UNIVERSIDADE FEDERAL DE SÃO CARLOS
PRÓ-REITORIA DE PÓS-GRADUAÇÃO

MESAILDE SOUZA DE OLIVEIRA MATIAS

BASE REFERENCIAL PARA O POVOAMENTO DE REPOSITÓRIOS INSTITUCIONAIS: COLETA AUTOMATIZADA DE METADADOS DA PLATAFORMA LATTES

SÃO CARLOS – SP
2015

UNIVERSIDADE FEDERAL DE SÃO CARLOS
PRÓ-REITORIA DE PÓS-GRADUAÇÃO

MESAILDE SOUZA DE OLIVEIRA MATIAS

BASE REFERENCIAL PARA O POVOAMENTO DE REPOSITÓRIOS INSTITUCIONAIS: COLETA AUTOMATIZADA DE METADADOS DA PLATAFORMA LATTES

Dissertação apresentada ao Programa de Pós-Graduação em Gestão de Organizações e Sistemas Públicos da Universidade Federal de São Carlos, para a obtenção do título de mestre.

Orientação: Prof. Dr. Roniberto Morato do Amaral

SÃO CARLOS – SP
2015

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

M433br

Matias, Mesailde Souza de Oliveira.
Base referencial para o povoamento de repositórios institucionais : coleta automatizada de metadados da Plataforma Lattes / Mesailde Souza de Oliveira Matias. -- São Carlos : UFSCar, 2015.
86 p.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2015.

1. Comunicação na ciência. 2. Repositórios institucionais. 3. Metadados. 4. Plataforma Lattes. 5. Acesso aberto à informação. I. Título.

CDD: 302.2 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Pró-Reitoria de Pós-Graduação
Programa de Pós-Graduação em Gestão de Organizações e Sistemas Públicos

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado da candidata Mesailde Souza de Oliveira, realizada em 14/08/2015:

Prof. Dr. Roniberto Morato do Amaral
UFSCar

Profa. Dra. Ariadne Chloe Mary Furnival
UFSCar

Prof. Dr. José Eduardo Santarem Segundo
USP

"Não vos amoldeis às estruturas deste mundo,
mas transformai-vos pela renovação da mente,
a fim de distinguir qual é a vontade de Deus:
o que é bom, o que Lhe é agradável, o que é perfeito.
(Bíblia Sagrada, Romanos 12, 2)

Resumo

Os repositórios institucionais abertos (RI) tornaram-se um dos principais instrumentos para o acesso aberto à produção científica global. Além da disponibilização de documentos científicos, os RI podem ser utilizados para os diversos interesses da instituição, promovendo armazenamento de suas documentações, preservação da memória institucional e também dando visibilidade mundial para sua produção. A implementação de um RI, no entanto, envolve uma série de ações, onde um dos grandes desafios é o seu povoamento. Visando contribuir para uma solução que minimize o esforço humano necessário para o povoamento do repositório, o objetivo geral deste trabalho foi desenvolver uma sistemática automatizada de extração de metadados provenientes da Plataforma Lattes realizando a inclusão dos dados em uma base referencial, considerada uma etapa intermediária no processo de povoamento do RI. Além de facilitar o processo de submissão, a base referencial poderá amparar efetivas políticas de divulgação e de incentivo ao autodepósito da produção científica. Trata-se de uma pesquisa de natureza qualitativa, desenvolvida por meio de pesquisa-ação, tendo como unidade-caso a Universidade Federal de São Carlos – UFSCar. Os resultados alcançados compreenderam o desenvolvimento de um conjunto de soluções automatizadas, necessárias à coleta, tratamento e importação dos metadados pela ferramenta DSpace. A solução de povoamento aqui proposta poderá ser implementada em qualquer outra instituição que possua ou não um RI, contribuindo fortemente para a maximização da comunicação científica ao automatizar parte do processo de implementação e manutenção de repositórios institucionais.

Palavras-chaves: Repositório institucional. Metadados. Plataforma Lattes. Acesso aberto.

Abstract

Open institutional repositories (IR) have became a major instrument for providing open access to global scientific production. In addition to providing scientific documents, IR can be used to fulfill institutional interests of the institution, such as promoting the storage of documentation, preserving institutional memory and also bringing worldwide exposure to its production. The implementation of an IR, however, depends on a number of activities, where one of the major challenges is populating it with data. In order to find solutions to improve the process of populating an IR, the aim of this work was to develop a system to automatically extract metadata from the Lattes Platform, inserting this metadata in a reference database, regarded as an intermediate step for populating the IR. Besides easing the submission process, the reference base will be able to support effective marketing campaigns to foster the auto-archival of scientific works. Our approach consists of a qualitative research, developed through action research, within the Federal University of São Carlos – UFSCar. The achieved results comprised the development of a collection of automated solutions essential for gathering and handling metadata to be imported by DSpace. The solution here proposed can be implemented by any institution, with or without an IR, strongly contributing to maximize scientific communication by automating part of the process of implementing and maintaining institutional repositories.

Key-words: Institutional repository. Metadata. Lattes platform. Open access.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

API	*Application Programmer Interface*: Interface para Programação de Aplicativos.
BSD	*Berkeley Software Distribution*: licença livre permissiva, cujo nome é o mesmo do sistema operacional no qual primeiramente foi empregada.
CAPTCHA	*Completely Automated Public Turing test to tell Computers and Humans Apart*: Teste de Turing público completamente automatizado para diferenciação entre computadores e humanos.
CRNI	*Corporation for National Research Initiatives*: Corporação para Iniciativas de Pesquisa Nacionais.
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico.
CRUESP	Conselho de Reitores das Universidades Estaduais Paulistas.
DIM	*Dspace Internal Metadata*: Metadados Internos do DSpace.
DNS	*Domain Name System*: Sistema de Nomes de Domínio.
ERP	*Enterprise Resource Planning*: Planejamento de Recursos Corporativos – termo genérico para um software formado por um conjunto de aplicativos integrados voltados para a gestão institucional.
Fiocruz	Fundação Oswaldo Cruz.
FSF	*Free Software Foundation*: Fundação do Software Livre.
GPL	*General Public License*: licença livre de *copyleft* criada pela FSF.
HTML	*HyperText Markup Language*: Linguagem de Marcação de HiperTexto.
HTTP	*HyperText Transfer Protocol*: Protocolo de Transferência de HiperTexto.
IBICT	Instituto Brasileiro de Informação em Ciência e Tecnologia.
ID CNPq	Identificador CNPq, código numérico que identifica univocamente uma pessoa na Plataforma Lattes.
ISSN	*International Standard Serial Number*: Número de Série Padrão Internacional.
JSON	*JavaScript Object Notation*: Notação de Objetos JavaScript.

JSONB	*JavaScript Object Notation – Binary*: JSON em formato binário.
LDAP	*Lightweight Directory Access Protocol*: Protocolo Leve de Acesso a Diretórios – permite organizar e autenticar o acesso a recursos da rede de forma hierárquica.
MCT	Ministério da Ciência e Tecnologia.
MIT	*Massachusetts Institute of Technology*: Instituto de Tecnologia de Massachusetts.
OA	*Open Archives*: Arquivos Abertos.
OAI	*Open Archives Initiative*: Iniciativa de Arquivos Abertos.
OAI-PMH	*OAI Protocol for Metadata Harvesting*: Protocolo OAI para Colheita de Metadados.
OpenDOAR	*Directory of Open Access Repositories*: Diretório de Repositórios de Acesso Aberto.
ORCID	*Open Researcher and Contributor ID*: Identificador Aberto para Pesquisadores e Colaboradores.
ORM	*Object-Relational Mapping*: Mapeamento Objeto-Relacional.
PLoS	*Public Library of Science*: Biblioteca Pública de Ciência.
REST	*REpresentational State Transfer*: Transferência de Estado Representacional – estilo arquitetural para a construção de serviços web escaláveis.
RI	Repositório Institucional.
RIS	*Research Information Systems*: formato de metadados, cujo nome é o mesmo da empresa que o criou.
RoMEO	*Rights MEtadata for Open archiving*: Metadados de Direitos Autorais para Arquivamento Aberto.
SciELO	*Scientific Electronic Library Online*: Biblioteca Científica Eletrônica Online.
SHERPA	*Securing a Hybrid Environment for Research Preservation and Access*: Assegurando um Ambiente Híbrido para Preservação e Acesso à Pesquisa.
SOAP	*Simple Object Access Protocol*: Protocolo Simples de Acesso a Objetos – protocolo de comunicação baseado em XML para troca de informações estruturadas na implementação de serviços Web.

SQL	*Structured Query Language*: Linguagem de Consulta Estruturada.
SWORD	*Simple Web-service Offering Repository Deposit*: Serviço Web Simples Fornecendo Depósito em Repositório – protocolo padronizado que permite que ferramentas externas realizem o depósito de itens em repositórios.
SWORDv2	Segunda versão do protocolo SWORD.
TLS	*Transport Layer Security*: Segurança para a Camada de Transporte – o protocolo mais utilizado para troca de dados criptografados na Internet.
UFBA	Universidade Federal da Bahia.
UFRGS	Universidade Federal do Rio Grande do Sul.
UFSCar	Universidade Federal de São Carlos.
UNESP	Universidade Estadual Paulista "Júlio de Mesquita Filho".
UNICAMP	Universidade Estadual de Campinas.
USP	Universidade de São Paulo.
UUID	*Universally Unique IDentifier*: Identificador Universalmente Único.
WoS	*Web of Science*: Teia da Ciência.
WSDL	*Web Services Description Language*: Linguagem para Descrição de Serviços Web.
WWW	*World Wide Web*: Teia Mundial.
X.509	Padrão de certificado digital que provê uma infraestrutura e uma hierarquia de confiança em chaves públicas.
XML	*Extensible Markup Language*: Linguagem de Marcação Extensível.

Sumário

1 Introdução

O modelo de comunicação científica tradicional estabeleceu-se na publicação de artigos e trabalhos científicos por revistas especializadas, onde o acesso a essas publicações se dava por meio de assinatura paga por parte das bibliotecas ou pesquisadores interessados. Esse modelo ainda hoje persiste e é uma das grandes entraves para que comunidade científica tenha acesso à informação produzida por ela mesma.

Partindo da necessidade de ter a maior visibilidade possível, a comunidade científica buscou outras alternativas que pudessem gerar maior acesso à sua produção, o que foi possibilitado principalmente graças aos avanços das tecnologias de informação e comunicação (TICs). O avanço das TICs proporcionou o desenvolvimento de uma sociedade regida pela informação , sociedade esta que Castells (1999) chama de "sociedade em rede":

> A sociedade em rede, em termos simples, é uma estrutura social baseada em redes operadas por tecnologias de comunicação e informação fundamentadas na microelectrônica e em redes digitais de computadores que geram, processam e distribuem informação a partir de conhecimento acumulado nos nós dessas redes. (CASTELLS, 1999, p. 20)

Esses avanços oportunizaram o surgimento de um movimento progressivo na direção do acesso aberto à informação científica. "A possibilidade de publicar eletronicamente o periódico científico e a preocupação com o acesso a essa publicação resultaram em uma série de iniciativas em todo o mundo" (NEVES, 2004, p.117).

O primeiro repositório aberto surgiu em 1991, no laboratório de Física Nuclear de Los Alamos, Novo México, nos EUA, e recebeu o nome de ArXiv. O ArXiv trouxe para a via digital o que já fazia parte da cultura dos pesquisadores das ciências exatas: o compartilhamento precoce de suas pesquisas, a fim de receber sugestões de seus pares antes mesmo do envio para as revistas.

Em 1999 ocorreu a *Santa Fé Convention / Open Archives Initiative* – OAI, em Santa Fé, Novo México, EUA. A convenção estabeleceu padrões de interoperabilidade entre os repositórios existentes, que na época já haviam se multiplicado, com o objetivo de que estes repositórios pudessem comunicar-se entre si, facilitando ainda mais a coleta de metadados.

Os repositórios de produção científica se popularizaram no mundo inteiro e tornaram-se nos dias de hoje indispensáveis para a preservação e visibilidade da informação científica. Nesse contexto, as bibliotecas digitais, os repositórios institucionais e os periódicos de acesso aberto são muito mais que ferramentas tecnológicas, são políticas públicas para a garantia do livre acesso à produção científica desenvolvida no país e no mundo.

No Brasil, o Instituto Brasileiro de Informação em Ciência e Tecnologia – IBICT, tem forte atuação na disseminação de repositórios abertos, proporcionando treinamentos e disponibilizando, em língua portuguesa, os principais softwares utilizados para implantação de repositórios no mundo. O "Manifesto Brasileiro de Apoio ao Acesso Livre à Informação Científica" faz parte das ações do IBICT no âmbito da iniciativa mundial pelo acesso aberto originada por meio da OAI.

Os repositórios digitais trazem significativos benefícios para a comunicação científica, porém sua implementação constitui desafio e ultrapassa o escopo técnico: os maiores entraves para a implantação de repositórios institucionais estão nas questões culturais e organizacionais das instituições. Viana, Márdero Arellano e Shintaku (2005, p. 8) apontam que, "assim como as novas tecnologias de informação estão sendo um desafio para as organizações, existe também um grande número de desafios relacionados com a habilidade dessas organizações para integrar o gerenciamento de materiais digitais na sua estrutura organizacional".

Dentre os principais desafios encontrados na implementação de repositórios institucionais, este trabalho tem seu foco em um dos mais importantes: o povoamento do repositório. É preciso, no entanto, ressaltar que todas as fases da implementação de um repositório precisam estar respaldadas por políticas institucionais bem definidas voltadas para a gestão desse repositório em suas mais variadas nuances.

Ley (2013) afirma que o povoamento implica em aspectos políticos e técnicos, e trata-se de uma fase contínua na implementação de um repositório, um vez que todo repositório precisa ser povoado de forma cumulativa e perene. A autora destaca a importância desta fase para a implementação eficiente de um RI:

> A política de povoamento, que Viana e Márdero Arellano (2006) nomeiam de "política para engajamento de pesquisadores/autores" e Foster e Gibbons (2005) de "recrutamento de conteúdo", pode sintetizar a razão de ser do sistema de informação, pois sem conteúdo ou com conteúdo incipiente, o repositório não cumpre o papel para o qual foi criado, de reunir, divulgar e promover a visibilidade da produção intelectual de dada instituição (LEY, 2013, p.74)

O povoamento de um repositório institucional exige dos profissionais envolvidos a recuperação das publicações nas mais diversas fontes de informação, o que gera significativa demanda de trabalho. No entanto, o Brasil possui um portal que concentra dados de pesquisadores de todo o país, trata-se da Plataforma Lattes, do CNPq – Conselho Nacional de Desenvolvimento Cientifico e Tecnológico, que, por meio do armazenamento dos currículos de pesquisadores, docentes e discentes de todo o país, possui dados da produção científica nacional. "Por sua riqueza de informações e sua crescente confiabilidade e abrangência, o currículo Lattes se tornou elemento indispensável e compulsório à análise de mérito e competência dos pleitos de financiamentos na área de ciência e tecnologia" (CNPq, 2015).

Com o objetivo de agilizar a fase de povoamento, alguns repositórios existentes no Brasil já coletam metadados de bases como a Web of Science (WoS) e a SciELO. Ainda assim, a Plataforma Lattes (CNPq) é uma fonte promissora de metadados. Vale ressaltar que a WoS indexa somente os periódicos mundialmente mais citados em suas respectivas áreas, e que a SciELO reúne apenas periódicos de acesso aberto da América Latina e do Caribe, ao passo que o Currículo Lattes congrega dados sobre todos os trabalhos de vários tipos do pesquisadores, os quais têm incentivo institucional e profissional para mantê-los atualizados no Lattes, independente da revista ou editora responsável pela publicação.

Além disso, cada Currículo Lattes é atrelado univocamente à identidade civil (CPF) de seu portador, que concorda com um termo responsabilizando-se legalmente pelas informações ali contidas. Isso significa que os metadados estão agrupados por autor já em sua fonte original, ao contrário do que ocorre em outras bases. Na WoS, a identificação de autoria é realizada a partir dos nomes adotados nos trabalhos por meio de algoritmos de aprendizagem de máquina e clusterização (CRL, 2015). Na SciELO, essa identificação depende do uso consistente do nome do autor, obedecendo sempre a um mesmo formato em todas as publicações.

No entanto, fatores tais como a falta de uma sistemática para a coleta, tratamento e importação dos metadados dificultam o aproveitamento do potencial da Plataforma Lattes. Desse modo, visando contribuir para maximizar a comunicação científica, o objetivo geral deste trabalho foi desenvolver uma sistemática automatizada de extração de metadados provenientes da Plataforma Lattes realizando a inclusão dos dados em uma base referencial, considerada uma etapa intermediária no processo de povoamento do RI. Dentre os objetivos específicos estão:

a) Pesquisar sobre os processos para implantação de um RI, dando ênfase ao processo de povoamento;

b) Analisar o padrão de metadados Dublin Core, utilizado pelo DSpace;

c) Testar possibilidades de customização dos metadados;

d) Estudar metodologias e ferramentas de extração dos dados do provedor escolhido (Lattes) e a posterior importação desses dados de maneira apropriada na ferramenta DSpace.

O método utilizado nesta pesquisa foi a pesquisa-ação. Esse método foi escolhido pois tem como foco a "pesquisa em ação", tendo como ideia principal a utilização do método científico para estudar a resolução de importantes problemas sociais ou organizacionais, diretamente em conjunto com aqueles que sofrem esses problemas. Ou seja, trata-se de uma abordagem participativa cujo principal objetivo é fazer com que a ação seja mais efetiva, enquanto simultaneamente constrói-se um corpo de conhecimento científico (COUGHLAN; COGHLAN, 2002).

A unidade caso foi a Universidade Federal de São Carlos – UFSCar. A Universidade

ainda não possui um repositório institucional oficial voltado para a sua produção científica. Portanto, o presente projeto pretende contribuir à viabilização dessa implementação por meio de um povoamento inicial do repositório com metadados extraídos da Plataforma Lattes.

As principais contribuições deste trabalho dizem respeito ao fato de que esta sistemática de carga automatizada de dados poderá ser utilizado como carga inicial não só na UFSCar, mas em qualquer outra instituição, provendo um primeiro ambiente de produção para um desenvolvimento subsequente mais elaborado do RI, possibilitando uma grande visibilidade inicial do repositório.

Galina Russell (2011) destaca que uma das principais finalidades de um RI é ser uma ferramenta para o autodepósito da produção científica, no entanto, a autora pontua que alcançar acesso aberto por meio do autodepósito é uma estratégia que não tem tido o êxito que se esperava. Todo o conjunto de ações burocráticas, que vão desde as licenças de direito autoral até a inserção de metadados, dentre outras, formam obstáculo para que pesquisadores depositem diretamente seus trabalhos. No entanto, é possível inferir que a preexistência dos metadados no repositório facilitará esta atividade, restando ao autor apenas o aceite das licenças e o depósito do texto completo.

Sobre o povoamento por meio de autodepósito, Kuramoto (2011) afirma ainda que:

> O povoamento de um repositório institucional (RI) é fruto de uma série de fatores, que se inicia pela sua gestão e continua com a adoção de uma política adequada e capaz de induzir os pesquisadores a autodepositar a sua produção científica. Dentre esses fatores, encontra-se um programa de marketing para sensibilizar e induzir os pesquisadores a autodepositar a sua produção científica. (KURAMOTO, 2011)

Outro fato bastante relevante nesta pesquisa é que a possibilidade de integração com ferramentas externas, que já são utilizadas pelos pesquisadores, facilitará a utilização do RI. Ao atualizar seu Currículo Lattes, todos os metadados serão carregados e atualizados no RI de modo automatizado, bastando que o pesquisador envie o arquivo contendo o texto completo do trabalho e aceite a licença, além do fato de que estes metadados poderão ser indexados por ferramentas externas de terceiros (por exemplo, buscadores), implicando em um eficiente meio de divulgação para que os autores busquem aumentar o impacto e a visibilidade de suas pesquisas.

Vale ressaltar que a *Budapest Open Access Initiative*[1] (BOAI, 2002) recomenda que "os repositórios digitais abertos devem estar em conformidade com as recomendações da OAI de modo que possa ter seu conteúdo localizado por motores de busca na Internet" (COSTA, 2014, p.42). A BOAI, em sua comemoração de 10 anos, recomendou ainda que os repositórios de acesso aberto devem proporcionar ferramentas, já existentes de forma gratuita, para converter

[1] <http://www.budapestopenaccessinitiative.org/translations/portuguese-translation>

os depósitos realizados em PDF em formatos legíveis por máquina como o XML[2] (BOAI, 2015).

Os processos executados neste trabalho podem ser considerados como etapas preliminares de um processo de Data Mining, ou, em português, Mineração de Dados. Fayyad, Piatetsky-Shapiro e Smyth (1996 apud NAVEGA, 2002) definem Data Mining como sendo "...o processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis". Navega (2002, p.1-2), no entanto, salienta que, apesar da existência de diversos algorítmos que realizam muito bem a função de descobrir "padrões válidos e novos", para encontrar os chamados "padrões valiosos", a interação de analistas humanos é fundamental, pois somente a interação humana é capaz de determinar o valor dos padrões encontrados.

O autor afirma ainda que existem muitos passos necessários para a realização de fato de uma mineração de dados, mas que os passos fundamentais para uma mineração bem sucedida devem começar a partir de uma fonte de dados na qual seja efetuada uma limpeza, eliminando inconsistências, preenchendo informações, removendo ruídos e redundâncias etc... para, somente depois, ter um repositório organizado que já será útil para a aplicação de diversos algorítimos de Data Mining. Ou seja, "encontrar padrões requer que os dados brutos sejam sistematicamente "simplificados" de forma a desconsiderar aquilo que é específico e privilegiar aquilo que é genérico" (NAVEGA, 2002, p.2).

Nesta pesquisa, os dados brutos extraídos da Plataforma Lattes passaram pelo processo de limpeza e, posteriormente, foram transcodificados para o formato Dublin Core qualificado, de modo que pudessem atender aos objetivos deste trabalho e que, além disso, pudessem ser alvo de possíveis minerações de dados futuras que possam prover padrões valiosos e indicadores que apoiem a gestão da Instituição.

Esta dissertação está organizada em cinco seções. A primeira compreende esta introdução. A segunda seção abrange o referencial teórico, contendo uma breve revisão da literatura acerca dos paradigmas da comunicação científica e repositórios institucionais, uma discussão de temas relacionados ao acesso à informação, interoperabilidade, padrões de metadados, implementação de repositórios (infraestrutura, politica de informação e ferramentas) e algumas iniciativas de sucesso no campo dos repositórios abertos no Brasil e no mundo. A terceira seção é dedicada ao desenvolvimento da pesquisa: método, unidade-caso – UFSCar, provedor de metadados – Plataforma Lattes, extração com scriptLattes, extração por meio de *web service*, tratamento dos dados extraídos, importação de itens no DSpace com o protocolo SWORDv2, e catálogo decisório de autoridade de nomes baseado na Plataforma Lattes. A quarta seção apresenta resultados tanto qualitativos como quantitativos alcançados por este projeto, seguidos por uma discussão das limitações e de diversas ideias para execução em trabalhos futuros. Por fim, a quinta seção conclui o trabalho, trazendo as considerações finais.

[2] EXtensible Markup Language

2 Repositórios Institucionais

2.1 Comunicação Científica

O sistema de comunicação científica foi tradicionalmente estabelecido em um modelo no qual a informação científica é publicada em periódicos e revistas especializadas acessadas por meio do pagamento de assinaturas. Desde a sua concepção, este modelo atendeu principalmente aos interesses de editores e publicadores que visavam o lucro crescente proporcionado pelas altas taxas das assinaturas de seus periódicos.

Esse movimento gerou a chamada "crise dos periódicos" em meados das décadas de 80 e 90, período no qual as altas taxas cobradas pelas editoras fizeram com que as bibliotecas já não tivessem mais condições financeiras de manter atualizadas suas coleções de periódicos.

Em meio a esta realidade, crescia também o dilema ético envolvido no limitado acesso ao conhecimento científico que este modelo de comunicação científica proporcionava até mesmo para a comunidade científica de autores de artigos que, em sua grande parte, tinha suas pesquisas financiadas pelo Estado.

> Trata-se de um modelo cujo maior beneficiário são os editores das revistas científicas, suportado pelos pesquisadores e pelo Estado, que, em última análise, mantém as assinaturas dessas revistas e, indiretamente, exige que seus pesquisadores tenham a notoriedade de publicar nessas revistas. (KURAMOTO, 2006, p. 93)

A busca pelo acesso livre às publicações científicas sempre fez parte deste cenário. Castells (1999) toma a revolução da tecnologia da informação como ponto de partida para analisar o novo paradigma tecnológico, organizado com base na tecnologia da informação, constituído principalmente nos Estados Unidos, que concretizou um novo estilo de produção, comunicação, gerenciamento e vida, ainda nas décadas de 60 e 70.

Marcondes e Sayão (2009) destacam o surgimento da Internet e da WWW, no final da década de 80, como os grandes impulsionadores do movimento de livre acesso como acontece atualmente. Um dos marcos deste movimento é a criação, em 1991, do primeiro repositório digital de *pre-prints*, no laboratório de física de Los Alamos, Novo México, EUA, coordenado pelo físico Paul Ginsparg. Os repositórios digitais, também denominados e-prints, surgem então como alternativas ao tradicional sistema de comunicação científica.

> A partir do início da década de 90 do século XX parcelas crescentes da comunidade acadêmica, incluindo aí associações de pesquisadores e de bibliotecas especializadas e acadêmicas, passam a buscar alter nativas ao crescente custo de assinaturas imposto pelos publicadores comerciais. Com

> o surgimento da Internet a alternativa torna-se clara. Era possível publicar na Internet, a um custo mínimo, com um alcance mundial e com uma rapidez entre a submissão do artigo e sua publicação consideravelmente maior. A lógica das cobranças de assinaturas, começa a ser confrontada por vários setores da comunidade acadêmica, a lógica do livre acesso. (MARCONDES; SAYÃO, 2009, p. 16)

Neste processo nasce a busca pelo acesso livre, que promove a disseminação de repositórios pelas principais instituições acadêmicas existentes. (MARCONDES; SAYÃO, 2009) destacam os principais marcos na trajetória pelo acesso livre:

> - Lançamento do ArXiv, em 1991 — primeiro repositório eletrônico, no laboratório de física nuclear de Los Alamos, Novo México, EUA;
> - Santa Fé Convention / Open Archives Initiative, em 1999. Santa Fé, Novo México, EUA — Propõe mecanismos tecnológicos de interoperabilidade entre esses repositórios eletrônicos para que o crescente número de repositórios que começa a se formar se torne um efetivo meio de comunicação científica;
> - Scholarly Publishing & Academic Resources Coalition (SPARC)– uma associação mundial de bibliotecas especializadas, através do manifesto Declaring Independence, 2001: "Please join me in DECLARING INDEPENDENCE from publishers and journals that do not serve the research community.";
> - Budapest Declaration, em 2001 — Evento do Open Society Institute;
> - Primeira Instituição Acadêmica a adotar o Livre acesso a sua produção, School of Electronic and Computer Science, Univ. de Southampton, 2001.
> - Declaração de Berlin, em 2003;
> - Declaração de Bethesda, 2003;
> - WSIS 2003, Declaração de Princípios (UNESCO), compromisso com livre acesso, item B3, 28;
> - Resolução da Câmara dos Comuns, no Reino Unido, em 2004;
> - Declaração de Salvador: Commitment to Equity, durante o ICML 2005 — Ninth World Congress on Health Information and Libraries, Salvador, Brasil;
> - Manifesto pelo Livre Acesso, Brasil, em 2005;
> - Projeto de Lei n. 1.120, em 2007, Política de livre acesso para o Brasil;
> - Decisão dos pesquisadores da Univ. de Harvard a favor do livre acesso, em 12 fev. 2008;
>
> (MARCONDES; SAYÃO, 2009, p. 18–19)

Suaiden (2006, p. 7) afirma que, no mundo contemporâneo, "a grande questão é como as pessoas terão amplo e livre acesso aos benefícios das tecnologias de informação e comunicação (TICs), de forma que a Internet, por sua apropriação social, seja um poderoso instrumento de educação, ciência e tecnologia, cultura e formação de cidadania".

Almeida (2006, p.29) ressalta a importância da comunicação científica como processo vital para o avanço e o desenvolvimento da ciência, uma vez que "é por seu intermédio

que ocorre a disseminação, a interação da comunidade científica e a legitimação pelos pares, consolidando assim a geração de novos conhecimentos".

As políticas públicas voltadas para a ampliação da comunicação científica têm hoje no Brasil muito mais meios e instrumentos. O acesso aberto, por intermédio dessas políticas, tem avançado de forma gradativa, principalmente depois da informatização de periódicos nacionais (via portal SciElo) e internacionais (via portal de periódicos da Capes) (SOUZA, 2003).

O portal SciElo – *Scientific Electronic Library Online* [1], presente na América Latina e no Caribe, inclui unicamente títulos de livre acesso ao texto completo em todas as suas versões. O portal foi resultado de um projeto de pesquisa da Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp), em parceria com o Centro LatinoAmericano e do Caribe de Informação em Ciências da Saúde (Bireme) (NEVES, 2004) .

Já o Portal de Periódicos da CAPES [2], criado em 2000, é um instrumento de política pública para subsidiar o acesso ao conhecimento científico financiado pelo Ministério da Educação através da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) (ALMEIDA; GUIMARÃES; ALVES, 2010). O Portal tem importância estratégica para o país e possui um dos maiores acervos mundiais por meio de contratos firmados com editoras internacionais, com o objetivo de proporcionar acesso a estas publicações, acesso este bastante oneroso aos cofres públicos brasileiros (KURAMOTO, 2006) .

A existência dos referidos portais coloca o Brasil em destaque no cenário mundial e cria uma situação de comodidade para os pesquisadores brasileiros, pricipalmente no acesso às publicações internacionais pagas que fazem parte do acervo do Portal de Periódicos da CAPES. Talvez por esta facilidade de acesso, muitas universidades brasileiras, mesmo nos dias de hoje, ainda não sentem a urgência do acesso aberto à produção científica, e, por este motivo, não colocam a construção de um RI como estratégica para a instituição. No entanto, Kuramoto destaca que:

> Ignorar esse movimento e não desenvolver qualquer ação no sentido de criar repositórios ou provedores de serviços para colheita dos metadados de publicações ou repositórios nacionais e internacionais significa continuar dependente das publicações científicas comerciais (KURAMOTO, 2006, p. 101).

Um importante marco pelo livre aberto à informação científica no Brasil foi o Manifesto Brasileiro de apoio ao Acesso Livre à Informação Científica (2005), o qual afirma que "a informação científica é o insumo básico para o desenvolvimento científico e tecnológico de uma nação. Trata-se de um processo contínuo em que a informação científica contribui para o desenvolvimento científico, e este, por sua vez, gera novos conteúdos realimentando todo o processo" (IBICT, 2005).

[1] <http://www.scielo.br>
[2] <http://www.periodicos.capes.gov.br>

Esse manifesto é resultado de um longo processo de mobilização mundial pelo livre acesso que teve como principal marco a *OPen Archives Initiative*(1999) . A Iniciativa de Arquivos Abertos e do Movimento pelo Livre Acesso à Informação Científica propõe o livre acesso à informação em todo o mundo e sua principal política para chegar a este objetivo é a construção de repositórios institucionais e a implantação de interoperabilidade entre estes repositórios por meio de princípios e padrões estabelecidos. É a partir dessa iniciativa que surgem os principais softwares para a construção e manutenção de repositórios no mundo.

Neste contexto, os Repositórios Institucionais (RI) têm ganhado mais e mais relevância nas instituições não só como forma de disponibilizar acesso à sua produção científica, mas também como forma de preservação da memória institucional, dentre outras diversas funcionalidades dessa importante ferramenta.

Os avanços da tecnologia da informação "permitiram o surgimento de redes de comunicação eletrônica, revolucionando os fluxos de informação, forma de acesso e troca de informações ampliando o espiral do conhecimento, graças a um novo parâmetro espaço-tempo possibilitado pelas tecnologias" (ROSA; TOUTAIN, 2009), sendo estas as relações que servem de base para a chamada Sociedade da Informação.

Esses avanços propiciaram o surgimento da Internet e dos sistemas de gestão que permitiram a criação dos chamados repositórios institucionais. Os repositórios digitais surgem então como um novo canal da comunicação científica propiciado pelas tecnologias da informação e comunicação.

O termo repositório em si é, de certa forma, genérico:

> Um repositório é um espaço – real ou virtual – para armazenamento de grande quantidade de alguma coisa – produtos, software, arquivos, dados, informações etc – visando principalmente à sua segurança e preservação. No ambiente virtual, os repositórios, nesse caso chamados repositórios digitais, tem sido cada vez mais utilizados também para a ampla divulgação do que é armazenado. (CONSERVA JR., 2014, p. 16–17)

Segundo Viana, Márdero Arellano e Shintaku (2005, p. 3), um repositório digital é uma forma de armazenamento de objetos digitais que tem a capacidade de manter e gerenciar material por longos períodos de tempo e prover o acesso apropriado. Sendo assim, pode-se entender um repositório institucional como um banco de dados da produção científica de uma determinada instituição disponível na web.

Porém, um repositório institucional significa muito mais que um simples banco de dados, Sayão e Marcondes (2009, p. 24) ressaltam que "a criação de repositórios institucionais compreende um grande número de atividades que ensejam aspectos políticos, legais, educacionais, culturais e alguns componentes técnicos importantes. O encaminhamento correto desses vários aspectos e de suas interrelações é que vai determinar o perfil do repositório e a sua aproximação aos objetivos fixados pela organização e, por fim, o sucesso do empreendimento".

Os autores afirmam ainda que "repositórios institucionais trazem agora para universidades e instituições de pesquisa a oportunidade de se fortalecerem institucionalmente a partir da visibilidade de sua produção acadêmica organizada e disponível, como um retrato fiel de sua instituição, a partir de seu repositório institucional" (SAYÃO; MARCONDES, 2009).

Os repositórios institucionais representam uma importante quebra de paradigma não só no acesso à produção científica, Viana, Márdero Arellano e Shintaku (2005) destacam que:

> Estabelecer um repositório institucional indica que a biblioteca esta mudando seu papel de custódia para contribuir ativamente na mudança do modelo de comunicação científica. As bibliotecas mantêm a responsabilidade de gerenciar e arquivar material impresso. Mas, a medida que o volume de material para pesquisa de acesso aberto em formato digital cresce, o papel e o valor das coleções impressas declinam proporcionalmente. (VIANA; MÁRDERO ARELLANO; SHINTAKU, 2005, p. 3)

O movimento pelo acesso aberto ganhou cada vez mais adeptos desde a criação dos primeiros repositórios institucionais. A iniciativa foi disseminada por diversos países e pelas mais diversas áreas do conhecimento, gerando, por parte do movimento pelo livre acesso, a necessidade de organizar-se politicamente e desenvolver sua própria tecnologia para promover interoperabilidade entre os repositórios existentes.

É com esse objetivo que ocorre em julho de 1999 a Santa Fé Convetion, uma reunião de gestores de repositórios de documentos científicos que cria a *Open Archives Initiative* (OAI). "No bojo desta iniciativa foi criado o padrão de metadados Dublin Core e o *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH), para propiciar a coleta automática e o reuso de metadados de repositórios abertos (*open archives*)" (WARNER, 2001 apud MARCONDES; SAYÃO, 2009, p. 15).

> Essa convenção definiu as especificações técnicas e os princípios administrativos para se estabelecer um mínimo, mas potencialmente alto, nível funcional de interoperabilidade entre esses repositórios. São necessários os seguintes componentes tidos como essenciais para um arquivo de e-prints:
>
> - mecanismo de submissão;
> - sistema de armazenamento a longo prazo;
> - uma política de gestão para a submissão e preservação de documentos;
> - uma interface aberta que permita terceiros coletar os metadados dos respectivos arquivos.
>
> (KURAMOTO, 2006, p. 94)

Desde então, a *Open Archives Initiative*[3] desenvolve e promove padrões de interoperabilidade que visam facilitar a disseminação eficiente de conteúdos.

[3] <http://www.openarchives.org>

No Brasil, cumpre destacar os esforços de duas instituições estratégicas que tem envolvimento direto na disseminação de repositórios institucionais: o CNPq e o IBICT. O CNPq surgiu como Conselho Nacional de Pesquisas e foi criado no início da década de 50, tendo entre suas atribuições "manter relação com instituições nacionais e estrangeiras para intercâmbio de documentação técnico-científica" (IBICT, 2015).

Foi por meio de proposta conjunta do CNPq e da Fundação Getúlio Vargas -FGV, que foi criado, em 1954, o Instituto Brasileiro de Bibliografia e Documentação (IBBD), que passou a integrar a estrutura organizacional do CNPq, porém, com a reorganização das atividades de ciência e tecnologia no país, nos anos 70, o Conselho Nacional de Pesquisas transforma-se em Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq. Da mesma maneira que o CNPq, o IBBD passa a ser chamado de Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), consolidando-se, então, como órgão que coordenaria, no Brasil, as atividades de informação em C&T (IBICT, 2015).

A intenção de criar um banco de registro dos currículos de pesquisadores brasileiros existia no CNPq desde meados dos anos 80, quando foi feita a primeira captação de dados de currículos, ainda em papel, com algumas etapas formalizadas em um sistema informatizado. Ao final dos anos 90, o CNPq desenvolveu um formulário eletrônico que, preenchido pelo pesquisador, deveria ser enviado em disquete para o CNPq, que os carregava na base de dados.

Nessa mesma época, o CNPq contratou os grupos universitários Stela, vinculado à Universidade Federal de Santa Catarina, e C.E.S.A.R, da Universidade Federal de Pernambuco, para que, juntamente com profissionais da empresa Multisoft, e técnicos das Superintendências de Informática e Planejamento, desenvolvessem uma única versão de currículo capaz de integrar as já existentes. Assim, em agosto de 1999, o CNPq lançou e padronizou o Currículo Lattes como sendo o formulário de currículo a ser utilizados no âmbito do Ministério da Ciência e Tecnologia e CNPq (CNPq, 2015).

Desde então, o Currículo Lattes tornou-se um padrão nacional compulsório no registro da vida pregressa e atual de estudantes e pesquisadores do país, sendo utilizado pelas principais universidades federais, institutos, centros de pesquisa e fundações de amparo à pesquisa dos estados como instrumento para a avaliação de pesquisadores, professores e alunos, de modo que a Plataforma Lattes constitui-se hoje como um grande repositório de dados de pesquisadores de todo o país:

> A Plataforma Lattes representa a experiência do CNPq na integração de bases de dados de Currículos, de Grupos de pesquisa e de Instituições em um único Sistema de Informações. Sua dimensão atual se estende não só às ações de planejamento, gestão e operacionalização do fomento do CNPq, mas também de outras agências de fomento federais e estaduais, das fundações estaduais de apoio à ciência e tecnologia, das instituições de ensino superior e dos institutos de pesquisa. Além disso, se tornou estratégica não só para as atividades de planejamento e gestão, mas também para a formu-

> lação das políticas do Ministério de Ciência e Tecnologia e de outros órgãos governamentais da área de ciência, tecnologia e inovação. (CNPq, 2015)

Já o IBICT tornou-se uma instituição estratégica voltada, dentre outras ações, para a criação de bibliotecas digitais e implantação de repositórios digitais em todas as universidades do Governo Federal e em diversas unidades de pesquisa do MCT, fornecendo inclusive treinamento técnico para este fim. O IBICT possui mais de 500 periódicos eletrônicos que visam não somente preservar a memória do patrimônio científico e tecnológico nacional, mas principalmente criar condições para o aumento da produção científica e a consequente visibilidade internacional, sendo também o depositário legal das teses e dissertações defendidas por brasileiros no exterior (IBICT, 2015).

2.2 Acesso à informação

2.2.1 Direito autoral

2.2.1.1 Proteção da propriedade intelectual e "uso justo"

Para garantir o amplo acesso a uma determinada obra, não basta que o autor a divulgue. É necessário ele dê explicitamente a permissão de cópia e redistribuição. A Convenção Internacional de Berne (RICKETSON, 1987), da qual 168 países (incluindo o Brasil) são signatários, garante que o direito autoral seja automático — não é necessário obtê-lo por meio de registro oficial, basta que uma ideia tenha sido reduzida a um formato tangível (e.g. uma folha de papel, um filme fotográfico, ou um arquivo de computador) para que exista proteção pelo direito autoral.

O direito autoral, por padrão, impede a cópia e a redistribuição da obra, integral ou parcialmente. A Convenção de Berne permite apenas que os países definam exceções para o denominado "uso justo" (*fair use*), que envolvem, por exemplo, a permissão para que um pequeno trecho de um artigo seja copiado para comentários e críticas, desde que citada a fonte, sem que seja necessária a autorização explícita do autor.

O Capítulo IV da lei 9610/98[4], que trata de direitos autorais no Brasil, em seu Artigo 46, afirma não constituir ofensa aos direitos autorais:

> VIII - a reprodução, em quaisquer obras, de pequenos trechos de obras preexistentes, de qualquer natureza, ou de obra integral, quando de artes plásticas, sempre que a reprodução em si não seja o objetivo principal da obra nova e que não prejudique a exploração normal da obra reproduzida nem cause um prejuízo injustificado aos legítimos interesses dos autores.

A cópia apenas de metadados de uma produção científica para fins de catalogação, tais como título, local e data de publicação, e até mesmo do texto do resumo (*abstract*),

[4] <http://www.planalto.gov.br/ccivil_03/leis/L9610.htm>

pode ser enquadrada como "uso justo", sendo desta forma amparada pelas exceções da lei de direitos autorais. Entretanto, a cópia integral do artigo (texto completo, ou *full text*), ou de grandes trechos do mesmo, é terminantemente proibida a menos que tenha sido autorizada pelo portador dos direitos autorais.

2.2.1.2 Contratos de editoras e transferência de direitos autorais

Ao publicar um artigo por meio de uma editora, o pesquisador sujeita-se ao contrato dessa editora, da mesma forma que um artista fica sujeito ao contrato de uma gravadora ao lançar um disco. Os contratos geralmente envolvem a transferência dos direitos autorais do artigo para a editora, o que implica na perda do direito do autor de escolher os termos de licenciamento. O autor deixa de ter o poder de definir uma licença livre para o artigo, e qualquer cópia do texto só pode ser realizada caso autorizada pelo contrato com a editora.

A severidade dos contratos varia de editora para editora. Algumas não permitem nenhum tipo de distribuição do texto além daquela realizada por si próprias, por meio de assinaturas pagas — nem ao menos rascunhos iniciais do texto podem ser disponibilizados para acesso. Outras são bastante liberais, e permitem que exatamente o mesmo arquivo disponibilizado pela editora seja reproduzido livremente.

2.2.1.3 Licenças livres

Com o desejo de uma crescente parcela de autores de produção intelectual de disponibilizar livremente suas obras para toda a população, acabaram consolidando-se as licenças livres (CARVER, 2007). Tratam-se de textos legais cuidadosamente projetados para assegurar que o acesso aberto a uma determinada obra seja preservado, facilitando com que autores leigos em legislação possam disponibilizá-la, e com que terceiros possam reaproveitá-la sem o risco de futuras ameaças legais.

Existem diversas modalidades de licenças livres. Algumas delas, denominadas "permissivas", permitem que uma obra seja copiada integral ou parcialmente com a única condição de que o autor original seja citado. Exemplos dessa modalidade de licença são a BSD, originária de projetos de software da Universidade da Califórnia em Berkeley, a MIT, proveniente do Instituto de Tecnologia de Massachusetts, e a Creative Commons Attribution.

Outra modalidade é a das licenças de *copyleft*[5], que exigem que trabalhos contendo partes substanciais da obra original sigam os mesmos termos de licenciamento. A lógica da licença de *copyleft* é a de perpetuar a filosofia de liberdade, exigindo que qualquer trabalho derivado de uma obra livre também seja livre. Exemplos dessa modalidade são a licença GPL, da Fundação do Software Livre (FSF), e a Creative Commons ShareAlike.

[5] O termo *copyleft* é um trocadilho com a palavra *copyright*. Em Português, *copyleft* seria equivalente a dizer "esquerdos autorais" em vez de "direitos autorais".

2.2.2 Projeto SHERPA/RoMEO

É um trabalho árduo demais para um pesquisador ler e entender os pormenores dos contratos com cada editora na qual este publica, incluindo termos que podem variar de revista para revista. Essa é uma grande barreira, portanto, quando se deseja incentivar que todo pesquisador disponibilize seus artigos por meio de repositórios institucionais ou de outros meios de acesso.

Com o objetivo de reduzir essa barreira, surgiu o projeto RoMEO, desenvolvido pela SHERPA, uma organização existente desde 2002, estabelecida pelas universidades de Nottingham, Edinburgh, Glasgow, Leeds, Oxford, Sheffield, Biblioteca Britânica, York, dentre outras (PINFIELD, 2002). Trata-se de um mecanismo de busca de revistas e editoras, que cataloga informações a respeito das permissões concedidas aos autores pelos respectivos contratos.

2.2.2.1 Sistema de cores para categorização de permissões de arquivamento

O projeto RoMEO concebeu o seguinte sistema de cores (JENKINS et al., 2007) para categorizar as permissões de arquivamento de trabalhos em repositórios de e-prints:

a) **Verde**: permitido o arquivamento de *pre-prints* **e** de *post-prints*.

b) **Azul**: permitido o arquivamento **somente** de *post-prints*.

c) **Amarelo**: permitido o arquivamento **somente** de *pre-prints*.

d) **Branco**: nenhuma permissão concedida formalmente (necessária autorização caso-a-caso da editora).

Na nomenclatura do projeto, os termos *pre-print* e *post-print* significam o seguinte:

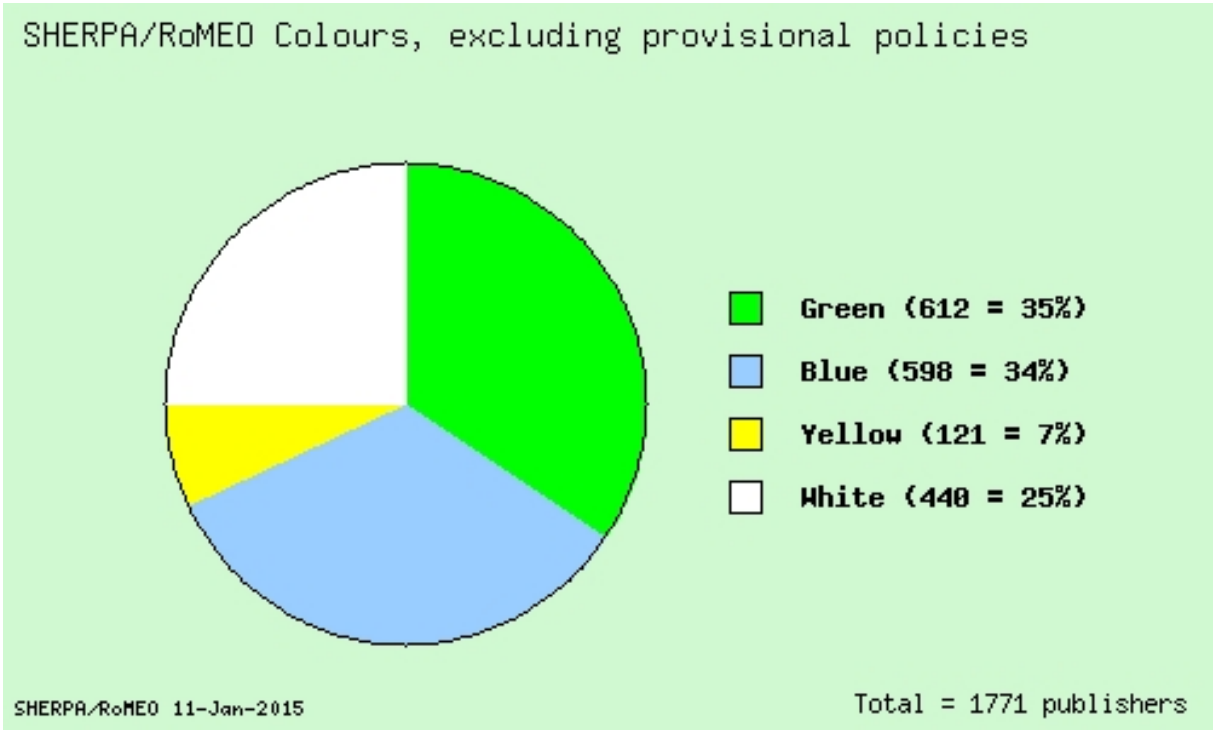
a) ***Pre-print***: Rascunho inicial do artigo, antes de passar pela revisão por pares. Contém o mesmo texto que (ou um texto anterior ao que) foi submetido para a revista, sem incluir nenhuma alteração que tenha sido sugerida pelos revisores.

b) ***Post-print***: Rascunho final do artigo, incluindo quaisquer mudanças sugeridas pelos revisores. O texto em si é o mesmo do artigo publicado pela editora, porém sem nenhum tipo de formatação ou editoração realizada pela mesma.

Ao passo que é evidente que o arquivamento de *post-prints* é crucial para ampliar o acesso ao resultado final da pesquisa, a importância do arquivamento dos *pre-prints* pode não estar aparente à primeira vista. Os *pre-prints* aumentam a qualidade da pesquisa antes que esta seja publicada, pois incentivam uma maior discussão durante o processo de preparação do texto. O arXiv[6], primeiro repositório digital, opera até hoje e é muito utilizado, principalmente por físicos, para receber uma ampla revisão informal e comentários preliminares da comunidade científica a respeito de seus artigos, antes de submetê-los para revistas (onde estes passarão por um processo formal de revisão por pares).

[6] <http://arxiv.org>

Figura 1 – Estatísticas do RoMEO em 11 de janeiro de 2015 sobre permissões concedidas por editoras para arquivamento de publicações. As editoras representadas pelas cores azul, verde e amarelo (75%) permitem algum tipo de arquivamento. As editoras em branco (25%) não possuem nenhuma política oficial que permita arquivamento.



Fonte: SHERPA/RoMEO (2015).

Como pode ser verificado na Figura 1, atualmente 35% das editoras catalogadas pelo RoMEO permitem o arquivamento tanto de *pre-prints* como de *post-prints*. Essa é a modalidade menos restrita de contrato, que felizmente tem se tornado a mais comum. O arquivamento somente de *post-prints* é permitido por 34% das editoras, e o somente de *pre-prints* é permitido por 7%. Um quarto (25%) das editoras não possuem nenhuma política oficial permitindo o arquivamento, exigindo desta forma que a permissão seja solicitada caso-a-caso para a editora.

2.2.2.2 Restrições adicionais e tempo de embargo

Em alguns casos, as editoras permitem o arquivamento se algumas restrições adicionais forem satisfeitas. O RoMEO também cataloga essas informações a respeito de cada editora e de cada revista, disponibilizando-as ao usuário de forma resumida.

Algumas editoras permitem que o *post-print* seja disponibilizado livremente em repositórios públicos após passado um determinado tempo que o artigo tenha sido publicado na revista. Esse é o denominado tempo de embargo. São comuns tempos de 6 meses a 4 anos, dependendo da editora e da revista.

O tempo de embargo também pode variar de acordo com o tipo de acervo onde o

trabalho será arquivado. Algumas editoras permitem arquivamento imediato (sem embargo) no repositório da instituição à qual o autor é filiado, porém exigem um tempo de embargo para arquivamento em outros repositórios (por exemplo o PubMed Central[7]). Outras permitem arquivamento sem embargo no site pessoal do autor, porém exigem um tempo de embargo para disponibilização por meio do repositório institucional.

Por esse motivo, uma boa implementação de repositório institucional deve suportar receber o texto completo do artigo e mantê-lo pendente, para que seja disponibilizado apenas após passado o tempo de embargo. O software DSpace, adotado neste trabalho, suporta esse recurso desde 2012 (DIGGORY et al., 2013). Desta forma, evita-se que o autor esqueça de disponibilizar o texto para o repositório, o que poderia ocorrer caso tivesse de esperar até a data do término do embargo para enviá-lo, e assegura-se que o texto estará disponível ao público tão breve quanto permitido.

2.2.3 Acesso aberto (*Open Access*) e a "via dourada"

O ato de disponibilizar artigos voluntariamente em repositórios abertos para que estes possam ser acessados livremente é apelidado de "via verde[8]" pelo movimento de acesso aberto (*open access*).

É crescente o número de revistas que adotam, por si próprias, um modelo de distribuição aberta. Essas revistas, denominadas de "revistas de acesso aberto", estão acessíveis a todos os usuários da Internet, e não somente a assinantes. Os artigos são publicados sob licenças livres, geralmente Creative Commons Attribution ou Creative Commons ShareAlike. Esse caminho para alcançar o acesso aberto é denominado de "via dourada" pelo movimento.

Harnad (2007 apud LEY, 2013) argumenta que a "via verde" é muito mais simples e eficaz, quando comparada à "via dourada", uma vez que a "via verde" trata da disponibilização de artigo por artigo, e não de toda a revista, como requer a "via dourada", que envolve operações muito mais elaboradas de editoração e publicação.

Existem revistas que adotam exclusivamente o modelo de acesso aberto, por exemplo as publicadas pela PLoS (*Public Library of Science*). Já em outras revistas, o autor pode optar ou não por publicar no modelo da "via dourada" — é o caso de muitas das revistas híbridas editadas pela Elsevier. Quando opta pela "via dourada", o autor precisa pagar uma taxa de processamento da publicação, que não é necessária caso o autor opte pelo modelo tradicional. Essa é uma maneira da editora custear suas operações, já que não arrecadará dinheiro com o acesso a artigos no modelo "via dourada".

Harnad (2008 apud LEY, 2013) defende que as universidades/instituições de pesquisa estabeleçam políticas/mandatos que tornem obrigatório o depósito de todas as versões finais

[7] <http://www.ncbi.nlm.nih.gov/pmc>

[8] É importante notar que a cor "verde" neste contexto nada tem a ver com o sistema de categorização do SHERPA/RoMEO.

de artigos com revisão por pares (postprints) de cada um de seus autores no repositório da instituição, imediatamente após sua aceitação para publicação. O acesso a esse depósito pode ser definido imediatamente como acesso aberto, se as condições/contratos de direitos autorais assim o permitirem, ou, caso contrário, pode ser definido como restrito/fechado, enquanto perdurarem as condições de embargo.

2.3 Interoperabilidade

Assim que a criação de repositórios digitais popularizou-se no meio acadêmico, houve a necessidade de padronizar políticas para criação de um RI para ampliar cada vez mais o acesso aberto à informação científica. Essa padronização permitiria que os repositórios existentes fossem "interoperáveis" entre si, ou seja, que, ao mesmo tempo, pudessem prover e coletar dados uns dos outros, ainda que utilizem diferentes softwares. Essa necessidade dá origem à *Open Archives Initiative* (OAI), que ocorreu no ano de 1999 com o objetivo principal de implantar uma política de interoperabilidade, dentre outros princípios, para a construção de repositórios digitais de acesso aberto à informação.

A partir da OAI, foi criado o protocolo OAI-PMH (*The Open Archives Initiative Protocol for Metadata Harvesting*) um protocolo baseado em HTTP e XML (OLIVEIRA; CARVALHO, 2009) que provê interoperabilidade entre repositórios por meio da coleta mútua de metadados. O protocolo OAI-PMH pode ser adaptado aos mais diversos padrões de metadados. Desse modo, os metadados tornam-se ferramenta fundamental para a consolidação da interoperabilidade entre os repositórios.

2.3.1 Formatos de metadados

Um registo de metadados é composto por um conjunto de atributos, ou elementos, necessário para descrever o recurso em questão. Por exemplo, um sistema de metadados comum nas bibliotecas — o catálogo da biblioteca — contém um conjunto de registros de metadados com elementos que descrevem um livro ou outro item de biblioteca: autor, título, data de criação ou publicação, a cobertura de assunto, bem como o número de chamadas especificando localização do item na prateleira (HILLMANN, 2005).

"Os benefícios da interoperabilidade bem sucedida estão amplamente documentados na literatura. Metadados dão significado semântico a dados aparentemente isolados, provendo um contexto ao processo de busca" (VIANA; MÁRDERO ARELLANO; SHINTAKU, 2005).

> O termo metadados frequentemente designa dados sobre dados, ou informação sobre informação. Os metadados estão categorizados em metadados descritivos, voltados para facilitar a descoberta, a identificação, a compreensão e a seleção de recursos; metadados administrativos que facilitam a gestão, o acesso e a preservação dos recursos digitais; e os metadados estru-

> turais que documentam a estrutura dos objetos e os relacionamentos entre objetos digitais. (SAYÃO; MARCONDES, 2009, p. 38–39)

Existem diversos padrões de metadados: Dublin Core, RIS, DIDL, ISO, MARC 21, METS, MODS, MTD-BR, BibTeX, dentre outros. O mais utilizado mundialmente, no âmbito dos repositórios digitais abertos, é o Dublin Core.

2.3.2 Dublin Core

Entre os esquemas[9] de metadados existentes, destaca-se o Dublin Core[10], que é utilizado em grande parte dos RI em todo o mundo.

> Existem vários esquemas de metadados com graus diferenciados de especificidade, porém o mais importante deles é o Dublin Core, considerado a língua franca para representação de recursos na web. O esquema é composto por 15 elementos que foram projetados para serem de simples compreensão e de fácil aplicação pelo próprio autor. Entretanto o esquema permite que diferentes comunidades façam adaptações e o customizem, adicionando elementos, redefinindo e ampliando a semântica deles, o que pode ser facilitado pelo uso de qualificadores, formando o que se chama de "perfil de aplicação" (SAYÃO; MARCONDES, 2009, p. 39).

O padrão de metadados Dublin Core é um padrão simples, mas eficaz, definido para descrever uma ampla gama de recursos de rede. Possui dois níveis: o simples e o qualificado. O Dublin Core simples compreende quinze elementos; o Dublin Core qualificado inclui três elementos adicionais (audiência, proveniência e detentor de direitos), bem como um grupo de refinamentos de elementos (também chamados qualificadores) que refinam a semântica dos elementos de maneiras que podem ser úteis na descoberta de recursos. A semântica do Dublin Core foi estabelecida por um grupo internacional e interdisciplinar de profissionais de biblioteconomia, ciência da computação, codificação de texto, comunidade de museus, e outras áreas afins (HILLMANN, 2005).

A Figura 2 apresenta um exemplo de arquivo no formato Dublin Core, descrevendo metadados de um artigo revisado por pares. A leitura do formato é bastante intuitiva — cada campo pertencente aos metadados é descrito por uma *tag* do tipo `<dcvalue>`.

2.3.3 RIS

O RIS é um formato de metadados desenvolvido originalmente pela empresa *Research Information Systems*. Esse formato é suportado por uma série de softwares para gerenciamento de referências bibliográficas. Os registros RIS podem ser obtidos a partir de diversos portais de acesso mantidos por editoras, tais como ScienceDirect, IEEE Xplore e SpringerLink.

---

[9] O termo "esquema" é uma tradução técnica do Latim *schema*, que significa "forma" ou "formato".

[10] <http://dublincore.org>

Figura 2 – Exemplo de metadados de um artigo revisado por pares no formato Dublin Core. Trata-se de um formato baseado em XML onde cada campo é descrito por uma *tag* <dcvalue>.

```
<dublin_core>
<dcvalue element="identifier" qualifier="systemid">101532xPUB304</dcvalue>
<dcvalue element="description" qualifier="pubcategory">C1 - Refereed journal articles
</dcvalue>
<dcvalue element="date" qualifier="issued">2010</dcvalue>
<dcvalue element="title" qualifier="publication">Quantifying nitrogen process rates
in a constructed wetland using natural abundance stable isotope signatures and stable
isotope amendment experiments</dcvalue>
<dcvalue element="description" qualifier="internalauthors">2</dcvalue>
<dcvalue element="description" qualifier="externalauthors">0</dcvalue>
<dcvalue element="title" qualifier="journalname">Journal of Environmental Quality
</dcvalue>
<dcvalue element="publisher" qualifier="none">American Society of Agronomy</dcvalue>
<dcvalue element="description" qualifier="volume">39</dcvalue>
<dcvalue element="description" qualifier="issue">6</dcvalue>
<dcvalue element="description" qualifier="startpage">2191</dcvalue>
<dcvalue element="description" qualifier="endpage">2199</dcvalue>
<dcvalue element="identifier"  qualifier="chrScopusID">2-s2.0-78650800393</dcvalue>
<dcvalue element="identifier" qualifier="uriconference">
10.2134/jeq2010.0067</dcvalue>
<dcvalue element="description" qualifier="comments">N/A</dcvalue>
<dcvalue element="contributor" qualifier="author">ERLER, Dirk</dcvalue>
<dcvalue element="contributor" qualifier="department">School of Environmental Science
&amp;amp; Management (ES&amp;amp;MDep)</dcvalue>
<dcvalue element="contributor" qualifier="author">EYRE, Bradley David</dcvalue>
<dcvalue element="contributor" qualifier="department">School of Environmental Science
&amp;amp; Management (ES&amp;amp;MDep)</dcvalue>
<dcvalue element="title" qualifier="none">Quantifying nitrogen process rates in a
constructed wetland using natural abundance stable isotope signatures and stable
isotope amendment experiments</dcvalue>
</dublin_core>
```

Fonte: Intersect Australia (2013).

Trata-se de um formato de texto puro, onde cada linha corresponde a um campo diferente. O significado do campo é dado por um identificador padronizado de duas letras que é colocado ao início da linha. Segue-se então um hífen e o valor do campo. A última linha de um registro RIS é sempre o campo ER, que não possui valor algum e significa *End of Reference* – Fim da Referência. A Figura 3 mostra um exemplo de entrada RIS contendo os metadados do mesmo artigo descrito pelos metadados Dublin Core apresentados como exemplo na Figura 2.

2.3.4 BibTeX

O BibTeX é um formato de metadados para citação de referências em trabalhos científicos escritos com o sistema de preparação de documentos LaTeX. O presente trabalho é um exemplo de documento escrito em LaTeX, e sua seção de referências foi gerada automaticamente

Figura 3 – Exemplo de metadados em formato RIS. Trata-se de um formato de texto simples, onde cada linha representa um campo.

```
TY  - JOUR
T1  - Quantifying Nitrogen Process Rates in a Constructed Wetland Using Natural
Abundance Stable Isotope Signatures and Stable Isotope Amendment Experiments
JO  - Journal of Environmental Quality
VL  - 39
IS  - 6
SP  - 2191
EP  - 2199
PY  - 2010/11//
AU  - Erler, Dirk V.
AU  - Eyre, Bradley D.
SN  - 0047-2425
DO  - 10.2134/jeq2010.0067
UR  - http://dl.sciencesocieties.org/publications/jeq/abstracts/39/6/2191
ER  -
```

Fonte: Elaborada pela autora.

de acordo com as regras da ABNT a partir de registros BibTeX.

O LaTeX é um software livre e tornou-se o padrão de fato para submissão de artigos para revistas e congressos em diversas áreas do conhecimento, principalmente em Física, Matemática, Estatística e Ciência da Computação, além de ser utilizado na editoração de livros e outros trabalhos científicos. Assim como ocorre com o RIS, um grande número de portais de acesso mantidos por editoras suporta exportar registros bibliográficos no formato BibTeX.

A Figura 4 apresenta um exemplo de entrada BibTeX contendo os metadados do mesmo artigo descrito pelos metadados Dublin Core apresentados como exemplo na Figura 2. Na mesma figura, é mostrada uma citação bibliográfica gerada automaticamente pelo software a partir dos metadados, seguindo as normas ABNT.

2.4 Implementação de repositórios

2.4.1 Politica de informação

A implementação de um repositório envolve, desde o seu início, a necessidade da elaboração de políticas e normas para depósito e publicação de conteúdos no mesmo.

Sayão e Marcondes (2009, p. 61) destacam que para o desenvolvimento de repositórios há a necessidade de definições de políticas institucionais. A definição de uma consistente política de informação é essencial para uma eficaz implementação de repositórios digitais. São essas políticas que irão regulamentar o funcionamento do RI, seu escopo/conteúdo, propósito, modos de depósito/submissão de documentos, direito do autor, dentre outras importantes

Figura 4 – Exemplo de metadados em formato BibTeX. A caixa abaixo dos metadados representa uma citação bibliográfica gerada automaticamente a partir dos metadados pelo software LaTeX.

```
@article{Erler20102191,
title = "Quantifying Nitrogen Process Rates in a Constructed Wetland Using Natural
Abundance Stable Isotope Signatures and Stable Isotope Amendment Experiments",
journal = "Journal of Environmental Quality",
volume = "39",
number = "6",
pages = "2191-2199",
year = "2010",
month = nov,
issn = "0047-2425",
doi = "10.2134/jeq2010.0067",
url = "http://dl.sciencesocieties.org/publications/jeq/abstracts/39/6/2191",
author = "Dirk V. Erler and Bradley D. Eyre",
}
```

ERLER, D. V.; EYRE, B. D. Quantifying nitrogen process rates in a constructed wetland using natural abundance stable isotope signatures and stable isotope amendment experiments. *Journal of Environmental Quality*, v. 39, n. 6, p. 2191–2199, nov. 2010. ISSN 0047-2425. Disponível em: <http://dl.sciencesocieties.org/publications/jeq/abstracts/39/6/2191>.

Fonte: Elaborada pela autora.

questões.

Todas as políticas voltadas para a informação no âmbito da implantação de RI perpassam pela política de povoamento do mesmo. O povoamento contempla questões que vão além do escopo técnico, um exemplo disso diz respeito às formas mais comuns de depósito/submissão de documentos em um repositório digital: realizadas/monitoradas por equipe capacitada responsável por esta atividade na instituição, ou o auto-arquivamento, realizado pelo próprio autor. Ambas as modalidades possuem limitações. No caso de ser depositado por equipe da instituição, existem as limitações referentes ao total de pessoas que podem ser dedicadas a esta tarefa. No caso de ser depositado pelo autor, a dificuldade se relaciona com toda a metodologia de submissão que compreende a uma série de passos, o que, muitas vezes, desencoraja os autores a depositarem seus trabalhos.

Ley (2013) afirma que:

> A política de povoamento, que se pretende eficaz, deve, pois, prever ações referentes aos itens expostos: tipologia documental; depósito voluntário ou compulsório; permissão/embargo de documentos; direito autoral; divulgação e marketing do sistema de informação, com vistas a pleitear um RI de ampla aceitação e povoado sistematicamente pela comunidade acadêmica a que se destina, servindo como aporte para a divulgação da produção científica da Universidade, de forma a agregar valor à instituição e a todos que contribuírem para sua divulgação, através da disponibilização de trabalhos no RI. (LEY, 2013, p.95)

Viana e Márdero Arellano (2006) listam importantes políticas que devem fazer parte da implementação de um RI, dentre as quais destacam-se: políticas de depósito/submissão de documentos; políticas relacionadas ao acesso à informação e políticas para envolvimento dos *stakeholders*.

Definir também uma política voltada para o acesso à informação é importante para definir os níveis e perfis de permissão e acesso no RI, de modo a implementar tanto o acesso aberto à comunidade, quanto o acesso restrito, voltado para documentos que só interessem a determinados setores da instituição.

As políticas para envolvimento dos *stakeholders* também são muito importantes. O termo *stakeholder*, que em inglês, significa, numa tradução literal, "a parte interessada" em algo, neste caso refere-se a toda a comunidade que tem interesse direto na construção e uso de um RI, ou seja, toda a comunidade envolvida na instituição: professores, alunos, gestores, etc. Essas políticas são fundamentais para definir estratégias no intuito de dirimir duvidas e contornar as dificuldades relativas a questão de direito do autor, hábitos e valores dos pesquisadores, barreiras tecnológicas etc.. Tratam-se de estratégias importantes para obter o comprometimento de todos os agentes institucionais (VIANA; MÁRDERO ARELLANO, 2006, p. 11).

Segundo Shearer (2003 apud LEY, 2013), o sucesso de um repositório é medido pelo uso do material nele contido e a atividade de depósito desse material está intimamente relacionada à satisfação percebida, constituindo uma das variáveis mais importantes na determinação da utilização do sistema de informação.Garantir o povoamento sistemático de um repositório não é uma tarefa simples e requer muitas ações de todos os envolvidos com a instauração do RI, que devem estar bem explicitadas na formulação da política de povoamento, para conduzir os ditames necessários ao sucesso da empreitada.

2.4.2 Infraestrutura

Para Viana e Márdero Arellano (2006, p. 3) a implementação de um RI pode ser realizada de uma forma simples: organizado numa estrutura hierarquizada que possa ser acessada via web e tenha metadados coletados através do protocolo OAI-PMH de modo que permita aos usuários, ao utilizarem qualquer mecanismo de busca da OAI, possam encontrar e recuperar o conteúdo do repositório.

O processo que os autores descrevem pode parecer simples, mas envolve diversas ações até que se chegue propriamente no ponto da implementação do RI. Sayão e Marcondes (2009, p. 31) destacam que a escolha adequada do software para implementar o RI deve ir além da escolha da melhor plataforma a ser utilizada, deve conter desde requisitos de hardware, software, até requisitos humanos, financeiros e metodológicos. Todas essas informações concorrem para que a instalação, plena operação e manutenção do software ocorram de forma efetiva.

- Hardware: disponibilidade dos requisitos mínimos de hardware – máquinas servidoras; memória, capacidade de armazenamento, processador etc.; infraestrutura de rede: roteadores, largura de banda etc.
- Software: disponibilidade dos software necessários: ambiente operacional; servidor Web, banco de dados, linguagens, ferramentas de indexação e de busca etc.
- Expertise técnica gerencial e metodológica: disponibilidade interna ou contratada de conhecimento e de experiência para a instalação e operação do software e para a gestão do sistema como um todo.
- Sustentabilidade: disponibilidade de fundos para planejar, implantar, operar e manter o repositório; deve ficar claro que os custos das plataformas de software e de hardware são geralmente previsíveis, entretanto os custos do planejamento total, da implementação, do treinamento da equipe, de eventuais obras e instalações, consultorias externas e de operação do sistema são dependentes do alcance e grau de sofisticação do projeto do repositório.

(SAYÃO; MARCONDES, 2009, p. 31–32)

Outro ponto fundamental destacado pelos autores diz respeito à Segurança da Informação. O sistema e a infraestrutura adequados devem assegurar a integridade física dos estoques de conteúdos digitais em casos de intervenções e acessos indevidos. Desse modo, é importante que o software escolhido ofereça controle de acesso, autenticação de usuários, níveis de permissão, bem como o uso de padrões como LDAP[11], X.509[12] e TLS[13].

Não há como falar de segurança da informação sem falar de *backup*, por isso é essencial que o sistema forneça métodos de *backup* e restauração de dados, bem como ofereça um consistente sistema de criptografia que proteja os dados em casos de transmissão de conteúdo.

2.4.3 Ferramentas para Criação de Repositórios

Existem dezenas de softwares (proprietários ou não) voltados para a implementação de repositórios digitais. Graças ao Movimento de Software Livre, existem hoje muitos softwares livres de alta qualidade para este fim.

Entre os diversos benefícios relacionados ao uso de software livre, destaca-se o fato de que, além da possibilidade de customização que este tipo de software oferece, existe toda uma comunidade de desenvolvedores mundiais comprometidos com a sua qualidade.

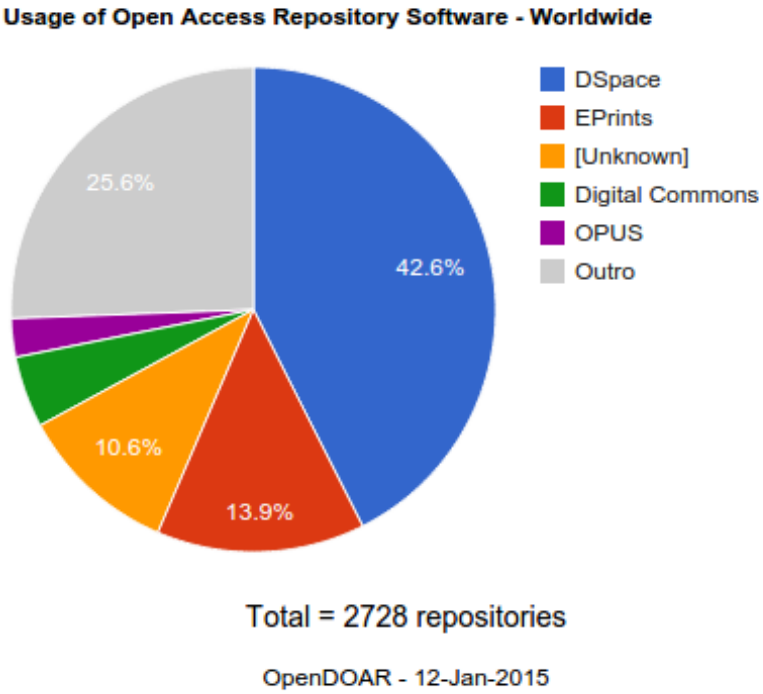
A OpenDOAR aponta que o software para implementação de repositórios digitais DSpace é o mais utilizado em todo o mundo, seguido pelo Eprints, como pode ser observado na Figura 5

[11] *Lightweight Directory Access Protocol* é um protocolo de rede que permite organizar e autenticar o acesso a recursos da rede de forma hierárquica.

[12] Na criptografia, X.509 é um padrão de certificado digital que provê uma infraestrutura e uma hierarquia de confiança em chaves públicas.

[13] *Transport Layer Security* é o protocolo mais utilizado para troca de dados criptografados na Internet.

Figura 5 – O DSpace e o Eprints são os softwares para repositórios digitais mais usados no mundo.



Fonte: OpenDOAR (2015).

2.4.3.1 DSpace

O DSpace[14] é o resultado de um esforço de desenvolvimento conjunto entre o Instituto Tecnológico de Massachusetts – MIT e a Hewlett Packard Corporation – HP iniciado em 2002. Foi desenvolvido como software de código aberto para gerenciar pesquisa, trabalhos acadêmicos e outros conteúdos publicados em um repositório digital, com o objetivo de prover armazenamento, acesso e preservação a longo prazo.

Como o número cada vez maior de usuários do DSpace, um grupo de instituições formou a Federação DSpace em 2004, que determinou a governança do desenvolvimento futuro para DSpace (DURASPACE, 2015).

Hoje é mantido pela DuraSpace[15], uma organização sem fins lucrativos que tem entre seus objetivos apoiar tecnologias livres que promovem o acesso durável e persistente a dados digitais.

Sayão e Marcondes (2009) destacam as principais características do DSpace:

- Características Técnicas
  - Ambiente Operacional – Unix, Linux, Windows

[14] <http://www.dspace.org>
[15] <http://www.duraspace.org>

  - Tecnologias usadas – Java, Tomcat Servlet Engine
  - Banco de Dados – PostgreSQL, MySQL, Oracle
  - Motor de Pesquisa – Lucene ou Google
  - Formatos aceitos – sem restrições
  - Extensível via Java API
- Padrões
  - Interoperabilidade – Protocolo OAI-PMH, Web Services, SRU/SRW
  - Esquema de metadados aceitos – Dublin Core qualificado
  - Identificadores – Handle System
  - Preservação digital – aderente ao modelo OAIS – Open Archive Information System; o software é focado no problema de preservação digital de longo prazo de materiais de pesquisa depositados.
  - Importação/exportação de dados
  - formato XML e padrão METS
- Características específicas
  - Implementa o conceito de comunidades
  - Voltado para repositórios institucionais
  - Foco em materiais para pesquisa e ensino.
  - Workflow para submissão de conteúdos
  - Interface web customizável

(SAYÃO; MARCONDES, 2009, p. 45)

O DSpace preserva e possibilita acesso fácil e aberto a todos os tipos de formatos digitais, incluindo texto, som, imagens e vídeos. Com uma crescente comunidade de desenvolvedores comprometidos com a expansão e melhoramento do software, cada DSpace instalado no mundo beneficia o próximo (DSPACE, 2015).

2.5 Iniciativas de sucesso

2.5.1 Apoio do IBICT

O IBICT[16] surgiu no início da década de 50 por meio da iniciativa da Unesco de sugerir à Fundação Getúlio Vargas (FGV) a criação de um centro nacional de bibliografia. Somente em 1976 passou a se chamar de Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), consolidando-se como órgão que coordenaria, no Brasil, as atividades de informação em C&T. Na década de 80, estabeleceu-se como Centro Brasileiro do ISSN e passou a ser o único membro no Brasil para atribuição do código ISSN.

O IBICT desenvolve uma série de ações voltadas para a consolidação do acesso aberto à produção científica no país.A formação e a capacitação dos recursos humanos para pesquisa na área de Ciência da Informação é um dos principais objetivos do Instituto, bem como ações voltadas para a criação de bibliotecas digitais e implantação de repositórios digitais (IBICT, 2015).Dentre as principais ações do IBICT, Kuramoto (2006) destaca:

[16] <http://www.ibict.br>

> 1. implantação da Biblioteca Digital de Teses e Dissertações usando o modelo OA;
> 2. absorção e transferência de conhecimentos sobre o modelo OA;
> 3. absorção, customização, divulgação e transferência de ferramentas de software para a construção de publicações eletrônicas e repositórios de acesso livre;
> 4. capacitação de técnicos quanto ao uso dessas ferramentas;
> 5. desenvolvimento e implantação do Portal de Repositórios e Publicações de Acesso Livre;
> 6. aquisição e distribuição de infra-estrutura tecnológica (hardware e software) às instituições de ensino superior e pesquisa para o desenvolvimento e implantação de repositórios institucionais e temáticos de acesso livre;
> 7. divulgação do Manifesto Brasileiro de Apoio ao Acesso Livre à Informação Científica (2005)
>
> (KURAMOTO, 2006, p. 101)

Dentre as atuações do IBICT destaca-se como marco no Brasil para o desenvolvimento do modelo Open Archives (OA) o Manifesto Brasileiro de Apoio ao Acesso Livre à Informação Científica[17] publicado em 2005.

> O Ibict, fortemente engajado em ações dessa natureza, lançou o "Manifesto Brasileiro de Apoio ao Acesso Livre à Informação Científica" (Ibict, 2005), integrando-se ao movimento que emana de diferentes países, tendo por base sobretudo a Declaração de Berlim e em harmonia com as idéias e ideais da International Federation of Lybrary Association and Institutions (Ifla) e da Organização para a Cooperação e Desenvolvimento Econômico (OCDE). O Manifesto tem por objetivo "mobilizar a comunidade científica e a sociedade brasileira em geral para se universalizar e democratizar a informação em ciência e tecnologia, condição fundamental para o desenvolvimento econômico e social de nosso país", bem como atuar como "forte agente de inclusão social". (SUAIDEN, 2006, p. 7)

2.5.2 Repositórios Nacionais

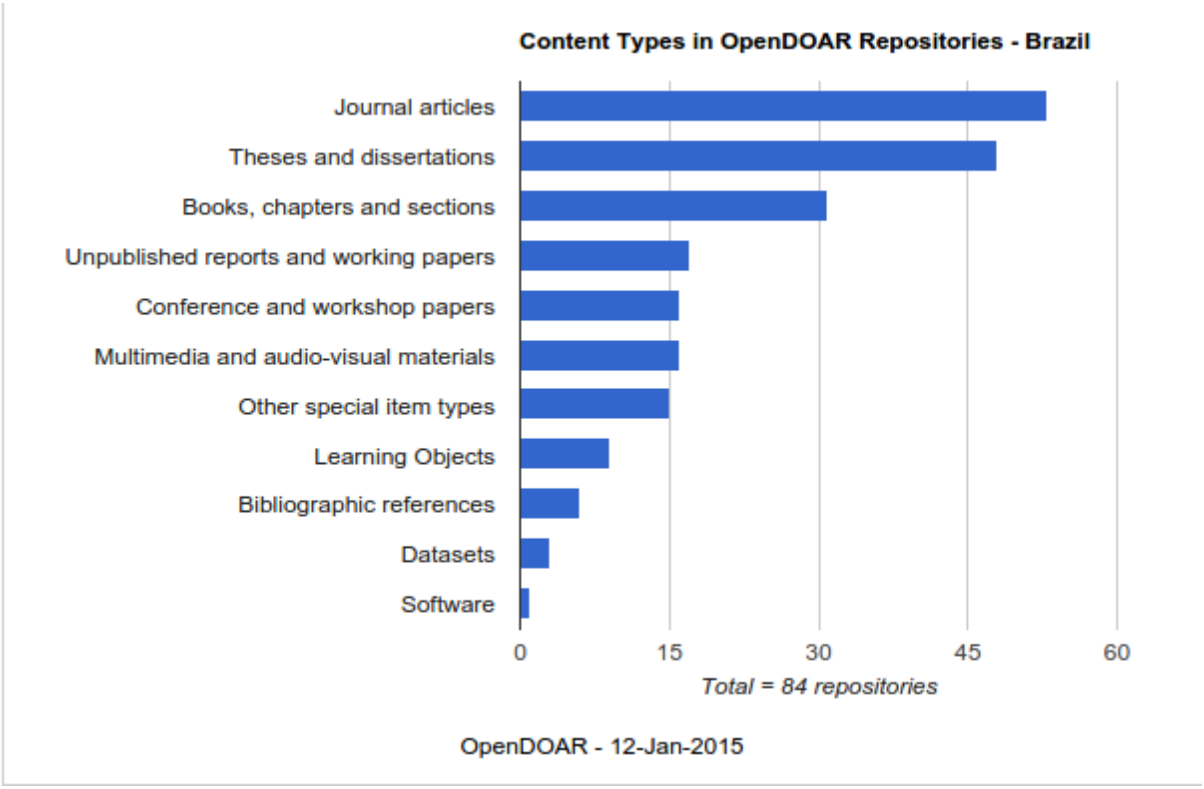
Esta seção destina-se a reunir algumas das principais iniciativas de RI nacionais e suas particularidades. Segundo o site da OpenDOAR[18], existem 84 repositórios abertos no Brasil, como pode ser visto na Figura 6.

2.5.2.1 UFBA

O RI[19] da Universidade Federal da Bahia – UFBA existe desde 2010 e foi desenvolvido com base no software livre DSpace. Trata-se de um RI com o propósito de divulgar a produção acadêmica desenvolvida no âmbito da UFBA, alinhado às suas políticas de informação, que

[17] <http://livroaberto.ibict.br/docs/Manifesto.pdf>
[18] <http://www.opendoar.org>
[19] <https://repositorio.ufba.br/ri>

Figura 6 – Existem 84 repositórios abertos no Brasil catalogados pelo OpenDOAR. Esses repositórios são destinados à preservação de diferentes tipos de acervos, sendo maioria aqueles destinados ao depósito de artigos científicos. Em segundo lugar ficam aqueles destinados a teses e dissertações.



Fonte: OpenDOAR (2015).

estão disponíveis para a consulta de qualquer pessoa e servem como base consistente e exemplo para implantação de um RI em qualquer instituição acadêmica do país.

Em consonância com o Movimento de Acesso Aberto e em conformidade com os anseios da comunidade científica mundial, o RI da UFBA possibilita a preservação e o acesso aberto à produção da universidade. Além disso, atende às recomendações governamentais de utilização de software livre, diminuindo o ônus de licenciamento. Agrega-se ao grande contingente de usuários do DSpace, composto por instituições mundialmente reconhecidas (UFBA, 2015).

2.5.2.2 CRUESP

O RI[20] do CRUESP (Conselho de Reitores das Universidades Estaduais Paulistas), lançado em outubro de 2013, foi concebido a partir da junção dos Repositórios Institucionais da USP[21], da UNESP[22] e da UNICAMP[23] (Biblioteca Digital da Produção Intelectual da USP, Repositório Institucional UNESP e Biblioteca Digital da Produção Intelectual e Científica da

[20] <http://cruesp.sibi.usp.br>
[21] <http://producao.usp.br>
[22] <http://repositorio.unesp.br>
[23] <http://reposip.unicamp.br>

UNICAMP). Todos estes utilizam a plataforma livre DSpace e adotam os padrões e normais internacionais de interoperabilidade (FERREIRA et al., 2013).

A integração dos repositórios se deu por meio do metabuscador Primo (um software proprietário), que permite ao usuário a realização da busca da produção do CRUESP a partir de uma única interface. Também permite a geração de indicadores do tipo: identificação das agências de fomento que mais subsidiam a pesquisa paulista, revistas mais requisitadas para publicação, co-autoria entre unidades, temas mais pesquisados, idiomas utilizados e ainda, texto em acesso aberto, restrito ou embargado.

O povoamento inicial deste repositório deu-se com artigos publicados em revistas indexadas pela Web of Science (WoS) e SciELO no período de 2008 a 2012 (FERREIRA et al., 2013).

2.5.2.3 UFRGS

O RI[24] da Universidade Federal do Rio Grande do Sul – UFRGS tem nome próprio: Lume. O nome foi escolhido por significar "manifestação de conhecimento, saber, luz e brilho".

É destinado às coleções digitais produzidas pela universidade e por outros documentos que são do interesse da mesma em centralizar sua preservação e difusão. Desse modo, o repositório tem por objetivo "reunir, preservar, divulgar e garantir o acesso confiável e permanente aos documentos acadêmicos, científicos, artísticos e administrativos gerados na Universidade, bem como às suas coleções históricas, e a outros documentos de relevância para a Instituição, que fazem parte de suas coleções, embora não produzidos por ela, maximizando a visibilidade e uso desses recursos" (UFRGS, 2015).

Os documentos digitais aceitos pelo repositório podem conter texto, imagem, vídeo e áudio, sendo a grande maioria de acesso aberto. No entanto, em alguns casos, o acesso é restrito à comunidade da UFRGS. O Lume também utiliza a plataforma livre DSpace. Os metadados utilizados para descrição dos documentos digitais seguem o padrão Dublin Core e o sistema CNRI Handle é usado para designar identificadores permanentes para cada documento disponível no repositório (UFRGS, 2015).

2.5.2.4 Fiocruz

O RI[25] da Fundação Oswaldo Cruz – Fiocruz também recebeu nome próprio: Arca. Este repositório foi criado em 2007 com o objetivo inicial de preservação da produção técnico-científica do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde – ICICT.

[24] <http://www.lume.ufrgs.br>
[25] <http://www.arca.fiocruz.br>

No ano de 2010 institucionalizou-se e teve seu escopo ampliado com o objetivo de dar maior visibilidade à produção intelectual por meio da informação proveniente dos institutos, escolas, centros, editoras e revistas da Fiocruz. Atualmente disponibiliza 16 comunidades: Casa de Oswaldo Cruz, Centro de Criação de Animais de Laboratório, Centro de Pesquisas Aggeu Magalhães, Centro de Pesquisas Gonçalo Moniz, Centro de Pesquisas René Rachou, Escola Nacional de Saúde Pública Sergio Arouca, Escola Politécnica de Saúde Joaquim Venâncio, Instituto Carlos Chagas, Instituto de Comunicação e Informação Científica e Tecnológica em Saúde, Instituto de Tecnologia em Fármacos, Instituto de Tecnologia em Imunobiológicos, Instituto Leônidas e Maria Deane, Instituto Nacional de Controle de Qualidade em Saúde, Instituto Nacional de Infectologia Evandro Chagas, Instituto Nacional de Saúde da Mulher, da Criança e do Adolescente Fernandes Figueira e Instituto Oswaldo Cruz.

As coleções de cada uma dessas comunidades são de natureza técnico-científica, tais como: artigos, livros, capítulos de livro, teses e dissertações, trabalhos de conclusão de curso, relatórios de pesquisa, relatórios institucionais, manuais, vídeos e a Revista RECIIS. Integra o grupo de repositórios brasileiros que utilizam a plataforma livre DSpace. É regido pela Política Institucional de Acesso Aberto à Produção Científica e Intelectual da Fundação Oswaldo Cruz, que visa garantir à sociedade o acesso gratuito, público e aberto ao conteúdo integral de toda obra intelectual produzida pela Fiocruz (HENNING et al., 2011).

2.5.3 Repositórios Internacionais

Nesta seção, são apresentados 3 importantes RI no mundo : o RI do *Massachusetts Institute of Technology* – MIT[26], o da *Harvard University*[27] e o repositório da universidade do Minho[28].

2.5.3.1 DSpace MIT

O repositório do MIT é descrito em sua página principal com um serviço de bibliotecas digitais para fornecer aos seus professores, pesquisadores e suas comunidades o suporte de armazenamento estável e de longo prazo para suas pesquisas, além maximizar a exposição do seu conteúdo para um público mundial.

O conteúdo do repositório inclui textos de conferências, imagens, artigos acadêmicos revisados por pares, pre-prints, relatórios técnicos, teses, documentos de trabalho, conjuntos de dados de pesquisa e muito mais. Sua coleção atual possui mais de 60.000 obras de alta qualidade e é reconhecido como um dos principais repositórios do mundo recebendo, em média, mais de 1 milhão de downloads por mês.

[26] <http://dspace.mit.edu>
[27] <https://osc.hul.harvard.edu/dash>
[28] <http://repositorium.sdum.uminho.pt>

Dentre os benefícios do repositório para os pesquisadores, o MIT destaca: evitar links quebrados ao citar sua pesquisa com URLs persistentes, proteger os dados, obter resultados de busca no Google, desfrutar de visibilidade mundial e distribuir rapidamente a sua investigação.

Vale lembrar a importante atuação do MIT para a disseminação do acesso aberto por meio de repositórios digitais tendo como grande colaboração a criação do DSpace, o software mais usado atualmente no mundo para a gestão de RI. Em março de 2000, a Hewlett-Packard Company (HP) concedeu US$ 1,8 milhão para as bibliotecas do MIT em uma colaboração de 18 meses para construir um repositório dinâmico para a produção intelectual em formatos digitais de organizações de investigação multidisciplinar. A HP Labs e a MIT Libraries lançaram o sistema mundialmente em 4 de novembro de 2002, sob os termos da licença de código aberto BSD, um mês após a sua introdução como um novo serviço das bibliotecas do MIT (SMITH et al., 2003).

Como um sistema de código aberto, o DSpace está disponível livremente para outras instituições que podem utilizá-lo como está, ou modificá-lo de acordo com suas necessidades.

2.5.3.2 DASH Harvard

O *Digital Access to Scholarship At Harvard* – DASH, é o serviço da *Harvard University* para compartilhar e preservar o suas pesquisas. Além dos artigos de periódicos acadêmicos (alvos de várias resoluções de acesso aberto da Harvard), a instituição incentiva que o DASH também seja utilizado para o auto-arquivamento dos manuscritos e materiais relacionados aos trabalhos dos pesquisadores (incluindo dados, imagens, áudio e arquivos de vídeo, etc.), uma vez que o serviço suporta uma variedade de formatos de arquivo (HARVARD, 2015).

Na página do DASH, a intituição ressalta ainda a visibilidade que o serviço oferece aos pesquisadores, em virtude da coleta dos metadados pelo Google Scholar e outros serviços de indexação, incentivando inclusive o depósito das primeiras versões dos trabalhos. O repositório é também descrito como uma estratégia para alcançar a missão da Universidade de compartilhar e preservar o conhecimento nela produzido.

A Harvard tem agora uma uma licença prévia não-exclusiva para distribuir os artigos publicados pelo corpo docente (nas unidades que possuem política de acesso aberto). Esses docentes precisam agir de acordo com a política de acesso aberto quando publicarem seus artigos, seja anexando um adendo ao acordo de publicação, ou obtendo uma renúncia de direitos por parte da editora. Feito isso, eles devem depositar a publicação no DASH (HARVARD, 2015). O DASH também utiliza o software DSpace.

2.5.3.3 RepositoriUM - Universidade do Minho

O RepositóriUM - repositório institucional da Universidade do Minho foi constituído com o objetivo de armazenar, preservar, divulgar e dar acesso à produção intelectual da Uni-

versidade do Minho em formato digital. Surgiu em meados de 2002 a ideia da criação de um repositório institucional destinado a toda produção intelectual da UMinho, mas somente em 2003, também utilizando o software DSpace, a instituição lançou o RepositoriUM (UMINHO, 2014).

Em sua fase piloto, o RepositoriUM foi povoado com teses e dissertações defendidas na Universidade do Minho nos anos anteriores, bem como de outros documentos (artigos, relatórios, *working papers*, etc.), onde apenas quatro comunidades experimentais o utilizavam.

A última fase do processo de criação do RepositóriUM ocorreu no dia 20 de Novembro de 2003. Nesse dia, e repositório foi aberto ao público, ficando acessível para toda a UMinho e para o público em geral. Na oportunidade, o RepositóriUM reunia 280 documentos e constituiu-se como a primeira instalação de destaque mundial de um repositório DSpace em língua portuguesa.

Entre os objetivos do RepositoriUM também estão o alcance e a visibilidade de sua produção científica e a preservação da memória institucional da Universidade do Minho. Neste ano de 2015, o repositório já registra, até o momento, mais de um milhão de downloads e de consultas em sua base de dados (UMINHO, 2014). A Universidade do Minho é participante ativa do movimento pelo acesso aberto à comunicação científica.

3 Método e desenvolvimento

3.1 Abordagem e tipologia da pesquisa

Para o desenvolvimento deste trabalho, foi realizada pesquisa de cunho bibliográfico com abordagem qualitativa sobre o tema, e também pesquisa exploratória das ferramentas escolhidas para implementação do projeto.

Segundo Terence e Filho (2006) a abordagem qualitativa numa pesquisa procura aprofundar a compreensão dos fenômenos estudados, sejam eles ações dos indivíduos, grupos ou organizações em seu ambiente e contexto social, interpretando-os sem se preocupar com representatividade numérica, generalizações estatísticas e relações lineares de causa e efeito.

A pesquisa bibliográfica foi de grande importância para o entendimento do escopo do projeto e para obter a fundamentação teórica necessária para o uso das ferramentas. Lima e Mioto (2007) destaca que "ao tratar da pesquisa bibliográfica, é importante destacar que ela é sempre realizada para fundamentar teoricamente o objeto de estudo, contribuindo com elementos que subsidiam a análise futura dos dados obtidos. Portanto, difere da revisão bibliográfica uma vez que vai além da simples observação de dados contidos nas fontes pesquisadas, pois imprime sobre eles a teoria, a compreensão crítica do significado neles existente".

Na fase de pesquisa exploratória, foram escolhidas as ferramentas com as quais o projeto seria desenvolvido: para a implementação de um repositório para a UFSCar, foi escolhida a plataforma DSpace. Collis e Hussey (2005) afirmam que em uma pesquisa do tipo exploratória, o foco é obter *insights* e familiaridade com a área do assunto para investigação mais rigorosa num estágio posterior, sendo que raramente fornece respostas conclusivas para problemas ou questões, mas indica qual pesquisa futura deve ser realizada.

Tabela 1 – Especificações de hardware e de software da máquina virtual contendo a instalação do DSpace utilizada para estudos.

Item	**Descrição**
Processador	Intel Xeon E5630 2.53GHz
Memória RAM	4 GB
Disco	70 GB
Sistema Operacional	Debian GNU/Linux versão "jessie"
Máquina virtual Java	OpenJDK 1.8.0_45
Servidor de aplicações	Tomcat 8.0.14
DSpace[1]	versão 5.3

[1] <https://github.com/dspace/dspace/tree/dspace-5_x>

Os primeiros passos após a instalação da ferramenta em local apropriado para estudo (máquina virtual fornecida pela Secretaria Geral de Informática da UFSCar – SIn, vide Tabela 1) foram:

a) Estudar as possibilidades de povoamento do repositório;

b) Estudar o padrão de metadados Dublin Core, utilizado pelo DSpace;

c) Estudar metodologias de exportação dos dados do provedor de dados escolhido;

d) Estudar técnicas de tratamento e de sincronização desses dados com o DSpace.

3.2 Desenvolvimento

A presente pesquisa está sendo desenvolvida com o apoio do NIT - Materiais[2] e da SIn/UFSCar e utiliza como procedimento metodológico a pesquisa-ação, uma das principais formas de abordagem qualitativa (TERENCE; FILHO, 2006). Em geral, a pesquisa-ação é um procedimento apropriado quando a pergunta da pesquisa refere-se a descrever e executar uma série de ações ao longo do tempo em determinado grupo, comunidade ou organização, de modo a compreender, enquanto membro de um grupo, como e por que a sua ação pode mudar ou melhorar o funcionamento de alguns aspectos de um sistema (COUGHLAN; COGHLAN, 2002).

Thiollent (2008 apud MIGUEL, 2007) define pesquisa-ação como um tipo de pesquisa de base empírica que é concebida e realizada em estreita associação com uma ação ou com a resolução de um problema coletivo e na qual os pesquisadores e participantes representativos da situação ou do problema estão envolvidos de modo cooperativo ou participativo.

A unidade-caso utilizada neste trabalho é a Universidade Federal de São Carlos – UFSCar. Fundada em 1968, e atualmente formada pelos campi São Carlos, Araras, Sorocaba e Lagoa do Sino (em Buri), é uma instituição de grande relevância no cenário nacional. Pode ser considerada a primeira Universidade Federal do Estado de São Paulo, uma vez que a Escola Paulista de Medicina (apesar de ter sido federalizada em 1956) foi elevada ao título de Universidade (UNIFESP) somente em 1994. Hoje, além do Instituto Federal de São Paulo (IFSP), é a única Instituição Federal de Ensino Superior presente no interior do Estado.

A UFSCar possui iniciativas incipientes envolvendo a criação de Repositórios Institucionais (RI), mas ainda não existe um RI oficial cuja finalidade seja catalogar a produção científica. A Secretaria de Educação a Distância (SEaD) possui um repositório digital destinado ao arquivamento de material didático, denominado Livre Saber – LiSa[3]. O LiSa é implementado com

[2] O NIT-Materiais é um núcleo de pesquisa ligado ao departamento de Engenharia de Materias da UFSCar, que atua na pesquisa de prospecção tecnológica e inteligência competitiva, suas metodologias, ferramentas e aplicações para suporte ao desenvolvimento sustentável de empresas, arranjos empresariais e instituições públicas <http://www.nit.ufscar.br>

[3] <http://livresaber.sead.ufscar.br>

a ferramenta DSpace, e é uma iniciativa de sucesso, sendo inclusive membro da Federação de Repositórios Educa Brasil – FEB[4]. Outra iniciativa de implantação de repositório parte da SIn – Secretaria Geral de Informática, que tem utilizado o software Alfresco[5] para a construção de um repositório voltado para o armazenamento de documentos administrativos dos diversos setores da universidade.

Vale ressaltar que, neste ano de 2015, a UFSCar montou, por meio da Portaria GR nº 1137/15, de 13 de fevereiro de 2015, um Grupo de Trabalho que tem dedicado esforços para a implantação do repositório oficial da instituição. O Grupo já concluiu a parte de elaboração de políticas do RI, as quais serão submetidas para análise do Conselho Universitário – ConsUni, órgão deliberativo máximo da universidade.

Esta pesquisa tem o objetivo de desenvolver uma sistemática de carga automatizada de metadados em repositórios institucionais, e a fonte escolhida para prover os metadados para essa carga foi a Plataforma Lattes (CNPq), uma vez que praticamente todo pesquisador brasileiro já está familiarizado com essa ferramenta, devido ao Currículo Lattes ser uma das principais formas de avaliação para concessão de bolsas e projetos de pesquisa por meio das agências de fomento do país, o que culmina no fato já mencionado neste trabalho sobre esta plataforma constituir-se hoje como o principal repositório de dados dos pesquisadores de todo o país.

Além de prover o primeiro ambiente de produção para desenvolvimentos subsequentes do RI, essa carga possibilitará uma grande visibilidade inicial do repositório. A existência de integração com ferramentas externas que já são utilizadas pelos pesquisadores facilitará a utilização do RI, uma vez que todos os metadados já estarão previamente copiados de seu Currículo Lattes. Outro fator importante é que os metadados do RI poderão ser indexados por ferramentas externas de terceiros (por exemplo, buscadores), implicando em um meio de divulgação bastante atrativo para que os autores busquem aumentar o impacto de suas pesquisas.

Os principais procedimentos realizados durante o desenvolvimento da pesquisa foram:

a) Instalação do DSpace;

b) Análise e modificação do código do scriptLattes;

c) Prospecção de uma metodologia de consulta ao *web service* da plataforma Lattes via protocolo SOAP, utilizando as linguagens Java e Python;

d) Desenvolvimento de um proxy[6] para compartilhar o acesso aos *web services* da plataforma Lattes;

e) Testes de importação (carga) de metadados no DSpace;

[4] <http://feb.ufrgs.br>
[5] <http://www.alfresco.com>
[6] Servidor intermediário

f) Desenvolvimento e integração com o DSpace de um sistema de autoridade de nomes baseado na Plataforma Lattes;

g) Desenvolvimento de scripts para extração, tratamento e sincronização de metadados da Plataforma Lattes com o DSpace.

3.2.1 Extração com scriptLattes

O primeiro mecanismo estudado como passível de ser utilizado para a extração de dados dos Currículos Lattes foi o scriptLattes[7] (MENA-CHALCO; CESAR JR., 2013). Trata-se de um script-robô escrito em linguagem Python, que interpreta os currículos formatados em HTML disponíveis publicamente no site da Plataforma Lattes, extraindo as informações de interesse. Inicialmente, o scriptLattes foi desenvolvido com o objetivo de gerar automaticamente listagens de publicações para sites de grupos ou de núcleos de apoio à pesquisa. Com o passar do tempo, foi sendo utilizado para extrair conjuntos maiores de dados e foi ganhando recursos de bibliometria.

O scriptLattes é capaz de exportar metadados de artigos em formato RIS, porém não contempla todos os campos especificados pelo Dublin Core, e nem todas as informações contidas nos currículos são incluídas no RIS. Porém, essa não seria uma grande barreira, já que o scriptLattes é livre e pode ser facilmente estendido.

Inicialmente, para esta pesquisa, foram realizadas algumas modificações no scriptLattes visando melhorar o desempenho da ferramenta, por exemplo com a atualização da biblioteca `urllib2`[8] para a `urllib3`[9], que suporta o recurso de *Keep-Alive* para conexões HTTP persistentes (MOGUL, 1995), possibilitando uma redução do tempo perdido em reconexões com o servidor. Entretanto, em conversas com autores do scriptLattes, estes demonstraram preocupação com a possibilidade de que o uso intensivo e constante da ferramenta pudesse causar transtornos junto ao CNPq.

Certo tempo depois, ao final do mês de abril de 2015, a Plataforma Lattes passou a incluir um CAPTCHA[10] para dificultar o acesso automático aos currículos (FAUSTO, 2015), citando em sua página principal uma preocupação com a "publicação indevida [de espelhos dos currículos] por sites não autorizados".

---

7 <http://scriptlattes.sf.net>

8 <https://docs.python.org/2/library/urllib2.html>

9 <https://pypi.python.org/pypi/urllib3>

10 *Completely Automated Public Turing test to tell Computers and Humans Apart*: Teste de Turing público completamente automatizado para diferenciação entre computadores e humanos, é um teste bastante utilizado na Internet para assegurar que não sejam realizados acessos automatizados a servidores, e geralmente é composto por imagens distorcidas de letras e números ou sons de difícil compreensão.

3.2.2 Extração por meio de *web service*

Mesmo antes da inclusão do CAPTCHA pelo CNPq (dezembro de 2014), ao procurar outras alternativas que não sobrecarregassem os servidores da Plataforma Lattes, verificou-se que esta dispõe de um convênio no qual fornece às Instituições de Nível Superior um acesso direto, por meio de um *web service*[11], para consulta de currículos em formato XML. Esses dados são completos — incluem todas as informações digitadas pelos pesquisadores ou buscadas automaticamente pelo Lattes. Como os dados consultados pelo *web service* estão em formato bruto, e são disponibilizados de forma oficial justamente com a finalidade de acesso automatizado por parte das instituições, o risco de sobrecarregar os servidores do CNPq torna-se praticamente nulo, de forma que optamos por investir nossos esforços nesta solução.

Em conversas na SIn/UFSCar, tomou-se conhecimento de que a UFSCar anteriormente já havia firmado convênio com o CNPq para acesso ao *web service* da Plataforma Lattes. Porém, isso havia ocorrido por orientação de uma empresa que prestava serviços para a universidade na época, com a qual o contrato havia sido encerrado há vários anos. O acesso ao *web service* era realizado pelo software proprietário fornecido por essa empresa, e portanto não havia muito conhecimento dos envolvidos a respeito do que consistia o convênio.

Uma vez esclarecida esta questão, a SIn/UFSCar reestabeleceu com sucesso o convênio com o CNPq para acesso aos *web services* da Plataforma Lattes. As credenciais de acesso foram concedidas, e foi implementado o software para consulta[12] a esses *web services* (Figura 7).

3.2.3 Proxy para acesso ao *web service* da Plataforma Lattes

A Plataforma Lattes é um banco de dados público mantido pelo governo que contém os currículos de pesquisadores brasileiros, e que pode ser acessado por qualquer um por meio de um navegador web. Metadados brutos dos currículos em formato XML também podem ser obtidos, mas o download automatizado (sem CAPTCHA) desses dados só é permitido oficialmente por meio de um serviço SOAP[13], que é disponibilizado somente para instituições brasileiras de pesquisa e ensino superior. No entanto, cada instituição só pode solicitar a liberação de acesso para um único endereço IP, por esse motivo, no âmbito deste trabalho, foi criado o cnpqwsproxy[14], um proxy cacheante[15] baseado em *OpenResty*[16] para os *web services* SOAP do CNPq – Plataforma Lates.

---

[11] Um *web service* é um conjunto de métodos (*web methods*) logicamente associados e chamados através de um servidor HTTP.

[12] <https://github.com/nitmateriais/cxf-repl>

[13] *Simple Object Access protocol* – SOAP, é um protocolo de comunicação baseado em XML para troca de informações estruturadas na implementação de *web services*.

[14] <https://github.com/nitmateriais/cnpqwsproxy>

[15] *Cache* é uma cópia local, ou próxima de quem requisita os dados, das informações mais acessadas dentre um conjunto de dados que está distante ou cujo acesso é lento.

[16] *OpenResty* é uma plataforma composta pelo nginx <http://nginx.org>, LuaJIT <http://luajit.org>, e alguns módulos de extensão.

Figura 7 – Exemplo de consulta ao *web service* da Plataforma Lattes via protocolo SOAP utilizando a linguagem Python.

```
import suds, base64, io, zipfile
# Conexão ao web service
c = suds.client.Client('file:WSCurriculo.wsdl')

# Exemplo de consulta do Identificador CNPq por nome e data de nascimento.
c.service.getIdentificadorCNPq(cpf='',nomeCompleto='NOME PESSOA',
                               dataNascimento='00/00/0000')
# Resulta em ⟶
'1234567890123456'

# Exemplo de consulta do Identificador CNPq por CPF.
c.service.getIdentificadorCNPq(cpf='12345678901',nomeCompleto='',dataNascimento='')
# Resulta em ⟶
'1234567890123456'

# Obtenção de currículo Lattes em arquivo XML compactado (zip).
xmlz = zipfile.ZipFile(io.BytesIO(base64.b64decode(
    c.service.getCurriculoCompactado(id='1234567890123456')
)))
# Descompactação do arquivo XML.
xmlz.read(xmlz.namelist()[0])
# Resulta em ⟶
"""
   <?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
   <CURRICULO-VITAE SISTEMA-ORIGEM-XML="LATTES_OFFLINE"
    DATA-ATUALIZACAO="27052013" HORA-ATUALIZACAO="094109"'
   [...]
"""
```

Fonte: Elaborada pela autora.

Dentre os principais benefícios alcançados pelo proxy, destacam-se:

a) Permite que a instituição gerencie sua própria listagem interna de endereços IP que podem acessar o serviço web.

b) Assegura que múltiplos aplicativos da mesma instituição acessando o serviço web não causem uma sobrecarga significativa nos servidores do CNPq, fazendo cache das respostas sempre que possível.

c) Preserva a compatibilidade com quaisquer aplicativos existentes. Mudar o endereço do serviço web no arquivo WSDL[17] ou sobrescrever a resposta do servidor DNS[18] usando o arquivo `/etc/hosts` (em um sistema compatível com UNIX) é suficiente para fazer com que um aplicativo preexistente utilize o proxy.

[17] *Web Services Description Language*: Linguagem para Descrição de Serviços Web.
[18] *Domain Name System*: Sistema de Nomes de Domínio.

3.2.4 Autoridade de Nomes – lattesAuthority

Em Biblioteconomia, catálogo decisório de autoridade de nomes (ou de identidade) é um instrumento que permite padronizar a forma como são registrados os autores das obras de um acervo, resolvendo imprecisões causadas pela existência de homônimos ou por divergências na forma de citar um autor. Essa padronização pode ser alcançada escrevendo-se o nome de um mesmo autor sempre exatamente da mesma forma, caso no qual é necessário, ainda, adotar alguma marcação especial para distinguir homônimos entre si, ou então por meio de códigos de identificação (numéricos ou alfanuméricos) únicos para cada autor.

Na literatura nacional, a pesquisa de Corrêa et al. (2012) é uma das poucas que abordam o tema em contexto similar ao deste trabalho, utilizando o Currículo Lattes como consulta para a criação de um catálogo decisório voltado para a padronização dos metadados, especialmente no que se refere ao nome dos autores, para facilitar o trabalho de catalogação de itens em um RI.

As autoras justificam a necessidade de criação de um catálogo decisório de autoridade por causa das inconsistências encontradas nas citações bibliográficas dos autores. Para evitar as inconsistências, a forma de escrever o nome de cada autor foi padronizada e registrada, de modo que a equipe que trabalha na revisão do repositório tenha políticas para se orientar, uma vez que "só se tem precisão na recuperação da informação quando não existem dúvidas em sua representação" (CORRÊA et al., 2012, p. 36).

No presente trabalho, identificamos um grande potencial para estender essa ideia e adotar a própria Plataforma Lattes como autoridade de nomes, utilizando uma modalidade de catalogação baseada em códigos de identificação numéricos para os autores. Isso deve-se aos seguintes fatores:

a) Todo currículo na Plataforma Lattes é identificado pelo ID CNPq, um identificador numérico único que, internamente, é ligado univocamente ao registro civil (CPF) do portador do currículo.

b) Existe na Plataforma Lattes uma funcionalidade denominada "identificar coautores", que insere links para os currículos dos coautores no currículo de quem utilizou o recurso. No XML obtido por meio do *web service*, esses links são representados pelo ID CNPq dos coautores.

Em resumo, pressupondo que o portador do currículo esteja listado como coautor de cada uma de suas próprias publicações, é possível atrelar ao menos um código identificador de autor (o do próprio portador do currículo) a cada trabalho. Para currículos que utilizem o recurso "identificar coautores" da Plataforma Lattes, também é possível atrelar os demais identificadores diretamente.

Para implementar o suporte à Plataforma Lattes como autoridade de nomes, tomou-se

como base uma integração já existente com o ORCID[19], introduzida na versão 5 do DSpace. Essa integração foi desenvolvida pela Universidade do Missouri (EUA) em parceira com a empresa Atmire[20], tradicional prestadora de serviços relacionados ao DSpace, em um programa patrocinado pela Fundação Alfred P. Sloan.

Neste trabalho, implementou-se um novo módulo de autoridade capaz de conviver lado a lado com o módulo ORCID original. O módulo Lattes reutiliza a mesma infraestrutura de base de dados Solr[21] e consulta a serviços REST[22] do módulo ORCID, porém:

a) Foi alocado o prefixo "`will be generated::lattes::`", seguindo a mesma convenção do original "`will be generated::orcid::`", para indicar um ID CNPq alimentado externamente, que esteja pendente de ser indexado pelo DSpace.

b) A interface do módulo Lattes com o indexador de autoridades do DSpace, que gera o identificador final preenchido no atributo "`authority`" dos metadados, foi adaptada para utilizar o formato "`lattes_`" seguido pelo ID CNPq, por exemplo "`lattes_0920066398294419`" para identificar o Prof. Dr. Targino de Araújo Filho. Antes da adaptação, todos os módulos de autoridade geravam identificadores aleatórios no formato UUID[23], similares a este: "`de305d54-75b4-431b-adb2-eb6b9e546014`". A preservação do ID CNPq no identificador final contribui para uma maior riqueza de informações disponíveis para colheita via OAI.

c) Foi desenvolvido um serviço REST em linguagem Python denominado lattesAuthority[24], que alimenta o módulo de autoridade com o nome completo, nomes alternativos em citações bibliográficas e instituição (obtida a partir do endereço profissional informado no Currículo Lattes) de uma pessoa. O serviço REST provê buscas por ID CNPq e por nome.

Além de acomodar adequadamente os identificadores CNPq extraídos dos Currículos Lattes, o módulo de autoridade permite que os autores sejam pesquisados por nome na interface de edição de metadados do DSpace, como mostra a Figura 8. Para implementar esse recurso, foi necessário utilizar a página de busca do servidor `buscatextual.cnpq.br`, que não é originalmente destinada ao acesso automatizado por software. Isso deve-se ao fato de que o *web service* do CNPq disponibiliza apenas uma busca por nome completo e data de nas-

[19] *Open Researcher and Contributor ID*: Identificador Aberto para Pesquisadores e Colaboradores, iniciativa que provê uma identificação única para autores, além de um cadastro de publicações similar a um Currículo Lattes internacional.

[20] <http://atmire.com>

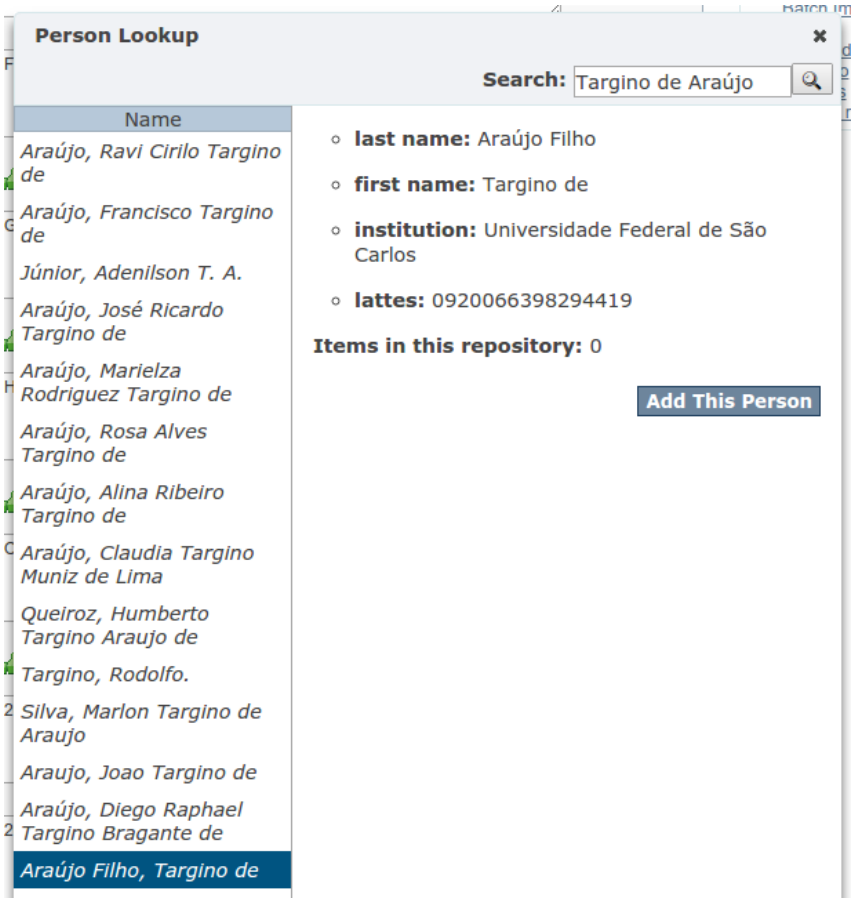
[21] Solr <http://lucene.apache.org/solr> é uma solução completa para indexação e busca de informações, construída sobre uma base de dados não-relacional extremamente escalável.

[22] *REpresentational State Transfer*: Transferência de Estado Representacional – estilo arquitetural para a construção de serviços web escaláveis.

[23] *Universally Unique IDentifier*: Identificador Universalmente Único, é um padrão que representa de forma alfanumérica um código identificador de 128 bits.

[24] <https://github.com/nitmateriais/lattesAuthority>

Figura 8 – A autoridade de nomes baseada na Plataforma Lattes desenvolvida no âmbito deste projeto foi totalmente integrada ao DSpace, e pode ser utilizada de forma independente dos demais módulos. Os autores podem ser pesquisados por nome, por meio da própria interface do DSpace.



Fonte: Elaborada pela autora.

cimento, mas não uma busca por parte de um nome, talvez em uma tentativa de precaver-se contra a obtenção do ID CNPq de homônimos por parte dos usuários do serviço.

Outra dificuldade enfrentada foi o fato do serviço `buscatextual.cnpq.br` não retornar o ID CNPq das pessoas encontradas pela busca, porém um outro identificador interno (em um formato similar a "`K4783652E6`") que o CNPq não garante que seja fixo para determinada pessoa. Como o acesso aos currículos por meio do identificador da busca textual só pode ser realizado por meio de página protegida por CAPTCHA, a forma encontrada para convertê-lo em um ID CNPq foi consultar um serviço interno de busca por nome exato (que, entretanto, não requer que seja fornecida a data de nascimento da pessoa) utilizado pela interface de edição do Currículo Lattes (servidor `wwws.cnpq.br`).

O ideal, no entanto, seria que caso esta ferramenta torne-se popular, o CNPq disponibilize a busca por nome diretamente em seu *web service*, sanando as dificuldades mencionadas. Devido ao grande número de consultas que precisam ser realizadas aos servidores do CNPq para efetuar uma busca por nome, foi implementada uma segunda camada de cache direta-

mente no serviço lattesAuthority, que armazena os resultados das buscas mais recentes, para o caso destas repetirem-se. Além disso, foi alterado o comportamento original da janela de busca de autores da interface do DSpace, fazendo com que ela efetue a busca apenas após um clique no botão, em vez de pesquisar continuamente conforme as letras são digitadas.

Os módulos de autoridade de nomes ORCID e Lattes podem estar ativos simultaneamente em uma mesma instalação do DSpace. A limitação existente é que apenas um dos módulos pode ficar acessível por meio da interface de busca por nome do DSpace.

3.2.5 Protocolo SWORDv2

Ao buscar uma tecnologia que pudesse prover o povoamento do DSpace a partir de um grande volume de dados extraídos da Plataforma Lattes, optou-se por utilizar o protocolo SWORDv2 nesta pesquisa.

SWORD – *Simple Web-service Offering Repository Deposit* – é um protocolo que permite, dentre outras coisas, interoperabilidade por meio de depósitos remotos entre repositórios e outros sistemas. É um protocolo baseado no *Atom Publishing Protocol*, um tipo de interface comum para depósito de itens em repositórios. Sua primeira versão foi criada em 2007 pelo JISC – *Joint Information Systems Comitte*, UK (ALLINSON; FRANÇÓIS; LEWIS, 2008).

O SWORD já está integrado a softwares de gestão de repositórios como o DSpace, Eprints e Fedora. No DSpace, o `dspace-swordv2` é o módulo que implementa a versão mais recente do protocolo (SWORDv2).

O SWORDv2 é uma ferramenta dinâmica que faz interface com diferentes sistemas, por exemplo, por meio dele pode-se implementar depósito de itens em um repositório diretamente a partir do software Microsoft Word, de um serviço de e-mail, ou até mesmo da rede social Facebook.

A versatilidade desse protocolo já é conhecida no meio acadêmico que estuda a implementação de repositórios abertos. Kuramoto (2011) já afirmava o potencial da ferramenta como interface de depósito entre repositórios e a Plataforma Lattes: "[...] considerando que o DSpace já tem integrado o protocolo SWORD, os RI poderiam ser interoperáveis com a plataforma Lattes. Isto, certamente, facilitaria ao pesquisador a atualização do seu curriculum Lattes e o povoamento do RI de sua instituição".

O SWORDv2, além de permitir o depósito dos metadados da Plataforma Lattes, poderá ser utilizado para a implementação automatizada do depósito do texto completo, nos casos em que isto for legalmente possível.

Para possibilitar o depósito dos itens em formato Dublin Core qualificado com suporte a controle de autoridade, o módulo SWORDv2 do DSpace foi modificado e estendido para

suportar o esquema XML DIM[25]. Originalmente, além do esquema Atom exigido pelo padrão SWORD, o módulo suporta apenas metadados expressos no esquema XML dcterms.

3.2.6 Conjunto de scripts synclattes

O synclattes[26] é um conjunto de scripts desenvolvidos ao longo desta pesquisa para extração, tratamento e sincronização de metadados do Currículo Lattes com o DSpace. Cada script executa uma etapa bem definida do processo, produzindo e gravando as informações para a próxima etapa em uma base de dados comum a todos os scripts.

A base de dados comum a todos os scripts foi construída com o PostgreSQL[27] versão 9.4, adotando uma abordagem semi-relacional. Os metadados são representados em formato JSONB[28], fornecendo a flexibilidade de um esquema livre. O PostgreSQL suportaria representar os metadados diretamente em XML, porém o formato JSONB pode ser indexado mais facilmente e permitir um maior desempenho. As demais informações da base de dados são representadas de forma relacional.

A Figura 9 apresenta uma representação conceitual, de alto nível, dessa base de dados. Ao Identificador CNPq é atribuída uma relação com a entidade pessoa do ERP[29] da UFSCar. Duas entidades básicas são manipuladas pelos scripts – item e revisão.

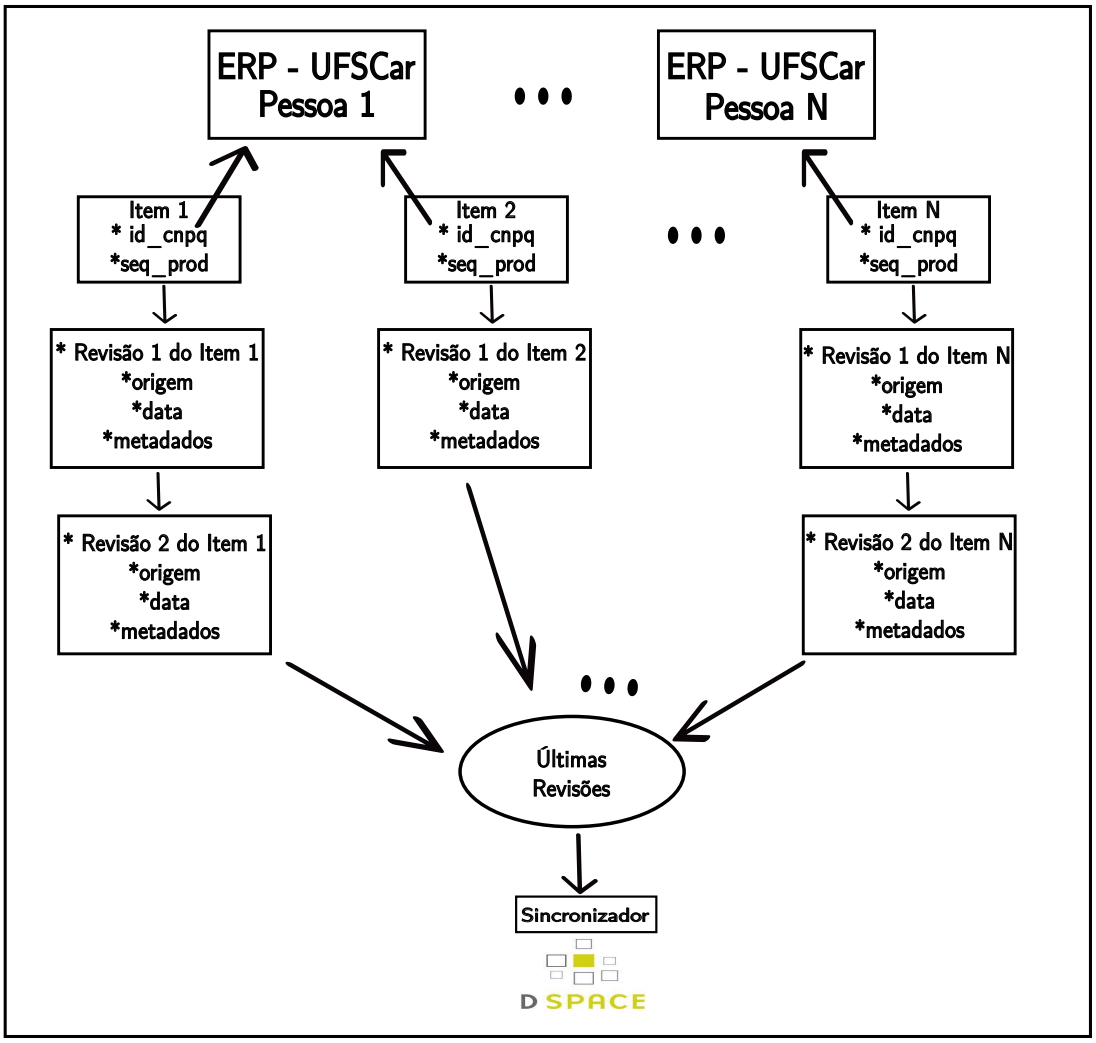
Um item representa uma entrada de produção científica existente no Currículo Lattes de certa pessoa. Sua chave primária[30] é composta pelo ID CNPq da pessoa e pelo atributo "sequência–produção" extraído do currículo em formato XML. Cada item pode ter mais de uma revisão. Cada revisão é, basicamente, uma versão dos metadados daquele item. Uma nova revisão é gerada toda vez que o autor editar as informações daquele item em seu Currículo Lattes, ou sempre que alguma ferramenta de tratamento alterar os metadados do item. O campo "origem" indica se a revisão é proveniente de uma alteração do usuário ou de alguma ferramenta de tratamento de dados.

A revisão mais recente de cada item é agregada a um conjunto denominado "últimas revisões" que, na base de dados, é implementado utilizando uma visão materializada (*materialized view*). Apenas as revisões pertencentes a esse conjunto são utilizadas na fase de sincronização. Entretanto, preservar as demais revisões na base de dados é importante para controle de histórico, depuração de falhas, e maior desacoplamento entre diferentes scripts.

Todos os scripts foram implementados em linguagem Python, utilizando a biblioteca

[25] *Dspace Internal Metadata*: Metadados Internos do DSpace, esquema XML que é utilizado internamente pelo DSpace para representar metadados.
[26] <https://github.com/nitmateriais/synclattes>
[27] <http://postgresql.org>
[28] *JavaScript Object Notation – Binary*: JSON representado em formato binário.
[29] ERP – *Enterprise Resource Planning* é um termo genérico para um software formado por um conjunto de aplicativos integrados voltados para a gestão institucional.
[30] Chave primária é o conjunto de campos que representam univocamente uma entrada em certa tabela.

Figura 9 – Diagrama conceitual da representação de dados utilizada pelo synclattes.



Fonte: Elaborada pela autora.

SQLAlchemy[31], que fornece tanto um ORM[32] flexível como uma linguagem de expressões para a composição de consultas SQL.

Os scripts, descritos em sua ordem natural de execução, são:

a) **extract**: Extrai os currículos a partir do *web service*, seleciona os dados de interesse com o auxílio da biblioteca lxml[33] e do padrão XPath. Transcodifica os dados para Dublin Core qualificado representado no formato JSONB. Alguns metadados que não são diretamente representáveis no Dublin Core, mas que são úteis para as etapas posteriores, são inseridos como elementos ou esquemas ocultos (cujos nomes são iniciados por *underline* para indicá-lo). Na Tabela 2 é possível verificar os campos do arquivo XML do Lattes que foram utilizados para a composição do arquivo no

31 <http://sqlalchemy.org>
32 *Object-Relational Mapping*: Mapeamento Objeto-Relacional é o nome dado a métodos utilizados para se mapear entidades da álgebra relacional em objetos da programação orientada ao objeto.
33 <http://lxml.de>

formato Dublin Core qualificado.

Tabela 2 – Mapeamento dos campos do XML do Lattes utilizados para a composição dos campos Dublin Core Qualificado no DSpace

XML – Currículo Lattes	**Dublin Core Qualificado**
<TRABALHO-EM-EVENTOS> ou <ARTIGO-PUBLICADO>	dc.type
<AUTORES NOME-COMPLETO-DO-AUTOR=❒/>	dc.contributor.author
<DADOS-BASICOS-DO-TRABALHO ou -ARTIGO ANO-DO-TRABALHO ou -ARTIGO=❒/>	dc.date.issued
<DADOS-BASICOS-DO-TRABALHO ou -ARTIGO DOI=❒/>	dc.identifier.uri
<DADOS-BASICOS-DO-TRABALHO ou -ARTIGO HOME-PAGE-DO-TRABALHO=❒/>	dc.identifier.uri
<DETALHAMENTO-DO-ARTIGO ISSN=❒/>	dc.identifier.issn
<DETALHAMENTO-DO-TRABALHO ISBN=❒/>	dc.identifier.isbn
<DETALHAMENTO-DO-TRABALHO ou -ARTIGO VOLUME=❒ FASCICULO=❒ PAGINA-INICIAL=❒ PAGINA-FINAL=❒/>	dc.identifier.citation
<DADOS-BASICOS-DO-TRABALHO ou -ARTIGO IDIOMA=❒/>	dc.language.iso
<DETALHAMENTO-DO-TRABALHO ou -ARTIGO TITULO-DOS-ANAIS-OU-PROCEEDINGS ou -DO-PERIODICO-OU-REVISTA=❒/>	dc.relation.ispartof
<PALAVRAS-CHAVE PALAVRA-CHAVE-n=❒/>	dc.subject
<AREA-DO-CONHECIMENTO NOME-x=❒/>	dc.subject.classification
<DADOS-BASICOS-DO-TRABALHO ou -ARTIGO TITULO-DO-TRABALHO ou -ARTIGO=❒/>	dc.title

b) **deduplicate**: Detecta duplicatas, ou seja, trabalhos que foram especificados no currículo de mais de um de seus coautores. No caso de DOI idêntico, dois trabalhos são sempre considerados duplicatas. No caso de DOI diferente, nunca são considerados. Quando não é especificado um DOI, é realizada uma busca por similaridade de títulos (OKAZAKI; TSUJII, 2010) utilizando alguma métrica *n-gram* configurada pelo usuário (por exemplo, a métrica *jaccard*). Caso os títulos forem muito similares, a lista de autores for minimamente similar, o ano de publicação for o mesmo, e as entradas pertencerem a currículos diferentes, estas são consideradas duplicatas. Atualmente, o nome do periódico ou congresso não faz parte da comparação, porém poderia ser incluído em trabalhos futuros.

c) **electmaindup**: Para cada grupo de duplicatas registrado na base de dados pelo script anterior, pontua e escolhe uma entrada principal, com base em um índice que leva em conta o vínculo do portador do currículo com a instituição (por exemplo, servidor ou estudante), a existência ou não de um DOI, o número de entradas de

autoridade (ID CNPq) especificadas para os coautores, a existência da "flag de relevância" do Currículo Lattes (que indica ser um dos cinco trabalhos mais relevantes do autor), e a variedade dos tipos de metadados especificados no currículo (palavras-chave, áreas de atuação, etc.).

d) **authoritymix**: Agrega à entrada principal de cada grupo de duplicatas as informações de autoridade (ID CNPq) especificadas em outras entradas do mesmo grupo. Para cada autor cujo ID CNPq não esteja especificado na entrada principal, procura alguma entrada secundária na qual ele esteja especificado, casando o nome do autor por meio da menor distância Levenshtein (1966).

e) **sync**: Sincroniza a entrada principal de cada grupo de duplicatas com o DSpace por meio do protocolo SWORDv2. Quando vários itens já existentes e previamente separados no repositório são mesclados, elege algum deles e remove (via *withdraw*) os demais. Recusa-se a remover automaticamente itens para os quais já tenha sido adicionado texto completo no DSpace. Quando itens previamente mesclados são separados, reaproveita itens que tenham sido criados e depois removidos (*reinstate*). As operações *withdraw* e *reinstate* não são suportadas pelo protocolo SWORDv2, motivo pelo qual foram implementadas utilizando a API REST do DSpace.

O comportamento dos scripts pode ser controlado também por meio das *flags* "`nofetch`" e "`nosync`", que desabilitam, respectivamente, a extração e a sincronia de determinado item. Desta forma, torna-se possível:

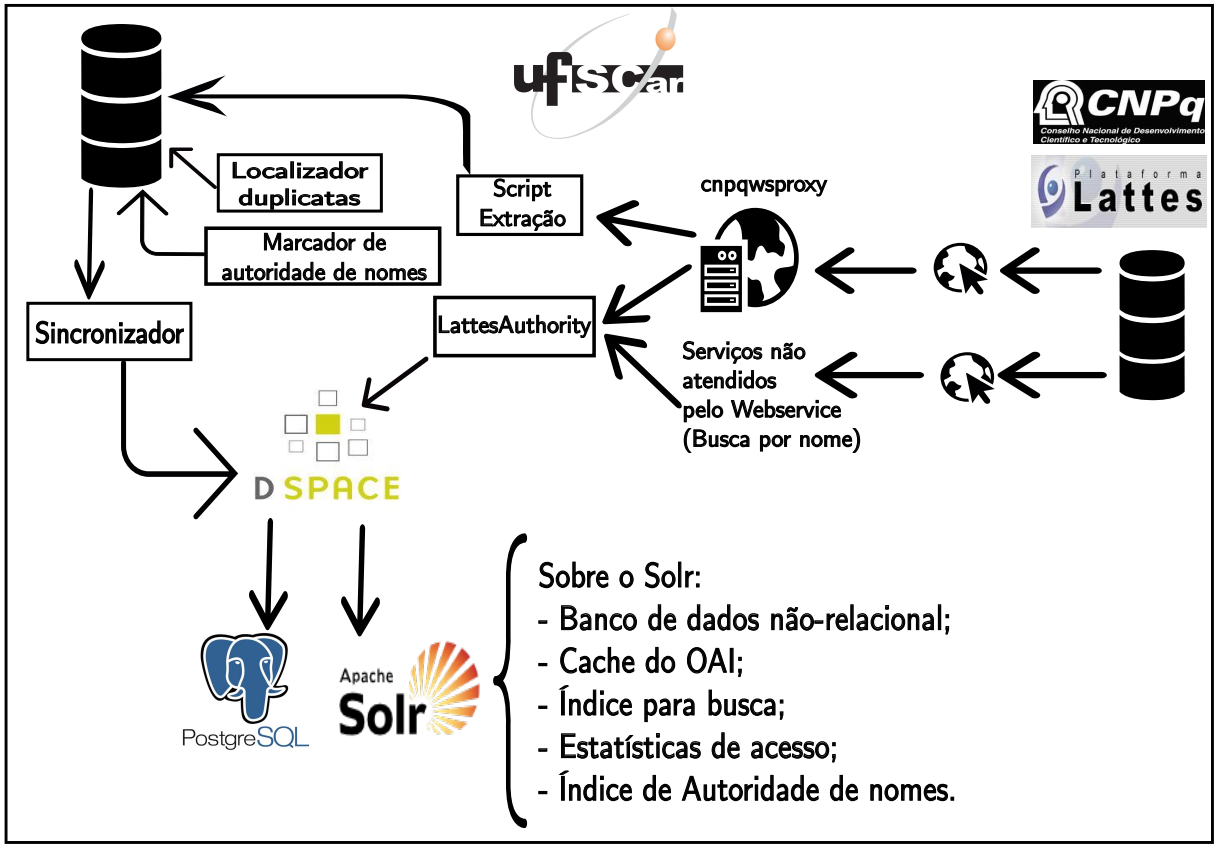
a) Que o autor indique itens pertencentes a seu Currículo Lattes que ele não deseje, por algum motivo, que sejam incluídos no RI.

b) Que o bibliotecário exclua da extração itens que sejam muito problemáticos devido a grandes inconsistências, nos casos em que não exista colaboração por parte do portador do currículo.

c) Que o bibliotecário congele o estado de um certo item no RI após sua revisão, impedindo que sofra alterações automáticas pelo serviço de sincronização com a Plataforma Lattes.

A Figura 10 apresenta o uma visão geral de toda a arquitetura de software desenvolvida ao longo desta pesquisa.

Figura 10 – Visão geral da arquitetura de microserviços implementada nesta pesquisa



Fonte: Elaborada pela autora.

4 Resultados

4.1 Resultados e discussão

Em 28 de julho de 2015, foram processados 1414 currículos de docentes ativos e aposentados. O tempo de extração, transcodificação e armazenamento das publicações em periódicos e em anais de eventos contidas nesses currículos foi de aproximadamente 18 minutos.

O total de entradas descrevendo publicações existentes nos currículos foi de 116960 registros. A pesquisa de duplicatas dentre esses registros levou cerca de 1 minuto para completar a etapa de localização de DOIs idênticos, e cerca de 20 minutos para realizar a busca por similaridade.

Apenas 12,8% dos registros possuíam DOI. O ideal seria que a maior parte dos trabalhos possuísse DOI cadastrado, tornando mais rápido e preciso o cruzamento de duplicatas. Assim, é necessário conscientizar os autores para que se atentem ao cadastro do DOI, e no caso de publicarem em periódicos ou congressos que não emitem DOI, orientar os editores para que busquem registrar-se junto a alguma agência conveniada à Fundação Internacional DOI.

Após a busca de duplicatas, 100346 registros foram considerados únicos. Dentre estes, 13081 possuíam uma ou mais duplicatas na base de dados, indicando que 13% das publicações são em colaboração entre mais de um docente da UFSCar. Ressalta-se, entretanto, que este número não leva em conta o momento em que a pessoa ingressou como docente, de forma que parte desta porcentagem pode refletir trabalhos de estudantes da UFSCar que publicaram em conjunto com seus orientadores e posteriormente tornaram-se docentes da instituição.

Em 3267 publicações (3,2% do total), o script marcador de autoridade de nomes (`authoritymix`) foi capaz de preencher identificadores de autoridade além daqueles que estavam originalmente presentes no currículo principal do grupo de duplicatas.

Após a sincronização da base referencial e execução do indexador de autoridades do DSpace, o módulo da autoridade Lattes catalogou 11766 identificadores de pessoas na base de dados Solr. Ou seja, os docentes da UFSCar correspondem a apenas 12% de todos os autores univocamente identificados e conectados a seus Currículos Lattes. Abrangendo um universo mais de 8 vezes maior que o grupo original de currículos extraídos, a autoridade de nomes baseada na Plataforma Lattes apresenta potencial de tornar-se representativa em nível nacional.

A Figura 11 mostra o resultado final desta pesquisa, um repositório piloto carregado com a base referencial produzida pelo synclattes, atualmente disponível no endereço <https://repositorion.ufscar.br>. A Figura 12 mostra o detalhe de um dos registros pertencentes à base,

Figura 11 – Página inicial do repositório piloto carregado com a base referencial.

Fonte: Elaborada pela autora.

da forma como é exibido pelo DSpace. É possível observar os diversos campos pertencentes ao esquema Dublin Core qualificado, além de ícones para fácil navegação e exploração de outros trabalhos dos autores, que são possíveis graças à implementação da autoridade de nomes.

A base de dados utilizada pelo synclattes, além de servir como fonte para alimentar o DSpace, pode ser acessada diretamente para gerar indicadores de produção científica da universidade, ou ainda, fornecer informações para tratamento automatizado com ferramentas externas, como por exemplo o Vantage Point[1], também visando a elaboração de indicadores. A representação dos metadados em formato JSONB torna bastante direta a construção de consultas SQL utilizando recursos do PostgreSQL versão 9.4, como exemplifica a Figura 13. Um exemplo de indicador de produção científica gerado a partir dessas informações é apresentado na Figura 14, que mostra um gráfico de barras totalizando a produção científica da universidade a cada ano a partir de 1966. Na mesma figura, os dados provenientes da base do synclattes

[1] <https://www.thevantagepoint.com>

Figura 12 – Exemplo de item inserido no DSpace pelas ferramentas desenvolvidas nesta pesquisa. Ao clicar no ícone de engrenagem ao lado do nome de um dos autores, é possível explorar todos os trabalhos do autor contidos no repositório. Ao clicar no ícone amarelo, o usuário é direcionado para o Currículo Lattes do autor.

Mostrar registro simples

dc.contributor.author	Milanez, Douglas Henrique	
dc.contributor.author	Conserva Junior, Antonio Carlos Alves	
dc.contributor.author	Amaral, Roniberto Morato do	
dc.contributor.author	Faria, Leandro Innocentini Lopes de	
dc.contributor.author	Gregolin, Jose Angelo Rodrigues	
dc.date.accessioned	2015-07-31T06:51:50Z	
dc.date.available	2015-07-31T06:51:50Z	
dc.date.issued	2014	
dc.identifier.citation	IV Encontro Brasileiro de Bibliometria e Cientometria, Recife: UFPE, v. 4, 2014	
dc.identifier.uri	https://repositorion.ufscar.br/handle/ufscar/43693	
dc.language.iso	por	
dc.relation.ispartof	IV Encontro Brasileiro de Bibliometria e Cientometria	
dc.subject	bibliometria	
dc.subject	bibliometrics	
dc.subject	inteligencia competitiva	
dc.subject	competitive intelligence	
dc.subject	gestão por competências	
dc.subject	competence-based management	
dc.title	Estudo dos termos de busca para recuperação de publicações científicas em nanocelulos	por
dc.type	conferenceObject	
dc.date.updated	2015-07-31T06:51:50Z	

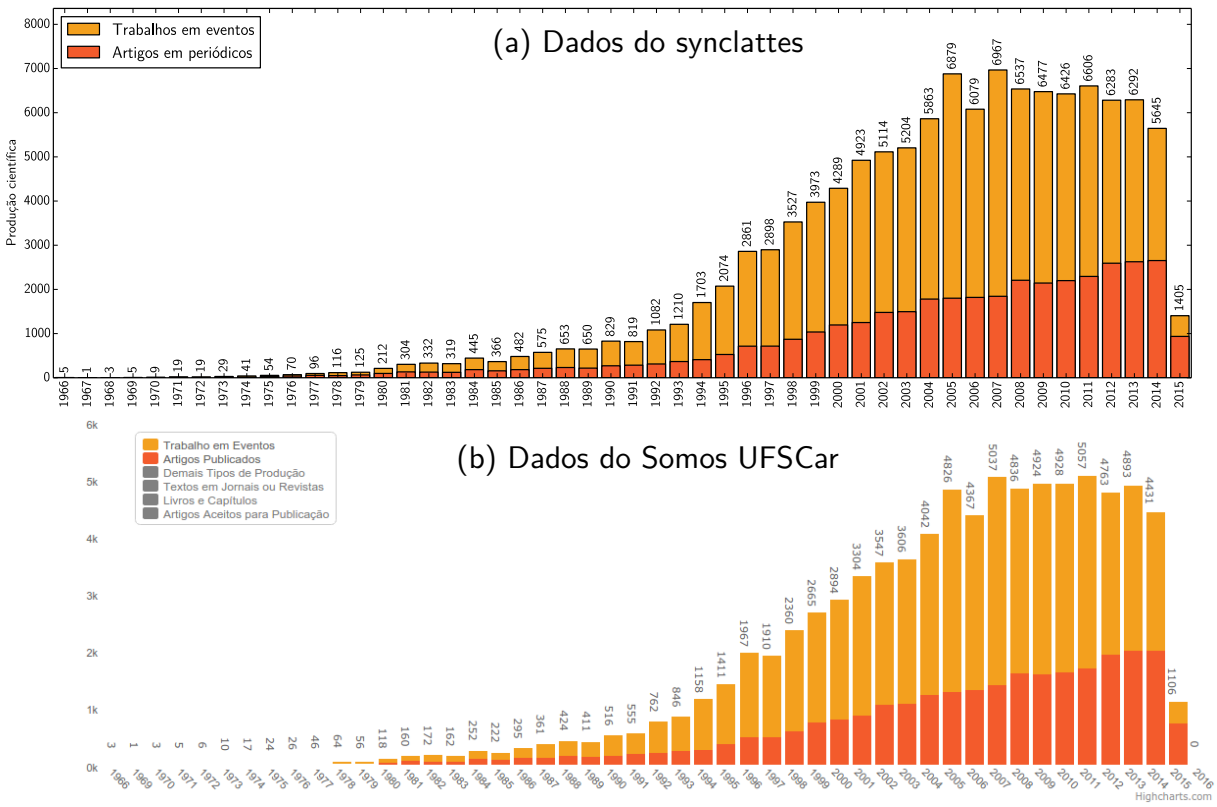
Fonte: Elaborada pela autora.

Figura 13 – Consultas SQL utilizadas para obter a quantidade total de trabalhos publicados por docentes da UFSCar por ano.

```
-- Obtém somente artigos em periódicos
select meta #>> '{dc,date,issued,0,value}', count(*)
from synclattes.last_revision
where meta #>> '{dc,type,"",0,value}' = 'article'
group by 1 order by 1 asc;

-- Obtém tanto artigos como trabalhos em eventos
select meta #>> '{dc,date,issued,0,value}', count(*)
from synclattes.last_revision
group by 1 order by 1 asc;
```

Figura 14 – Comparação entre total de publicações por ano obtidas pelo projeto synclattes em comparação com dados do Somos UFSCar.



Fonte: (a) Elaborada pela autora. (b) <http://somos.ufscar.br>

são comparados com dados do Somos UFSCar[2]. Os dados do synclattes são mais completos, apresentando, em geral, um maior número de publicações por ano, provavelmente devido à inclusão dos docentes aposentados na lista de currículos extraídos. Entretanto, é possível observar que a tendência geral de subida e queda das barras é a mesma.

Outra contribuição deste trabalho foi a execução de uma correção no código do conversor XOAI para DIM, necessária para que o identificador de autoridade fosse apresentado corretamente nos dados em formato DIM retornados pelo protocolo OAI. Vale ressaltar que, além do próprio XOAI bruto, o formato DIM é o único atualmente disponível no DSpace capaz de comportar esses identificadores. A Figura 15 mostra uma comparação entre o Repositório BDPI da USP, que ainda não recebeu a correção, e o repositório piloto da UFSCar. Devido à falha, o atributo "`authority`" do primeiro autor aparece repetido em todos os outros autores na resposta do protocolo OAI. A correção para esta falha já está disponível no repositório GitHub do NIT–Materiais[3], e será em breve enviada ao projeto DSpace oficial.

4.2 Limitações e trabalhos futuros

O trabalho desenvolvido no âmbito deste projeto é apenas um primeiro passo. Para alcançar os objetivos de divulgação da filosofia do acesso aberto, e da ampla disponibilização do trabalho científico conduzido na universidade, é necessário criatividade e inovação. A visão a longo prazo, e o encaminhamento fornecido para trabalhos futuros é tão importante quanto o desenvolvimento presente.

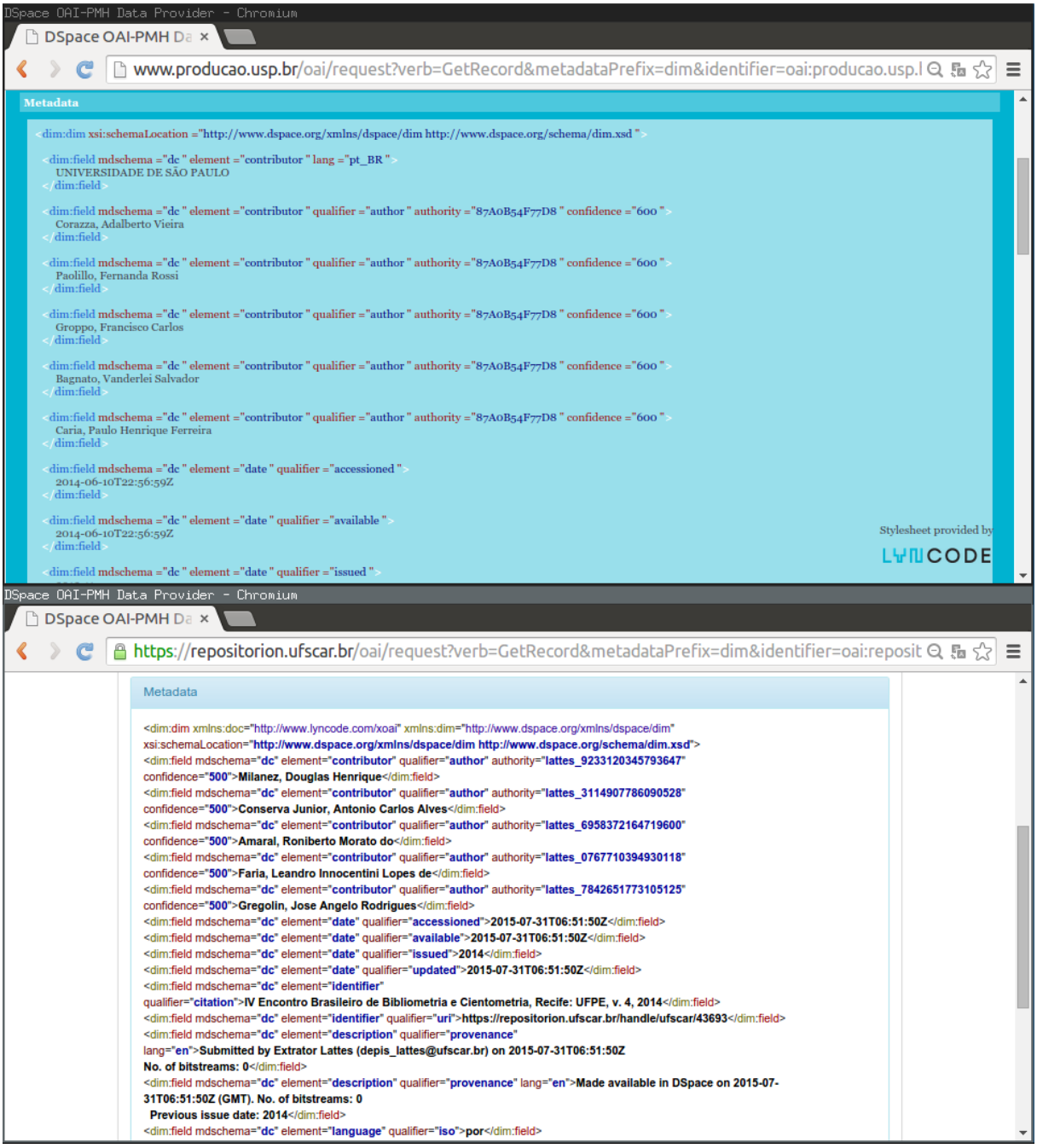
4.2.1 Integração com outros sistemas da universidade

Como este projeto foi elaborado com base na criação de um protótipo de RI para a UFSCar, todos os procedimentos foram executados pensando na integração do RI aos sistemas já existentes na instituição. Essa integração foi possível graças a uma parceria firmada com o DePIS – Departamento de Planejamento e Implantação de Sistemas – da Secretaria Geral de Informática. A solução aqui implementada já está integrada ao ERP (*Enterprise Resource Planning*) da instituição no sentido de catalogar as informações na base de dados em um formato que permite o cruzamento das informações pessoais de cada autor com o seu identificador do Currículo Lattes (ID CNPq) e com seus registros de produção científica.

A integração poderá ser reforçada configurando-se o DSpace para autenticar os acessos por meio do servidor LDAP da UFSCar. Isso permitirá que os autores façam o *login* no DSpace com as mesmas credenciais que já utilizam para acessar o ERP e outros sistemas da instituição. Além disso, como os usuários cadastrados no LDAP estarão automaticamente visíveis para o DSpace, será possível preencher o campo `On-Behalf-Of` ("em nome de") do protocolo

[2] <http://somos.ufscar.br>
[3] <https://github.com/nitmateriais/DSpace>

Figura 15 – Comparação entre respostas ao verbo OAI `GetRecord` do repositório BDPI da USP e do repositório piloto UFSCar. Devido a correções realizadas no âmbito desta pesquisa, o identificador de autoridade passa a ser colhido corretamente utilizando o esquema DIM.



Fonte: Elaborada pela autora.

SWORDv2 com o nome de usuário do portador do Currículo Lattes que originou aqueles metadados, identificando sua proveniência pelos mecanismos internos do próprio DSpace (que gera um metadado `dc.description.provenance` armazenando essa informação).

Ainda no quesito integração entre RI e ERP, outro recurso interessante que pode ser implementado é a atribuição automática dos itens a comunidades e coleções que sejam dependentes do vínculo dos autores com a instituição. Por exemplo, um artigo em coautoria entre pesquisadores do Departamento de Ciência da Informação (DCI), Departamento de Computação (DC) e Departamento de Física (DF) poderia ser incluso simultaneamente em três comunidades, uma para cada um desses departamentos. Os relacionamentos da base de dados do ERP da UFSCar referentes a esses vínculos já estão modelados no código produzido ao longo deste trabalho. Entretanto, como não existe suporte no protocolo SWORDv2 ao depósito de um mesmo item em mais de uma coleção simultaneamente (de forma que resulte em um único identificador para o item), esse recurso ainda não foi implementado. Duas alternativas para fazê-lo seriam:

a) Construir um esquema XML estendido a ser fornecido no documento enviado por meio do protocolo SWORDv2, contendo um elemento capaz de expressar coleções adicionais às quais o item pertença, e realizar as devidas alterações no módulo SWORDv2 do DSpace para o suporte a esse novo esquema.

b) Adicionar ao módulo REST do DSpace um novo método capaz de modificar as coleções às quais determinado item pertença. Esse método seria chamado após o depósito do item por meio do SWORDv2.

4.2.2 Potencialização da base referencial como meio de divulgação do Repositório Institucional

Foi alcançado o objetivo de criar uma sistemática de carga automatizada, contínua e perene para extração de metadados da Plataforma Lattes, compondo uma base referencial para povoamento inicial de um RI. No entanto, é importante lembrar que, apesar de a base referencial populada no DSpace já ser por si só uma ferramenta de grande utilidade para a visualização e exploração dos dados originalmente contidos nos Currículos Lattes, trata-se apenas de uma etapa transitória para o povoamento efetivo do RI, que culminará com a inclusão dos textos completos associados aos metadados já existentes. Em outras palavras, na ausência dos textos completos, a base referencial não pode ser considerada ainda, por definição, um repositório de acesso aberto.

A base referencial aqui construída, além de amparar o povoamento no sentido de facilitar o depósito por meio do preenchimento automático dos metadados, consiste em um ativo estratégico para a divulgação do RI. Todo autor sente-se mais impelido a compartilhar os seus trabalhos por certa via de publicação quando percebe que há visibilidade e demanda

de acesso, ou seja, quando existe um interesse geral de consultá-los por meio dessa via.

Essa prerrogativa já é explorada pela rede social acadêmica ResearchGate[4], que utiliza a mesma estratégia de popular-se automaticamente com uma base referencial de cada autor, mesmo que este ainda não tenha disponibilizado o texto completo de seus trabalhos. O ResearchGate coleta, então, estatísticas de acesso a cada uma dessas entradas bibliográficas, posteriormente utilizando-as para compelir os autores a depositar o texto completo, por meio de e-mails e mensagens que demonstram o quanto o trabalho está sendo encontrado por meio da ferramenta.

Como o DSpace coleta estatísticas detalhadas de acesso a cada um dos itens em sua base de dados Solr, é possível sumarizar esses dados e incluí-los em mensagens destinadas aos autores, incentivando-os a depositar o texto completo dos respectivos itens com base em informações concretas, que demonstrarão o aumento de visibilidade a ser proporcionado por meio dessa ação. Ao mesmo tempo em que essa estratégia maximiza o benefício para os autores, maximiza também a qualidade do RI e sua utilidade para o público geral, priorizando a disponibilização das obras mais procuradas, tanto diretamente por meio do RI como através de buscadores, como o Google.

4.2.3 Integração com o SHERPA/RoMEU

Uma dificuldade recorrente dos autores é no que diz respeito aos direitos autorais e contratos de publicação com editoras. Com vistas a essa questão, o DSpace implementou, a partir da versão 5, uma integração com a API (HANSEN, 2012) do projeto SHERPA/RoMEU. Esse recurso, quando habilitado, mostra um resumo dos direitos do autor ao lado do campo de submissão do texto completo, com base no ISSN da revista informado nos metadados.

Existem diversas oportunidades de melhoria para esse recurso. Por exemplo, o preenchimento do tempo de embargo ainda não é realizado automaticamente a partir do SHERPA/RoMEU. Essa informação é exibida de forma textual para o usuário, que deve copiá-la manualmente para o campo adequado durante a submissão. No entanto, implementar o preenchimento automático não é uma tarefa trivial pois a informação não está catalogada de forma completamente estruturada na base do SHERPA/RoMEU.

Para algumas revistas, o tempo de embargo é parte da lista de restrições, enquanto que em outras ele é listado junto às condições para autodepósito. Em alguns casos, não está disponível a informação exata para determinada revista, mas apenas uma condição geral da editora (por exemplo, "tempo de embargo variando de 12 a 48 meses", no caso da política atual da editora Elsevier), ficando a cargo do usuário pesquisar a restrição específica daquela revista. Nesses casos, ainda seria possível preencher automaticamente o embargo com o tempo máximo adotado pela editora, porém essa medida atrasaria desnecessariamente a abertura ao

[4] <http://www.researchgate.net>

acesso de muitos dos artigos depositados. Desta forma, melhorias neste quesito dependem não somente de trabalho no código do DSpace, mas também de colaboração diretamente junto ao projeto SHERPA/RoMEU.

4.2.4 Integração com serviços de edição colaborativos

Uma ideia surgida[5] em discussões com o DePIS – Departamento de Planejamento e Implantação de Sistemas, da Secretaria Geral de Informática, foi a de futuramente implementar um serviço institucional de edição colaborativa de artigos científicos, com base no software livre ShareLaTeX. Uma integração com esse serviço teria um grande potencial para promover a inserção de texto completo no RI, uma vez que o ShareLaTeX armazena todo o histórico de edição do trabalho, podendo facilmente recuperar tanto suas versões *pre-print* como *post-print*.

A última edição realizada no ShareLaTeX poderia ser, a princípio, considerada como a versão *post-print* do trabalho, assumindo que os autores tenham utilizado a ferramenta até o último momento, para incluir no artigo as sugestões dos revisores. A versão *pre-print* poderia ser identificada como tal pelos próprios autores antes da submissão à revista, ou posteriormente detectada por uma heurística simples — geralmente ocorre um hiato, um intervalo de tempo durante o qual não há edições, enquanto os autores esperam a resposta dos revisores.

Tendo acesso às versões *pre-print* e *post-print* do trabalho, e resolvidas as questões de direitos por meio de uma integração automatizada com o SHERPA/RoMEU, o depósito do texto completo no RI poderia ser reduzido a um simples aceite ou consentimento do autor.

Outra grande vantagem da utilização do ShareLaTeX seria a possibilidade de arquivar não somente o arquivo em formato PDF para leitura, mas também o código fonte do trabalho na linguagem LaTeX, que é importante para a preservação digital do acervo (por tratar-se de formato baseado em texto puro), para a fácil adaptação do *layout* do trabalho para leitura em outros dispositivos (como *e-readers*), ou mesmo para promover a acessibilidade ao conteúdo para deficientes visuais.

[5] Comunicação pessoal de Erick Lazaro Melo, da SIn/UFSCar, que propôs a ideia da criação do serviço de edição colaborativo e de sua integração com este trabalho.

5 Considerações finais

A carga automatizada de metadados da Plataforma Lattes como método de povoamento de um RI apresenta-se como alternativa ágil e viável para qualquer instituição interessada, independente de esta já ter ou não um RI implementado. No entanto, a continuidade dessa implementação estará diretamente ligada à definição das políticas de informação da instituição.

Apesar de existir o scriptLattes como forma alternativa para a extração desses dados, a extração direta da base de dados da Plataforma Lattes disponibiliza o acesso completo aos dados que, posteriormente, podem ser tratados de acordo com o interesse da instituição. Não é foco deste trabalho, mas é importante citar que, além desse tipo de acesso, a Plataforma também fornece possibilidade de espelhamento dos dados voltada às fundações estaduais de apoio à pesquisa, que consiste na disponibilização integral dos dados da Plataforma Lattes e dos currículos atualizados diariamente, para replicação em base espelho da fundação.

É preciso também ressaltar o caráter original da presente pesquisa. Existem registros de pesquisas envolvendo extração de dados da Plataforma Lattes, no entanto, nenhuma com o propósito e complexidade do presente trabalho. Algumas pesquisas baseiam-se em resultados do ScriptLattes (MENA-CHALCO; CESAR JR., 2013), ou até mesmo no desenvolvimento de outras ferramentas baseadas neste script (FARIAS; VARGAS; BORGES, 2012), porém não há nenhum outro registro de pesquisa que tenha utilizado o *web service* oficial da Plataforma Lattes para popular um RI.

Outro fator de originalidade da presente pesquisa é o fato desta prover infraestrutura para realizar o povoamento contínuo e perene de um RI, a partir dos dados que o pesquisador atualizar em seu Currículo Lattes. Ou seja, somente a primeira extração e povoamento se dará em um banco de dados limpo, sendo que as próximas cargas incrementarão os dados já existentes com correções ou adição de novos dados que o pesquisador tenha incluído. O objetivo é atualizar o repositório periodicamente para que este possa assimilar as atualizações realizadas no Currículo Lattes, por meio de um controle de versão que manterá a confiabilidade e a integridade dos dados.

Este projeto também contribuirá para o aumento da precisão dos dados inseridos no Currículo Lattes por parte dos pesquisadores que, ao verificar seus perfis de publicações no RI, serão orientados a realizar quaisquer correções necessárias no Lattes, caso algum metadado possua erros. Além disso, os pesquisadores serão ainda mais incentivados a incluir o DOI de seus trabalhos, ampliando ainda mais a facilidade de acesso à produção intelectual da instituição. É importante destacar que nesta pesquisa foi verificado apenas se o DOI está formatado corretamente, mas não foi verificado se o link no Lattes está funcionando, uma vez

que o próprio Lattes verifica a integridade do DOI no momento em que o registro é salvo. No entanto, verificamos alguns casos raros em que o DOI existente no currículo não estava mais funcional e havia sido renomeado para outro apontando para a mesma publicação original, apesar disto ir contra o princípio básico do DOI, que é o de ser um identificador imutável. A busca por similaridade de títulos pode ajudar a encontrar estes casos e a saná-los, muitas vezes diretamente nos currículos.

O DSpace é capaz de gerar algumas estatísticas de indicadores muito relevantes a partir dos dados extraídos da Plataforma Lattes. No entanto, o banco de dados implementado no âmbito desta pesquisa também poderá ser utilizado para gerar diversas outras estatísticas e indicadores, como por exemplo indicadores de rede de colaboração entre departamentos da UFSCar, e até mesmo de redes de colaboração externas, cuja implementação será fortemente beneficiada pelos identificadores providos pelo serviço de autoridade de nomes criado neste projeto.

Essa base de dados poderá servir também como alicerce para a construção de um Sistema de Gestão de Informação Científica (CRIS – Current Research Information System), fornecendo informações a respeito de projetos, publicações, agências financiadoras, dentre outras, que poderão, uma vez sistematizadas, contribuir para a tomada de decisão dos gestores institucionais (AMANTE et al., 2014).

Todos os scripts e ferramentas criados neste trabalho foram referenciados no decorrer do texto com seus respectivos endereços na Internet. O conjunto de ferramentas utilizadas nesta sistemática pode ser encontrado na página da organização NIT–Materiais no GitHub[1]. Todo o código que lida com a base de dados específica da UFSCar está contido dentro de um mesmo pacote denominado "`ufscar`" contido no projeto synclattes, visando facilitar a adaptação das ferramentas para qualquer outra instituição.

[1] <https://github.com/nitmateriais>

Referências

ALLINSON, J.; FRANÇÓIS, S.; LEWIS, S. SWORD: Simple Web-service Offering Repository Deposit. *Ariadne*, n. 54, jan. 2008. Disponível em: <http://www.ariadne.ac.uk/issue54/allinson-et-al>. Acesso em: 10 Jul. 2015. Citado na página 62.

ALMEIDA, E. C. E. de. *O Portal de Periódicos da Capes: estudo sobre a sua evolução e utilização* . Dissertação (Mestrado) — Universidade de Brasília – UNB, 2006. Citado na página 28.

ALMEIDA, E. C. E. de; GUIMARÃES, J. A.; ALVES, I. T. G. Dez anos do Portal de Periódicos da Capes: histórico, evolução e utilização. Revista Brasileira de Pós-Graduação – Capes, v. 7, n. 13, p. 218–246, nov. 2010. Disponível em: <http://ojs.rbpg.capes.gov.br/index.php/rbpg/article/viewFile/194/188>. Acesso em: 02 jul. 2015. Citado na página 29.

AMANTE, M. J. et al. *Cardernos BAD*, n. 2, p. 83–93, 2014. ISSN 0007-9421. Disponível em: <http://www.bad.pt/publicacoes/index.php/cadernos/article/view/1183>. Acesso em: 4 Ago. 2015. Citado na página 80.

BUDAPEST OPEN ACCESS INITIATIVE. *Budapest Open Access Initiative*. 2002. Disponível em: <http://www.budapestopenaccessinitiative.org/translations/portuguese-translation>. Acesso em: 02 Abr. 2015. Citado na página 24.

BUDAPEST OPEN ACCESS INITIATIVE. *Dez anos da Iniciativa de Budapeste em Acesso Aberto: a abertura como caminho a seguir*. 2015. Disponível em: <http://www.budapestopenaccessinitiative.org/boai-10-translations/portuguese-brazilian-translation>. Acesso em: 02 Abr. 2015. Citado na página 25.

CARVER, B. W. Share and share alike: Understanding and enforcing open source and free software licenses. *Berkeley Technology Law Journal*, v. 20, n. 1, p. 443–481, 2007. Disponível em: <http://scholarship.law.berkeley.edu/btlj/vol20/iss1/46>. Acesso em: 11 jan. 2015. Citado na página 34.

CASTELLS, M. *A Sociedade em Rede: A Era da Informação, Sociedade e Cultura*. 3. ed. São Paulo: Paz e Terra, 1999. Citado 2 vezes nas páginas 21 e 27.

CENTER FOR RESEARCH LIBRARIES. *Database: Web of Science*. 2015. Disponível em: <http://adat.crl.edu/databases/about/web_of_science>. Acesso em: 31 jul. 2015. Citado na página 23.

COLLIS, J.; HUSSEY, R. *Pesquisa em administração: um guia prático para alunos de graduação e pós-graduação*. 2. ed. Porto Alegre: Bookman, 2005. ISBN 85-363-0419-7. Citado na página 53.

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO. *A Criação – Portal CNPq*. 2015. Disponível em: <http://cnpq.br/a-criacao>. Acesso em: 21 jan. 2015. Citado 3 vezes nas páginas 22, 32 e 33.

CONSERVA JR., A. C. A. *Aproveitamento de dados bibliográficos de publicação científica indexada na base Web of Science e Currículo Lattes para a criação de repositório institucional da UFSCar*. 2014. 45 p. Citado na página 30.

CORRÊA, T. P. P. et al. Implementação do Repositório Institucional da Universidade Federal do Rio Grande: uma visão através do catálogo decisório de autores. Revista ACB: Biblioteconomia em Santa Catarina, v. 17, n. 1, p. 27–41, jun. 2012. Disponível em: <http://eprints.rclis.org/18001/>. Acesso em: 10 Abr. 2015. Citado na página 59.

COSTA, M. A visibilidade no Google Scholar dos repositórios digitais de acesso aberto brasileiros e portugueses. In: . [S.l.: s.n.], 2014. p. 41–53. Citado na página 24.

COUGHLAN, P.; COGHLAN, D. Action research for operations management. *International Journal of Operations & Production Management*, v. 22, n. 2, p. 220–240, 2002. Disponível em: <http://dx.doi.org/10.1108/01443570210417515>. Acesso em: 21 jan. 2015. Citado 2 vezes nas páginas 23 e 54.

DIGGORY, M. et al. Embargo. In: DURASPACE. *DSpace 4.x Documentation*. 2013. Disponível em: <https://wiki.duraspace.org/pages/viewpage.action?pageId=34640846>. Acesso em: 11 jan. 2015. Citado na página 37.

DSPACE. *About DSpace*. 2015. Disponível em: <http://www.dspace.org/introducing>. Acesso em: 12 jan. 2015. Citado na página 46.

DURASPACE. *DuraSpace: Committed to our digital future*. 2015. Disponível em: <http://www.duraspace.org/history>. Acesso em: 12 jan. 2015. Citado na página 45.

FARIAS, L. R. de; VARGAS, A. P.; BORGES, E. N. Um sistema para análise de redes de pesquisa baseado na Plataforma Lattes. In: *Anais da VIII Escola Regional de Banco de Dados*. [s.n.], 2012. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/erbd/2012/003.pdf>. Acesso em: 10 Abr. 2015. Citado na página 79.

FAUSTO, S. S. de. *Captcha nos CVs Lattes*. 2015. Disponível em: <https://web.archive.org/web/20150804174148/https://social.stoa.usp.br/sibelefausto/blog/captcha-nos-cvs-lattes>. Acesso em: 4 Ago. 2015. Citado na página 56.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. *Commun. ACM*, ACM, New York, NY, USA, v. 39, n. 11, p. 27–34, nov. 1996. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/240455.240464>. Citado na página 25.

FERREIRA, S. M. S. P. et al. Repositório da produção científica CRUESP: Mais do que um consórcio, um trabalho integrado USP, UNESP e UNICAMP. In: *IV Conferência Luso-brasileira de Acesso Aberto*. São Paulo: [s.n.], 2013. [Pôster]. Disponível em: <http://www.producao.usp.br/handle/BDPI/43861>. Acesso em: 11 jan. 2015. Citado na página 49.

GALINA RUSSELL, I. La visibilidad de los recursos académicos. Una revisión crítica del papel de los repositorios institucionales y el acceso abierto. *Investigación Bibliotecológica*, v. 25, n. 53, p. 159–183, jan. 2011. ISSN 0187-358 X. Citado na página 24.

HANSEN, D. Understanding and making use of academic authors' open access rights. *Journal of Librarianship and Scholarly Communication*, Pacific University Library, Forest Grove, v. 1, n. 2, p. eP1050, set. 2012. ISSN 2162-3309. Disponível em: <http://dx.doi.org/10.1016/j.acalib.2007.09.020>. Acesso em: 12 jan. 2015. Citado na página 76.

HARNAD, S. The green road to open access: A leveraged transition. In: ANNA, G. (Ed.). *The culture of periodicals from the perspective of the electronic age*. L'Harmattan, 2007. p. 99–105. Disponível em: <http://eprints.soton.ac.uk/265753>. Acesso em: 21 jan. 2015. Citado na página 37.

HARNAD, S. *How to integrate university and funder open access mandates*. [S.l.], 2008. Disponível em: <http://eprints.soton.ac.uk/265265>. Acesso em: 21 jan. 2015. Citado na página 37.

HARVARD UNIVERSITY. *About DASH*. 2015. Disponível em: <https://osc.hul.harvard.edu/dash/about>. Acesso em: 02 Abr. 2015. Citado na página 51.

HENNING, P. C. et al. Repositório institucional da Fiocruz – ARCA: Manual de tratamento de objetos digitais. In: *II Conferência Luso-brasileira de Acesso Aberto*. [s.n.], 2011. [Pôster]. Disponível em: <http://arca.icict.fiocruz.br/handle/icict/3699>. Acesso em: 12 jan. 2015. Citado na página 50.

HILLMANN, D. *Using Dublin Core*. 2005. Disponível em: <http://dublincore.org/documents/usageguide>. Acesso em: 11 jan. 2015. Citado 2 vezes nas páginas 38 e 39.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA. *Manifesto Brasileiro de apoio ao Acesso Livre à Informação Científica*. 2005. Disponível em: <http://livroaberto.ibict.br/docs/Manifesto.pdf>. Acesso em: 13 jan. 2015. Citado na página 29.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA. *Sobre o IBICT*. 2015. Disponível em: <http://www.ibict.br/sobre-o-ibict/apresentacao>. Acesso em: 11 jan. 2015. Citado 2 vezes nas páginas 32 e 46.

INTERSECT AUSTRALIA. *Dublin Core refereed journal 2nd example*. 2013. Disponível em: <https://github.com/IntersectAustralia/dspacetobpress/blob/master/Library/dc-refereed-journal-2.xml>. Acesso em: 13 jan. 2015. Citado na página 40.

JENKINS, C. et al. RoMEO studies 8: self-archiving: The logic behind the colour-coding used in the Copyright Knowledge Bank. *Program*, v. 41, n. 2, p. 124–133, 2007. ISSN 0033-0337. Disponível em: <http://eprints.nottingham.ac.uk/580>. Acesso em: 11 jan. 2015. Citado na página 35.

KURAMOTO, H. Informação científica: Proposta de um novo modelo para o Brasil. *Ci. Inf.*, Brasília, v. 35, n. 2, p. 91–102, maio 2006. ISSN 0100-1965. Disponível em: <http://dx.doi.org/10.1590/S0100-19652006000200010>. Acesso em: 11 jan. 2015. Citado 5 vezes nas páginas 27, 29, 31, 46 e 47.

KURAMOTO, H. *Como a TI poderia auxiliar no povoamento dos RI*. 2011. Disponível em: <https://web.archive.org/web/20150804174252/http://kuramoto.blog.br/2011/10/31/como-a-ti-poderia-auxiliar-no-povoamento-dos-ri/>. Acesso em: 4 Ago. 2015. Citado 2 vezes nas páginas 24 e 62.

LEVENSHTEIN, V. I. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, v. 10, n. 8, p. 707–710, 1966. ISSN 1028-3358. Disponível em: <https://gitlab.doc.ic.ac.uk/wm613/inidividual-project/raw/4fbe43f863b229b50ca13bf5f59eb029ad71f6b6/reading/litreview/levenshtein66.pdf>. Acesso em: 10 Abr. 2015. Citado na página 66.

LEY, M. D. L. M. G. *Diretrizes para a proposição de política de povoamento de Repositório Institucional: O contexto da Universidade Federal Fluminense (UFF)*. 242 p. Dissertação (Mestrado) — Universidade Federal Fluminense, Niterói, jun. 2013. Disponível em: <http://www.ci.uff.br/ppgci/arquivos/Dissert/2013/MARIA%20DULCE%20LAGOEIRO%20M%20GAUDIE%20LEY.pdf>. Acesso em: 18 jan. 2015. Citado 4 vezes nas páginas 22, 37, 42 e 43.

LIMA, T. C. S.; MIOTO, R. C. T. Procedimentos metodológicos na construção do conhecimento científico: a pesquisa bibliográfica. *Revista Katálysis*, v. 10, p. 37–45, 2007. ISSN 1414-4980. Disponível em: <http://dx.doi.org/10.1590/S1414-49802007000300004>. Acesso em: 21 jan. 2015. Citado na página 53.

MARCONDES, C. H.; SAYÃO, L. F. À guisa de introdução: Repositórios institucionais e livre acesso. In: SAYÃO, L. et al. (Ed.). *Implantação e gestão de repositórios institucionais: políticas, memória, livre acesso e preservação*. Salvador: EDUFBA, 2009. p. 365. ISBN 978-85-232-0655-0. Disponível em: <https://repositorio.ufba.br/ri/handle/ufba/473>. Acesso em: 11 jan. 2015. Citado 3 vezes nas páginas 27, 28 e 31.

MENA-CHALCO, J. P.; CESAR JR., R. M. Prospecção de dados acadêmicos de currículos Lattes através de scriptLattes. In: *Bibliometria e Cientometria: Reflexões teóricas e interfaces*. São Carlos: Pedro & João, 2013. p. 109–128. ISBN 978-85-7993-117-8. Disponível em: <http://professor.ufabc.edu.br/~jesus.mena/publications/pdf/capitulo-livro-scriptlattes-2013.pdf>. Acesso em: 12 jan. 2015. Citado 2 vezes nas páginas 56 e 79.

MIGUEL, P. A. C. Estudo de caso na engenharia de produção: estruturação e recomendações para sua condução. *Production*, v. 17, p. 216–229, 2007. ISSN 0103-6513. Disponível em: <http://dx.doi.org/10.1590/S0103-65132007000100015>. Acesso em: 21 jan. 2015. Citado na página 54.

MOGUL, J. C. The case for persistent-connection HTTP. *SIGCOMM Comput. Commun. Rev.*, ACM, New York, v. 25, n. 4, p. 299–313, out. 1995. ISSN 0146-4833. Disponível em: <http://dx.doi.org/10.1145/217391.217465>. Acesso em: 12 jan. 2015. Citado na página 56.

NAVEGA, S. Princípios Essenciais do Data Mining. In: INTELLIWISE RESEARCH AND TRAINING. *Anais do Infoimagem* . Cenadem, 2002. Disponível em: <http://www.intelliwise.com/reports/i2002.pdf>. Acesso em: 16 jul. 2015. Citado na página 25.

NEVES, T. M. G. das. Livre acesso à publicação acadêmica. *Ciência da Informação [online]*, v. 33, n. 3, p. 116–121, 2004. ISSN 1518-8353. Disponível em: <http://dx.doi.org/10.1590/S0100-19652004000300014>. Acesso em: 02 jul. 2015. Citado 2 vezes nas páginas 21 e 29.

OKAZAKI, N.; TSUJII, J. Simple and efficient algorithm for approximate dictionary matching. In: *Proceedings of the 23rd International Conference on Computational Linguistics*.

Beijing, China: Association for Computational Linguistics, 2010. p. 851–859. Disponível em: <http://www.aclweb.org/anthology/C10-1096>. Acesso em: 04 Abr. 2015. Citado na página 65.

OLIVEIRA, R. R. de; CARVALHO, C. L. de. *Implementação de Interoperabilidade entre Repositórios Digitais por meio do Protocolo OAI-PMH*. [S.l.], 2009. Citado na página 38.

OPENDOAR. *OpenDOAR - Directory of Open Access Repositories*. 2015. Disponível em: <http://www.opendoar.org/find.php>. Acesso em: 11 jan. 2015. Citado 2 vezes nas páginas 45 e 48.

PINFIELD, S. Open archiving in UK universities. In: *2nd Workshop on the Open Archives Initiative (OAI): Gaining independence with e-prints archives and OAI (OAI2)*. Geneva: CERN, 2002. [Apresentação]. Disponível em: <http://eprints.rclis.org/4548>. Acesso em: 11 jan. 2015. Citado na página 35.

RICKETSON, S. *The Berne Convention for the Protection of Literary and Artistic Works: 1886–1986*. London: Centre for Commercial Law Studies, Queen Mary College Kluwer, 1987. ISBN 978-05-824-6368-4. Citado na página 33.

ROSA, F. G.; TOUTAIN, L. B. Apresentação. In: SAYÃO, L. et al. (Ed.). *Implantação e gestão de repositórios institucionais: Políticas, memória, livre acesso e preservação*. Salvador: EDUFBA, 2009. p. 365. ISBN 978-85-232-0655-0. Disponível em: <https://repositorio.ufba.br/ri/handle/ufba/473>. Acesso em: 11 jan. 2015. Citado na página 30.

SAYÃO, L. F.; MARCONDES, C. H. Software livres para repositórios institucionais: alguns subsídios para a seleção. In: SAYÃO, L. et al. (Ed.). *Implantação e gestão de repositórios institucionais: Políticas, memória, livre acesso e preservação*. Salvador: EDUFBA, 2009. p. 365. ISBN 978-85-232-0655-0. Disponível em: <https://repositorio.ufba.br/ri/handle/ufba/473>. Acesso em: 11 jan. 2015. Citado 8 vezes nas páginas 30, 31, 39, 41, 43, 44, 45 e 46.

SHEARER, K. Institutional repositories: Towards the identification of critical success factors. *Canadian Journal of Information and Library Science*, v. 27, n. 3, p. 89–108, set. 2003. Disponível em: <http://dspace.ucalgary.ca/handle/1880/43357>. Acesso em: 20 jan. 2015. Citado na página 43.

SHERPA/RoMEO. *Statistics for the 1771 publishers in the RoMEO database*. Nottingham, 2015. Disponível em: <http://www.sherpa.ac.uk/romeo/statistics.php>. Acesso em: 11 jan. 2015. Citado na página 36.

SMITH, M. et al. DSpace an open source dynamic digital repository. *D-Lib Magazine*, The Magazine of Digital Library Research, v. 9, n. 13, jan. 2003. ISSN 1082-9873. Disponível em: <http://www.dlib.org/dlib/january03/smith/01smith.html>. Acesso em: 04 Abr. 2015. Citado na página 51.

SOUZA, C. "Estado do campo"da pesquisa em políticas públicas no Brasil. *Revista Brasileira de Ciências Sociais*, v. 18, n. 51, fev. 2003. ISSN 1806-9053. Disponível em: <http://dx.doi.org/10.1590/S0102-69092003000100003>. Acesso em: 02 jul. 2015. Citado na página 29.

SUAIDEN, E. Dimensão e perspectivas sociais do acesso livre à informação. *Ci. Inf.*, v. 35, n. 2, p. 7–8, maio 2006. ISSN 0100-1965. Disponível em: <http:

//dx.doi.org/10.1590/S0100-19652006000200001>. Acesso em: 11 jan. 2015. Citado 2 vezes nas páginas 28 e 47.

TERENCE, A. C. F.; FILHO, E. E. Abordagem quantitativa, qualitativa e a utilização da pesquisa-ação nos estudos organizacionais. In: *XXVI Encontro Nacional de Engenharia de Produção*. Fortaleza: ABEPRO, 2006. Disponível em: <http://www.abepro.org.br/biblioteca/enegep2006_tr540368_8017.pdf>. Acesso em: 21 jan. 2015. Citado 2 vezes nas páginas 53 e 54.

THIOLLENT, M. *Metodologia da Pesquisa-Ação*. 16. ed. São Paulo: Cortez, 2008. 132 p. ISBN 978-85-249-1170-5. Citado na página 54.

UNIVERSIDADE DO MINHO. *Sobre o RepositóriUM*. 2014. Disponível em: <https://repositorium.sdum.uminho.pt/about/about.htm>. Acesso em: 02 Abr. 2015. Citado na página 52.

UNIVERSIDADE FEDERAL DA BAHIA. *Sobre o Repositório Institucional da Universidade Federal da Bahia*. 2015. Disponível em: <https://repositorio.ufba.br/ri/about/about.jsp>. Acesso em: 11 jan. 2015. Citado na página 48.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL. *Repositório digital LUME – UFRGS*. 2015. Disponível em: <http://www.lume.ufrgs.br/apresentacao>. Acesso em: 11 jan. 2015. Citado na página 49.

VIANA, C. L. M.; MÁRDERO ARELLANO, M. Á. Repositórios institucionais baseados em DSpace e EPrints e sua viabilidade nas instituições acadêmico-científicas. In: *XIV Seminário Nacional de Bibliotecas Universitárias*. Salvador: [s.n.], 2006. Disponível em: <http://eprints.rclis.org/8834>. Acesso em: 12 jan. 2015. Citado na página 43.

VIANA, C. L. M.; MÁRDERO ARELLANO, M. Á.; SHINTAKU, M. Repositórios institucionais em ciência e tecnologia: Uma experiência de customização do DSpace. In: *II Simpósio Internacional de Bibliotecas Digitais*. São Paulo: [s.n.], 2005. Disponível em: <http://eprints.rclis.org/7168>. Acesso em: 11 jan. 2015. Citado 4 vezes nas páginas 22, 30, 31 e 38.

WARNER, S. Exposing and harvesting metadata using the OAI metadata harvesting protocol: A tutorial. *eprint arXiv:cs/0106057*, jun. 2001. Disponível em: <http://arxiv.org/abs/cs/0106057>. Acesso em: 11 jan. 2015. Citado na página 31.