertinente como um arquivo de páginas *Web* conforme vemos em demais países.

Outra questão importante concerne ao papel fundamental que os usuários executam no arquivamento da *Web*. Se por um lado, os usuários podem participar ativamente na construção das coleções *Web*, como observamos em algumas coleções criadas pelo grupo de trabalho CDG do IIPC e pelo grupo WAHR da Universidade de *Waterloo* com parcerias com as Universidades de *Western* e de *York* no Canadá, por outro lado, os usuários igualmente são atores essenciais na "sobrevivência" da própria iniciativa de arquivo da *Web* e suas coleções desenvolvidas. A criação do arquivo da *Web* e seus objetivos e o estabelecimento do processo de arquivamento da *Web* (isto é, a política de seleção, a coleta, o armazenamento, a disponibilização dos *sites*

etc.) devem-se pautar sempre nos usuários pretendidos. O sucesso desse processo e do arquivo da *Web* está atrelado a se estes correspondem as necessidades e os desejos do seu público-alvo atual e futuro (aplicação de metodologias de estudos de usuários). Também o uso das coleções será um fator dominante para a comprovação da qualidade e do êxito do processo e seu produto (isto é, o arquivo da *Web*) como do alcance dos objetivos, além da obtenção de financiamento para o seguimento das atividades do arquivo da *Web* e a sua sustentabilidade em longo prazo.

De fato, o arquivamento da *Web* não é perfeito e os arquivos da *Web* são incompletos em termos de integralidade dos *sites* arquivados. Isto decorre não apenas devido à natureza da *Web* e as deficiências das tecnologias de arquivamento da *Web* existentes, mas ainda pelo viés de seleção adotado nas políticas de arquivos da *Web* como, por exemplo, as diretrizes seguidas por rastreadores *Web* que definem o nível de profundidade de coleta dos *sites* e quais conteúdos serão capturados, repercutindo assim na qualidade final das coleções *Web* disponibilizadas. Por sua vez, as decisões de seleção trazem algum grau de subjetividade e deliberação, o que pode ser evitado com a justificação explícita dos critérios em uma política de seleção, seja baseado nos recursos técnicos, humanos e financeiros disponíveis, nos usuários, na opinião de curadores e especialistas etc., tornando o arquivamento e o seu produto final (as coleções *Web*) coerentes.

Portanto, temos a necessidade de estudos futuros sobre os impactos sociais, acadêmicos, de memória etc. das políticas e critérios de seleção utilizados em iniciativas da área, buscando-se responder se estes atendem as demandas institucionais ou nacionais, se deixam de preservar algo de significante valor hoje e no futuro, se apoiam as necessidades do público alvo (usuários) entre outras coisas. Além disto, as futuras investigações poderiam abranger como definir a integralidade do material *Web* arquivado; ou quais critérios devemos adotar para determinar o armazenamento do material no arquivo da *Web*; ou como analisar o reuso do material arquivado para se averiguar a inclusão ou a exclusão de um *site* ou página *Web*; além da identificação de critérios de seleção de conteúdos para o arquivamento da *Web* dentro do campo da Ciência da Informação ou, melhor, mais aplicáveis ao contexto de bibliotecas, de arquivos e/ou de museus.

REFERÊNCIAS

À PROCURA da felicidade. Direção: Gabriele Muccino. Produção: Todd Black; Jason Blumenthal; James Lassiter; Will Smith; Steve Tisch e Teddy Zee. Roteiro: Steven Conrad. Culver City, United States: Columbia Pictures, 2006. 1 DVD (118 min). Baseando no livro "The Pursuit of Happyness", de Chris Gardner.

ADAMS, Kimberly; FERNANDEZ, Sasha. Digital archivists race to preserve ukrainian heritage. **Marketplace Tech**, [Saint Paul, Minnesota], 11 March 2022. Disponível em: https://www.marketplace.org/shows/marketplace-tech/digital-archivists-race-to-preserve-ukrainian-heritage/. Acesso em: 7 jun. 2023.

ALEMNEH, Daniel Gelaw; HASTINGS, Samantha Kelly. Exploration of adoption of preservation metadata in cultural heritage institutions: case of PREMIS. **Proceedings of the American Society for Information Science and Technology**, v. 47, n. 1, p. 1-8, Nov./Dec. 2010. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/epdf/10.1002/meet.14504701187. Acesso em: 7 jun. 2023.

ALLEN INSTITUTE FOR AI. Semantic Scholar. About us. Resources. **Frequently Asked Questions**. [Seattle, Washington, United States], [2021a]. Acesso em: https://www.semanticscholar.org/faq#index-size. Acesso em: 7 jun. 2023.

ALLEN INSTITUTE FOR AI. Semantic Scholar. Research. Datasets. CORD-19: COVID-19 Open Research Dataset. **About CORD-19**. [Seattle, Washington, United States], [2021b]. Disponível em: https://www.semanticscholar.org/cord19/about. Acesso em: 7 jun. 2023.

ALLEN INSTITUTE FOR AI. **The COVID-19 Open Research Dataset (CORD-19)**. [*S. l.*]: GitHub, c2023. Disponível em: https://github.com/allenai/cord19. Acesso em: 5 ago. 2023.

ALLISON-BUNNELL, Jodi. Review of encoded archival description tag library: version EAD3. **Journal of Western Archives**, v. 7, n. 1, p. 1-4, 2016. Disponível em: https://digitalcommons.usu.edu/westernarchives/vol7/iss1/6/. Acesso em: 7 jun. 2023.

ALMEIDA, Maurício Barcellos; CENDÓN, Beatriz Valadares; SOUZA, Renato Rocha. Metodologia para implantação de programas de preservação de documentos digitais a longo prazo. **Encontros Bibli:** R. Eletr. Bibliotecon. Ci. Inf., Florianópolis, SC, v. 17, n. 34, p. 103-130, maio./ago. 2012. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2012v17n34p103/22622. Acesso em: 7 jun. 2023.

ALNOAMANY, Yasmin; WEIGLE, Michele C.; NELSON, Michael L. Detecting off-topic pages within timemaps in web archives. **Int J Digit Libr**, v. 17, n. 3, p. 203-221, September 2016. Disponível em: https://link.springer.com/article/10.1007/s00799-016-0183-5. Acesso em: 7 jun. 2023.

ALVES, Rachel Cristina Vesú. **Metadados como elementos do processo de catalogação**. 2010. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, SP, 2010. Disponível em: https://www.marilia.unesp.br/Home/Pos-Graduacao/CienciadaInformacao/Dissertacoes/alves_rachel.pdf. Acesso em: 7 jun. 2023.

ALVES, Rachel Cristina Vesú. Metadados para representação e recuperação da informação em ambiente Web. *In*: MARINGELLI, Isabel Cristina Ayres da Silva. (org.). **IV Seminário Serviços de Informação em Museus**: informação digital como patrimônio cultural. São Paulo: Pinacoteca de São Paulo, 2017. p. 95-106. Disponível em: http://www.getty.edu/publications/intrometadata/. Acesso em: 7 jun. 2023.

ALVES, Rachel Cristina Vesú; SANTOS, Plácida Leopoldina Ventura Amorim da Costa. **Metadados no domínio bibliográfico**. Rio de Janeiro: Intertexto, 2013. 196 p.

AMERICAN COUNCIL FOR TECHNOLOGY (ACT). ACT Collaboration & Transformation (C&T) Shared Interest Group (SIG). **Best practices study of social media records policies**. Washington, DC: ACT, Mar. 2011. 31 p. Disponível em: https://www.actiac.org/system/files/Best%20Practices%20of%20Social%20Media%20Records%20Policies%20-%20CT%20SIG%20-%2003-31-11%20(3).pdf. Acesso em: 27 dez. 2019.

ANDRADE, Ricardo; BORGES, Jussara; JAMBEIRO, Othon. Digitalizando a memória de Salvador: nossos presente e passado têm futuro. **Perspect. ciênc. inf.**, Belo Horizonte, v. 11, n. 2, p. 243-254, maio/ago. 2006. Disponível em: https://www.scielo.br/j/pci/a/zbhV63MJwf7S8M7hvWTs5kz/?format=pdf&lang=pt. Acesso em: 7 jun. 2023.

ANTOUN, Naira. The archive of the revolution: a living archive. **ArabLit, ArabLit Quarterly**: a magazine of arabic literature in translation, [*S. l.*], 30 Sept. 2012. Disponível em: https://arablit.org/2012/09/30/the-archive-of-the-revolution-a-living-archive/. Acesso em: 7 jun. 2023.

ARAUJO, Priscilla Mara Bermudes. **Preservação digital e os periódicos científicos eletrônicos brasileiros em Ciência da Informação**. 2015. Dissertação (Mestre em Ciência da Informação) – Escola de Comunicação, Universidade Federal do Rio de Janeiro, Instituto Brasileiro de Informação em Ciência e Tecnologia, Rio de Janeiro, RJ, 2015. Disponível em: https://ridi.ibict.br/bitstream/123456789/857/1/PriscillaDisserta%c3%a7%c3%a3oFinal01.pdf . Acesso em: 7 jun. 2023.

ARCHAEOLOGY DATA SERVICE. About. **Our Work**. York, United Kingdom, [2021]. Disponível em: https://archaeologydataservice.ac.uk/about/ourWork.xhtml. Acesso em: 7 jun. 2023.

ARCHAEOLOGY Data Service. York, United Kingdom, c2023. Disponível em: https://archaeologydataservice.ac.uk/. Acesso em: 7 jun. 2023.

ARCHIVE SOCIAL. **Social media records in Australia**. Durham, North Carolina, c2021. Disponível em: https://archivesocial.com/social-media-records/australia/. Acesso em: 7 jun. 2023.

ARCHIVE SOCIAL. **The Obama White House Social Media Archive**. Durham, North Carolina, c2023. Disponível em: https://archivesocial.com/whitehouse/. Acesso em: 7 jun. 2023.

ARCOMAN, John. SourceForge. Projects. **ARCOMEM**: semantic and social web crawling. Wiki page. [*S. l.*], 2016. Disponível em: https://sourceforge.net/projects/arcomem/. Acesso em: 7 jun. 2023.

ARQUIVO NACIONAL (Brasil). **Dicionário brasileiro de terminologia arquivística**. Rio de Janeiro: Arquivo Nacional, 2005. 232 p. (Publicações Técnicas, 51). Disponível em: http://www.arquivonacional.gov.br/images/pdf/Dicion_Term_Arquiv.pdf. Acesso em: 7 jun. 2023.

ARQUIVO NACIONAL (Brasil). **Política de preservação digital**. Versão 2. [Rio de Janeiro, RJ], dez. 2016. 33 p. Disponível em: http://www.arquivonacional.gov.br/images/conteudo/servicos_ao_governo/Programas_e_Projetos/AND_Politica_Preservacao_Digital_v2.pdf. Acesso em: 7 jun. 2023.

ARQUIVO.PT. [*S. l.*], 2023. Disponível em: https://arquivo.pt/. Acesso em: 7 jun. 2023.

ARQUIVO.PT. **Metadados acerca dos conteúdos**. [*S. l.*], ago. 2018. Disponível em: https://sobre.arquivo.pt/pt/recomendacoes/metadados-acerca-dos-conteudos/. Acesso em: 7 jun. 2023.

ARQUIVO.PT. Sobre. Sobre.arquivo.pt. Perguntas frequentes. **Recolha de conteúdos**. [*S. l.*], jul. 2021a. Disponível em: https://sobre.arquivo.pt/pt/ajuda/recolha-e-arquivo-de-conteudos/. Acesso em: 7 jun. 2023.

ARQUIVO.PT. Sobre.arquivo.pt. Acerca. **Termos e condições**. [*S. l.*], abr. 2022a. Disponível em: https://sobre.arquivo.pt/pt/acerca/termos-e-condicoes/. Acesso em: 7 jun. 2023.

ARQUIVO.PT. Sobre.arquivo.pt. Colabore. **Recomendações para a publicação na web de informação preservável**. [*S. l.*], jul. 2022b. Disponível em: https://sobre.arquivo.pt/pt/colabore/recomendacoes/. Acesso em: 7 dez. 2022.

ARQUIVO.PT. Sobre.arquivo.pt. Perguntas frequentes. **Informações gerais**. [*S. l.*], abr. 2021b. Disponível em: https://sobre.arquivo.pt/pt/ajuda/o-que-e-o-arquivo-pt/#qe-faq-2085. Acesso em: 7 jun. 2023.

ARTEFACTUAL SYSTEMS. **Archivematica**. [*S. l.*], [2023?]. Disponível em: https://www.archivematica.org/pt-br/. Acesso em: 7 jun. 2023.

AULER, Décio; BAZZO, Walter Antonio. Reflexões para a implementação do movimento CTS no contexto educacional brasileiro. **Ciência & Educação**, Bauru, SP, v. 7, n. 1, p. 1-13, 2001. Disponível em: https://www.scielo.br/j/ciedu/a/wJMcpHfLgzh53wZrByRpmkd/?format=pdf&lang=pt. Acesso em: 7 jun. 2023.

AUSTRALIAN War Memorial. [Canberra, Australia], c2023. Disponível em: https://www.awm.gov.au/. Acesso em: 11 ago. 2023.

AUSTRALIAN WAR MEMORIAL. Australian War Memorial Research Centre. **Australian War Memorial, PANDORA selection guidelines**. [*S. l.*], [2020?]. 4 p. Disponível em: https://pandora.nla.gov.au/guidelines/Pandora_Selection_Guidelines_AWM.doc. Acesso em: 7 jun. 2023.

AZEVEDO, Ana Cristina Vaz de *et al*. Divulgação científica em contexto de inteligência artificial através do Instagram. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, [*S. l.*], v. 9, n. 1, 2022. Disponível em: https://proceedings.sbmac.emnuvens.com.br/sbmac/article/view/4012/4063. Acesso em: 22 abr. 2023.

BACA, Murtha. (ed.). **Introduction to metadata**. 3rd ed. Los Angeles, California: Getty Publications, c2016. 92 p. Disponível em: http://www.getty.edu/publications/intrometadata/. Acesso em: 7 jun. 2023.

BACON, Francis. **Novum Organum ou verdadeiras indicações acerca da interpretação da natureza**: nova atlântida. Tradução José Aluysio Reis de Andrade. 2 ed. São Paulo: Abril Cultural, 1979. 39 p. (Os Pensadores).

BAGGIO, Claudia Carmem; FLORES, Daniel. Estratégias, critérios e políticas para preservação de documentos digitais em arquivos. **Ci. Inf.**, Brasília, DF, v. 41, n. 2/3, p. 58-71, maio/dez. 2012. Disponível em: http://revista.ibict.br/ciinf/article/view/1336/1515. Acesso em: 7 jun. 2023.

BAILEY, Jefferson; LACALLE, Maria. Don't warc away: preservation metadata and web archives. *In*: AMERICAN LIBRARY ASSOCIATION (ALA) ANNUAL CONFERENCE, 16., June 2015, San Francisco, California. **Proceedings** […]. San Francisco, California: ALA, 2015. p. 1-46. Disponível em: https://connect.ala.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=ad8ddd39-943d-4041-b627-1274e45cfba9. Acesso em: 7 jun. 2023.

BAKER, Fran; BUTLER, Phil.; GREEN, Ben. **Carcanet Press Email Preservation Project**. Manchester, UK: University of Manchester, May 2012. 21 p. Disponível em: https://www.research.manchester.ac.uk/portal/files/33713849/FULL_TEXT.PDF. Acesso em: 1 out. 2021.

BANOS, Vangelis *et al*. CLEAR: a credible method to evaluate website archivability. *In*: INTERNATIONAL CONFERENCE ON PRESERVATION OF DIGITAL OBJECTS (iPRES), 10, May c2013, Lisboa, Portugal. **Proceedings** […]. Lisboa, Portugal: iPRES, 2010. p. 9-18. Disponível em: http://purl.pt/24107/1/iPres2013_PDF/iPres2013-Proceedings.pdf. Acesso em: 7 jun. 2023.

BANOS, Vangelis. **ArchiveReady**: website archivability evaluation tool. [*S. l.*], [c2017]. Disponível em: http://archiveready.com/. Acesso em: 7 jun. 2023.

BARBEDO, Francisco; CORUJO, Luís; SANT'ANA, Mário. **Recomendações para a produção de planos de preservação digital**. Versão 2.1. Lisboa: Direção-Geral de Arquivos (DGARQ), nov. 2011. 111 p. Disponível em: https://arquivos.dglab.gov.pt/wp-content/uploads/sites/16/2014/02/Recomend_producao_PPD_V2.1.pdf. Acesso em: 7 jun. 2023.

BARDIN, Laurence. **Análise de conteúdo**. Lisboa: Edições 70, 1977.

BARDIN, Laurence. **Análise de conteúdo**. Lisboa: Edições 70, 2009.

BARDIN, Laurence. **Análise de conteúdo**. São Paulo: Edições 70, 2016.

BARRETO, Paloma da Silva *et al*. Zika e microcefalia no Facebook da Fiocruz: a busca pelo diálogo com a população e a ação contra os boatos sobre a epidemia. **Reciis** – Rev Eletron Comun Inf Inov Saúde, Rio de Janeiro, v. 14, n. 1, p. 18-33, 2020. Disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/1792/2332. Acesso em: 22 abr. 2023.

BATES, Jo. The politics of data friction. **Journal of Documentation**, [*S. l.*], v. 74, n. 2, August 2017. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/JD-05-2017-0080/full/html. Acesso em: 7 jun. 2023.

BAUCHSPIES, Wenda K.; CROISSANT, Jennifer; RESTIVO, Sal. **Science, technology, and society**: a sociological approach. Malden, Massachusetts: Blackwell Publishing, c2006.

BAUCOM, Erin. Planning and implementing a sustainable digital preservation program. **Library Technology Reports**, Chicago, v. 55, n. 6, Aug./Sept. 2019. Disponível em: https://www.alastore.ala.org/content/planning-and-implementing-sustainable-digital-preservation-program. Acesso em: 7 jun. 2023.

BEAGRIE, Neil. Preservation, Trust and continuing access for e-journals. **DPC Technology Watch Report 13-04**, [*S. l.*], p. 1-43, Sept. c2013. Disponível em: https://www.dpconline.org/docs/technology-watch-reports/924-dpctw13-04/file. Acesso em: 7 jun. 2023.

BENTANCOURTI, Silva Silvia Maria Puentes; ROCHA, Rafael Port da. Metadados de qualidade e visibilidade na comunicação científica. **Enc. Bibli**: R. Eletr. Bib. Ci. Inf., Florianópolis, v. 17, n. esp. 2 – III SBCC, p.82-101, 2012. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2012v17nesp2p82/23571. Acesso em: 7 jun. 2023.

BENTLEY HISTORICAL LIBRARY. **Bentley historical library web archives**: collection development policy. Version 5.0. Ann Arbor, United States: University of Michigan, February 2016. 11 p. Disponível em: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/94163/BHL_webArchivesCollectingPolicy_20160203.pdf?sequence=10&isAllowed=y. Acesso em: 7 jun. 2023.

BENTLEY HISTORICAL LIBRARY. **University of Michigan Bentley Historical Library**. [San Francisco, California]: Archive-It, [2023?]. Disponível em: https://archive-it.org/organizations/934. Acesso em: 7 jun. 2023.

BERNARDO, Wanderley Marques; NOBRE, Moacyr Roberto Cuce; JATENE, Fábio Biscegli. A prática clínica baseada em evidências. Parte II – buscando as evidências em fontes de informação. **Rev Assoc Med Bras**, São Paulo, SP, v. 50, n. 1, p. 104-108, 2004. Disponível em: https://www.scielo.br/j/ramb/a/WgCzqZ5n8ZyjpNCd7nxF5VQ/?lang=pt. Acesso em: 9 maio 2023.

BIBLARZ, Dora *et al*. **Guidelines for a collection development policy using the conspectus model**. [*S. l.*]: International Federation of Library Associations and Institutions, 2001. 11 p. Disponível em: https://www.ifla.org/files/assets/acquisition-collection-development/publications/gcdp-en.pdf. Acesso em: 7 jun. 2023.

BIBLIOTECA DE CATALUNYA. **Patrimoni Digital de Catalunya (PADICAT)**: L'Arxiu Web de Catalunya. Coneix-nos. PMF. [Barcelona, España], c2011. Disponível em: https://www.padicat.cat/ca. Acesso em: 7 jun. 2023.

BIBLIOTECA NACIONAL DE CHILE. Biblioteca Nacional Digital de Chile. **Archivo de la Web chilena**. Preguntas frecuentes. [Santiago, Chile], [2022]. Disponível em: http://archivoweb.bibliotecanacionaldigital.cl/. Acesso em: 7 jun. 2023.

BIBLIOTECA NACIONAL DE CHILE. Biblioteca Nacional Digital de Chile. **Archivo de la Web chilena**. [Santiago, Chile], [2023?]. Disponível em: http://archivoweb.bibliotecanacionaldigital.cl/. Acesso em: 7 jun. 2023.

BIBLIOTHÈQUE NATIONALE DE FRANCE. Accueil. **Archives de l'internet**. Paris, France, c2022a. Disponível em: https://www.bnf.fr/fr/archives-de-linternet. Acesso em: 7 jun. 2023.

BIBLIOTHÈQUE NATIONALE DE FRANCE. **Archives de l'internet**. Paris, France, c2023. Disponível em: https://www.bnf.fr/fr/archives-de-linternet. Acesso em: 7 jun. 2023.

BIBLIOTHÈQUE NATIONALE DE FRANCE. Accueil. Collaborer. Déposer. **Qu'est-ce que le dépôt légal?** Paris, France, c2022b. Disponível em: https://www.bnf.fr/fr/quest-ce-que-le-depot-legal. Acesso em: 7 jun. 2023.

BIBLIOTHÈQUE NATIONALE DE FRANCE. Accueil. **La BnF archive le web du coronavirus**. Paris, France, c2022c. Disponível em: https://www.bnf.fr/fr/la-bnf-archive-le-web-du-coronavirus. Acesso em: 7 jun. 2023.

BIBLIOTHÈQUE NATIONALE DU LUXEMBOURG. **L'archivage du web**: possibilités et limites. [Kirchberg, Luxembourg], [2021]. 2 p. Disponível em: https://www.webarchive.lu/wp-content/uploads/2021/03/Luxembourg-Web-Archive-Possibilités-et-Limites.pdf. Acesso em: 7 jun. 2023.

BIBLIOTHÈQUE NATIONALE DU LUXEMBOURG. **Luxembourg web archive.** [Kirchberg, Luxembourg], [c2023?]. Disponível em: https://www.webarchive.lu/. Acesso em: 7 jun. 2023.

BIBLIOTHÈQUE NATIONALE DU LUXEMBOURG. **Luxembourg web archive.** What we do. How it works. Faq. [Kirchberg, Luxembourg], [c2022]. 2 p. Disponível em: https://www.webarchive.lu/. Acesso em: 7 jun. 2023.

BIG DATA. *In*: WIKIPEDIA: the free encyclopedia. [San Francisco, CA: Wikimedia Foundation], 2022. Disponível em: https://en.wikipedia.org/wiki/Big_data. Acesso em: 7 jun. 2023.

BOERES, Sonia Araújo de Assis. **Competências necessárias para equipes de profissionais de preservação digital**. 2017. Tese (Doutorado em Ciência da Informação) - Faculdade de Ciência da Informação, Universidade de Brasília, Brasília, 2017. Disponível em: https://repositorio.unb.br/bitstream/10482/24354/1/2017_SoniaAraujodeAssisBoeres.pdf. Acesso em: 15 dez. 2021.

BOERES, Sonia Araujo de Assis; FARIA, Ana Carolina Cintra. A preservação digital na biblioteca central da Universidade de Brasília. **Ci. Inf.**, Brasília, DF, v. 41, n. 1, p.175-183, jan./abr. 2012. Disponível em: http://revista.ibict.br/ciinf/article/view/1363/1542. Acesso em: 7 jun. 2023.

BOERES, Sonia Araújo de Assis; MÁRDERO ARELLANO, Miguel Ángel. Políticas e estratégias de preservação de documentos digitais. *In*: ENCONTRO NACIONAL DE CIÊNCIA DA INFORMAÇÃO – CINFORM, 6., 2005, Salvador, BA. **Anais...** Salvador, BA: UFBA, 2005. p. 1-15. Disponível em: http://www.cinform-anteriores.ufba.br/vi_anais/docs/SoniaMiguelPreservacaoDigital.pdf. Acesso em: 7 jun. 2023.

BONDI, André B. Characteristics of scalability and their impact on performance. *In*: INTERNATIONAL WORKSHOP ON SOFTWARE AND PERFORMANCE (WOSP), 2., Sept. 2000, Ottawa, Canada. **Proceedings** […]. Ottawa, Canada: WOSP, 2000. p. 195-203. Disponível em: https://www.researchgate.net/profile/Andre_Bondi/publication/221556521_Characteristics_of_Scalability_and_Their_Impact_on_Performance/links/5523fbeb0cf22e181e730469/Characteristics-of-Scalability-and-Their-Impact-on-Performance.pdf. Acesso em: 7 jun. 2023.

BORGMAN, Christine L. Bibliometrics and scholarly communication: editor's introduction. **Communication Research**, [*S. l.*], v. 16, n. 5, p. 583-599, 1989. Disponível em: https://journals.sagepub.com/doi/10.1177/009365089016005002. Acesso em: 22 abr. 2023.

BORKO, Harold. Information science: what is it? **American Documentation**, [*S. l.*], v. 19, n. 1, p. 3-5, 1968. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/asi.5090190103. Acesso em: 9 maio 2023.

BOTÉ, Juanjo; FERNANDEZ-FEIJOO, Belen; RUIZ, Silvia. Digital preservation cost: a cost accounting approach. **The Learning Organization**, [*S. l.*], v. 20, n. 6, p. 419-432, Sept. 2013. Disponível em: https://www.researchgate.net/publication/263478773_Digital_preservation_cost_A_cost_accounting_approach. Acesso em: 7 jun. 2023.

BOUDREZ, Filip. **Filing and archiving e-mail**. Antwerp, Belgium: Expertisecentrum DAVID vzw, 2006. 47 p. Disponível em: http://www.expertisecentrumdavid.be/docs/filingArchiving_email.pdf. Acesso em: 7 jun. 2023.

BRAGG, Molly *et al*. **The Web archiving life cycle model.** WhitePaper. March 2013. Disponível em: https://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf. Acesso em: 7 jun. 2023.

BREXIT. *In*: WIKIPEDIA: the free encyclopedia. [San Francisco, CA: Wikimedia Foundation], 2022. Disponível em: https://en.wikipedia.org/wiki/Brexit. Acesso em: 7 jun. 2023.

BRITISH LIBRARY. Information for publishers. **About legal deposit**. [London, United Kingdom], [c2022]. Disponível em: https://www.bl.uk/legal-deposit/about-legal-deposit. Acesso em: 7 jun. 2023.

BRITISH LIBRARY. Press Office. Press releases. **Some sort of record seemed vital.' British Library acquires the archive of Wendy Cope**. [London, United Kingdom], Apr. 2011. Disponível em: https://www.bl.uk/press-releases/2011/april/some-sort-of-record-seemed-vital-british-library-acquires-the-archive-of-wendy-cope. Acesso em: 18 de out. 2021.

BROOKER, Phillip; BARNETT, Julie; CRIBBIN, Timothy. Doing social media analytics. **Big Data & Society**, [*S. l.*], v. 3, n. 2, p. 1-12, July/Dec. 2016. Disponível em: https://journals.sagepub.com/doi/pdf/10.1177/2053951716658060. Acesso em: 7 jun. 2023.

BROUSSARD, Meredith; BOSS, Katherine. Saving data journalism: new strategies for archiving interactive, born-digital news. **Digital Journalism**, [*S. l.*], v. 6, n. 9, p. 1206-1221, 2018. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/21670811.2018.1505437. Acesso em: 7 jun. 2023.

BROWN, Adrian. **Archiving websites**: a practical guide for information management professionals. London: Facet Publishing, c2006. 227 p. Disponível em: https://books.google.com.br/books/about/Archiving_Websites.html?id=DeTgAAAAMAAJ&redir_esc=y. Acesso em: 7 jun. 2023.

BROWN, Adrian. Selecting storage media for long-term preservation. **The National Archives Digital Preservation Guidance Note 2,** [London], Issue 2, p. 1-7, Aug. 2008. Disponível em: https://cdn.nationalarchives.gov.uk/documents/selecting-storage-media.pdf. Acesso em: 7 jun. 2023.

BRÜGGER, Niels. **Archiving websites**: general considerations and strategies. Århus, Denmark: The Centre for Internet Research, January 2005. 64 p. Disponível em: https://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/cfis_b__ger/nb_archiving.pdf. Acesso em: 7 jun. 2023.

BRÜGGER, Niels. **The archived Web**: doing history in the digital age. Cambridge, Massachusetts: MIT Press, c2018. 185 p. Disponível em: https://direct.mit.edu/books/book/4215/The-Archived-WebDoing-History-in-the-Digital-Age. Acesso em: 7 jun. 2023.

BRÜGGER, Niels. Web archiving: between past, present, and future. *In*: CONSALVO, Mia; ESS, Charles (Ed.). **The Handbook of Internet Studies**. [Hoboken, Nova Jersey]: Blackwell Publishing, c2011. p. 24-42. Disponível em: https://onlinelibrary.wiley.com/doi/epdf/10.1002/9781444314861.ch2. Acesso em: 7 jun. 2023.

BRÜGGER, Niels. Web historiography and internet studies: challenges and perspectives. **New Media & Society**, [*S. l.*], v. 15, n. 5, p. 752-764, 2012. Disponível em: https://journals.sagepub.com/doi/abs/10.1177/1461444812462852. Acesso em: 7 jun. 2023.

BRÜGGER, Niels; FINNEMANN, Niels Ole. The web and digital humanities: theoretical and methodological concerns. **Journal of Broadcasting & Electronic Media**, [*S. l.*], v. 57, n. 1, p. 66–80, March 2013. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/08838151.2012.761699. Acesso em: 7 jun. 2023.

BUENO, Wilson Costa. Comunicação científica e divulgação científica: aproximações e rupturas conceituais. **Inf. Inf.**, Londrina, v. 15, n. esp., p. 1-12, 2010. Disponível em: https://aprender.ead.unb.br/pluginfile.php/191603/mod_resource/content/1/COMUNICA%C3%87%C3%83O%20CIENT%C3%8DFICA%20E%20DIVULGA%C3%87%C3%83O.pdf. Acesso em: 7 jun. 2021.

BUENO, Wilson Costa. **Jornalismo científico no Brasil**: os compromissos de uma prática dependente. 1985. Tese (Doutorado em Ciências da Comunicação) – Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, SP, 1985.

BULLOCK, Alison. Preservation of digital information: issues and current status. **Network Notes**, National Library of Canada, Ottawa, n. 60, Apr. 1999. Disponível em: http://epe.lac-bac.gc.ca/100/202/301/netnotes/netnotes-h/notes60.htm. Acesso em: 15 dez. 2021.

BURNAP, Peter *et al*. COSMOS: towards an integrated and scalable service for analysing social media on demand. **International Journal of Parallel, Emergent and Distributed Systems**, [*S. l.*], v. 30, n. 2, p. 80-100, 2015. Disponível em: https://www.tandfonline.com/doi/full/10.1080/17445760.2014.902057. Acesso em: 7 jun. 2023.

BYRNE, Helena. **Web archiving Rio 2016**: the story so far. International Internet Preservation Consortium (IIPC), [*S. l.*], Sept. 2016. Disponível em: https://netpreserveblog.wordpress.com/2016/09/02/web-archiving-rio-2016-the-story-so-far/. Acesso em: 7 jun. 2023.

CADAVID, Jhonny Antonio Pabo´n. Evolution of legal deposit in new zealand: from print to digital heritage. **International Federation of Library Associations and Institutions**, v. 43, n. 4, p. 379-390, 2017. Disponível em: https://journals.sagepub.com/doi/10.1177/0340035217713763. Acesso em: 7 jun. 2023.

CALIFORNIA DIGITAL LIBRARY. Services and Projects. Publishing, Archives, and Digitization. **Web Archiving Activities**. Oakland, California, May 2021. Disponível em: https://cdlib.org/services/pad/webarchiving/. Acesso em: 7 jun. 2023.

CAMPOS, Fernanda Maria. Informação digital: um novo património a preservar. **Cadernos BAD**, Lisboa, n. 2, p. 8-14, 2002. Disponível em: https://publicacoes.bad.pt/revistas/index.php/cadernos/article/view/861/860. Acesso em: 7 jun. 2023.

CANTARA, Linda. METS: the metadata encoding and transmission standard. **Cataloging & Classification Quarterly**, Philadelphia, v. 40, n. 3/4, p. 237-253, 2005. Disponível em: https://www.tandfonline.com/doi/abs/10.1300/J104v40n03_11. Acesso em: 7 jun. 2023.

CAPLAN, Priscilla. **Understanding PREMIS**. [Washington, DC]: Library of Congress Network Development and MARC Standards Office, 2017. 22 p. Disponível em: https://www.loc.gov/standards/premis/understanding-premis-rev2017.pdf. Acesso em: 7 jun. 2023.

CARDIFF UNIVERSITY. Social data science lab. **COSMOS**. Blog. Cosmos tutorials. Cardiff, United Kingdom, [2022]. Diponível em: http://socialdatalab.net/cosmos. Acesso em: 7 jun. 2023.

CARDIFF UNIVERSITY. Social data science lab. **COSMOS**. Cardiff, United Kingdom, [2023?]. Diponível em: http://socialdatalab.net/cosmos. Acesso em: 7 jun. 2023.

CARIBÉ, Rita de Cássia do Vale. **Comunicação científica para o público leigo no Brasil**. 2011. Tese (Doutorado em Ciência da Informação) – Faculdade de Ciência da Informação, Universidade de Brasília, Brasília, 2011. Disponível em: https://repositorio.unb.br/bitstream/10482/9003/1/2011_RitadeC%c3%a1ssiadoValeCarib%c3%a9.pdf. Acesso em: 18 jun. 2021.

CARIBÉ, Rita de Cássia do Vale. Comunicação científica: reflexões sobre o conceito. **Inf. & Soc.**:Est., João Pessoa, v. 25, n. 3, p. 89-104, set./dez. 2015. Disponível em: https://www.researchgate.net/profile/Rita-Caribe/publication/292198040_Scientific_communication_Reflections_on_the_concept/links/583821ec08ae3a74b49cccd4/Scientific-communication-Reflections-on-the-concept.pdf. Acesso em: 7 jun. 2023.

CARVALHO, Lidiane dos Santos; LIMA, Clóvis Ricardo Montenegro de; MACÊDO, Wânia Cristina Morais de. A comunicação científica em tempos de pandemia do Covid-19: preprints, informação válida e ciência rápida. **ASKLEPION**: Informação em Saúde, Rio de Janeiro, v. 2, edição especial, p. 141-161, 2022. Disponível em: https://asklepionrevista.info/asklepion/article/view/59/126. Acesso em: 22 abr. 2023.

CASTELFRANCHI, Yurij *et al*. As opiniões dos brasileiros sobre ciência e tecnologia: o 'paradoxo' da relação entre informação e atitudes. **História, Ciências, Saúde – Manguinhos**, Rio de Janeiro, v. 20, supl., p. 1163-1183, nov. 2013. Disponível em: https://www.scielo.br/j/hcsm/a/7JGKDbkgfn5XBLTg8TzRC9S/?lang=pt&format=pdf. Acesso em: 7 jun. 2023.

CASTRO, Fabiano Ferreira de Castro. **Elementos de interoperabilidade na catalogação descritiva**: configurações contemporâneas para a modelagem de ambientes informacionais digitais. 2012. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, SP, 2012. Disponível em: https://www.marilia.unesp.br/Home/Pos-Graduacao/CienciadaInformacao/Dissertacoes/Castro,%20F.F._doutorado_CI_2012.pdf. Acesso em: 7 jun. 2023.

CASTRO, Fabiano Ferreira de. COVID-19: um olhar pelas lentes dos metadados. **folha de rosto**: Revista de Biblioteconomia e Ciência da Informação, v. 6, n. 2, p. 58-69, maio/ago. 2020. Disponível em: https://brapci.inf.br/index.php/res/download/149991. Acesso em: 7 jun. 2023.

CASTRO, Fabiano Ferreira de; SILVEIRA, Júlio César Tauil. Mapeamento de padrões de metadados de preservação digital em cloud services. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB, 19., 2018, Londrina, PR. **Anais...** Londrina, PR: UEL, 2018. p. 5183-5204. Disponível em: http://enancib.marilia.unesp.br/index.php/XIXENANCIB/xixenancib/paper/viewFile/1003/1697. Acesso em: 7 jun. 2023.

CAVALCANTE, Ricardo Bezerra; CALIXTO, Pedro; PINHEIRO, Marta Macedo Kerr. Análise de conteúdo: considerações gerais, relações com a pergunta de pesquisa, possibilidades e limitações do método. **Inf. & Soc.**: Est., João Pessoa, v. 24, n. 1, p. 13-18, jan./abr. 2014. Disponível em: https://periodicos.ufpb.br/ojs/index.php/ies/article/view/10000/10871. Acesso em: 7 maio 2023.

CENTER FOR RESEARCH LIBRARIES. Archiving & preservation. **Digital preservation**. Certification and assessment of digital repositories. Digital preservation metrics. TRAC metrics. Chicago, United States, [2022?]. Disponível em: https://www.crl.edu/archiving-preservation/digital-archives. Acesso em: 7 jun. 2023.

CENTRO DE GESTÃO E ESTUDOS ESTRATÉGICOS. **Percepção pública da C&T no Brasil, 2019**: resumo executivo. Brasília, DF: 2019. 23p. Disponível em: https://www.cgee.org.br/documents/10195/734063/CGEE_resumoexecutivo_Percepcao_pub_CT.pdf. Acesso em: 26 jan. 2022.

CHAN, Lois Mai; ZENG, Marcia Lei. Metadata interoperability and standardization: a study of methodology part i: achieving interoperability at the schema level. **D-Lib Magazine**, [*S. l.*], v. 12, n. 6, June c2006. Disponível em: http://www.dlib.org/dlib/june06/chan/06chan.html. Acesso em: 7 jun. 2023.

CHEN, Mingyu; REILLY, Michele. Implementing METS, MIX, and DC for sustaining digital preservation at the University of Houston Libraries. **Journal of Library Metadata**, [*S. l.*], v. 11, n. 2, p. 83-99, May 2011. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/19386389.2011.570662. Acesso em: 7 jun. 2023.

CLOCKSS Archive. [Stanford, California], c2023. Disponível em: https://clockss.org/. Acesso em: 7 jun. 2023.

CLOCKSS ARCHIVE. About. **Why CLOCKSS?** Frequently asked questions (faqs). Triggered content. [Stanford, California], c2022. Disponível em: https://clockss.org/about/. Acesso em: 7 jun. 2023.

COCCIOLO, Anthony. Community archives in the digital era: a case from the lgbt community. **PDT&C,** v. 45, n. 4, p. 157–165, 2016. Disponível em: https://www.degruyter.com/document/doi/10.1515/pdtc-2016-0018/html. Acesso em: 7 jun. 2023.

COLUMBIA UNIVERSITY LIBRARIES. About our collections. **Web archives at Columbia**. Web resources collection program. Guidelines for preservable websites. Website owner faq. [New York, United States], c2021a. Disponível em: https://library.columbia.edu/collections/web-archives.html. Acesso em: 7 jun. 2023.

COLUMBIA UNIVERSITY LIBRARIES. **Web archives at Columbia**. [New York, United States], [c2023?]. Disponível em: https://library.columbia.edu/collections/web-archives.html. Acesso em: 7 jun. 2023.

COLUMBIA UNIVERSITY LIBRARIES. Columbia university archives. **Web archives**. New York, Sept. 2021b. Disponível em: https://library.columbia.edu/libraries/cuarchives/resources/webarchives.html. Acesso em: 7 jun. 2023.

CONSELHO NACIONAL DE ARQUIVOS (Brasil). Câmara Técnica de Documentos Eletrônicos – CTDE. **Glossário**: Documentos Arquivísticos Digitais. Versão 8. Rio de Janeiro, RJ, 2020. 50 p. Disponível em: https://www.gov.br/conarq/pt-br/assuntos/camaras-tecnicas-setoriais-inativas/camara-tecnica-de-documentos-eletronicos-ctde/glosctde_2020_08_07.pdf. Acesso em: 7 jun. 2023.

CONSELHO NACIONAL DE ARQUIVOS (Brasil). Câmara Técnica de Documentos Eletrônicos – CTDE. **Diretrizes para a gestão arquivística do correio eletrônico corporativo**. Rio de Janeiro, RJ, 2012. 35 p. Disponível em: https://www.gov.br/conarq/pt-br/centrais-de-conteudo/publicacoes/Correio_eletronico_completo_2.pdf. Acesso em: 7 jun. 2023.

COOLUTILS. Products. **Total Outlook Converter Pro**. [*S. l.*], c2022. Disponível em: https://www.coolutils.com/TotalOutlookConverter. Acesso em: 7 jun. 2023.

CORDEIRO, Alexander Magno *et al*. Revisão sistemática: uma revisão narrativa. **Comunicação Científica**, v. 34, n. 6, p. 428-431, nov./dez. 2007. Disponível em: https://www.scielo.br/j/rcbc/a/CC6NRNtP3dKLgLPwcgmV6Gf/?format=pdf&lang=pt. Acesso em: 7 jun. 2023.

CORETRUSTSEAL. **Certification**. Why certification. Hague, Netherlands, c2022. Disponível em: https://www.coretrustseal.org/why-certification/. Acesso em: 7 jun. 2023.

COSTA, Miguel; GOMES, Daniel; SILVA, Mário J. The evolution of web archiving. **International Journal on Digital Libraries**, v. 18, n. 3, p. 191-205, Sept. 2017. Disponível em: https://link.springer.com/article/10.1007/s00799-016-0171-9. Acesso em: 7 jun. 2023.

COUNCIL ON LIBRARY AND INFORMATION RESOURCES (CLIR). **The Future of Email Archives**: A Report from the Task Force on Technical Approaches for Email Archives. Washington, DC, Aug. 2018. 120 p. Disponível em: https://www.clir.org/wp-content/uploads/sites/6/2018/08/CLIR-pub175.pdf. Acesso em: 7 jun. 2023.

CROFT, Nahali. We've Got Mail: Email Preservation at a Small, Private University. **Journal for the Society of North Carolina Archivists**, v. 13, p. 65-90, 2016. Disponível em: https://kb.gcsu.edu/lib/2/. Acesso em: 7 jun. 2023.

DANTAS, Luiz Felipe Santoro; DECCACHE-MAIA, Eline. O retorno da era do áudio: analisando os podcasts de divulgação científica. **Revista de Ensino de Ciências e Matemática**, [*S. l.*], v. 13, n. 4, p. 1-25, 2022. Disponível em: https://revistapos.cruzeirodosul.edu.br/index.php/rencima/article/view/3730/1810. Acesso em: 22 abr. 2023.

DAPPERT, Angela *et al*. Describing and preserving digital object environments. **New Review of Information Networking**, Philadelphia, v. 18, n. 2, p. 106-173, Oct. 2013. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/13614576.2013.842494. Acesso em: 7 jun. 2023.

DAPPERT, Angela; ENDERS, Markus. Digital preservation metadata standards. **Information Standards Quarterly (ISQ)**, v. 22, n. 2, p. 4-13, spring 2010. Disponível em: http://www.loc.gov/standards/premis/FE_Dappert_Enders_MetadataStds_isqv22no2.pdf. Acesso em: 7 jun. 2023.

DAVIES, Sarah R. STS and science communication: reflecting on a relationship. **Public Understanding of Science**, [*S. l.*], v. 31, n. 3, p. 305-313, 2022. Disponível em: https://journals.sagepub.com/doi/epub/10.1177/09636625221075953. Acesso em: 22 abr. 2023.

DAY, Michael. **Collecting and preserving the world wide web**: a feasibility study undertaken for the jisc and wellcome trust. Version 1. February 2003a. Disponível em: http://www.ukoln.ac.uk/preservation/web-archiving/index.html. Acesso em: 7 jun. 2023.

DAY, Michael. **Preserving the fabric of our lives:** a survey of *web* preservation initiatives. *In*: INTERNATIONAL CONFERENCE ON THEORY AND PRACTICE OF DIGITAL LIBRARIES: RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES (ECDL), 7., 2003, Trondheim, Norway. **Proceedings** […]. Trondheim, Norway: ECDL, 2003b. p. 461-472. Disponível em: https://link.springer.com/chapter/10.1007/978-3-540-45175-4_42. Acesso em: 7 jun. 2023.

DAY, Michael. The long-term preservation of web content. *In*: MASANÈS, Julien. (Ed.). **Web archiving**. Berlin: Springer, c2006. p. 177-199. Disponível em: https://link.springer.com/book/10.1007/978-3-540-46332-0. Acesso em: 7 jun. 2023.

DET KONGELIGE BIBLIOTEK. Find materiale. Samlinger. **Netarkivet**. [København, Danmarks], [2023?]. Disponível em: https://www.kb.dk/find-materiale/samlinger/netarkivet. Acesso em: 5 fev. 2023.

DET KONGELIGE BIBLIOTEK. Find materiale. Samlinger. Netarkivet. **Begivenhedsindsamlinger.** [København, Danmarks], [2022a]. Disponível em: https://www.kb.dk/en/find-materials/collections/netarkivet/event-collections. Acesso em: 5 fev. 2023.

DET KONGELIGE BIBLIOTEK. Find materiale. Samlinger. **Netarkivet**. Forskningsadgang. Samarbejde om webarkiver. Om indsamling af internetmateriale. [København, Danmarks], [2022b]. Disponível em: https://www.kb.dk/en/find-materials/collections/netarkivet. Acesso em: 5 fev. 2023.

DET KONGELIGE BIBLIOTEK. Om os. **Digital pligtaflevering**. [København, Danmarks], [2022c]**.** Disponível em: https://www.kb.dk/en/about-us/digital-legal-deposit. Acesso em: 5 fev. 2023.

DI PRETORO, Emmanuel; GEERAERT; Friedel. Behind the scenes of web archiving: metadata of harvested websites. Archives et Bibliothèques de Belgique – Archief – En Bibliotheekwezen in Belgie; Archief, in press, trust an Undertanding: The value of metadata en a digitally joined-up world. 2019. Disponível em: https://hal.archives-ouvertes.fr/hal-02124714/document. Acesso em: 7 jun. 2023.

DIGITAL LIBRARY FEDERATION. **<METS> Metadata Encoding and Transmission Standard**: primer and reference manual. Version 1.6. [Washington, DC], 2010. 144 p. Disponível em: https://www.loc.gov/standards/mets/METSPrimer.pdf. Acesso em: 7 jun. 2023.

DIGITAL PRESERVATION COALITION. **Digital preservation handbook**. 2th ed. [Glasgow], c2015. Disponível em: https://www.dpconline.org/handbook. Acesso em: 7 jun. 2023.

DIGITAL PRESERVATION COALITION. **Metadata**. [Glasgow, Scotland], [2018a]. 2 p. (Digital Preservation Topical Note, 5). Disponível em: https://www.dpconline.org/docs/knowledge-base/1866-dp-note-5-metadata/file. Acesso em: 7 jun. 2023.

DIGITAL PRESERVATION COALITION. **Personal digital archiving**. [Glasgow, Scotland], [2018b]. 2 p. (Digital Preservation Topical Note, 6). Disponível em: https://www.dpconline.org/docs/knowledge-base/1867-dp-note-6-personal-digital-archiving/file. Acesso em: 7 jun. 2023.

DIGITAL PRESERVATION COALITION. **Preserving email**. [Glasgow, Scotland], [2018c]. 2 p. (Digital Preservation Topical Note, 7). Disponível em: https://www.dpconline.org/docs/knowledge-base/1868-dp-note-7-preserving-email/file. Acesso em: 7 jun. 2023.

DIGITAL PRESERVATION COALITION. **Preserving social media**. [Glasgow, Scotland], [2018d]. 2 p. (Digital Preservation Topical Note, 8). Disponível em: https://www.dpconline.org/docs/knowledge-base/1869-dp-note-8-preserving-social-media/file. Acesso em: 7 jun. 2023.

DIGITAL PRESERVATION COALITION. **Preserving the authoritative record**. [Glasgow, Scotland], [2018e]. 2 p. (Digital Preservation Topical Note, 9). Disponível em: https://www.dpconline.org/docs/knowledge-base/1860-dp-note-9-preserving-the-authoritative-record/file. Acesso em: 7 jun. 2023.

DIGITAL PRESERVATION COALITION. **Preserving the web**. [Glasgow, Scotland], [2018f]. 2 p. (Digital Preservation Topical Note, 10). Disponível em: https://www.dpconline.org/docs/knowledge-base/1861-dp-note-10-preserving-the-web/file. Acesso em: 7 jun. 2023.

DIGITAL PRESERVATION COALITION; ARTEFACTUAL SYSTEMS. **Preserving email**. [Glasgow, Scotland], July. c2021. 11 p. (DPC Technology Watch Guidance Note) Disponível em: https://www.dpconline.org/docs/technology-watch-reports/2472-preserving-email/file. Acesso em: 7 jun. 2023.

DIGITAL PRESERVATION. *In*: WIKIPEDIA: the free encyclopedia. [San Francisco, CA: Wikimedia Foundation], 2022. Disponível em: https://en.wikipedia.org/wiki/Digital_preservation. Acesso em: 7 jun. 2023.

DIGITALE archivering in/voor Vlaamse instellingen en diensten (DAVID). [*S. l.*], 2005. Disponível em: http://www.edavid.be/davidproject/eng/index.htm. Acesso em: 15 ago. 2023.

DOCUMENTING The Now. [*S. l.*], [2023?b]. Disponível em: https://www.docnow.io/. Acesso em: 7 jun. 2023.

DOCUMENTING The Now. **About**. Tools. [*S. l.*], [2021]. Disponível em: https://www.docnow.io/. Acesso em: 7 jun. 2023.

DOCUMENTING The Now. **DocNow**. [*S. l.*], [2023?a]. Disponível em: https://www.docnow.io/docnow-app/. Acesso em: 7 jun. 2023.

DOLLAR, Charles M.; ASHLEY, Lori J. Long-term digital preservation. *In*: SMALLWOOD, Robert F. (Ed.) **Information governance**: concepts, strategies and best practices. Hoboken, New Jersey: John Wiley & Sons, Inc., 2012. p. 315-347. Disponível em: https://onlinelibrary.wiley.com/doi/book/10.1002/9781118433829. Acesso em: 16 ago. 2023.

DOOLEY, Jackie M. *et al*. Developing web archiving metadata best practices to meet user needs. **Journal of Western Archives**, [Provo], v. 8, n. 2, p. 1-14, 2017. Disponível em: https://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1079&context=westernarchives. Acesso em: 7 jun. 2023.

DOOLEY, Jackie; BOWERS, Kate. **Descriptive metadata for web archiving**: recommendations of the oclc research library partnership web archiving metadata working group. Dublin, Ohio: Online Computer Library Center (OCLC) Research, Feb. c2018. 53 p. Disponível em: https://www.oclc.org/content/dam/research/publications/2018/oclcresearch-wam-recommendations.pdf. Acesso em: 7 jun. 2023.

DOORN, Peter; TJALSMA, Heiko. Introduction: archiving research data. **Arch Sci**, v. 7, p. 1-20, Sept. 2007. Disponível em: https://link.springer.com/content/pdf/10.1007/s10502-007-9054-6.pdf. Acesso em: 7 jun. 2023.

DOUGHERTY, Meghan *et al*. **Researcher engagement with web archives:** state of the art. Joint Information Systems Committee Report. [*S. l.*], August 2010. 38 p. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1714997. Acesso em: 7 jun. 2023.

DUBLIN CORE METADATA INITIATIVE. About DCMI. **DCMI History**. [*S. l.*], June c2020a. Disponível em: https://dublincore.org/about/history/. Acesso em: 7 jun. 2023.

DUBLIN CORE METADATA INITIATIVE. DCMI Usage Board. Specifications. **DCMI Metadata Terms**. [*S. l.*], Jan. 2020b. Disponível em: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/. Acesso em: 7 jun. 2023.

DUBLIN CORE METADATA INITIATIVE. DCMI Usage Board. Specifications. **Dublin Core Metadata Element Set, Version 1.1:** reference description. [*S. l.*], June 2012. Disponível em: http://www.dublincore.org/documents/dces. Acesso em: 7 jun. 2023.

DUBLIN CORE METADATA INITIATIVE. Dublin Core Usage Committee. **DCMI Qualifiers**. [*S. l.*], July 2000. Disponível em: https://dublincore.org/specifications/dublin-core/dcmes-qualifiers/. Acesso em: 7 jun. 2023.

DURANTI, Luciana. The long-term preservation of thedigital heritage: a case study of universities institutional repositories. **JLIS.it.**, Macerata, v. 1, n. 1, p. 157-168, Giugno/June 2010. Disponível em: https://www.jlis.it/article/view/12/21. Acesso em: 14 jan. 2022.

EDWARDS, Paul N. **A vast machine**: computer models, climate data, and the politics of global warming. Cambridge, Massachusetts: MIT Press, March 2010. 518 p.

EDWARDS, Paul N. et al. Science friction: Data, metadata, and collaboration. **Social Studies of Science**, Thousand Oaks, Califórnia, v. 41, n. 5, p. 667-690, August c2011. Disponível em: https://journals.sagepub.com/doi/abs/10.1177/0306312711413314. Acesso em: 7 jun. 2023.

EESTI RAHVUSRAAMATUKOGU. **Eesti veebiarhiivis**. [Tallinn, Eesti], [2023?]. Disponível em: https://veebiarhiiv.digar.ee/. Acesso em: 11 ago. 2023.

EESTI RAHVUSRAAMATUKOGU. **Veebiarhiiv**. Tallinn, Eesti, 2022. Disponível em: https://www.nlib.ee/et/veebiarhiiv. Acesso em: 7 jun. 2023.

EIDSON, Jennifer G.; ZAMON, Christina J. EAD twenty years later: a retrospective of adoption in the early twenty-first century and the future of ead. **The American Archivist**, v. 82, n. 2, p. 303-330, 2019. Disponível em: https://americanarchivist.org/doi/pdf/10.17723/aarc-82-02-02. Acesso em: 29 abr. 2020.

EÍTO-BRUN, Ricardo. A metadata infrastructure for a repository of civil engineering records: eac-cpf as a cornerstone for content publishing. **Journal of Archival Organization**, v. 12, n. 1-2, p. 62-76, 2015. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/15332748.2015.999502. Acesso em: 7 jun. 2023.

ELECTRONIC DISCOVERY. *In*: WIKIPEDIA: a enciclopédia livre. [San Francisco, California: Wikimedia Foundation], 2019. Disponível em: https://pt.wikipedia.org/wiki/Electronic_discovery. Acesso em: 7 jun. 2023.

ENCRYPTOMATIC LLC. **Pst Viewer Pro**. [Fargo, North Dakota], [c2023?]. Disponível em: https://www.pstviewer.com/. Acesso em: 7 jun. 2023.

ENCRYPTOMATIC LLC. **Pst Viewer Pro**. Advanced features. [Fargo, North Dakota], c2022. Disponível em: https://www.pstviewer.com/pst-viewer-features.html. Acesso em: 7 jun. 2023.

ENDERS, Markus. A METS based information package for long term accessibility of web archives. *In*: INTERNATIONAL CONFERENCE ON PRESERVATION OF DIGITAL OBJECTS (iPRES), 7., Sept. 2010, Vienna, Austria. **Proceedings** […]. Vienna, Austria: iPRES, 2010. p. 31-39. Disponível em: https://ipres-conference.org/ipres10/papers/enders-70.pdfhttps://ipres-conference.org/ipres10/papers/enders-70.pdf. Acesso em: 7 jun. 2023.

EUROPEAN COMMISSION. Law. Law by topic. Data protection. **Data protection in the EU**. [*S. l.*], [2022]. Disponível em: https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en. Acesso em: 7 jun. 2023.

EXPERTISECENTRUM DAVID vzw (eDAVID). [*S. l.*], 2014. Disponível em: http://www.edavid.be/index.php. Acesso em: 15 ago. 2023.

FARIAS, Juliana Pinheiro; ARAÚJO, Luiza Martins de Santana; EVANGELISTA, Raimunda Lima. Percepções da importância da preservação digital. **RICI: R.Ibero-amer. Ci. Inf.**, Brasília, v. 11, n. 1, p. 200-218, jan./abr. 2018. Disponível em: https://periodicos.unb.br/index.php/RICI/article/view/8475/7062. Acesso em: 7 jun. 2023.

FELT, Ulrike *et al*. **The handbook of science and technology studies**. 4th ed. Cambridge, Massachusetts: The MIT Press, c2017.

FERREIRA, Lisiane Braga; MARTINS, Marina Rodrigues; ROCKEMBACH, Moisés. Usos do arquivamento da Web na comunicação científica. **PRISMA.COM**, v. 36, p. 78-98, 2018. Disponível em: https://ojs.letras.up.pt/index.php/prismacom/article/view/3927/3676. Acesso em: 7 jun. 2023.

FERREIRA, Sueli Mara Soares Pinto; GADELHA, Zacharias; GAMBA, Camila. Digitalização e preservação digital: a experiência do Sistema Integrado de Bibliotecas da Universidade de São Paulo (SIBiUSP). **Ci. Inf.**, Brasília, DF, v. 41, n. 1, p. 143-149, jan./abr. 2012. Disponível em: http://revista.ibict.br/ciinf/article/view/1360/1539. Acesso em: 7 jun. 2023.

FIGUEIRA, Ana Cristina Peixoto; BEVILAQUA, Diego Vaz. Podcasts de divulgação científica: levantamento exploratório dos formatos de programas brasileiros. **Reciis** – Revista Eletrônica de Comunicação, Informação & Inovação em Saúde, Rio de Janeiro, v. 16, n. 1, p. 120-138, jan./mar. 2022. Disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/2427/2505. Acesso em: 22 abr. 2023.

FLEMING-MAY, Rachel. Scholarly communication: a concept analysis. **Journal of Documentation**, [*S. l.*], 2023. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/JD-09-2022-0197/full/html. Acesso em: 25 abr. 2023.

FOOT, Kirsten A.; SCHNEIDER, Steven M. **Web campaigning**. Foreword by Michael Cornfield. Cambridge, United States: The MIT Press, c2006. 288 p. (Acting With Technology, 6). Disponível em: https://direct.mit.edu/books/book/3818/Web-Campaigning. Acesso em: 7 jun. 2023.

FORMENTON, Danilo. *et al*. Os padrões de metadados como recursos tecnológicos para a garantia da preservação digital. **Biblios**, Pittsburgh, n. 68, p. 82-95, jul. 2017. Disponível em: http://www.scielo.org.pe/pdf/biblios/n68/a06n68.pdf. Acesso em: 7 jun. 2023.

FORMENTON, Danilo. **Identificação de padrões de metadados para preservação digital**. 2015. 102 f. Dissertação (Mestrado em Ciência, Tecnologia e Sociedade) – Centro de Educação e Ciências Humanas, Universidade Federal de São Carlos, São Carlos, 2015. Disponível em: https://repositorio.ufscar.br/bitstream/handle/ufscar/7221/DissDF.pdf?sequence=1&isAllowed=y. Acesso em: Acesso em: 7 jun. 2023.

FORMENTON, Danilo; GRACIOSO, Luciana de Souza; CASTRO, Fabiano Ferreira de. Revisitando a preservação digital na perspectiva da ciência da informação: aproximações conceituais. **Rev. digit. bibliotecon. cienc. inf.**, Campinas, SP, v. 13, n. 1, p. 170-191, jan./abr. 2015. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/1587/pdf_91. Acesso em: 7 jun. 2023.

FREITAS, Maria Helena. Considerações acerca dos primeiros periódicos científicos brasileiros. **Ci. Inf.**, Brasília, v. 35, n. 3, p. 54-66, set./dez. 2006. Disponível em: https://www.scielo.br/j/ci/a/RRqQp5h4xm5FSn7dSK99gTG/?lang=pt&format=pdf. Acesso em: 24 jun. 2021.

FUNDAÇÃO BIBLIOTECA NACIONAL (Brasil). **Política de preservação digital**. Rio de Janeiro: Fundação Biblioteca Nacional, 2020. 25 p. Disponível em: http://bndigital.bn.gov.br/wp-content/uploads/2021/01/politica_de_preservacao_digital_FBN_web.pdf. Acesso em: 17 jun. 2023.

GALLOTTI, Monica Marques Carvalho; BORGES, Maria Manuel. Uso do Twitter e Facebook na comunicação científica de doutorandos em ciência da informação na Península Ibérica e Brasil. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 20., 2019, Florianópolis. **Anais...** Florianópolis: Associação

Nacional de Pesquisa e Pós-Graduação em Ciência da Informação, 2019. Disponível em: https://conferencias.ufsc.br/index.php/enancib/2019/paper/view/1182/944. Acesso em: 22 abr. 2023.

GALRÃO, Ana Filomena. Estudo de caso em arqueologia digital: o gabinete da área de Sines. **Páginas A&B, Arquivos e Bibliotecas**, n. especial, p. 99-114, 2017. Disponível em: https://ojs.letras.up.pt/index.php/paginasaeb/article/view/2658/2446. Acesso em: 7 jun. 2023.

GARVEY, William D.; GRIFFITH, Belver C. Communication, the essence of science, Apêndice A, B. *In*: GARVEY, W.D. **Communication**: the essence of science. Oxford: Pergamon Press, 1979. p. 299.

GEORGE WASHINGTON UNIVERSITY LIBRARIES. **Social feed manager**: helping researchers and archivists build social media collections. About. Documentation. Washington, D.C., United States, c2021. Disponível em: https://gwu-libraries.github.io/sfm-ui/. Acesso em: 7 jun. 2023.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 5. ed. São Paulo: Atlas, 2010. 184 p.

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2012. 200 p.

GILLILAND, Anne J. Setting the stage. *In*: BACA, Murtha. (ed.). **Introduction to metadata**. 3rd ed. Los Angeles, California: Getty Publications, c2016. 92 p. Disponível em: http://www.getty.edu/publications/intrometadata/. Acesso em: 19 ago. 2020.

GOMES, Daniel. Preservar a web: um desafio ao alcance de todos. **Actas dos Congressos Nacionais de Bibliotecários, Arquivistas e Documentalistas**, Lisboa, n. 10, p. 1-9, 2010. Disponível em: https://publicacoes.bad.pt/revistas/index.php/congressosbad/article/view/158. Acesso em: 7 jun. 2023.

GOMES, Daniel; FREITAS, Sérgio; SILVA, Mário J. Design and selection criteria for a national web archive. *In*: EUROPEAN CONFERENCE ON RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES – ECDL, 10., September 2006, Alicante, Spain. **Proceedings** […]. Alicante, Spain: ECDL, 2006. p. 196-207. Disponível em: https://link.springer.com/chapter/10.1007/11863878_17. Acesso em: 7 jun. 2023.

GRÁCIO, José Carlos Abbud. Modelo para elaboração de políticas de preservação digital de documentos de arquivo por instituições de ensino superior: o caso da Unesp. **Reciis** – Rev Eletron Comun Inf Inov Saúde, v. 14, n. 3, p. 563-579, jul./set. 2020. Disponível em: https://www.arca.fiocruz.br/bitstream/icict/43728/2/6.pdf. Acesso em: 7 jun. 2023.

GRÁCIO, José Carlos Abbud. **Preservação digital na gestão da informação**: um modelo processual para as instituições de ensino superior. São Paulo: Cultura Acadêmica, c2012. 214 p. Disponível em: https://repositorio.unesp.br/bitstream/handle/11449/113727/ISBN9788579833335.pdf?sequence=1&isAllowed=y. Acesso em: 7 jun. 2023.

GRÁCIO, José Carlos Abbud; FADEL, Bárbara; VALENTIM, Marta Lígia Pomim. Preservação digital nas instituições de ensino superior: aspectos organizacionais, legais e técnicos. **Perspect. ciênc. inf.**, Belo Horizonte, v. 18, n. 3, p. 111-129, jul./set. 2013. Disponível em: https://www.scielo.br/j/pci/a/XnvBfYVhjnpzxWPQ79NwFCb/?format=pdf&lang=pt. Acesso em: 7 jun. 2023.

GUENTHER, Rebecca S. MODS: the metadata object description schema. **Portal**: Libraries and the Academy, v. 3, n. 1, p. 137-150, Jan. 2003. Disponível em: https://muse.jhu.edu/article/38558/pdf. Acesso em: 7 jun. 2023.

GUENTHER, Rebecca Squire; DAPPERT, Angela; PEYRARD, Sébastien. An introduction to the PREMIS data dictionary for digital preservation metadata. *In*: GUENTHER, Rebecca Squire; DAPPERT, Angela; PEYRARD, Sébastien. **Digital preservation metadata for practitioners**. Cham, Switzerland: Springer, Dec. c2016. p. 23-36. Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-43763-7_3. Acesso em: 7 jun. 2023.

GUENTHER, Rebecca; MYRICK, Leslie. Archiving web sites for preservation and access: MODS, METS and MINERVA. **Journal of Archival Organization**, v. 4, n. 1/2, p. 141-166, 2007. Disponível em: https://www.tandfonline.com/doi/abs/10.1300/J201v04n01_08. Acesso em: 7 jun. 2023.

GUL, Sumeer; MAHAJAN, Iram; ALI, Asifa. The growth and decay of urls citation: a case of an online library & information science journal. **Malaysian Journal of Library & Information Science**, Kuala Lumpur, v. 19, n. 3, p. 27-39, 2014. Disponível em: https://www.researchgate.net/publication/268871495_The_growth_and_decay_of_URLs_citation_A_case_of_an_online_Library_Information_Science_journal. Acesso em: 7 jun. 2023.

GULKA, Juliana Aparecida; SILVEIRA, Lúcia da. Revisão de metadados para confiabilidade de artigos publicados em acesso aberto. *In*: Congresso Brasileiro de Biblioteconomia e Documentação (CBBD), 28., out. 2019, Vitória, ES. **Anais...** Vitória, ES: CBBD, 2019. p. 1-6. Disponível em: https://anaiscbbd.emnuvens.com.br/anais/article/view/2080. Acesso em: 7 jun. 2023.

HABING, Thomas G. **ECHO Dep METS profile for web site captures**. [*S. l.*], 2006. Disponível em: https://www.loc.gov/standards/mets/profiles/00000016.html. Acesso em: 7 jun. 2023.

HACKETT, Edward J. *et al*. **The handbook of science and technology studies**. 3th ed. Cambridge, Massachusetts: The MIT Press, c2008.

HALLGRIMSSON, Thorsteinn. Access and finding aids. *In*: MASANÈS, Julien. (Ed.). **Web archiving**. Berlin: Springer, c2006. p. 130-151. Disponível em: https://link.springer.com/book/10.1007/978-3-540-46332-0. Acesso em: 7 jun. 2023.

HANZO. [*S. l.*], c2022. Disponível em: https://www.hanzo.co/. Acesso em: 7 jun. 2023.

HARPER, Corey A. Dublin Core Metadata Initiative: beyond the element set. **Information Standards Quarterly (ISQ)**, v. 22, n. 1, p. 19-28, winter 2010. Disponível em: https://www.niso.org/sites/default/files/stories/2019-11/FE_DCMI_Harper_isqv22no1.pdf. Acesso em: 7 jun. 2023.

HARVARD LIBRARY. Preservation services. Digital collections. **Web archiving**. Cambridge, Massachusetts, c2022. Disponível em: https://preservation.library.harvard.edu/web-archiving. Acesso em: 7 jun. 2023.

HATHITRUST. [Ann Arbor, Michigan], c2023. Disponível em: https://www.hathitrust.org/. Acesso em: 7 jun. 2023.

HATHITRUST. About. Welcome to HathiTrust! Help. **Help – General**. [Ann Arbor, Michigan], [2022]. Disponível em: https://www.hathitrust.org/help_general. Acesso em: 7 jun. 2023.

HAYASHI, Carlos Roberto Massao; FERREIRA JR, Amarílio. A comunidade científica em educação: uma abordagem crítica. **Série-Estudos** - Periódico do Mestrado em Educação da UCDB, Campo Grande, n. 21, p. 11-27, jul./dez. 2006. Disponível em: https://www.serie-estudos.ucdb.br/serie-estudos/article/view/258/113. Acesso em: 7 jun. 2023.

HAYASHI, Maria Cristina Piumbato Innocentini; HAYASHI, Carlos Roberto Massao; FURNIVAL, Ariadne Chloe Mary. Ciência, Tecnologia e Sociedade: apontamentos preliminares sobre a constituição do campo no Brasil. *In*: SOUZA, Cidoval Morais de; HAYASHI, Maria Cristina Piumbato Innocentini. **Ciência, Tecnologia e Sociedade**: enfoques teóricos e aplicados. São Carlos, SP: Pedro & João Editores, 2008. p. 29-88.

HEDSTROM, Margaret. Digital preservation: a time bomb for digital libraries. **Computers and the Humanities**, Netherlands, v. 31, p. 189-202, 1998. Disponível em: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/42573/?sequence=1. Acesso em: 7 jun. 2023.

HEDSTROM, Margaret. Digital preservation: problems and prospects. **Digital Library Network (DLnet)**, n. 20, Mar. 2001. Disponível em: http://www.dl.slis.tsukuba.ac.jp/DLjournal/No_20/1-hedstrom/1-hedstrom.html. Acesso em: 7 jun. 2023.

HIGGINS, Sarah. Digital curation: the emergence of a new discipline. **The International Journal of Digital Curation**, v. 6, n. 2, 2011. Disponível em: http://www.ijdc.net/index.php/ijdc/article/view/184. Acesso em: 7 jun. 2023.

HOCKX-YU, Helen. The past issue of the Web. *In*: INTERNATIONAL WEB SCIENCE CONFERENCE – WebSci, 3., June 2011, Koblenz, Germany. **Proceedings** […]. Koblenz, Germany: Association for Computing Machinery, 2011. p. 15-17. Disponível em: https://dl.acm.org/doi/10.1145/2527031.2527050. Acesso em: 7 jun. 2023.

HUSSAIN, Abid; VATRAPU, Ravi. Social Data Analytics Tool (SODATO). *In*: TREMBLAY, Monica Chiarini *et al*. (Eds.). **Advancing the impact of design science**: moving from theory to practice. 9th International Conference On Design Science Research In Information Systems (DESRIST) 2014, Miami, Florida, United States, May 22-24, 2014. Proceedings. [Switzerland]: Springer, 2014. p. 368-372. (Lecture Notes in Computer Science – LNCS, v. 8463). Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-06701-8_27. Acesso em: 7 jun. 2023.

INFORMASUS-UFSCAR. [*S. l.*], c2023. Disponível em: https://informasus.ufscar.br/. Acesso em: 5 ago. 2023.

INFORMASUS-UFSCAR. Institucional. Sobre o Informasus. **Sobre o projeto**. [*S. l.*], c2021. Disponível em: https://www.informasus.ufscar.br/sobre-o-projeto/. Acesso em: 7 jun. 2023.

INNARELLI, Humberto Celeste. Os dez mandamentos da preservação digital: uma brevíssima introdução. *In*: SEMINÁRIO SERVIÇOS DE INFORMAÇÃO EM MUSEUS, 2., 2012, São Paulo. **Anais...** São Paulo: Pinacoteca do Estado de São Paulo, 2014. p. 317-325. Disponível em: http://biblioteca.pinacoteca.org.br:9090/bases/biblioteca/322818.pdf. Acesso em: 7 jun. 2023.

INNARELLI, Humberto Celeste. Preservação digital e seus dez mandamentos. *In*: SANTOS, Vanderlei Batista dos; INNARELLI, Humberto Celeste; SOUSA, Renato Tarciso Barbosa de. (Org.). **Arquivística**: temas contemporâneos: classificação, preservação digital, gestão do conhecimento. 3. ed. Brasília: Senac, 2009. p. 21-71.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA. Tecnologias para Informação. Rede Brasileira de Serviços de Preservação Digital – CARINIANA. **Portal da Rede Cariniana**. Institucional. Cariniana. Brasília, DF, Nov. 2016. Acesso em: https://cariniana.ibict.br/index.php/institucional/cariniana. Disponível em: 7 jun. 2023.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA. **Cariniana, Rede Brasileira de Serviços de Preservação Digital**. Brasília, DF, c2022. Disponível em: https://cariniana.ibict.br/. Acesso em: 7 jun. 2023.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS (IFLA) Study Group on the Functional Requirements for Bibliographic Records. Functional requirements for bibliographic records: final report. **UBCIM Publications - New Series**, vol. 19. München: K. G. Saur, 1998. 136 p. Disponível em: https://www.ifla.org/files/assets/cataloguing/frbr/frbr.pdf. Acesso em: 7 jun. 2023.

INTERNATIONAL INTERNET PRESERVATION CONSORTIUM (IIPC). About IIPC. IIPC members. **California digital library**. [*S. l.*], c2021a. Disponível em: https://netpreserve.org/about-us/members/california-digital-library/. Acesso em: 7 jun. 2023.

INTERNATIONAL INTERNET PRESERVATION CONSORTIUM (IIPC). About the IIPC. IIPC members. **Hanzo archives limited**. [*S. l.*], c2021b. Disponível em: http://netpreserve.org/about-us/members/hanzo-archives-limited/. Acesso em: 7 jun. 2023.

INTERNATIONAL INTERNET PRESERVATION CONSORTIUM (IIPC). About IIPC. Working groups. **Content development working group**. [*S. l.*], c2023. Disponível em: https://netpreserve.org/about-us/working-groups/content-development-working-group/. Acesso em: 7 jun. 2023.

INTERNATIONAL INTERNET PRESERVATION CONSORTIUM (IIPC). **Web archiving**. About archiving. [*S. l.*], c2022a. Disponível em: https://netpreserve.org/web-archiving/. Acesso em: 7 jun. 2023.

INTERNATIONAL INTERNET PRESERVATION CONSORTIUM (IIPC). **Web archiving**. Case studies. [*S. l.*], c2022b. Disponível em: https://netpreserve.org/web-archiving/case-studies/. Acesso em: 7 jun. 2023.

INTERNATIONAL INTERNET PRESERVATION CONSORTIUM. Web archiving. **Tools & software**. [*S. l.*], c2022c. Disponível em: https://netpreserve.org/web-archiving/tools-and-software/. Acesso em: 7 jun. 2023.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 15489-1:2001**: information and documentation: records management: part 1: general. Geneva: ISO, 2001. Disponível em: https://www.iso.org/standard/31908.html. Acesso em: 7 jun. 2023.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 16439:2014**: Information and documentation: methods and procedures for assessing the impact of libraries, Geneva: ISO, 2014. Disponível em: https://www.iso.org/standard/56756.html. Acesso em: 7 jun. 2023.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO/TR 14873:2013**: Information and documentation: statistics and quality issues for web archiving, Geneva: ISO, 2013. Disponível em: https://www.iso.org/obp/ui/#iso:std:iso:tr:14873:ed-1:v1:en. Acesso em: 7 jun. 2023.

INTERNATIONAL STANDARD SERIAL NUMBER PORTAL. **Keepers registry**. [Paris, France], c2023. Disponível em: https://keepers.issn.org/. Acesso em: 7 jun. 2023.

INTERNATIONAL STANDARD SERIAL NUMBER PORTAL. Keepers registry. **The keepers**. [Paris, France], c2022. Disponível em: https://keepers.issn.org/keepers. Acesso em: 7 jun. 2023.

INTERNET Archive. [San Francisco, California, United States], [2023?]. Disponível em: https://archive.org/. Acesso em: 5 ago. 2023.

INTERNET ARCHIVE. **Archive-it**. [San Francisco, California, United States], 2014a. Disponível em: https://archive-it.org/. Acesso em: 5 ago. 2023.

INTERNET ARCHIVE. Archive-it. **About archive-it**. [San Francisco, California], [2021a]. Disponível em: https://archive-it.org/blog/learn-more/. Acesso em: 7 jun. 2023.

INTERNET ARCHIVE. Archive-it. Archive-it help center. Archive-it user guide. Collections. **Add, edit, and manage your metadata**. [*S. l.*], 2021b. Disponível em: https://support.archive-it.org/hc/en-us/articles/208332603-Add-edit-and-manage-your-metadata. Acesso em: 7 jun. 2023.

INTERNET ARCHIVE. Archive-it. Projects. Projects and programs. **Spontaneous event collections**. [San Francisco, California], [2022a]. Disponível em: https://archive-it.org/blog/projects/spontaneous-events/. Acesso em: 7 jun. 2023.

INTERNET ARCHIVE. **Heritrix**. [*S. l.*]: GitHub, c2023. Disponível em: https://github.com/internetarchive/heritrix3/wiki. Acesso em: 5 ago. 2023.

INTERNET ARCHIVE. Internet archive APIs. About archive.org metadata. **Internet archive metadata**. [San Francisco, California], Dec. 2018. Disponível em: https://archive.org/services/docs/api/metadata-schema/index.html. Acesso em: 7 jun. 2023.

INTERNET ARCHIVE. **The wayback machine**. Wayback machine & web archiving. Wayback machine general information. [San Francisco, California], c2022b. Disponível em: https://help.archive.org/help/wayback-machine-general-information/. Acesso em: 7 jun. 2023.

INTERNET ARCHIVE. **Wayback machine**. [San Francisco, California], 2014b. Disponível em: http://web.archive.org/. Acesso em: 5 ago. 2023.

ISSN INTERNATIONAL CENTER. Services. Online services. **Keepers registry**. [Paris, France], [2022]. Disponível em: https://www.issn.org/services/online-services/keepers-registry/. Acesso em: 7 jun. 2023.

ITHAKA. JSTOR. About JSTOR. **Journals**. What's in JSTOR. New York, United States, c2022a. Disponível em: https://about.jstor.org/librarians/journals/. Acesso em: 7 jun. 2023.

ITHAKA. **JSTOR**. New York, United States, c2023a. Disponível em: https://www.jstor.org/. Acesso em: 7 jun. 2023.

ITHAKA. Portico. For participants. **Why portico**. Coverage. Triggered content. New York, United States, c2022b. Disponível em: https://www.portico.org/why-portico/. Acesso em: 7 jun. 2023.

ITHAKA. **Portico**. New York, United States, c2023b. Disponível em: https://www.portico.org/. Acesso em: 7 jun. 2023.

JONES, Maggie. e-Journals: archiving and preservation. Briefing paper. JISC, March 2007. Disponível em: http://www.jisc.ac.uk/publications/publications/pub_ejournalspreservationbp. Acesso em: 7 jun. 2023.

KA'AI-MAHUTA, Rachael. Preserving indigenous voices: web archiving in Aotearoa/New Zealand. **Interaction Design and Architecture(s) Journal** - IxD&A, n. 41, p. 24-30, 2019. Disponível em: http://www.mifav.uniroma2.it/inevent/events/idea2010/doc/41_2.pdf. Acesso em: 7 jun. 2023.

KHAN, Muzammil; RAHMAN, Arif Ur. A systematic approach towards web preservation. **Information Technology and Libraries**, v. 38, n. 1, p. 71-90. 2019. Disponível em: https://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf. Acesso em: 7 jun. 2023.

KIM, Heejung; LEE, Hyewon. Development of metadata elements for intensive web archiving. **Journal of the Korean Society for Information Management**, Songdo, South Korea, v. 24, n. 2, p. 143-160, June 2007. Disponível em: https://www.koreascience.or.kr/article/JAKO200727500195825.pdf. Acesso em: 7 jun. 2023.

KOERBIN, Paul. National Web Archiving in Australia: representing the comprehensive. *In*: GOMES, Daniel et al. (Ed.). **The Past Web**: exploring web archives. London, UK, c2021. p. 23-32. Disponível em: https://link.springer.com/book/10.1007/978-3-030-63291-5. Acesso em: 7 jun. 2023.

KONINKLIJKE BIBLIOTHEEK. Organisatie. Onderzoek & expertise. **Digitaal magazijn & duurzame opslag**. Den Haag, Nederlands, [c2021]. Disponível em: https://www.kb.nl/organisatie/onderzoek-expertise/digitaal-magazijn-duurzame-opslag. Acesso em: 7 jun. 2023.

KONINKLIJKE BIBLIOTHEEK. Over ons. Projecten. **Nieuw e-Depot**. Den Haag, Nederlands, [2023?]. Disponível em: https://www.kb.nl/over-ons/projecten/nieuw-e-depot. Acesso em: 7 jun. 2023.

KUMAR, D. Vinay; KUMAR, B. T. Sampath. Recovery of vanished urls: comparing the efficiency of internet archive and google. **Malaysian Journal of Library & Information Science**, Kuala Lumpur, v. 22, n. 2, p. 31-43, Aug. 2017. Disponível em: https://www.researchgate.net/publication/318316951_Recovery_of_vanished_URLs_Comparing_the_efficiency_of_Internet_Archive_and_Google. Acesso em: 7 jun. 2023.

KUSHKOWSKI, Jeffrey D. Web citation by graduate students: a comparison of print and electronic theses. **portal:** Libraries and the Academy, Baltimore, v. 5, n. 2, p. 259-276, April 2005. Disponível em: https://muse.jhu.edu/article/181570. Acesso em: 7 jun. 2023.

LANDSBÓKASAFN ÍSLANDS HÁSKÓLABÓKASAFN. **Íslenska vefsafnið**. Algengar spurningar. Um vefsafnið. Reykjavík, Ísland, [2022]. Disponível em: https://vefsafn.is/. Acesso em: 7 jun. 2023.

LANDSBÓKASAFN ÍSLANDS HÁSKÓLABÓKASAFN. **Íslenska vefsafnið**. Reykjavík, Ísland, [2023?]. Disponível em: https://vefsafn.is/. Acesso em: 7 jun. 2023.

LAVOIE, Brian; GARTNER, Richard. Preservation metadata. 2nd edition. **DPC Technology Watch Report 13-03**, p. 1-36, May c2013. Disponível em: https://www.dpconline.org/docs/technology-watch-reports/894-dpctw13-03/file. Acesso em: 7 jun. 2023.

LEROY-TERQUEM, Mélanie. Archiver le web d' aujourd'hui: pour les générations de demain. **Chroniques de la bnf**, Paris, n. 86, p. 26, sept./déc. 2019. Disponível em: http://chroniques.bnf.fr/pdf/chroniques_86.pdf. Acesso em: 7 jun. 2023.

LIBRARY AND ARCHIVES CANADA. **Guidelines on file formats for transferring information resources of enduring value**. [Ottawa], Feb. 2015. 25 p. Disponível em: https://www.bac-lac.gc.ca/eng/services/government-information-resources/guidelines/Documents/file-formats-irev.pdf. Acesso em: 7 jun. 2023.

LIBRARY OF CONGRESS. **Development of the Encoded Archival Description DTD**. Dec. 2013. Disponível em: http://www.loc.gov/ead/eaddev.html. Acesso em: 7 jun. 2023.

LIBRARY OF CONGRESS. **Library of Congress collections policy statements supplementary guidelines**: web archiving. [Washington, DC], July 2022a. 5 p. Disponível em: https://www.loc.gov/acq/devpol/webarchive.pdf. Acesso em: 7 jun. 2023.

LIBRARY OF CONGRESS. **METS**: an overview & tutorial. [Washington, DC], Mar. 2017. Disponível em: https://www.loc.gov/standards/mets/METSOverview.v2.html. Acesso em: 7 jun. 2023.

LIBRARY OF CONGRESS. **MODS user guidelines**. MODS elements and attributes. Version 3. [Washington, DC], Aug. 2018a. Disponível em: http://www.loc.gov/standards/mods/userguide/. Acesso em: 7 jun. 2023.

LIBRARY OF CONGRESS. **MODS**: uses and features. [Washington, DC], Feb. 2016. Disponível em: http://www.loc.gov/standards/mods/mods-overview.html. Acesso em: 7 jun. 2023.

LIBRARY OF CONGRESS. Programs. **Web archiving**. [Washington, DC], [2023a]. Disponível em: https://www.loc.gov/programs/web-archiving/. Acesso em: 7 jun. 2023.

LIBRARY OF CONGRESS. Programs. **Web archiving**. About this program. Frequently asked questions. [Washington, DC], [2022b]. Disponível em: https://www.loc.gov/programs/web-archiving/about-this-program/. Acesso em: 7 jun. 2023.

LIBRARY OF CONGRESS. Programs. Web archiving. About this program. Web archives. **Collections with web archives**. [Washington, DC], [2021]. Disponível em: https://www.loc.gov/web-archives/collections/?st=gallery. Acesso em: 7 jun. 2023.

LIBRARY OF CONGRESS. Programs. **Web archiving**. About this program. Web archiving collaborations. Frequently asked questions. Glossary. [Washington, DC], [2022?c]. Disponível em: https://www.loc.gov/programs/web-archiving/about-this-program/. Acesso em: 7 jun. 2023.

LIBRARY OF CONGRESS. Programs. Web archiving. **For researchers**. [Washington, DC], [2022d]. Disponível em: https://www.loc.gov/programs/web-archiving/for-researchers/. Acesso em: 7 jun. 2023.

LIBRARY OF CONGRESS. Programs. Web archiving. **For site owners**. [Washington, DC], [2022?e]. Disponível em: https://www.loc.gov/programs/web-archiving/for-site-owners/. Acesso em: 7 jun. 2023.

LIBRARY OF CONGRESS. Standards. **<ead>**, Encoded Archival Description, official site. [Washington, DC], January 2023b. Disponível em: https://www.loc.gov/ead/. Acesso em: 20 jul. 2023.

LIBRARY OF CONGRESS. Standards. **AudioMD and videoMD**: technical metadata for audio and video. [Washington, DC], Oct. 2011. Disponível em: https://www.loc.gov/standards/amdvmd/. Acesso em: 7 jun. 2023.

LIBRARY OF CONGRESS. Standards. **MADS**: metadata authority description schema: official web site. [Washington, DC], May 2018b. Disponível em: https://www.loc.gov/standards/mads/. Acesso em: 7 jun. 2023.

LIBRARY OF CONGRESS. Standards. **MARCXML**: MARC 21 xml schema: official web site. [Washington, DC], June 2020a. Disponível em: https://www.loc.gov/standards/marcxml/. Acesso em: 7 jun. 2023.

LIBRARY OF CONGRESS. Standards. **METS**, Metadata Encoding & Transmission Standard, official Web site. [Washington, DC], 2023c. Disponível em: https://www.loc.gov/standards/mets/. Acesso em: 20 jul. 2023.

LIBRARY OF CONGRESS. Standards. **MODS**, Metadata Object Description Schema, official Web site. [Washington, DC], 2023d. Disponível em: https://www.loc.gov/standards/mods/. Acesso em: 20 jul. 2023.

LIBRARY OF CONGRESS. Standards. **NISO metadata for images in xml schema**: technical metadata for digital still images standard: official web site. [Washington, DC], Nov. 2015. Disponível em: https://www.loc.gov/standards/mix/. Acesso em: 7 jun. 2023.

LIBRARY OF CONGRESS. Standards. **PREMIS**, preservation metadata, maintenance activity. [Washington, DC], June 2023e. Disponível em: https://www.loc.gov/standards/premis/. Acesso em: 20 jul. 2023.

LIBRARY OF CONGRESS. Standards. **TextMD**: technical metadata for text: official web site. [Washington, DC], Aug. 2020b. Disponível em: https://www.loc.gov/standards/textMD/. Acesso em: 7 jun. 2023.

LIBRARY OF CONGRESS. Standards. **VRA Core a data standard for the description of images and works of art and culture**. [Washington, DC], 2023f. Disponível em: https://www.loc.gov/standards/vracore/. Acesso em: 20 jul. 2023.

LIBRARY OF VIRGINIA. Virginia Memory. Collections. **Kaine Email Project @ Lva**. Look Under The Hood. Virginia, United States, c2016. Disponível em: https://www.virginiamemory.com/collections/kaine/under-the-hood. Acesso em: 7 jun. 2023.

LIMA, Fábio Rogério Batista; SANTOS, Plácida Leopoldina V. A. C.; SANTARÉM SEGUNDO, José Eduardo. Padrão de metadados no domínio museológico. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 21, n. 3, p. 50-69, jul./set. 2016. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/2639/1789. Acesso em: 18 maio 2020.

LONG, Andrew Stawowczyk. **Long-term preservation of web archives**: experimenting with emulation and migration methodologies. [*S. l.*]: International Internet Preservation Consortium (IIPC), Dec. 2009. 54 p. Disponível em: https://www.ltu.se/cms_fs/1.67312!/file/LongtermPresOfWebArchivesOsv.pdf. Acesso em: 7 jun. 2023.

LÓPEZ CEREZO, José Antonio. Ciência, tecnologia e sociedade: o estado da arte na Europa e nos Estados Unidos. *In*: SANTOS, Lucy Woellner dos. *et al*. (Org.). **Ciência, tecnologia e sociedade**: o desafio da interação. Londrina, PR: IAPAR, 2002. p. 3-38.

LUBAS, Rebecca L.; JACKSON, Amy S.; SCHNEIDER, Ingrid. Using VRA Core 4.0. *In*: LUBAS, Rebecca L.; JACKSON, Amy S.; SCHNEIDER, Ingrid. **The metadata manual**: a practical workbook. Oxford, UK: Chandos Publishing, 2013. p. 135-164. Disponível em: Acesso em: https://www.elsevier.com/books/the-metadata-manual/lubas/978-1-84334-729-3. Acesso em: 7 jun. 2023.

LUNA, Sergio Vasconcelos de. **Planejamento de pesquisa**: uma introdução. São Paulo: Educ, 1997. 108 p.

LUZ, Charlley dos Santos; MARINGELI, Isabel Cristina Ayres da Silva. Política de preservação digital: caso Pinacoteca de São Paulo. **Perspectivas em Ciência da Informação**, v. 23, n. 2, p. 189-200, abr./jun. 2018. Disponível em: https://www.scielo.br/j/pci/a/vDFbVMFKGd6hN8FqjZxt8rt/?format=pdf&lang=pt. Acesso em: 7 jun. 2023.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de metodologia científica**. 8. ed. atual. São Paulo: Atlas, 2017. 368 p.

MÁRDERO ARELLANO, Miguel Ángel. Cariniana: uma rede nacional de preservação digital. **Ci. Inf.**, Brasília, DF, v. 41, n. 1, p. 83-91, jan./abr. 2012. Disponível em: http://revista.ibict.br/ciinf/article/view/1354/1533. Acesso em: 7 jun. 2023.

MÁRDERO ARELLANO, Miguel Ángel. **Critérios para a preservação digital da informação científica**. 2008. Tese (Doutorado em Ciência da Informação) - Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília, 2008. Disponível em: https://repositorio.unb.br/bitstream/10482/1518/1/2008_MiguelAngelMarderoArellano.pdf. Acesso em: 15 dez. 2021.

MARTIN NEIRA, Juan Ignacio; TRILLO DOMÍNGUEZ, Magdalena; OLVERA LOBO, María Dolores. Comunicación científica tras la crisis del COVID-19: estrategias de publicación en TikTok en el tablero transmedia. **Revista Latina de Comunicación Social**, v. 81, p. 109-132, 2023. Disponível em: https://nuevaepoca.revistalatinacs.org/index.php/revista/article/view/1841/4104. Acesso em: 22 abr. 2023.

MARTÍNEZ, José M. **MPEG-7 overview (version 10)**. ISO/IEC JTC1/SC29/WG11 N6828. Coding of moving pictures and audio. Palma de Mallorca, Spain, Oct. 2004. 76 p. Disponível em: https://mpeg.chiariglione.org/standards/mpeg-7. Acesso em: 7 jun. 2023.

MASANÈS, Julien. Selection for web archives. *In*: MASANÈS, Julien. **Web archiving**. Berlin: Springer, c2006a. p. 71-91. Disponível em: https://link.springer.com/book/10.1007/978-3-540-46332-0. Acesso em: 7 jun. 2023.

MASANÈS, Julien. Towards continuous web archiving: first results and an agenda for the future. **D-Lib Magazine**, v. 8, n. 12, Dec. 2002. Disponível em: http://www.dlib.org/dlib/december02/masanes/12masanes.html. Acesso em: 7 jun. 2023.

MASANÈS, Julien. Web archiving methods and approaches: a comparative study. **Library Trends**, v. 54, n. 1, p. 72-90, Summer 2005. Disponível em: Acesso em: https://muse.jhu.edu/article/193226. Acesso em: 7 jun. 2023.

MASANÈS, Julien. **Web archiving**. Berlin: Springer, c2006b. 234 p. Disponível em: https://link.springer.com/book/10.1007/978-3-540-46332-0. Acesso em: 7 jun. 2023.

MASANÈS, Julien. Web archiving: issues and methods. *In*: MASANÈS, Julien. **Web archiving**. Berlin: Springer, c2006c. p. 1-53. Disponível em: https://link.springer.com/book/10.1007/978-3-540-46332-0. Acesso em: 7 jun. 2023.

MASSARANI, Luisa; MOREIRA, Ildeu de Castro. Divulgación de la ciencia: perspectivas históricas y dilemas permanentes. **Quark**, Barcelona, n. 32, p. 30-35, abr./jun. 2004. Disponível em: http://quark.prbb.org/32/032030.pdf. Acesso em: 12 jul. 2021.

MAYERNIK, Matthew S. Metadata accounts: achieving data and evidence in scientific research. **Social Studies of Science**, v. 49, n. 5, p. 732-757, 2019. Disponível em: https://journals.sagepub.com/doi/10.1177/0306312719863494. Acesso em: 7 jun. 2023.

MCCALLUM, Sally H. An introduction to the metadata object description schema (MODS). **Library Hi Tech**, v. 22, n. 1, p. 82-88, 2004. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/07378830410524521/full/html. Acesso em: 7 jun. 2023.

MCDONOUGH, Jerome P. METS: standardized encoding for digital library objects. **International Journal on Digital Libraries**, v. 6, n. 2, p. 148-158, April 2006. Disponível em: https://link.springer.com/article/10.1007/s00799-005-0132-1. Acesso em: 7 jun. 2023.

MEADOWS, Arthur Jack. **A Comunicação científica**. Brasília, DF: Briquet de Lemos, 1999. 268 p.

MEDEIROS, Jean Maicon Rickes; COSTA, Maria Conceição da. Divulgação científica nas redes sociais: estudos sobre o uso de redes sociais na C&T. *In*: SIMPÓSIO NACIONAL DE CIÊNCIA, TECNOLOGIA E SOCIEDADE, 7., 2017, Brasília, DF. **Anais…** Brasília, DF: UNB, 2017. p. 1-13. Disponível em: http://esocite2017.com.br/anais/beta/trabalhoscompletos/gt/13/esocite2017_gt13_jeanMaiconRickesMedeiros.pdf. Acesso em: 15 jan. 2022.

MELO, Jonas Ferrigolo. **Arquivamento dos websites do governo federal brasileiro**: preservação do domínio gov.br. 2020. Dissertação (Mestrado em Comunicação e Informação) - Programa de Pós-Graduação em Comunicação e Informação, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, 2020. Disponível em: https://lume.ufrgs.br/bitstream/handle/10183/210671/001115375.pdf?sequence=1&isAllowed=y. Acesso em: 7 jun. 2023.

MELO, Jonas Ferrigolo; ROCKEMBACH, Moisés. Arquivabilidade de websites para preservação digital: estudo a partir da área da saúde. **Reciis** – Rev Eletron Comun Inf Inov Saúde, v. 14, n. 3, p. 529-545, 2020. Disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/2116/2371. Acesso em: 7 jun. 2023.

MEMENTO PROJECT. **Time travel**. [*S. l.*], [2023]. Disponível em: http://timetravel.mementoweb.org/. Acesso em: 7 jun. 2023.

MEMENTO PROJECT. **Time travel**. About the Time Travel Service. [*S. l.*], 2016. Disponível em: http://timetravel.mementoweb.org/about/. Acesso em: 7 jun. 2023.

MIRANDA, João; GOMES, Daniel. Trends in web characteristics. Trends in web characteristics. *In*: LATIN AMERICAN WEB CONGRESS – LA-WEB, 11., 2009, Merida, Mexico. **Proceedings** […]. Merida, Mexico: Institute of Electrical and Electronic Engineers (IEEE) Computer Society, c2009. p. 146-153. Disponível em: https://ieeexplore.ieee.org/document/5341605. Acesso em: 7 jun. 2023.

MOGHADDAM, Golnessa Galyani. Preserving scientific electronic journals: a study of archiving initiatives. **The Electronic Library**, v. 26, n. 1, p. 83-96, 2008. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/02640470810851761/full/html. Acesso em: 7 jun. 2023.

MOORE, Ray. We passed! great result from coretrustseal accreditation. *In*: ADS Blog. [*S. l.*], 19 June 2020. Disponível em: https://archaeologydataservice.ac.uk/blog/2020/06/we-passed/. Acesso em: 7 jun. 2023.

MOREIRA, Ildeu de Castro; MASSARANI, Luisa. Aspectos históricos da divulgação científica no Brasil. *In*: MASSARANI, Luisa; MOREIRA, Ildeu de Castro; BRITO, Fatima. (org.). **Ciência e público**: caminhos da divulgação científica no Brasil. Rio de Janeiro: Casa da Ciência – Centro Cultural de Ciência e Tecnologia da Universidade Federal do Rio de Janeiro, 2002. p. 43-64. Disponível em: http://www.museudavida.fiocruz.br/images/Publicacoes_Educacao/PDFs/cienciaepublico.pdf. Acesso em: 7 jun. 2023.

MUELLER, Suzana P. M.; CARIBÉ, Rita de Cássia do Vale. Comunicação científica para o público leigo: breve histórico. **Inf. Inf.**, Londrina, v. 15, n. esp, p. 13-30, 2010. Disponível em: https://repositorio.unb.br/bitstream/10482/13202/1/ARTIGO_ComunicacaoCientificaPublico.pdf. Acesso em: 13 jun. 2021.

MUELLER, Suzana Pinheiro Machado. Literatura científica, comunicação científica e ciência da informação. *In*: TOUTAIN, Lídia Maria Batista Brandão. (Org.). **Para entender a ciência da informação**. Salvador, BA: EDUFBA, 2007. p. 125-144. Disponível em: https://repositorio.ufba.br/ri/bitstream/ufba/145/1/Para%20entender%20a%20ciencia%20da%20informacao.pdf. Acesso em: 7 jun. 2023.

MURRAY, Kathleen R.; HSIEH, Inga K. Archiving web-published materials: a needs assessment of librarians, researchers, and content providers. **Government Information Quarterly**, v. 25, n. 1, p. 66-89, 2007. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0740624X07000354. Acesso em: 7 jun. 2023.

NAJAR, Jaffer Kabir; WANI, Javaid Ahmad. Digital preservation: an overview. **Library Philosophy and Practice**, [Lincoln], p. 1-16, Sept. 2019. Disponível em: https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=6113&context=libphilprac. Acesso em: 7 jun. 2023.

NARODNA IN UNIVERZITETNA KNJIŽNICA. **Spletni arhiv**. [Ljubljana, Slovenija], c2023. Disponível em: https://arhiv.nuk.uni-lj.si/. Acesso em: 7 jun. 2023.

NARODNA IN UNIVERZITETNA KNJIŽNICA. **Spletni arhiv**. O arhivu. Za ustvarjalce vsebin. [Ljubljana, Slovenija], c2022. Disponível em: https://arhiv.nuk.uni-lj.si/. Acesso em: 7 jun. 2023.

NARODNE IN UNIVERZITETNE KNJIŽNICE. **Pravilnik o vrstah in izboru elektronskih publikacij za obvezni izvod**. Uradnem listu Republike Slovenije (RS), št. 90/2007 z dne 5.10.2007. Ljubljana, Slovenija, 2007. Disponível em: https://www.uradni-list.si/glasilo-uradni-list-rs/vsebina?urlid=200790&stevilka=4422. Acesso em: 7 jun. 2023.

NATIONAL ARCHIVES OF AUSTRALIA. **Long-term file formats**. Canberra, c2019. Disponível em: http://www.naa.gov.au/information-management/managing-information-and-records/preserving/long-term-file-formats.aspx#section16. Acesso em: 28 ago. 2019.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. Literature. PubMed Central (PMC). About PMC. **PMC FAQs**. Bethesda, Maryland, Sept. 2019. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/about/faq/. Acesso em: 7 jun. 2023.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. Literature. **PubMed Central (PMC)**. Bethesda, Maryland, [2023]. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/. Acesso em: 7 jun. 2023.

NATIONAL DIET LIBRARY. **Web Archiving Project (WARP)**. [Chiyoda City, Tokyo], [2023?]. Disponível em: https://warp.da.ndl.go.jp/. Acesso em: 11 ago. 2023.

NATIONAL DIET LIBRARY. **Web Archiving Project (WARP)**. About us: let's warp to the past of the web!. Frequently asked questions about warp. [Chiyoda City, Tokyo], c2013. Disponível em: https://warp.da.ndl.go.jp/. Acesso em: 16 nov. 2022.

NATIONAL DIET LIBRARY. **Web Archiving Project (WARP)**. Recommended. Mechanism of web archiving. [Chiyoda City, Tokyo], 2014. Disponível em: https://warp.da.ndl.go.jp/contents/reccommend/mechanism/index_en.html. Acesso em: 7 jun. 2023.

NATIONAL DIGITAL STEWARDSHIP ALLIANCE. **Glossary**. [United States], 2013. Disponível em: https://ndsa.org/glossary/. Acesso em: 7 jun. 2023.

NATIONAL INFORMATION STANDARDS ORGANIZATION. **Understanding metadata**. Bethesda: NISO Press, c2004. 16 p. Disponível em: https://www.lter.uaf.edu/metadata_files/UnderstandingMetadata.pdf. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY BOARD. **Web archive Singapore**. [*S. l.*], c2023. Disponível em: https://eresources.nlb.gov.sg/webarchives/landing-page. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY BOARD. **Web archive Singapore**. Frequently asked questions. [*S. l.*], c2022. Disponível em: https://eresources.nlb.gov.sg/webarchives/landing-page. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF AUSTRALIA. About PANDORA. Selection guidelines. National Library of Australia. **Collection policy for archiving websites**. Canberra, Australia, 2018. Disponível em: https://pandora.nla.gov.au/selectionguidelines.html. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF AUSTRALIA. **Australian web archive**. Archived websites. [Canberra, Australia]: Trove, [2023?a]. Disponível em: https://webarchive.nla.gov.au/collection. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF AUSTRALIA. Collections. Building our collections. **Australian Web Archive**. Canberra, Australia, [2022a]. Disponível em: https://www.nla.gov.au/collections/building-our-collections/australian-web-archive. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF AUSTRALIA. **Digital preservation policy**. 4th ed. [Canberra], Feb. 2013. Disponível em: https://www.nla.gov.au/policy-and-planning/digital-preservation-policy. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF AUSTRALIA. **PANDORA Australia's web archive**. About PANDORA. Selection guidelines. Frequently asked questions about PANDORA. Canberra, Australia, 2020. Disponível em: http://pandora.nla.gov.au/about.html. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF AUSTRALIA. **PANDORA, Australia's web archive**. Canberra, Australia, [2023?b]. Disponível em: http://pandora.nla.gov.au/. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF AUSTRALIA. Policy and planning. **Collection development policy**. What we collect. Canberra, Australia, [2022b]. Disponível em: https://www.nla.gov.au/about-us/corporate-documents/policy-and-planning/collection-development-policy/what-we-collect. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF IRELAND. Collections. Digital. **Web archive**. [Dublin, Ireland], c2023. Disponível em: https://www.nli.ie/en/web_archive.aspx. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF IRELAND. Collections. Digital. **Web archive**. Web archive faq & resources. [Dublin, Ireland], c2022. Disponível em: https://www.nli.ie/en/web_archive.aspx. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF IRELAND**. Web archive, cartlann ghréasáin**. [Dublin, Ireland], [20-?]. 9 p. Disponível em: https://www.nli.ie/getAttachment.aspx?Id=504f8c97-1aec-4370-b7d5-b184f6472e47. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF ISRAEL. Research and teach. **Israeli internet archive**. FAQ about the internet archive. [Jerusalem, Israel], 2022. Disponível em: https://www.nli.org.il/en/research-and-teach/internet-archive. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF ISRAEL. Research and teach. **Israeli internet archive**. [Jerusalem, Israel], 2023. Disponível em: https://www.nli.org.il/en. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF KOREA. **Online Archiving & Searching Internet Sources (OASIS)**. Oasis Overview. Frequently Asked Questions. [Seoul, South Korea], c2006. Disponível em: https://nl.go.kr/oasis/. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF LUXEMBOURG. **Luxembourg web archive**. How it works. Restrictions. [Kirchberg, Luxembourg], [2022]. Disponível em: https://www.webarchive.lu/how-it-works/. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF MEDICINE. **NLM web collecting and archiving**. [Bethesda, Maryland], 2023. Disponível em: https://www.nlm.nih.gov/webcollecting/index.html. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF MEDICINE. **NLM web collecting and archiving**. NLM web collecting and archiving faqs. [Bethesda, Maryland], 2019. Disponível em: https://www.nlm.nih.gov/webcollecting/index.html. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF NEW ZEALAND. Collections. A-Z of our collections. **New zealand web archive**. [Wellington, New Zealand], [2022?a]. Disponível em: https://natlib.govt.nz/collections/a-z/new-zealand-web-archive. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF NEW ZEALAND. Collections. A-Z of our collections. **Alexander Turnbull Library Collections**. [Wellington, New Zealand], [2023a]. Disponível em: https://natlib.govt.nz/collections/a-z/new-zealand-web-archive. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF NEW ZEALAND. **Metadata standards framework**: preservation metadata (revised). Wellington, New Zealand: National Library of New Zealand, June 2003. 50 p. Disponível em: https://natlib.govt.nz/files/digital-preservation/metaschema-revised.pdf. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF NEW ZEALAND. **New zealand web archive**. [Wellington, New Zealand], [2023?b]. Disponível em: https://natlib.govt.nz/collections/a-z/new-zealand-web-archive. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF NEW ZEALAND. Our services for publishers and authors. **Web harvesting**. General information for Publishers. New Zealand Web Archive. Whole of domain web harvest. [Wellington, New Zealand], [2022?b]. Disponível em: https://natlib.govt.nz/publishers-and-authors/web-harvesting. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF NEW ZEALAND. Our services for publishers and authors. **Legal deposit.** What's legal deposit?. [Wellington, New Zealand], [2022?c]. Disponível em: https://natlib.govt.nz/publishers-and-authors/legal-deposit/whats-legal-deposit. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF SCOTLAND. Guides. Publishers. **Web harvesting**. Frequently asked questions for webmasters. [Edinburgh, United Kingdom], c2022. Disponível em: https://www.nls.uk/guides/publishers/web-harvesting/. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF THE NETHERLANDS; NATIONAL LIBRARY OF NEW ZEALAND. **Web Curator Tool**: open-source workflow management for selective web archiving. WTC. Features. WCT features. [*S. l.*], 2020. Disponível em: https://webcuratortool.org/features. Acesso em: 7 jun. 2023.

NATIONAL LIBRARY OF THE NETHERLANDS; NATIONAL LIBRARY OF NEW ZEALAND. **Web Curator Tool**: open-source workflow management for selective web archiving. [*S. l.*], c2021. Disponível em: https://webcuratortool.org/. Acesso em: 7 jun. 2023.

NATIONAL MUSEUM OF AUSTRALIA. **Digital preservation and digitisation policy**. Version 2.2. Canberra, Aug. 2012. 11 p. Disponível em: https://www.nma.gov.au/about/corporate/plans-policies/policies/digital-preservation-and-digitisation. Acesso em: 15 dez. 2021.

NELSON, Michael L. A plan for curating "obsolete data or resources". Position paper for the UNC/NSFWorkshop "CuratingforQuality: Ensuring Data Quality to Enable New Science", September 10-11, 2012. Disponível em: https://arxiv.org/pdf/1209.2664. Acesso em: 7 jun. 2023.

NETWERK DIGITAAL ERFGOED. Nationaal Register Webarchieven. **Frequently asked questions**. [*S. l.*], [2021?]. Disponível em: https://www.registerwebarchieven.nl/faq. Acesso em: 7 jun. 2023.

NEW YORK ART RESOURCES CONSORTIUM. NYARC discovery. **Web archiving**. Frequently asked questions. [*S. l.*], [2022]. Disponível em: https://nyarc.org/initiatives/web-archiving. Acesso em: 7 jun. 2023.

NEW YORK ART RESOURCES CONSORTIUM. NYARC discovery. **Web archiving**. [*S. l.*], [2023?]. Disponível em: https://nyarc.org/initiatives/web-archiving. Acesso em: 7 jun. 2023.

NIU, Jinfang. An overview of web archiving. **D-Lib Magazine**, v. 18, n. 3/4, March/April 2012a. Disponível em: https://www.dlib.org/dlib/march12/niu/03niu1.html. Acesso em: 7 jun. 2023.

NIU, Jinfang. Functionalities of web archives. **D-Lib Magazine**, v. 18, n. 3/4, March/April 2012b. Disponível em: http://www.dlib.org/dlib/march12/niu/03niu2.html#:~:text=Abstract,data%20mining%2C%20and%20website%20reconstruction. Acesso em: 7 jun. 2023.

NORTH CAROLINA DEPARTMENT OF NATURAL AND CULTURAL RESOURCES. Resources. Records management. TOMES project. Online training. TOMES software. **TOMES software overview**. North Carolina, United States, [2018]. Disponível em: https://files.nc.gov/ncdcr/TOMES/20181127_TOMESsoftwareoverview.pdf. Acesso em: 7 jun. 2023.

NORTH CAROLINA DEPARTMENT OF NATURAL AND CULTURAL RESOURCES. Resources. Records management. **TOMES project**. North Carolina, United States, [2023?]. Disponível em: https://www.dncr.nc.gov/things-know/records-management/transforming-online-mail-embedded-semantics-tomes. Acesso em: 7 jun. 2023.

OFFICE OF THE NATIONAL COORDINATOR FOR HEALTH INFORMATION TECHNOLOGY. ISA Content. Vocabulary/Code Set/Terminology. COVID-19 Novel Coronavirus Pandemic. **COVID-19 Novel Coronavirus Pandemic**. United States, Sept. 2021a. Disponível em: https://www.healthit.gov/isa/covid-19. Acesso em: 7 jun. 2023.

OFFICE OF THE NATIONAL COORDINATOR FOR HEALTH INFORMATION TECHNOLOGY. **About the ISA**. Structure. Timeline and Comment Process. FAQs. United States, [2021b]. Disponível em https://www.healthit.gov/isa/about-isa. Acesso em: 7 jun. 2023.

OFFICE OF THE NATIONAL COORDINATOR FOR HEALTH INFORMATION TECHNOLOGY. **Interoperability Standards Advisory (ISA)**. United States, [2023?]. Disponível em https://www.healthit.gov/isa/. Acesso em: 7 jun. 2023.

OHIO Library and Information Network (OhioLINK). Columbus, Ohio, c2023. Disponível em: https://www.ohiolink.edu/. Acesso em: 7 jun. 2023.

OHIO LIBRARY AND INFORMATION NETWORK (OhioLINK). **OhioLINK electronic journal center**. Columbus, Ohio, c2022. Disponível em: https://www.ohiolink.edu/content/ohiolink_electronic_journal_center. Acesso em: 7 jun. 2023.

OLIVEIRA, Denize Cristina de., Análise de conteúdo temático-categorial: uma proposta de sistematização. **Rev. Enferm. UERJ**, Rio de Janeiro, v. 16, n. 4, p. 569-576, 2008. Disponível em: http://files.bvs.br/upload/S/0104-3552/2008/v16n4/a569-576.pdf. Acesso em: 7 maio 2023.

ONTARIO COUNCIL OF UNIVERSITY LIBRARIES. Services. Scholars portal. **Journals**. [Ontario, Canada], [2022]. Disponível em: https://scholarsportal.info/. Acesso em: 7 jun. 2023.

ONTARIO COUNCIL OF UNIVERSITY LIBRARIES. Services. **Scholars portal**. [Ontario, Canada], [2023?]. Disponível em: https://scholarsportal.info/. Acesso em: 7 jun. 2023.

ORSZÁGOS SZÉCHÉNYI KÖNYVTÁR. **OSZK Webarchívum**. Budapest, Magyarország, c2023. Disponível em: https://webarchivum.oszk.hu/. Acesso em: 7 jun. 2023.

ORSZÁGOS SZÉCHÉNYI KÖNYVTÁR. **OSZK Webarchívum**. Felhasználóknak. Tájékoztató a honlapról. Gyakran Ismételt Kérdések. Tartalomgazdáknak. Technikai tudnivalók az archiválásról. Ajánlások robot- és archívumbarát webhelyekhez. Újságíróknak. Alapinformációk és -adatok. Budapest, Magyarország, c2022. Disponível em: https://webarchivum.oszk.hu/. Acesso em: 7 jun. 2023.

ÖSTERREICHISCHEN NATIONALBIBLIOTHEK. **Webarchiv österreich**. [Wien, Republik Österreich], c2023. Disponível em: https://webarchiv.onb.ac.at/#. Acesso em: 7 jun. 2023.

ÖSTERREICHISCHEN NATIONALBIBLIOTHEK. **Webarchiv österreich**. FAQ. [Wien, Republik Österreich], [2022]. Disponível em: https://webarchiv.onb.ac.at/#. Acesso em: 7 jun. 2023.

OURY, Clement; POLL, Roswitha. Counting the uncountable: statistics for web archives. **Performance Measurement and Metrics**, v. 14, n. 2, p. 132-141, 2013. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/PMM-05-2013-0014/full/html. Acesso em: 7 jun. 2023.

PALA, Francesca. Lo standard EAD3 per la codifica dei dati archivistici: qualche novità e molte conferme. **JLIS.it**, Macerata, v. 8, n. 3, p. 148-176, Sept. 2017. Disponível: https://www.jlis.it/article/view/12407/11294. Acesso em: 1 maio 2020.

PAN AMERICAN HEALTH ORGANIZATION (PAHO). **Institutional Repository for Information Sharing (IRIS)**. Washington, D.C., [2023?]. Disponível em: https://iris.paho.org/. Acesso em: 14 ago. 2023.

PAN AMERICAN HEALTH ORGANIZATION (PAHO). PAHO's institutional repository. **Repositorio institucional de la OPS**: política general. Washington, D.C., United States, [2021]. Disponível em: https://www3.paho.org/hq/index.php?option=com_content&view=article&id=14914:repositorio-institucional-de-la-ops-politica-general&Itemid=72446&lang=en. Acesso em: 7 jun. 2023.

PEARSON, David; DEL POZO, Nick. **Explaining pres actions**: a working document. [Canberra]: National Library of Australia, Nov. 2009. 42 p. Disponível em: https://www.nla.gov.au/content/explaining-pres-actions-a-working-document. Acesso em: 7 out. 2019.

PENDSE, Liladhar R. Archiving the russian and east european lesbian, gay, bisexual, and transgender web, 2013: a pilot project. **Slavic & East European Information Resources**, v. 15, n. 3, p. 182-196, 2014. Disponível em: https://www.tandfonline.com/doi/full/10.1080/15228886.2014.930973. Acesso em: 7 jun. 2023.

PENNOCK, Maureen. Web-Archiving. **DPC Technology Watch Report 13-01**, p. 1-45, Mar. c2013. Disponível em: https://www.dpconline.org/docs/technology-watch-reports/865-dpctw13-01-pdf/file. Acesso em: 7 jun. 2023.

PINHEIRO, Lena Vania Ribeiro. Constituição epistemológica e social da comunicação científica no Brasil. *In*: PINHEIRO, Lena Vania Ribeiro.; OLIVEIRA, Eloisa da Conceição Príncipe de. (Orgs.). **Múltiplas facetas da comunicação e divulgação científicas**: transformações em cinco séculos. Brasília, DF: Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), 2012. p. 115-148. Disponível em: http://livroaberto.ibict.br/bitstream/1/711/1/M%c3%baltiplas%20facetas%20da%20comunica%c3%a7%c3%a3o%20e%20divulga%c3%a7%c3%a3o%20cient%c3%adficas.pdf. Acesso em: 7 jun. 2023.

PINHEIRO, Lena Vania Ribeiro; FERREZ, Helena Dodd. **Tesauro brasileiro de ciência da informação**. Rio de Janeiro; Brasília, DF: Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), 2014. 384 p. Disponível em: http://sitehistorico.ibict.br/publicacoes-e-institucionais/tesauro-brasileiro-de-ciencia-da-informacao-1/copy_of_TESAUROCOMPLETOFINALCOMCAPA24102014.pdf. Acesso em: 7 jun. 2023.

PINHEIRO, Nilcéia Aparecida Maciel; SILVEIRA, Rosemari Monteiro Castilho Foggiatto; BAZZO, Walter Antonio. Ciência, tecnologia e sociedade: a relevância do enfoque CTS para o contexto do ensino médio. **Ciência & Educação**, Bauru, SP, v. 13, n. 1, p. 71-84, 2007. Disponível em: https://www.scielo.br/j/ciedu/a/S97k6qQ6QxbyfyGZ5KysNqs/?format=pdf&lang=pt. Acesso em: 7 jun. 2023.

POST, Colin. Building a living, breathing archive: a review of appraisal theories and approaches for web archives. **PDT&C**, v. 46, n. 2, p. 69-77, 2017. Disponível em: https://webcache.googleusercontent.com/search?q=cache:QjwIkhA_sDMJ:https://libres.uncg.edu/ir/uncg/f/C_Post_Building_2017.pdf&cd=1&hl=pt-BR&ct=clnk&gl=br. Acesso em: 7 jun. 2023.

PREMIS EDITORIAL COMMITTEE. **PREMIS data dictionary for preservation metadata.** Version 3.0. [*S. l.: s. n.*], Nov. 2015. 273 p. Disponível em: https://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf. Acesso em: 7 jun. 2023.

PRICE, Derek John de Solla. **Little science, big science**. New York, United States: Columbia University Press, 1986.

PROM, Christopher J. Preserving email. 2nd edition. **DPC Technology Watch Report**, v. 19, p. 1-51, May. c2019. Disponível em: https://www.dpconline.org/docs/technology-watch-reports/2159-twr19-01/file. Acesso em: 7 jun. 2023.

PROM, Christopher J. Preserving email. **DPC Technology Watch Report**, v. 11, p. 1-51, Dec. c2011. Disponível em: https://www.dpconline.org/docs/technology-watch-reports/739-dpctw11-01-pdf/file. Acesso em: 7 jun. 2023.

RADINSKY, Kira; HORVITZ, Eric. Mining the web to predict future events. *In*: ASSOCIATION FOR COMPUTING MACHINERY (ACM) INTERNATIONAL CONFERENCE ON WEB SEARCH AND DATA MINING, 6., February 2013, Rome, Italy. **Proceedings** […]. New York, United States: ACM, 2013. p. 255-264. Disponível em: http://erichorvitz.com/future_news_wsdm.pdf. Acesso em: 7 jun. 2023.

REIS, Luiz Claudio Rezende; SÁ, Maria Irene da Fonseca e. Big data: um novo campo de atuação para bibliotecários. **Prisma.com**, Portugal, n. 41, p. 231-250, 2020. Disponível em: https://ojs.letras.up.pt/index.php/prismacom/article/view/6752/6243. Acesso em: 7 jun. 2023.

REYNOLDS, Emily. **Web archiving use cases**. Washington, DC: Library of Congress, UMSI, ASB13, March, 2013. Disponível em: https://netpreserve.org/resources/IIPC_archive-UseCases_Final.pdf. Acesso em: 7 jun. 2023.

REZENDE, Laura Vilela Rodrigues; MARTINS, Dalton Lopes. Experiências e desafios para a preservação digital de mídias sociais. **Investigación Bibliotecológica: archivonomía, bibliotecología e información**, v. 33, n. 80, p. 31-56, 2019. Disponível em: http://www.scielo.org.mx/pdf/ib/v33n80/2448-8321-ib-33-80-31.pdf. Acesso em: 7 jun. 2023.

REZENDE, Laura Vilela Rodrigues; MARTINS, Dalton Lopes. Iniciativas científicas de arquivamento e preservação de conteúdos em mídias sociais: panorama atual. **RICI: R.Ibero-amer. Ci. Inf.**, Brasília, DF, v. 11, n. 1, p. 219-236, jan./abr. 2018. Disponível em: https://periodicos.unb.br/index.php/RICI/article/view/8538/7115. Acesso em: 7 jun. 2023.

RHIZOME.ORG. **Conifer**. About. [New York, United States], [2022]. Disponível em: https://conifer.rhizome.org/_faq. Acesso em: 7 jun. 2023.

RHIZOME.ORG. **Conifer**. [New York, United States], [2023?]. Disponível em: https://conifer.rhizome.org/. Acesso em: 7 jun. 2023.

RILEY, Jenn. **Understanding metadata**: what is metadata, and what is it for? Baltimore, Maryland: National Information Standards Organization (NISO), c2017. 45 p. Disponível em: https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf. Acesso em: 9 ago. 2020.

ROCHE, Xavier. **HTTrack**: website copier. About. [*S. l.*], c2022. Disponível em: https://www.httrack.com/. Acesso em: 7 jun. 2023.

ROCKEMBACH, Moisés. Arquivamento da web no contexto das humanidades digitais: da produção a preservação da informação digital. **Liinc em Revista**, Rio de Janeiro, v. 15, n. 1, p. 131-139, maio 2019. Disponível em: http://revista.ibict.br/liinc/article/view/4578/4142. Acesso em: 7 jun. 2023.

ROCKEMBACH, Moisés. Arquivamento da web: estudos de caso internacionais e o caso brasileiro. **RDBCI**: Rev. Digit. Bibliotecon. Cienc. Inf., Campinas, SP, v. 16, n. 1, p. 07-24, jan./abr. 2018. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8648747/pdf. Acesso em: 21 jun. 2022.

ROCKEMBACH, Moisés; PAVÃO, Caterina Marta Groposo. Políticas e tecnologias de preservação digital no arquivamento da Web. **RICI**: R.Ibero-amer. Ci. Inf., Brasília, DF, v. 11, n. 1, p. 168-182, jan./abr. 2018. Disponível em: https://lume.ufrgs.br/bitstream/handle/10183/175153/001066630.pdf?sequence=1&isAllowed=y. Acesso em: 7 jun. 2023.

ROCKEMBACH, Moisés; SERRANO, Anabela. Climate change and web archives: an ibero-american study based on the portuguese and brazilian contexts**. Records Management Journal**, v. 31, n. 3, p. 222-239, 2021. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/RMJ-11-2020-0039/full/html. Acesso em: 7 jun. 2023.

RODRIGUES, Vander Luis Duarte; ROCKEMBACH, Moisés. Arquivos da web como fonte historiográfica. **RDBCI**: Rev. Dig. Bibliotec e Ci. Info., Campinas, SP, v. 19, 2021. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8663680/26626. Acesso em: 7 jun. 2023.

ROLLASON-CASS, SYLVIE; REED, Scott. Living movements, living archives: selecting and archiving web content during times of social unrest. **New Review of Information Networking**, v. 20, p. 241–247, 2015. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/13614576.2015.1114839. Acesso em: 7 jun. 2023.

ROTHENBERG, Jeff. **Avoiding technological quicksand**: finding a viable technical foundation for digital preservation. A Report to the Council on Library and Information Resources. Washington, DC: Commission on Preservation and Access and Council on Library and Information Resources, Jan. 1999. 35 p. Disponível em: https://clir.wordpress.clir.org/wp-content/uploads/sites/6/pub77.pdf. Acesso em: 15 dez. 2021.

ROTHER, Edna Terezinha. Revisão sistemática x revisão narrativa. **Acta Paul Enferm.**, São Paulo, SP, v. 20, n. 2, p. V-VI, 2007. Disponível em: https://www.scielo.br/j/ape/a/z7zZ4Z4GwYV6FR7S9FHTByr/. Acesso em: 9 maio 2023.

ROWELL, Chelcie Juliet; KREWER, Drew. Preservation metadata for complex digital objects. A Report of the ALCTS PARS Preservation Metadata Interest Group Meeting. American Library Association Annual Conference, San Francisco, June 2015. **Technical Services Quarterly**, v. 33, n. 2, p. 179-183, Mar. 2016. Disponível em: https://www.tandfonline.com/doi/full/10.1080/07317131.2016.1135003. Acesso em: 7 jun. 2023.

RUEST, Nick; MILLIGAN, Ian. An open-source strategy for documenting events: the case study of the 42nd canadian federal election on Twitter. **Code4Lib Journal**, n. 32, Apr. 2016. Disponível em: https://journal.code4lib.org/articles/11358. Acesso em: 7 jun. 2023.

SADAT-MOOSAVI, Ali; ISFANDYARI-MOGHADDAM, Alireza; TAJEDDINI, Oranus. Accessibility of online resources cited in scholarly lis journals: a study of Emerald isi-ranked journals. **Aslib Proceedings**: New Information Perspectives, [Bingley], v. 64, n. 2, p. 178-192, 2012. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/00012531211215196/full/html. Acesso em: 7 jun. 2023.

SAMOUELIAN, Mary; DOOLEY, Jackie. **Descriptive metadata for web archiving**: review of harvesting tools. Dublin, Ohio: Online Computer Library Center (OCLC) Research, Feb. c2018. 23 p. Disponível em: https://www.oclc.org/content/dam/research/publications/2018/oclcresearch-wam-harvesting-tools.pdf. Acesso em: 7 jun. 2023.

SANTOS, Gildenir Carolino; PASSOS, Rosemary; SAE, Marcos Dario. A preservação digital dos periódicos científicos produzidos na Unicamp: um relato de experiência. **Ci. Inf.**, Brasília, DF, v. 41, n. 1, p. 150-159, jan./abr., 2012. Disponível em: http://revista.ibict.br/ciinf/article/view/1361/1540. Acesso em: 7 jun. 2023.

SANTOS, Henrique Machado dos; FLORES, Daniel. Os impactos da obsolescência tecnológica frente à preservação de documentos digitais. **Brazilian Journal of Information Science: research trends**, v. 11, n. 2, p. 28-37, jun. 2017a. Disponível em: https://revistas.marilia.unesp.br/index.php/bjis/article/view/5550/4511. Acesso em: 15 dez. 2021.

SANTOS, Henrique Machado dos; FLORES, Daniel. Preservação de documentos arquivísticos digitais: reflexões sobre as estratégias de emulação. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 20, n. 43, p. 3- 19, maio/ago. 2015. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2015v20n43p3/30007. Acesso em: 7 jun. 2023.

SANTOS, Henrique Machado dos; FLORES, Daniel. Preservação de documentos arquivísticos digitais: reflexões sobre as estratégias de migração. **PRISMA.COM**, Porto, n. 37, p. 42-54, 2018. Disponível em: https://ojs.letras.up.pt/index.php/prismacom/article/view/4707/4395. Acesso em: 7 jun. 2023.

SANTOS, Henrique Machado dos; FLORES, Daniel. Preservação de documentos digitais: reflexões sobre as estratégias de refrescamento. **Revista Brasileira de Biblioteconomia e Documentação**, v. 13, n. 2, p. 31-41, jul./dez. 2017b. Disponível em: https://rbbd.febab.org.br/rbbd/article/view/449/641. Acesso em: 7 jun. 2023.

SANTOS, Lucy Woellner dos; ICHIKAWA, Elisa Yoshie. CTS e a participação pública na ciência. *In*: SANTOS, Lucy Woellner dos. *et al*. (org.). **Ciência, tecnologia e sociedade**: o desafio da interação. Londrina, PR: IAPAR, 2002. p. 239-271.

SANTOS, Thayse Natália Cantanhede. Curadoria digital e preservação digital: cruzamentos conceituais. **Rev. Digit. Bibliotecon. Cienc. Inf.**, Campinas, SP, v. 14, n.3, p. 450-464, set./dez. 2016. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/rdbci/article/view/8646336/pdf. Acesso em: 7 jun. 2023.

SANTOS, Vanderlei Batista dos. Arquivamento web: legislação correlata. **Rev. Bras. Presev. Digit. / Braz. J. Preserv. Digit.**, Campinas, SP, v. 1, p. 1-11, 2020. Disponível em: https://econtents.bc.unicamp.br/inpec/index.php/rebpred/article/view/14800/9790. Acesso em: 7 jun. 2023.

SARACEVIC, Tefko. Ciência da informação: origem, evolução e relações. **Perspec. Ci. Inf.**, Belo Horizonte, MG, v. 1, n. 1, p. 41-62, jan./jun. 1996. Disponível em: https://periodicos.ufmg.br/index.php/pci/article/view/22308/17916. Acesso em: 9 maio 2023.

SAYÃO, Luís Fernando. Uma outra face dos metadados: informações para a gestão da preservação digital. **Enc. Bibli**: R. Eletr. Bibliotecon. Ci. Inf., Florianópolis, v. 15, n. 30, p. 1-31, 2010. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2010v15n30p1/19527. Acesso em: 7 jun. 2023.

SAYÃO, Luis Fernando; SALES, Luana Farias. Curadoria digital: um novo patamar para preservação de dados digitais de pesquisa. **Inf. & Soc.:**Est., João Pessoa, v. 22, n. 3, p. 179-191, set./dez. 2012. Disponível em: https://www.icict.fiocruz.br/sites/www.icict.fiocruz.br/files/Curadoria%20digital_Luis%20Fernando%20Sayao.pdf. Acesso em: 16 ago. 2023.

SCHÄFER, Murilo Billig; CONSTANTE, Sônia Elisabete. A preservação da informação digital nos arquivos das IFES da Região Sul do Brasil. **RICI**: R. Ibero-amer. Ci. Inf., Brasília, DF, v. 6, n. 1, p. 44-67, jan./jul. 2013. Disponível em: https://periodicos.unb.br/index.php/RICI/article/view/1776/1566. Acesso em: 7 jun. 2023.

SCHÄFER, Murilo Billig; CONSTANTE, Sônia Elisabete. Políticas e estratégias para a preservação da informação digital. **Ponto de Acesso**, Salvador, v. 6, n. 3, p. 108-140, dez. 2012. Disponível em: https://periodicos.ufba.br/index.php/revistaici/article/view/6449/4817. Acesso em: 7 jun. 2023.

SCHAFER, Valérie *et al*. Paris and Nice terrorist attacks: exploring twitter and web archives. **Media, War & Conflict**, v. 12, n. 2, p 153-170, 2019. Disponível em: https://journals.sagepub.com/doi/10.1177/1750635219839382. Acesso em: 7 jun. 2023.

SCHNEIDER, J. et al. Appraising, processing, and providing access to email in contemporary literary archives. **Archives and Manuscripts**, v. 47, n. 3, p. 305–326, 2019. Disponível em: https://www-tandfonline.ez31.periodicos.capes.gov.br/doi/full/10.1080/01576895.2019.1622138. Acesso em: 3 out. 2021.

SCHNEIDER, Josh *et al*. ePADD: computational analysis software facilitating screening, browsing, and access for historically and culturally valuable email collections. **D-Lib Magazine**, [S. l.], v. 23, n. 5/6, May/June 2017. Disponível em: http://www.dlib.org/dlib/may17/schneider/05schneider.html. Acesso em: 7 jun. 2023.

SCHWEIZERISCHE NATIONALBIBLIOTHEK. Fachinformationen. e-Helvetica. **Websites**. FAQ zu Webarchivierung. [Bern, Schweiz], [2022]. Disponível em: https://www.nb.admin.ch/snl/de/home/fachinformationen/e-helvetica/webarchiv-schweiz.html. Acesso em: 7 jun. 2023.

SCHWEIZERISCHE NATIONALBIBLIOTHEK. **Webarchiv schweiz**. [Bern, Schweiz], 2023. Disponível em: https://www.e-helvetica.nb.admin.ch/. Acesso em: 11 ago. 2023.

SENANDER III, Mathew. Converting vra core records to marc records: a study in crosswalking. **Library Philosophy and Practice**, Lincoln, Dec. 2013. Disponível em: https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=2601&context=libphilprac. Acesso em: 7 jun. 2023.

SERRANO, Jode. How to stop ukrainian websites from vanishing during war. **Gizmodo**, New York, 4 February 2022. Disponível em: https://gizmodo.com/how-sucho-stops-ukrainian-websites-vanishing-in-russias-1848737441. Acesso em: 7 jun. 2023.

SETON HALL UNIVERSITY LIBRARIES. Web archiving policy. **SHU libraries university web archive collecting policy**. South Orange, United States, [2022?]. Disponível em: https://library.shu.edu/library/WebArchCollPolicy. Acesso em: 7 jun. 2023.

SEVERINO, Antônio Joaquim. **Metodologia do trabalho científico**. 24. ed. rev. e atual. São Paulo: Cortez, 2016. 320 p.

SHIOZAKI, Ryo; EISENSCHITZ, Tamara. Role and justification of web archiving by national libraries: a questionnaire survey. **Journal of Librarianship and Information Science**, v. 41, n. 2, p. 90-107, June 2009. Disponível em: https://journals.sagepub.com/doi/pdf/10.1177/0961000609102831. Acesso em: 7 jun. 2023.

SIEBRA, Sandra de Albuquerque *et al*. Curadoria digital: além da questão da preservação digital. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB, 14., 2013, Londrina, PR. **Anais...** Florianópolis, SC: UFSC, 2013. p. 1-20. Disponível em: https://brapci.inf.br/index.php/res/download/185251. Acesso em: 7 jun. 2023.

SIFE, Alfred S.; BERNARD, Ronald. Persistence and decay of web citations used in theses and dissertations available at the Sokoine National Agricultural Library, Tanzania. **International Journal of Education and Development using Information and Communication Technology (IJEDICT)**, Bridgetown, v. 9, n. 2, p. 85-94, 2013. Disponível em: https://files.eric.ed.gov/fulltext/EJ1071354.pdf. Acesso em: 7 jun. 2023.

SIFE, Alfred Said; LWOGA, Edda Tandi. Retrieving vanished web references in health science journals in East Africa. **Information and Learning Science**, [*S. l.*], v. 118, n. 7/8, p. 385-392, 2017. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/ILS-04-2017-0030/full/html. Acesso em: 7 jun. 2023.

SIGNORI, Barbara; BÄTTIG, Yvonne. **Webarchiv Schweiz**: merkblatt erschliessen. Version 2.1. Zürich, Schweiz: Schweizerischen Nationalbibliothek, Nov. 2017. 9 p. Disponível em: https://www.nb.admin.ch/dam/snl/de/dokumente/e-helvetica/normen_und_regelwerke/merkblatt-erschliessen.pdf.download.pdf/merkblatt-erschliessen.pdf. Acesso em: 7 jun. 2023.

SILVA JUNIOR, Laerte Pereira da; MOTA, Valéria Gameleira da. Políticas de preservação digital no Brasil: características e implementações. **Ci. Inf.**, Brasília, DF, v. 41, n. 1, p.51-64, jan./abr. 2012. Disponível em: http://revista.ibict.br/ciinf/article/view/1351/1530. Acesso em: 7 jun. 2023.

SILVA, Edna Lúcia da; MENEZES, Estera Muszkat. **Metodologia da pesquisa e elaboração de dissertação**. 4. ed. rev. e atual. Florianópolis: Universidade Federal de Santa Catarina (UFSC), 2005. 139 p. Disponível em: https://www.researchgate.net/publication/312125489_Metodologia_da_Pesquisa_e_Elaboracao_de_Dissertacao. Acesso em: 7 jun. 2023.

SILVA, Elaine da; VALENTIM, Marta Lígia Pomim. Avaliação da aplicação do método 'análise de conteúdo' em pesquisa sobre processos de gestão da informação e do conhecimento como subsídios para a geração de inovação. **Inf. Inf.**, Londrina, v. 24, n. 1, p. 326-355, jan./abr. 2019. Disponível em: https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/31957/pdf. Acesso em: 7 maio 2023.

SILVA, Victória de Abreu; SILVA, Márcio Bezerra da. Metadados para preservação digital de dados abertos: um estudo de identificação. **Biblios**, Pittsburgh, n. 78, p. 44-60, jul. 2020. Disponível em: http://biblios.pitt.edu/ojs/index.php/biblios/article/view/793/383. Acesso em: 7 jun. 2023.

SIMPSON, Joel. Email archiving systems interoperability. **Harvard Library Report**. [Cambridge, Massachusetts], July 2016. Disponível em: https://dash.harvard.edu/bitstream/handle/1/28682572/HL_Email_Archiving_Systems_Interoperability_Report_2016.pdf?sequence=3&isAllowed=y. Acesso em: 7 jun. 2023.

SISMONDO, Sergio. **An introduction to science and technology studies**. 2th ed. Malden, Massachusetts: Wiley-Blackwell, c2010.

SITE MAP. *In*: WIKIPEDIA: the free encyclopedia. [San Francisco, CA: Wikimedia Foundation], 2022. Disponível em: https://en.wikipedia.org/wiki/Site_map. Acesso em: 7 jun. 2023.

SKINNER, Katherine; SCHULTZ, Matt. **A guide to distributed digital preservation**. Atlanta, Georgia: Educopia Institute, c2010. 137 p. Disponível em: https://metaarchive.org/wp-content/uploads/2017/03/A_Guide_to_Distributed_Digital_Preservation_0.pdf. Acesso em: 7 jun. 2023.

SMITHSONIAN INSTITUTION ARCHIVES. **DArcMail**. [*S. l.*]: GitHub, 2021. Disponível em: https://github.com/Smithsonian/DArcMail. Acesso em: 12 ago. 2023.

SMITHSONIAN INSTITUTION ARCHIVES. What we do. Digital curation. Project highlights. **Email preservation**. Washington, DC, [2022?]. Disponível em: https://siarchives.si.edu/what-we-do/digital-curation/email-preservation-darcmail. Acesso em: 7 jun. 2023.

SMITHSONIAN INSTITUTION ARCHIVES; DOLLAR CONSULTING. **Archival preservation of smithsonian web resources**: strategies, principles, and best practices. [Washington, DC], July 2001. 51 p. Disponível em: https://siarchives.si.edu/sites/default/files/pdfs/dollar_report.pdf. Acesso em: 7 jun. 2023.

SOCIETY OF AMERICAN ARCHIVISTS. **Dictionary of Archives Terminology**. Checksum. [Chicago], c2022a. Disponível em: https://dictionary.archivists.org/entry/checksum.html. Acesso em: 7 jun. 2023.

SOCIETY OF AMERICAN ARCHIVISTS. **Dictionary of Archives Terminology**. Content standard. [Chicago], c2022b. Disponível em: https://dictionary.archivists.org/entry/content-standard.html. Acesso em: 7 jun. 2023.

SOCIETY OF AMERICAN ARCHIVISTS. **Dictionary of Archives Terminology**. Controlled vocabulary. [Chicago], c2022c. Disponível em: https://dictionary.archivists.org/entry/controlled-vocabulary.html. Acesso em: 7 jun. 2023.

SOCIETY OF AMERICAN ARCHIVISTS. **Dictionary of archives terminology**. Creative commons. [Chicago], c2022d. Disponível em: https://dictionary.archivists.org/entry/creative-commons.html. Acesso em: 7 jun. 2023.

SOCIETY OF AMERICAN ARCHIVISTS. **Dictionary of Archives Terminology**. Dark archives. [Chicago], c2022e. Disponível em: https://dictionary.archivists.org/entry/dark-archives.html. Acesso em: 7 jun. 2023.

SOCIETY OF AMERICAN ARCHIVISTS. **Dictionary of archives terminology**. Data mining. [Chicago], c2022f. Disponível em: https://dictionary.archivists.org/entry/data-mining.html. Acesso em: 7 jun. 2023.

SOCIETY OF AMERICAN ARCHIVISTS. **Dictionary of Archives Terminology**. DSpace. [Chicago], c2022g. Disponível em: https://dictionary.archivists.org/entry/dspace.html. Acesso em: 7 jun. 2023.

SOCIETY OF AMERICAN ARCHIVISTS. **Dictionary of Archives Terminology**. Internet. [Chicago], c2022h. Disponível em: https://dictionary.archivists.org/entry/internet.html. Acesso em: 7 jun. 2023.

SOCIETY OF AMERICAN ARCHIVISTS. **Dictionary of Archives Terminology**. Open access. [Chicago], c2022i. Disponível em: https://dictionary.archivists.org/entry/open-access.html. Acesso em: 7 jun. 2023.

SOCIETY OF AMERICAN ARCHIVISTS. **Dictionary of Archives Terminology**. Open Archives Initiative. [Chicago], c2022j. Disponível em: https://dictionary.archivists.org/entry/open-archives-initiative.html. Acesso em: 7 jun. 2023.

SOCIETY OF AMERICAN ARCHIVISTS. **Dictionary of archives terminology**. Web archives. [Chicago], c2022k. Disponível em: https://dictionary.archivists.org/entry/web-archives.html. Acesso em: 7 jun. 2023.

SOCIETY OF AMERICAN ARCHIVISTS. Technical Subcommittee for Encoded Archival Standards. **Encoded Archival Description Tag Library**: version EAD3 1.1.1. Chicago, Dec. 2019. 422 p. Disponível em: https://www.loc.gov/ead/EAD3taglib/tl_ead3.pdf. Acesso em: 7 jun. 2023.

SØRENSEN, Mikis Seth; HAVE, Ulrich Karstoft. **NetarchiveSuite**. [*S. l.*], 2020. Disponível em: https://sbforge.org/display/NAS/NetarchiveSuite. Acesso em: 6 jan. 2022.

SOUZA, Arthur Heleno Lima Rodrigues de *et al.* O modelo de referência OAIS e a preservação digital distribuída. **Ciência da Informação**, Brasília, DF, v. 41, n. 1, p. 65-73, jan./abr. 2012. Disponível em: https://revista.ibict.br/ciinf/article/view/1352/1531. Acesso em: 7 jun. 2023.

STAATSBIBLIOTHEK ZU BERLIN. Welcome to the EAC-CPF homepage. **About**. [Berlin, Germany], Oct. 2017. Disponível em: https://eac.staatsbibliothek-berlin.de/. Acesso em: 7 jun. 2023.

STANFORD LIBRARIES. Libraries. Special Collections. **Managing email**. Stanford, California: Stanford University, [c2022a]. Disponível em: https://library.stanford.edu/spc/university-archives/managing-university-records/managing-email. Acesso em: 7 jun. 2023.

STANFORD LIBRARIES. Projects. **Web archiving**. Collection development. Frequently asked questions. Archivability. Policy. stanford.edu archiving. Use cases. Stanford, California: Stanford University, [c2022?b]. Disponível em: https://library.stanford.edu/projects/web-archiving. Acesso em: 7 jun. 2023.

STANFORD LIBRARIES. Projects. **Web archiving**. Stanford, California: Stanford University, [2023?a]. Disponível em: https://library.stanford.edu/projects/web-archiving. Acesso em: 7 jun. 2023.

STANFORD LIBRARIES. Research support. Projects & innovations. **ePADD**. Stanford, California: Stanford University, [c2022c]. Disponível em: https://library.stanford.edu/projects/epadd. Acesso em: 7 jun. 2023.

STANFORD UNIVERSITY. **Collections**. [*S. l.*], [2023?a]. Disponível em: https://epadd.stanford.edu/epadd/collections. Acesso em: 7 jun. 2023.

STANFORD UNIVERSITY. Research support. Projects & innovations. LOCKSS. **About**. Why LOCKSS? What is LOCKSS? Preservation principles. Frequently asked questions. Stanford, California, [2021]. Disponível em: https://www.lockss.org/about. Acesso em: 7 jun. 2023.

STANFORD UNIVERSITY. Research support. Projects & innovations. **LOCKSS**. Stanford, California, [2023?b]. Disponível em: https://www.lockss.org/. Acesso em: 7 jun. 2023.

STANFORD UNIVERSITY. Stanford Mobile and Social Computing Research Group. MobiSocial Computing Laboratory. **Muse:** revive precious memories using email. Tip Sheet. IRE-2012 tip sheet for journalists. Muse: A tool for working with email archives. [*S. l.*], [2012?]. Disponível em: https://mobisocial.stanford.edu/muse/tipsheet.html. Acesso em: 7 jun. 2023.

STANFORD UNIVERSITY. Stanford Mobile and Social Computing Research Group. MobiSocial Computing Laboratory. **Muse:** revive precious memories using email. [*S. l.*], [2023?c]. Disponível em: https://mobisocial.stanford.edu/muse/. Acesso em: 7 jun. 2023.

STANFORD UNIVERSITY. **Stanford Web Archive Portal (SWAP)**. [*S. l.*], [2023?d]. Disponível em: https://swap.stanford.edu/was/. Acesso em: 7 jun. 2023.

STATE LIBRARY OF NEW SOUTH WALES. Public Library Services. **What Is Digital Preservation**? Sydney, Australia, 2022. Disponível em: https://www.sl.nsw.gov.au/public-library-services/digital-practice-guidelines-public-libraries/digital-preservation. Acesso em: 16 ago. 2023.

STATE LIBRARY OF NEW SOUTH WALES. **State Library of New South Wales PANDORA selection guidelines**. [*S. l.*], 2013. 2 p. Disponível em: https://pandora.nla.gov.au/guidelines/2013%20Pandora%20selection%20criteria.docx. Acesso em: 12 out. 2022.

STIMULUS SOFTWARE. **MailArchiva**. [*S. l.*], [2023]. Disponível em: https://mailarchiva.com/. Acesso em: 7 jun. 2023.

STIMULUS SOFTWARE. **MailArchiva**. Product. Features. Resources. Online help. End-user docs. MailArchiva help center. Frequently asked questions (faq). [*S. l.*], c2020. Disponível em: https://mailarchiva.com/. Acesso em: 7 jun. 2023.

STIRLING, Peter; CHEVALLIER, Philippe; ILLIEN, Gildas. Web archives for researchers: representations, expectations and potential uses. **D-Lib Magazine**, v. 18, n. 3/4, March/April 2012. Disponível em: http://www.dlib.org/dlib/march12/stirling/03stirling.html. Acesso em: 7 jun. 2023.

STREAMING MEDIA. *In*: WIKIPEDIA: the free encyclopedia. [San Francisco, CA: Wikimedia Foundation], 2022. Disponível em: https://en.wikipedia.org/wiki/Streaming_media. Acesso em: 7 jun. 2023.

TAINACAN. Documentação. Wiki. **Wiki do Tainacan**. [*S. l.*], [2021]. Disponível em: https://tainacan.github.io/tainacan-wiki/#/pt-br/?id=wiki-do-tainacan. Acesso em: 7 jun. 2023.

TAJEDDINI, Oranus *et al*. Death of web citations: a serious alarm for authors. **Malaysian Journal of Library & Information Science**, Kuala Lumpur, v. 16, n. 3, p. 17-29, Dec. 2011. Disponível em: https://www.researchgate.net/publication/227344592_Death_of_web_citations_a_serious_alarm_for_authors. Acesso em: 7 jun. 2023.

TARGINO, Maria das Graças. Comunicação científica: uma revisão de seus elementos básicos. **Informação & Amp; Sociedade**: Estudos, v. 10, n. 2, p. 1-27, 2000. Disponível em: https://periodicos.ufpb.br/ojs/index.php/ies/article/view/326/248. Acesso em: 7 jun. 2023.

THE NATIONAL ARCHIVES. **Guidance on cloud storage and digital preservation**. 2th ed. Surrey, Mar. c2015. 39 p. Disponível em: https://cdn.nationalarchives.gov.uk/documents/archives/Preserving-Digital-CloudStorage-Guidance_March-2015.pdf. Acesso em: 7 jun. 2023.

THE NATIONAL ARCHIVES. **Operational Selection Policy OSP27**: UK Central Government Web Estate. London, Apr. c2014. 9 p. Disponível em: https://www.nationalarchives.gov.uk/documents/information-management/osp27.pdf. Acesso em: 7 jun. 2023.

THE NATIONAL ARCHIVES. **UK government web archive**. [Richmond, United Kingdom], [2023?a]. Disponível em: https://www.nationalarchives.gov.uk/webarchive/. Acesso em: 7 jun. 2023.

THE NATIONAL ARCHIVES. UK government web archive. **About the UK government web archive**. Glossary of web archiving terms. [Richmond, United Kingdom], [2022?a]. Disponível em: https://www.nationalarchives.gov.uk/webarchive/about/. Acesso em: 7 jun. 2023.

THE NATIONAL ARCHIVES. UK government web archive. **Archive a website or social media channel in the UK Government Web Archive**. How to archive a website with us. Archive a website which isn't closing. Archive a website which is closing. How to archive social media channels. How to check your archived website. How to make your website archive compliant. How to request removal of content from the UK Government Web Archive. [Richmond, United Kingdom], [2022b]. Disponível em: https://www.nationalarchives.gov.uk/webarchive/archive-a-website-or-social-media-channel-in-the-uk-government-web-archive/. Acesso em: 7 jun. 2023.

THE NATIONAL ARCHIVES. UK government web archive. **Find websites in the UK government web archive**. Limitations of the UK government web archive. Legal information about re-using UK government web archive content. [Richmond, United Kingdom], [2022?c]. Disponível em: https://www.nationalarchives.gov.uk/webarchive/find-a-website/. Acesso em: 7 jun. 2023.

THE NATIONAL ARCHIVES. UK government web archive. **How to use the web archive**. [Richmond, United Kingdom], [2021]. Disponível em: https://www.nationalarchives.gov.uk/webarchive/information/. Acesso em: 7 jun. 2023.

THE NATIONAL ARCHIVES. UK government web archive. **Social media archive search**. [Richmond, United Kingdom], [2023?b]. Disponível em: https://webarchive.nationalarchives.gov.uk/social/search/. Acesso em: 7 jun. 2023.

THE NATIONAL ARCHIVES. UK government web archive. **Twitter Archives**. [Richmond, United Kingdom], [2022d]. Disponível em: https://webarchive.nationalarchives.gov.uk/twitter/. Acesso em: 7 jun. 2023.

THE NATIONAL ARCHIVES. **Web archiving guidance**. [*S. l.*], c2011. 15 p. Disponível em: https://cdn.nationalarchives.gov.uk/documents/information-management/web-archiving-guidance.pdf. Acesso em: 7 jun. 2023.

THOMAZ, Katia de Padua. **A preservação de documentos eletrônicos de caráter arquivístico**: novos desafios, velhos problemas. 2004. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2004. Disponível em: https://repositorio.ufmg.br/bitstream/1843/VALA-68ZRKF/1/doutorado___katia_de_padua_thomaz.pdf. Acesso em: 7 jun. 2023.

THOMAZ, Katia de Padua; SOARES, Antonio José. A preservação digital e o modelo de referência Open Archival Information System (OAIS). **DataGramaZero - Revista de Ciência da Informação**, Rio de Janeiro, RJ, v. 5, n. 1, fev. 2004. Disponível em: https://www.brapci.inf.br/index.php/res/download/45229. Acesso em: 7 jun. 2023.

THOMSON, Sara Day. Preserving social media. **DPC Technology Watch Report 16-01**, p. 1-42, Feb. c2016. Disponível em: https://www.dpconline.org/docs/technology-watchreports/1486-twr16-01/file. Acesso em: 7 jun. 2023.

THORPE, Clarissa. Trash to treasure: retro computer, software collection helps National Library access digital pieces. **Australian Broadcasting Corporation** (**ABC) News**, Sydney, 19 June, 2015. Disponível em: https://www.abc.net.au/news/2015-06-20/collecting-retro-computer-technology-to-save-digital-treasures/6560494. Acesso em: 7 jun. 2023.

TRENCH, Brian. Internet: turning science communication inside-out? *In*: BUCCHI, Massimiano; TRENCH, Brian (Edit.). **Handbook of public communication of science and technology**. London: Routledge, 2008. p. 185-198. Disponível em: https://moodle.ufsc.br/pluginfile.php/1485212/mod_resource/content/1/Handbook-of-Public-Communication-of-Science-and-Technology.pdf. Acesso em: 7 jun. 2023.

TRUMAN, Gail. **Web archiving environmental scan**. Harvard Library Report. [Cambridge, Massachusetts]: Harvard University, Jan. 2016. 83 p. Disponível em: https://dash.harvard.edu/bitstream/handle/1/25658314/print_HL_web_archiving_env_scan_2017.pdf?sequence=4&isAllowed=y. Acesso em: 7 jun. 2023.

TUCK, John. Web Archiving in the UK: Cooperation, Legislation and Regulation. **Liber Quarterly**: The Journal of the Association of European Research Libraries, v. 18, n. 3/4, p. 357-365, December 2008. Disponível em: https://liberquarterly.eu/article/view/10539/11227. Acesso em: 7 jun. 2023.

UK Web Archive. [*S. l.*], [2023]. Disponível em: https://www.webarchive.org.uk/en/ukwa/. Acesso em: 7 jun. 2023.

UK WEB ARCHIVE. **About us**. Frequently asked questions. Technical information. [*S. l.*], [2022?]. Disponível em: https://www.webarchive.org.uk/en/ukwa/about. Acesso em: 7 jun. 2023.

UNITED STATES. 107th United States Congress. H.R.2458. E-government act of 2002. **Public law 107–347**. [Washington, D.C., United States], Dec. 17 2002. Disponível em: https://www.congress.gov/107/plaws/publ347/PLAW-107publ347.pdf. Acesso em: 7 jun. 2023.

UNIVERSIDADE DE SÃO PAULO. Instituto de Psicologia. Serviços. Biblioteca. Pesquisar e publicar em Psicologia. **Revisão de literatura**. [*S. l.*], [2022] Disponível em: https://www.ip.usp.br/site/biblioteca/revisao-de-literatura/. Acesso em: 7 jun. 2023.

UNIVERSIDADE ESTADUAL DE CAMPINAS. Procuradoria Geral. **Resolução GR-017/2011, de 29 de junho de 2011**. Estabelece diretrizes e define procedimentos para a gestão, a preservação e o acesso contínuo aos documentos arquivísticos digitais da Universidade Estadual de Campinas. Campinas, SP, 1 jul. 2011. Disponível em: https://www.pg.unicamp.br/norma/3057/0. Acesso em: 7 jun. 2023.

UNIVERSIDADE ESTADUAL PAULISTA. **Política de preservação digital para documentos de arquivo da Unesp**. Versão 1.0. São Paulo, dez. 2017. 14 p. Disponível em: https://www2.unesp.br/portal#!/noticia/33100/politica-de-preservacao-digital-para-documentos-de-arquivo. Acesso em: 7 jun. 2023.

UNIVERSIDADE FEDERAL DE SÃO CARLOS. Programa de Pós-Graduação em Ciência, Tecnologia e Sociedade (PPGCTS). Programa. **Linhas de pesquisa**. São Carlos, São Paulo, [2023]. Disponível em: https://www.ppgcts.ufscar.br/apresentacao/linhas-de-pesquisa-1/linhas-de-pesquisa. Acesso em: 30 abr. 2023.

UNIVERSITY OF CALIFORNIA. California digital library. Services and projects. University of California Curation Center (UC3). Merritt. **Merritt**: A trusted, cost-effective digital preservation repositor. [Oakland, California], c2022. Disponível em: https://merritt.cdlib.org/. Acesso em: 7 jun. 2023.

UNIVERSITY OF CALIFORNIA. **Merritt**: A trusted, cost-effective digital preservation repositor. [Oakland, California], c2023. Disponível em: https://merritt.cdlib.org/. Acesso em: 7 jun. 2023.

UNIVERSITY OF TEXAS AT SAN ANTONIO LIBRARIES. Collections. **Archived websites.** [San Antonio, United States], [2023?a]. Disponível em: https://lib.utsa.edu/specialcollections/collections/websites. Acesso em: 7 jun. 2023.

UNIVERSITY OF TEXAS AT SAN ANTONIO LIBRARIES. **Web archives policy**. [San Antonio, United States], August 2020. 6 p. Disponível em: https://lib.utsa.edu/specialcollections/sites/specialcollections/files/2020-09/WebArchives_Policy_2020-08-20.pdf. Acesso em: 7 jun. 2023.

UNIVERSITY OF TEXAS AT SAN ANTONIO LIBRARIES. **Web archiving methods and collection guidelines**. [San Antonio, United States], 2016. 3 p. Disponível em: https://lib.utsa.edu/files/default/Special%20Collections/UTSAWebArchivingMethodsAndCollectionGuidelines_2016-03.pdf. Acesso em: 7 jun. 2023.

UNIVERSITY OF TEXAS AT SAN ANTONIO LIBRARIES. **Web archiving**. [San Antonio, United States], 2023b. Disponível em: https://libguides.utsa.edu/webarchiving. Acesso em: 7 jun. 2023.

UNIVERSITY OF TEXAS AT SAN ANTONIO LIBRARIES. **Web archiving**. What we collect. FAQs. [San Antonio, United States], Aug 2022. Disponível em: https://libguides.utsa.edu/webarchiving/UTSA. Acesso em: 7 jun. 2023.

UNIVERSITY OF WATERLOO. **Web Archives for Historical Research (WAHR) group**. Waterloo, Canada, [2023?]. Disponível em: https://uwaterloo.ca/web-archive-group/. Acesso em: 7 jun. 2023.

UNIVERZITNÁ KNIŽNICA V BRATISLAVE. **Digitálne pramene**. [Bratislava, Slovensko], [2023?]. Disponível em: https://www.webdepozit.sk/. Acesso em: 7 jun. 2023.

UNIVERZITNÁ KNIŽNICA V BRATISLAVE. Digitálne pramene. **Archívy a katalógy DIP**. Sprístupňovanie archívu. [Bratislava, Slovensko], [2022?a]. Disponível em: https://www.webdepozit.sk/archivy-a-katalogy-dip/. Acesso em: 7 jun. 2023.

UNIVERZITNÁ KNIŽNICA V BRATISLAVE. Digitálne pramene. Kontakty. **Časté otázky**. [Bratislava, Slovensko], [2022?b]. Disponível em: https://www.webdepozit.sk/. Acesso em: 7 jun. 2023.

UNIVERZITNÁ KNIŽNICA V BRATISLAVE. Digitálne pramene. **Systém DIP**. [Bratislava, Slovensko], [2022?c]. Disponível em: https://www.webdepozit.sk/system-dip/. Acesso em: 7 jun. 2023.

UNIVERZITNÁ KNIŽNICA V BRATISLAVE. Digitálne pramene. **WWW pramene**. Archivácia. Profil archívu WWW. Kampane www. [Bratislava, Slovensko], [2022?d]. Disponível em: https://www.webdepozit.sk/www-pramene/. Acesso em: 7 jun. 2023.

VALENTE, Maria Esther; CAZELLI, Sibele; ALVES, Fátima. Museus, ciência e educação: novos desafios. **História, Ciências, Saúde – Manguinhos**, Rio de janeiro, v. 12 (suplemento), p. 183-203, 2005. Disponível em: https://www.scielo.br/j/hcsm/a/8kBtsgnNggwkjCVYwwFCsGS/?format=pdf&lang=pt. Acesso em: 7 jun. 2023.

VALERIO, Palmira Moriconi; PINHEIRO, Lena Vania Ribeiro. Da comunicação científica à divulgação. **TransInformação**, Campinas, SP, v. 20, n. 2, p. 159-169, maio/ago. 2008. Disponível em: https://www.scielo.br/j/tinf/a/jXWgggxgBhXfsT57JDVbghp/?format=html&lang=pt. Acesso em: 7 jun. 2023.

VEIKKOLAINEN, Petteri; LAGER, Lassi. Long-term preservation of the finnish web archive. *In*: INTERNATIONAL INTERNET PRESERVATION CONSORTIUM (IIPC) GENERAL ASSEMBLY, 10., April 2016, Reykjavik, Iceland. **Proceedings** […]. Reykjavik, Iceland: IIPC, 2016. p. 195-203. Disponível em: https://web.archive.org/web/20170317132034/http://netpreserve.org/sites/default/files/GA04-HEKLA-Petteri_Veikkolainen_%26_Lassi_Lager.pdf. Acesso em: 7 jun. 2023.

VELLUCCI, Sherry L. Metadata and authority control. **Library Resources & Technical Services (LRTS)**, [Chicago], v. 44, n. 1, p. 33-43, Jan. 2000. Disponível em: https://journals.ala.org/index.php/lrts/article/view/5136. Acesso em: 7 jun. 2023.

VENLET, Jessica *et al*. **Descriptive metadata for web archiving**: literature review of user needs. Dublin, Ohio: Online Computer Library Center (OCLC) Research, Feb. c2018. 48 p. Disponível em: https://www.oclc.org/content/dam/research/publications/2018/oclcresearch-wam-literature-review-user-needs.pdf. Acesso em: 7 jun. 2023.

VISHWASRAO, Saket. **Performance evaluation of web archiving through in-memory page cache**. 2017. Thesis (Masters of Science in Computer Engineering) - Faculty of the Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 2017. Disponível em: https://vtechworks.lib.vt.edu/bitstream/handle/10919/78252/Vishwasrao_SD_T_2017.pdf?sequence=1. Acesso em: 7 jun. 2023.

VISUAL RESOURCES ASSOCIATION. **An introduction to VRA Core**. VRA Core 4.0 introduction. [*S. l.*], Oct. 2014. 2 p. Disponível em: https://www.loc.gov/standards/vracore/VRA_Core4_Intro.pdf. Acesso em: 7 jun. 2023.

VISUAL RESOURCES ASSOCIATION. **VRA Core 4.0 element description**. [*S. l.*], May 2007. 37 p. Disponível em: https://www.loc.gov/standards/vracore/VRA_Core4_Element_Description.pdf. Acesso em: 7 jun. 2023.

VLASSENROOT, Eveline *et al*. Web archives as a data resource for digital scholars. **International Journal of Digital Humanities**, [*S. l.*], v. 1, n. 1, p. 85-111, March 2019. Disponível em: https://link.springer.com/article/10.1007/s42803-019-00007-7. Acesso em: 19 jan. 2023.

WEB PAGE. *In*: WIKIPEDIA: the free encyclopedia. [San Francisco, CA: Wikimedia Foundation], 2022d. Disponível em: https://en.wikipedia.org/wiki/Web_page. Acesso em: 7 jun. 2023.

WEBARCHIV SCHWEIZ. **Glossar**. Version 1.6. Eidgenössisches Departement des Innern EDI. Bundesamt für Kultur BAK. Schweizerische Nationalbibliothek NB. [Bern, Switzerland], Februar 2016. 6 p. Disponível em: https://www.nb.admin.ch/dam/snl/en/dokumente/e-helvetica/normen_und_regelwerke/webarchiv-schweiz-glossar.pdf.download.pdf/webarchiv-schweiz-glossar.pdf. Acesso em: 7 jun. 2023.

WEBSTER, Peter. How fast does the web change and decay? Some evidence. **Web Archives for Historians**, [*S. l.*], 20 March 2015. Disponível em: https://webarchivehistorians.org/2015/03/. Acesso em: 7 jun. 2023.

WEBSTER, Peter. **How researchers use the archived web**. [Glasgow, Scotland], April c2020. 4 p. (DPC Technology Watch Guidance Note). Disponível em: https://www.dpconline.org/docs/technology-watch-reports/2263-twgn-20-01-how-researchers-use-the-archived-web-webster/file. Acesso em: 7 jun. 2023.

WILLER, Mirna. *et al*. Selective archiving of Web resources: a study of processing costs. **Program: electronic library and information systems**, v. 42, n. 4, p. 341-364, Sept. 2008. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/00330330810912043/full/html. Acesso em: 7 jun. 2023.

WITTER, Geraldina Porto. Pesquisa bibliográfica, pesquisa documental e busca da informação. **Estudos de psicologia**, Campinas, SP, v. 7, n. 1, p. 5-30, jan. /jul. 1990.

WORLD BANK GROUP. **Web archives**. [Washington, United States], c2023. Disponível em: https://www.worldbank.org/en/webarchives. Acesso em: 7 jun. 2023.

WORLD BANK GROUP. **Web archives**. How to use. FAQs. [Washington, United States], c2022. Disponível em: https://www.worldbank.org/en/webarchives. Acesso em: 7 jun. 2023.

ZENG, Marcia Lei.; QIN, Jian. **Metadata**. New York, United States: Neal-Schuman Publishers, June 2008. 365 p.

ZIMAN, John Michael. **A força do conhecimento**. Belo Horizonte, MG: Itatiaia, 1981. 380 p.

ZIMAN, John Michael. **Conhecimento público**. Belo Horizonte, MG: Itatiaia, 1979. 167 p.