UNIVERSIDADE FEDERAL DE SÃO CARLOS (UFSCAR)
CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS (CECH)

ANDRIELI CRISTINA BOTÁCIO

DADOS LINGUÍSTICOS E INICIATIVA *LINKING OPEN DATA*

SÃO CARLOS/SP
2020

ANDRIELI CRISTINA BOTÁCIO

DADOS LINGUÍSTICOS E INICIATIVA *LINKING OPEN DATA*

Trabalho de Conclusão de Curso apresentado como requisito parcial para a obtenção do título de Bacharel em Biblioteconomia e Ciência da Informação pela Universidade Federal de São Carlos.

Orientador(a): Profa. Dra. Ana Carolina Simionato Arakaki

Agência Financiadora: Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) – Processo 2018/05101-3

SÃO CARLOS/SP
2020

Botácio, Andrieli Cristina
B657d Dados linguísticos e iniciativa Linking Open Data / Andrieli Cristina Botácio. – 2020.
71 f.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal de São Carlos, São Carlos, 2020.

1. Linked Data. 2. Linking Open Data 3. Dados linguísticos. I. Título.

CDD 020

DADOS LINGUÍSTICOS E INICIATIVA *LINKING OPEN DATA*

ANDRIELI CRISTINA BOTÁCIO

Trabalho de Conclusão de Curso apresentado como requisito parcial para a obtenção do título de Bacharel em Biblioteconomia e Ciência da Informação pela Universidade Federal de São Carlos.

Aprovada em: 09/12/2020.

BANCA EXAMINADORA

_____

Profa. Dra. Ana Carolina Simionato Arakaki
Departamento de Ciência da Informação da UFSCar

_____

Profa. Dra. Paula Regina Dal´Evedove
Departamento de Ciência da Informação da UFSCar

_____

Prof. Dr. Rogério Aparecido Sá Ramalho
Departamento de Ciência da Informação da UFSCar

AGRADECIMENTOS

A Deus, pela vida e força que me concede.

Aos meus pais, Maria José e Pedro Botácio, e meu irmão, Antônio Marcos Botácio, pelo apoio em cada um de meus passos, pelos valores que me transmitiram e pela motivação constante.

À Profa. Dra. Ana Carolina Simionato Arakaki, minha orientadora, pela paciência e auxílio na execução deste trabalho.

Aos membros da banca examinadora, Profa. Dra. Paula Regina Dal´Evedove e Prof. Dr. Rogério Aparecido Sá Ramalho, pelas valiosas contribuições.

Aos docentes do Departamento de Ciência da Informação (DCI) da UFSCar, pelo comprometimento na transmissão de conhecimento.

Aos servidores técnicos-administrativos do referido departamento, Artur Protter Gouvea, Mônica Guimarães Pithon Calio e Renan Vasconcelos Ribeiro, por toda a ajuda concedida desde o início desta graduação.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pelo financiamento desta pesquisa (Processo nº 2018/05101-3).

RESUMO

Tim Berners-Lee propôs a *Web* Semântica para uma melhor recuperação da informação. Nesse contexto encontram-se os princípios *Linked Data* por meio dos quais se estabelece a conexão de dados na *Web*, fazendo com que os agentes de *software* e as pessoas possam trabalhar de maneira cooperativa, alcançando os objetivos de modo eficiente. Pensando nisso, foi analisada uma iniciativa que é um exemplo da aplicação do *Linked Open Data*, que é o *Linking Open Data* (LOD), que traz dados publicados em formato de dados ligados. A pesquisa se dedica a descrever e analisar os *datasets* linguísticos presentes nessa iniciativa. Desse ponto, incita-se como problemática de pesquisa: o que é identificado dentro das ligações e nos *datasets* linguísticos na iniciativa *Linking Open Data*? Tendo como foco tal problemática, o objetivo é mapear os *datasets* correspondentes à categoria 'Linguística' inseridos na iniciativa *Linking Open Data*. Trata-se de uma pesquisa exploratória e qualitativa e de natureza teórico-aplicada, abordando como tema principal o mapeamento dos *datasets* linguísticos no *Linking Open Data*. Primeiramente, fez-se uma investigação teórica e prática acerca da identificação dos *datasets* linguísticos e as tecnologias empregadas na ligação desses dados; depois a investigação foi direcionada para a análise da categoria Linguística da iniciativa *Linking Open Data*. Os resultados obtidos mostram quais tipos de *datasets* e tecnologias da *Web* Semântica encontram-se em cada uma das sete categorias de dados linguísticos: *Corpora*; *Lexicons and Dictionaries*; *Terminologies, Thesauri and Knowledge Bases*; *Linguistic Resource Metadata*; *Linguistic Data Categories*; *Typological Databases*; *Other*. Conclui-se que a iniciativa *Linking Open Data* cumpre de modo satisfatório a sua função, mostrando a viabilidade da conexão de dados abertos, por meio das tecnologias prescritas. Quanto aos dados linguísticos de tal iniciativa, nota-se que são de extrema relevância e empregam as tecnologias conforme o que se exige em cada categoria.

Palavras-chave: *Web*. *Web* Semântica. *Linked Data*. *Linking Open Data*. Dados linguísticos.

ABSTRACT

Tim Berners-Lee proposed the Semantic Web for better information retrieval. In this context, there are the Linked Data principles through which the data connection on the Web is established which the main objective is to generate meaning to the Web pages. And, by utilizing, this causes with which the software agents and people could cooperate with each other to reach their goals in an efficient manner. The project consists of working with an initiative that is an example of the application of Linked Open Data, which is the Linking Open Data (LOD), which brings data published in linked data format. The research will focus on describing and analyzing the linguistic datasets present in this initiative. From this point, it is incited as a research problem: what is identified in the links and linguistic datasets in the Linking Open Data initiative? Focusing on this problem, the objective is to map the datasets corresponding to the 'Linguistics' category inserted in the Linking Open Data initiative. It is an exploratory and qualitative research and of a theoretical-applied nature, addressing as main theme the mapping of linguistic datasets in Linking Open Data. First, a theoretical and practical investigation was carried out on the identification of the linguistic data sets and the technologies used in the connection of these data; after the investigation was directed to the analysis of the Linguistics category of the Linking Open Data initiative. The results obtained show that types of datasets and technologies of the Semantic Web are found in each of the seven categories of linguistic data: Corpora; Lexicons and Dictionaries; Terminologies, Thesauri and Knowledge Bases; Linguistic Resource Metadata; Linguistic Data Categories; Typological Databases; Other. It is concluded that the Linking Open Data initiative satisfactorily fulfills its function, showing the viability of the open data connection, through the prescribed technologies. As for the linguistic data of such an initiative, it is noted that they are extremely relevant and employs the technologies according required in each category.

Keywords: Web. Semantic Web. Liked Data. Linking Open Data. Linguistic data.

LISTA DE GRÁFICOS

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

Pode-se dizer que a *World Wide Web* é uma plataforma que possibilita produzir, disseminar e recuperar informações, e para acessá-la é necessário estar conectado à *Internet*. Embora esses termos *Web* e *Internet* possam ser confundidos, cada um designa a uma proposta diferente. A *Web* foi criada por Tim Berners-Lee, que a desenvolveu na década de 1990 com o propósito de possibilitar que os colegas de pesquisa pudessem trocar documentos sobre aquilo que estavam estudando com maior facilidade. De maneira estrutural, a *Internet* foi desenvolvida no período da Guerra Fria, em que era preciso um sistema que proporcionasse a troca de informações entre os computadores. (SOUZA; ALVARENGA, 2004). Com o desenvolvimento da *Internet* na Guerra Fria o objetivo era trocar informações para vencê-la, porém atualmente também é utilizada para troca de informações, mas agora com o propósito de construir um conhecimento e compartilhá-lo.

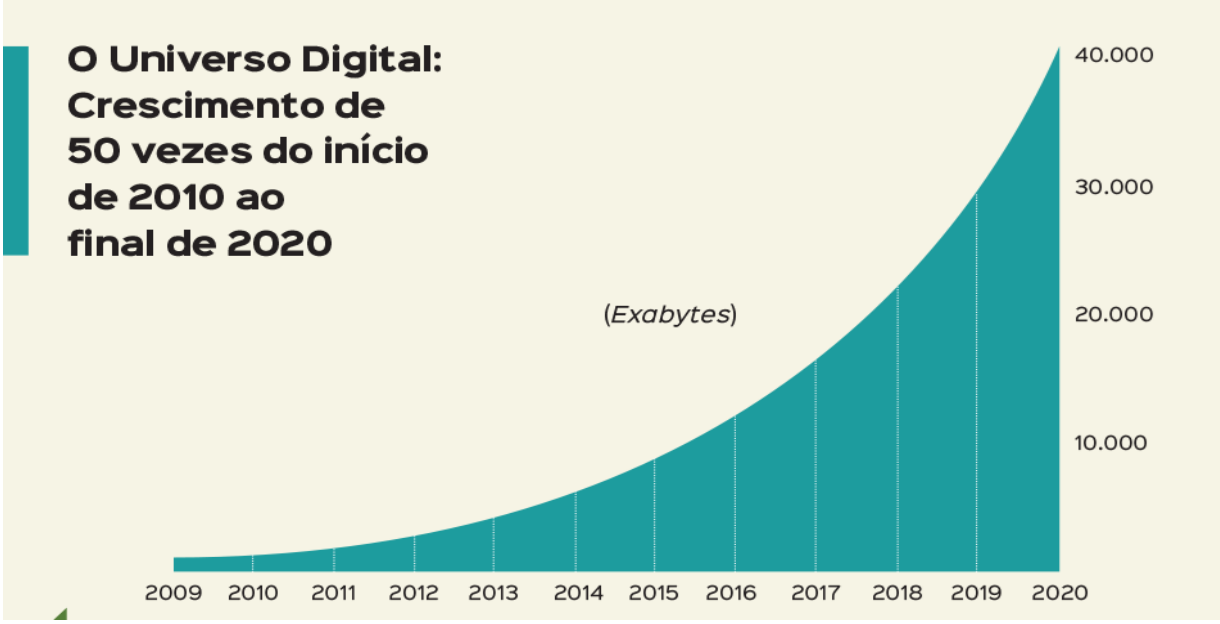
Na década de 90, Tim Berners-Lee propõe a *Web* Semântica para uma melhor recuperação da informação. Pode ser entendida como uma extensão da *Web* que por meio de tecnologias e padrões permite que o conhecimento seja compartilhado e reutilizado (GONÇALVES; JACYNTHO, 2020). Nesse contexto encontra-se o *Linked Data* que pressupõe a "[...] utilização da *Web* para publicar e interligar dados de forma direta através de *links* semânticos para que pessoas e máquinas possam explorá-los, tornando mais fácil a descoberta de dados relacionados." (GONÇALVES; JACYNTHO, 2020, p. 2).

Também, nesse cenário, há a noção de dados abertos. Segundo a *Open Definition* (2017, tradução nossa), "[...] qualquer pessoa pode livremente acessar, usar, modificar e compartilhar para qualquer finalidade (sujeito, no máximo, a requisitos que preservem a proveniência e a abertura)". Além disso, os dados abertos constituem-se como uma forma de democratizar o acesso à informação, eliminando as barreiras. A partir disso, é possível estabelecer as normas para que um dado seja considerado aberto (WHAT IS OPEN? 2018, tradução nossa), que consistem em três propriedades: disponibilidade e acesso, reuso e redistribuição, e participação universal. A primeira condição define que a disponibilidade dos dados não deve haver custos, além de poderem ser modificados; 'reuso e redistribuição', deve ser possível reutilizar e redistribuir os dados além de poder combiná-los com outros; e 'participação universal', lembra que todos devem ser capazes de trabalhar com os dados, não havendo restrições. Obedecendo às três propriedades, o direito de construir um conhecimento, por meio de informações levantadas, será ainda mais garantido.

É muito importante que se incentive cada vez mais a produção de dados abertos, pois ao observar os membros do G7[1] (grupo composto pela França, Alemanha, Estados Unidos, Reino Unido, Japão, Itália e Canadá), percebe-se que seu crescimento dá-se pela geração e consumo de dados, visto que são qualificados como a 'sociedade da informação', porque 70% da constante do Produto Interno Bruto (PIB) dependem de bens intangíveis. (ISOTANI; BITTENCOURT, 2015). Isso ocorre, porque ao ter acesso aos dados, consequentemente virá a informação, e é justamente ela que possibilitará ter o conhecimento necessário para traçar os objetivos, que mostrarão o que falta e o que deve ser feito para transformar-se em uma nação de sucesso.

A figura 1, mostra que o número de dados aumentou e tende a continuar em crescimento:

Figura 1 – Crescimento do número de dados digitais



Fonte: Isotani e Bittencourt (2015, p. 24).

Dessa forma, a figura 1 revela que até 2020 haverá 40.000 *exabytes* de dados digitais, por isso é imprescindível aprimorar as técnicas para que essa quantidade imensa de dados seja adequadamente utilizada. É importante ter a enorme quantidade de dados disponíveis de forma padronizada, acessível e gerenciável. Além da disponibilização dos dados, as relações entre eles também devem ser estabelecidas, seguindo os princípios do *Linked data*. (WORLD WIDE WEB CONSORTIUM, 2015). Assim sendo são sugeridas as tecnologias adequadas

[1] Grupo composto pelas nações mais poderosas do mundo.

para publicação dos dados que permitem conectá-los de modo que se forme uma rede de informação.

Os *datasets* podem ser derivados de diversos processos e temáticas de diferentes áreas do conhecimento. Inserido aos princípios *Linked data* e ao movimento *Open Data*, o projeto *Linking Open Data* é uma forma de tornar visível a rede de ligações entre os *datasets*. A rede em forma de 'nuvem' é mantida por Andrejs Abele e John McCrae, do *Insight Centre for Data Analytics* da *NUI Gateway* e a versão original foi desenvolvida pelos pesquisadores Richard Cyganiak e Anja Jentzsch (ABELE *et al*., 2020).

Entre uma das temáticas categorizadas por Abele e McCrae (2020) encontra-se a que corresponde aos *datasets* linguísticos, que compreendem em dados sobre os vocabulários de idiomas, índices de idiomas, como também, correspondem a sintaxe de linguagens tecnológicas, sendo fundamentais para subsidiar a construção de vocabulários e padrões de metadados, possibilitando o acesso, uso e reuso desses dados e *datasets*.

A Linguística é a ciência da linguagem. Ela passou a configurar-se como ciência a partir do "Curso de Linguística Geral" que foi conduzido por Saussure e publicado por seus alunos em 1916; nele a língua foi definida como objeto de estudo da Linguística, por meio da distinção que se estabelece entre língua e fala (GIACOMELLI; SOBRAL, 2016). De acordo com Mollica e Gonzalez (2011, p. 5), a Linguística contribui com a Ciência da Informação "[...] por meio de orientações paradigmáticas diversas, acerca de como as línguas naturais enunciam os sistemas convencionados e os modos pelos quais se dão os processos de construção, apropriação, processamento e interpretação das linguagens artificiais".

Contextualiza-se assim a importância desses *datasets* que apresentam a formalização de uma sintaxe tecnológica, no entanto, a nuvem de *datasets* não oferece uma identificação e uma aproximação real do que essa categoria vem vinculando seus dados. Tal circunstância, poderia beneficiar a outros *datasets* para se vincularem aos dados linguísticos, ou mesmo, criar uma identificação de outros setores e campos do conhecimento para ligação de seus dados. Desse ponto, incita-se como problemática de pesquisa: o que é identificado dentro das ligações e nos *datasets* linguísticos na iniciativa *Linking Open Data*?

1.1 Objetivos

O objetivo geral é mapear os *datasets* correspondentes a categoria 'Linguística' inseridos na iniciativa *Linking Open Data*.

Os objetivos específicos são:

- Descrever os *datasets* linguísticos da iniciativa *Linking Open Data*;
- Analisar as potencialidades descritivas para a categoria ‘Linguística’ da iniciativa *Linking Open Data*, colaborando com a iniciativa na identificação de possíveis subcategorias;
- Verificar as tecnologias utilizadas pelos *datasets* linguísticos descritos.

1.2 Justificativa

Partindo dos estudos que foram possíveis devido à disciplina optativa “Tecnologias de representação de conteúdos informacionais”, no curso de Biblioteconomia e Ciência da Informação da Universidade Federal de São Carlos (UFSCar) e da integralização do curso de Letras pela aluna, pode-se perceber que é necessário fazer uma melhor estruturação de dados, para que os usuários possam recuperar toda a informação que precisam de modo mais fácil e amplo. Com esse estudo buscou-se colaborar com a Ciência da Informação de duas formas: considerando a proximidade da Linguística e da Ciência da Informação alguns *datasets* descritos podem ser úteis à Ciência da Informação; e ao explicitar as tecnologias que estão sendo utilizadas para representação de recursos informacionais demonstra-se que seu uso possibilita uma recuperação da informação mais eficiente.

Além disso, houve uma motivação pessoal que foi o interesse em trabalhar com dados, para que, dessa forma, seja possível alcançar um maior conhecimento da área, que possibilitará, posteriormente, o desenvolvimento de outros projetos, permitindo o aperfeiçoamento no tema, o qual é de extrema relevância para a sociedade, pois como ressalta Beal (2004) a informação e conhecimento são cada vez mais valiosos. A raridade da informação está no fato de que com ela pode-se fazer previsões, compreender os fenômenos que ocorrem, respondendo aos questionamentos que podem surgir.

Ademais, percebe-se que é necessário o desenvolvimento de mais pesquisas, visto que no Brasil, os dados disponibilizados publicamente, não se fazem de forma estruturada e nem “[...] seguindo esquemas e padrões de metadados reconhecidos internacionalmente. Também não houve nenhuma preocupação em estabelecer ontologias ou aplicar tecnologias da *Web* Semântica, para que pudessem ser disponibilizados e integrados ao [*Linking Open Data*] LOD.” (SANTARÉM SEGUNDO, 2015, p. 17). Com este estudo procurou-se incentivar a estruturação correta desses dados.

Por esse motivo, o desenvolvimento dessa pesquisa justifica-se pela necessidade de um melhor aproveitamento da enorme quantidade de dados que estão disponíveis na *Web*.

Proporcionando o conhecimento de forma mais estruturada e rápida para a sociedade, tendo o acesso à informação facilitado, concedendo-lhe alcançar o conhecimento almejado, que é algo inestimável.

1.3 Procedimentos metodológicos

Diante do propósito investigativo, a pesquisa configura-se como exploratória e qualitativa e de natureza teórico-aplicada, abordando como tema principal de pesquisa o mapeamento dos *datasets* linguísticos no *Linking Open Data*.

Nesse sentido, em um primeiro momento foi feita uma investigação teórica e prática acerca da identificação dos *datasets* linguísticos e das tecnologias empregadas na ligação desses dados. Em um segundo momento, direcionou-se a investigação para análise da categoria Linguística da iniciativa *Linking Open Data*.

Para a contextualização teórica, foram utilizadas fontes bibliográficas como fundamentação para os resultados, e por essa razão, a pesquisa refere-se a uma pesquisa bibliográfica. Gil (2008, p. 44) conceitua que "A pesquisa bibliográfica é desenvolvida com base em material já elaborado, constituído principalmente de livros e artigos científicos.". Além disso, a pesquisa teve caráter exploratório, com a finalidade de proporcionar a familiaridade com a área de estudo e a sua delimitação (GIL, 2008), o que dará base teórica para a construção da pesquisa.

Foram consultadas fontes primárias, secundárias e terciárias nas bases de dados da Ciência da Informação, tais como Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI), *Library and Information Science Abstracts* (LISA), bem como periódicos da área de Ciência da Informação no Portal de Periódicos da CAPES. Além disso, consultaram-se as bases internacionais *Web of Science, Scopus e Scientific Electronic Library Online* (SciELO).

Para contemplar as publicações pertinentes ao tema, as estratégias de busca utilizadas consistiram nos seguintes termos "dados abertos", "*Linking Open Data*", "*Linked data*", "metadados" e "representação da informação" em português, inglês e espanhol. O período de busca contemplou publicações a partir do ano 2000 (período em que se observa o início do crescimento das discussões acerca da temática) até 2018.

De acordo com os materiais e métodos expostos, o plano de trabalho para a execução deste Trabalho de Conclusão de Curso foi baseado em cinco etapas, conforme segue:

1ª etapa: Levantamento bibliográfico e seleção do material obtido - identificação das publicações que puderam criar embasamento teórico sobre a temática de pesquisa, por meio do levantamento bibliográfico que foi realizado em nível nacional e internacional, levando em consideração materiais impressos e digitais.

2ª etapa: Leitura, interpretação, análise e sistematização das informações – leitura do material bibliográfico selecionado para o desenvolvimento da base teórica, bem como a sua prática na identificação dos *datasets* no *Linking Open Data*, abrangendo assim o primeiro objetivo específico dessa pesquisa.

3ª etapa: Análise e estabelecimento das características fundamentais extraídas da literatura – análise das principais características encontradas na literatura sobre o tema para elucidação do problema de pesquisa, contemplando o segundo objetivo específico dessa pesquisa.

4ª etapa: Sistematização do estudo exploratório – identificação e análise dos *datasets* linguísticos inseridos na iniciativa *Linking Open Data* e das tecnologias utilizadas, abrangendo assim objetivo geral dessa pesquisa e terceiro objetivo específico.

5ª etapa: Elaboração e redação final da pesquisa – desenvolvimento do Trabalho de Conclusão de Curso (TCC) para a defesa. Essa etapa não se constitui unicamente do momento final da pesquisa, mas sim, de um processo construtivo e contínuo conjunto da orientadora e aluna.

1.4 Estrutura do trabalho

O presente trabalho está estruturado do seguinte modo:

- **Capítulo 1** – Contextualiza o tema e a pesquisa, apresentando os objetivos, justificativa, procedimentos metodológicos, plano de trabalho e cronograma.
- **Capítulo 2** – Exibe as relações que se estabelecem entre *Linked Data* e *Linked Open Data*, além da importância dos metadados no ambiente digital.
- **Capítulo 3** – Apresenta as tecnologias semânticas e o princípio de “5 estrelas” como forma de avaliar os dados.
- **Capítulo 4** – Revela quais *datasets* compõem os dados linguísticos da Iniciativa *Linking Open Data* (LOD), por meio de um mapeamento linguístico e da identificação das tecnologias empregadas nos dados linguísticos.
- **Capítulo 5** – Considerações finais sobre a pesquisa realizada.

2 *LINKED DATA* E *LINKED OPEN DATA*

Para a melhor utilização dos dados, vê-se a necessidade da semântica na *Web*, como frisado por Tim Berners-Lee em maio de 1994 na *First International Conference on World Wide Web*. Ele esclareceu que o modo como os documentos na *Web* é apresentado aos usuários, impedia que as máquinas pudessem compreender o significado, visto que o uso de hipertextos não possibilita às máquinas identificar formas diferentes e significados distintos entre uma relação e outra (ISOTANI; BITTENCOURT, 2015). O hipertexto trata-se de um texto estruturado por meio de um conjunto interligado de elementos informacionais (LAUFER, 2015). Nessa *Web*, para o ser humano fica claro, por exemplo, por que um texto está conectado a uma determinada imagem. No entanto, isso é incompreensível para as máquinas (WOOD *et al*, 2014), visto que não é concedida uma informação, mediante a construção de uma estrutura semântica, para que ela possa extrair significado do que está sendo interligado.

Por isso, a *Web* Semântica tem como objetivo promover um melhor tratamento informacional dos dados digitais, além de preocupar-se com o usuário proporcionando-lhe uma busca e recuperação com maior eficiência, por meio da abertura de dados, promovendo a troca de informação e conhecimento (SANTARÉM SEGUNDO; SIMIONATO, 2016). Essa melhora no tratamento informacional dá-se na medida em que, como é proposto pela *Web* Semântica, seja conferido "sentido" aos dados. Com isso a recuperação da informação tornar-se-á mais eficaz, pois a máquina conseguirá "entender" o que o usuário busca, para poder retornar-lhe resultados precisos. Além de acarretar conhecimento, visto que os dados estarão conectados entre si, formando uma rede, que leva ao aprofundamento da informação, contribuindo para a geração do conhecimento. Entendendo a recuperação da informação como o encontro de um material que satisfaça uma necessidade informacional, geralmente, essa atividade era própria de profissionais da informação, mas atualmente, uma infinidade de pessoas participa desse processo ao realizar uma pesquisa na *Web*, por exemplo (MANNING; RAGHAVAN; SCHÜTZE, 2008).

Para o alcance de tal finalidade, há o desenvolvimento do *Linked Data* que apresenta quais tecnologias devem ser utilizadas. E ao fazer uso dessas tecnologias é perceptível sua influência nas questões de organização e representação de informações, quando se pensa no ambiente digital (NININ, 2018). A recomendação de quais tecnologias serão necessárias, é algo de extrema importância, pois deve haver uma padronização para representar e organizar

as informações digitais, para que todos possam ter acesso a elas, e ao mesmo tempo entender como estão estruturadas.

A conexão de dados propicia a:

> [...] a geração de novos dados, apresentação de resultados, relação com outros grupos de dados, aumento do conhecimento para tomadas de decisão, novos modelos de dados gerados a partir do relacionamento e cruzamento de dados [....] além da geração de novos modelos mentais de apresentação da informação de forma a facilitar o acesso dos dados pela sociedade civil. (SANTARÉM SEGUNDO, 2015, p. 5).

Ao realizar uma representação em um sistema de informação, são empregados metadados que podem ser definidos como elementos representativos de uma determinada entidade, fazendo com que um recurso informacional se torne único, favorecendo sua recuperação (ALVES, 2010). Ou seja, os metadados são uma espécie de "explicação" sobre os dados que são inseridos. Como frisado por Ninin (2018, p. 40), eles são fundamentais para a troca recíproca de dados:

> No ambiente digital, em especial no escopo da proposta *Web* Semântica, [...] metadados [...] são os elementos "chave" para a descrição e manutenção de recursos digitais. Ao mesmo tempo, a estruturação padronizada desses metadados é determinante para as questões de interoperabilidade, como apontado pelos padrões de formato e intercâmbio de dados.

Com isso percebe-se o quão significativo são os metadados, uma vez que são responsáveis por estabelecer uma "comunicação" entre os sistemas, economizando tempo, posto que possibilita a reutilização dos dados ao fornecer "[...] informações adicionais sobre os dados, para ajudar desenvolvedores de aplicações e usuários finais a entender melhor o significado dos dados publicados, o seu conteúdo, a sua estrutura." (LAUFER, 2015, p. 20).

De acordo com Rozsa, Dutra e Nhacuongue (2017, p. 35), o *Linked Data*, também conhecido como *Web* de Dados (SERRA; SANTARÉM SEGUNDO, 2017), é uma ferramenta moderna, responsável por satisfazer "[...] as necessidades de organizar, armazenar, disseminar, e acessar a informação [...]". E é exatamente nesse momento em que os metadados são utilizados, como forma de organizar o modo de descrever a informação, para que seja possível recuperá-la, em razão de que sem uma descrição padronizada e detalhada torna-se inviável encontrar aquilo que se busca.

O objetivo desta ferramenta (*Linked Data*), é incentivar as organizações, por meio do desenvolvimento de guias, a disponibilizar de modo gratuito o conteúdo que geram, na forma

de triplas *Resource Description Framework* (RDF)[2], e usando *Uniform Resource Identifier* (URI)[3] para nomear os itens (SERRA; SANTARÉM, 2017). Por isso pode-se dizer que na *Web* tradicional o rastreamento de documentos acontece por intermédio de links de hipertexto, e na *Web* de dados, isso ocorre por meio dos links RDF (BIZER *et al*., 2008). Algo interessante no *Linked Data*, é que não apenas diz como tudo deve ser feito, mas disponibiliza os dados, como modo de mostrar que isso é executável, que esse é o caminho que a informação deve seguir para melhor atender seus usuários, indo além dos links de hipertexto, ligando os dados anteriormente, no momento de representá-los.

Diferenciando *Linked Data* de *Linked Open Data*, tem-se o seguinte: o *Linked Data*, literalmente, refere-se a dados ligados ou conectados. Em contrapartida o *Linked Open Data* apresenta dois conceitos, que podem ser assim definidos:

> [...] *linked* define a capacidade de um dado publicado na *Web* se conectar facilmente com informações relacionadas e, a partir destas ligações, este dado (*data*) poder ser acessado por computadores e pessoas. O segundo conceito estabelece que os dados que serão relacionados (*linkados*) e reutilizados devem ser abertos e livres de restrições de direitos autorais. (SERRA; SANTARÉM, 2017, p. 11).

Nota-se que o *Linked Open Data* é um passo a mais, quando comparado com o *Linked Data*, pois além de os dados estarem conectados, seu acesso é aberto, permitindo que a informação ultrapasse a barreira dos direitos autorais, fazendo com que o conhecimento esteja ao alcance de todos.

Nessa perspectiva encontra-se a ideia de democratização do acesso à informação. Segundo Oliveira (2013), a Constituição da República Federativa do Brasil de 1988 prevê o direito à informação, além disso a autora destaca que por meio das tecnologias de informação e comunicação ampliam-se as possibilidades de produção e disponibilização da informação.

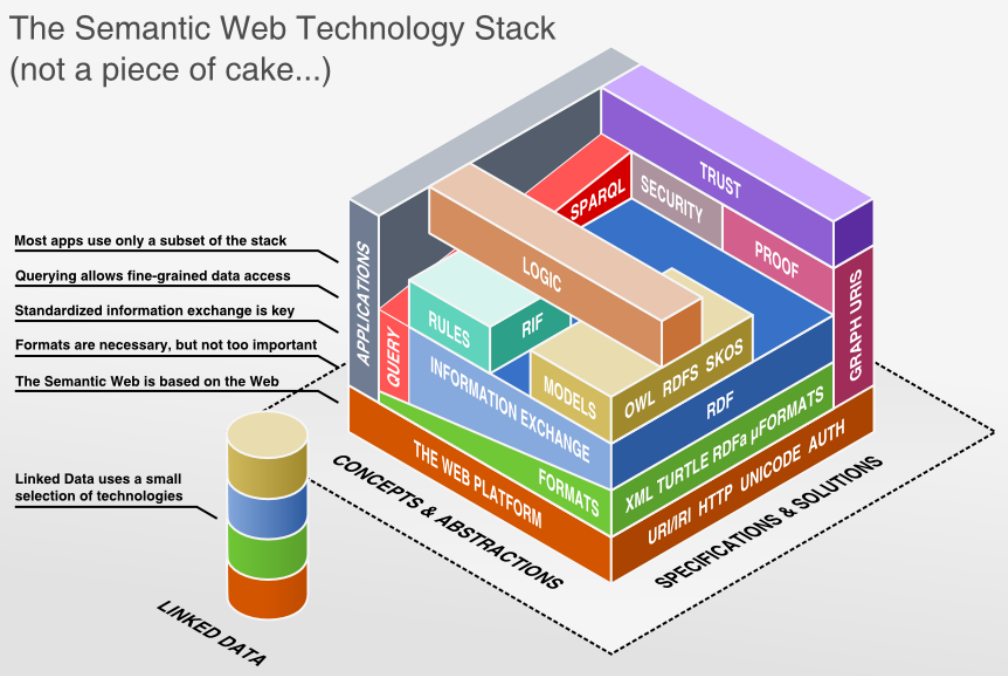
Sendo assim, a Iniciativa *Linking Open Data* ao colocar em prática os princípios do *Linked Open Data* viabiliza o acesso aos *datasets*, fazendo com que a informação esteja ao alcance de todos.

[2] O RDF é um modelo de dados que descreve recursos informacionais por meio de triplas – assunto, predicado e objeto.
[3] O URI é um identificador de recursos informacionais presentes na Internet.

3 TECNOLOGIAS SEMÂNTICAS

Em 2001, no artigo *The semantic Web: a new of Web content that is meaningful to computers will unleash a revolution of new possibilities*, Tim Berners-Lee, Hendler, e Lassila (2001) descrevem o funcionamento da *Web* Semântica. Para tornar mais fácil a visualização Tim Berners-Lee: "[...] propôs um modelo em camadas, conhecido como 'bolo de noiva' ou 'pirâmide da *Web* Semântica', descrevendo os recursos e as linguagens para a *Web* Semântica." (ISOTANI; BITTENCOURT, 2015, p. 29):

Figura 2 – Tecnologias semânticas



Fonte: Nowack (2009).

A *Web* é composta por diferentes camadas e em cada uma delas é preciso utilizar determinadas tecnologias para que os dados possam se relacionar e atribuir semântica aos dados. Na "Plataforma *Web*" tem-se URI/IRI e o *Hypertext Transfer Protocol* (HTTP); em "Formatos" aparece o *eXtensible Markup Language* (XML), *Terse RDF Triple Language* (TURTLE) e *Resource Description File in Attributes* (RDFa); em "Troca de informações" encontra-se o RDF; nos "Modelos" há *Ontology Web Language* (OWL), RDF *Schema* (RDF-S) e o *Simple Knowledge Organization System* (SKOS); por fim na camada "Consulta" localiza-se o *SPARQL Protocol and RDF Query Language* (SPARQL).

O URI/IRI é um identificador que permite tornar único um determinado recurso informacional (ARAÚJO; SOUZA, 2011). Utiliza-se para diferenciar um recurso de qualquer outro, para que sua recuperação seja unívoca. Além disso, caracteriza-se por três aspectos:

“Uniformidade: usar um único tipo de recurso; recurso: tudo aquilo que um URI pode identificar; identificador: informação que serve para identificar e diferenciar um recurso de outro.” (ISOTANI; BITTENCOURT, 2015, p. 57). Percebe-se que o URI é responsável por identificar um recurso informacional, tornando-o único, frente a tantos outros recursos que se tem acesso atualmente.

O meio pelo qual o usuário tem acesso a um dado recurso é o HTTP. Trata-se de um “[...] protocolo de comunicação para acesso aos documentos.” (LAUFER, p.2015). A linguagem XML permite a criação das próprias *tags* para registrar como se configuram as páginas da *Web* e suas seções textuais, possibilitando sua estruturação. TURTLE é um formato para representação de recursos informacionais descritos em RDF. O RDFa possibilita a inclusão de metadados nas páginas *Web* (LAUFER, 2015).

O RDF é a tecnologia que mais influencia a W*eb* Semântica, sendo considerado “[...] um modelo de dados que permite a representação do conhecimento em forma de rede [...] relaciona um recurso, chamado de sujeito, por meio de uma propriedade, ou predicado, a um valor, também referido como objeto.” (TADINI; CONEGLIAN; SANTARÉM SEGUNDO, 2017, p. 5). É considerado a tecnologia de maior influência, pois é o encarregado de descrever os recursos informacionais, por meio de triplas: assunto, predicado e objeto. Além disso, faz uso da semântica do *Dublin Core* para fornecer elementos de título, autor e editor de uma descrição (BROOKS, 2008), sendo responsável por representar e transmitir metadados (SERRA; SANTARÉM SEGUNDO, 2017), além de conectar recursos, tendo por objetivo “[...] criar uma rede de informações a partir de dados distribuídos.” (SANTARÉM SEGUNDO; CONEGLIAN, 2016, p. 220). O *Dublin Core* é um esquema de metadados, específico para recursos da *Web*; e juntamente com todas as outras tecnologias desempenha o importante papel de proporcionar a criação dessa rede, capaz de ampliar o conhecimento, resolvendo o problema “[...] da explosão de nomenclaturas diferentes e as várias situações em que a interpretação dos dados de maneira unívoca não é possível.” (SOUZA; ALVARENGA, 2004, p. 135). Portanto, o RDF será a tecnologia utilizada para fazer a descrição dos recursos informacionais e que possibilitará a conexão com outros dados, sendo imprescindível para a criação da *Web* de dados que levará a informação em forma de rede, visando a construção do conhecimento, que se estenderá cada vez mais, conforme a decisão do usuário.

De acordo com Isotani e Bittencourt (2015, p. 68), o RDF-S “[...] é um vocabulário para modelagem de dados que amplia a expressividade do RDF [...]”, pois propicia a descrição dos recursos e das relações estabelecidas entre eles.

Como a OWL é utilizada para construir ontologias, faz-se necessário compreender o que é uma ontologia. A ontologia, termo emprestado da filosofia, pretende solucionar situações em que os usuários procuram por informações com diferentes identificadores, mas com significados semelhantes. A ontologia, no contexto de *Web* e inteligência artificial, é um arquivo que define a relação entre termos, sendo seus dois modos mais típicos a taxonomia (define classes de objetos e as relações entre eles) e as regras de inferência (buscam deduzir as relações entre diferentes termos). (BERNERS-LEE; HENDLER; LASSILA, 2001). A ontologia é o que promoverá maior eficiência e eficácia na recuperação de dados, pois contém um vocabulário padronizado, além de ser reconhecido internacionalmente, assegurando a construção de: "[...] uma relação organizada entre termos dentro de um domínio, favorecendo a possibilidade de contextualizar os dados, tornando mais eficiente e facilitando o processo de interpretação dos dados pelas ferramentas de recuperação da informação." (SANTARÉM SEGUNDO, 2015, p. 7). O RDF-S e a OWL são considerados vocabulários básicos "[...] que apresentam diferentes níveis de expressividade e podem ser utilizados na definição de novos vocabulários." (ROZSA; DUTRA; NHACUONGUE, 2017, p. 39). Especificamente a OWL é uma linguagem de construção de ontologias, que permite a contextualização, tornando possível a conceituação de domínios e processos em um computador (SANTARÉM SEGUNDO; CONEGLIAN, 2016). Considera-se que ontologia é aquela que "[...] define os termos usados para representar uma área do conhecimento por meio de definições e conceitos básicos, de diferentes domínios e legíveis por máquina." (SANTARÉM SEGUNDO; SIMIONATO, 2016, p. 5).

O SKOS é visto como uma linguagem de construção de ontologias simplificada, muito utilizada para representação de tesauros na *Web*. Conforme afirmam Ramalho, Vidotti e Fujita (2007, p. 9) o SKOS faz uso de um

> [...] conjunto de propriedades descritas em *Resource Description Framework*, RDF, a partir de classes RDFS, que podem ser utilizadas para expressar o conteúdo e a estrutura de um esquema de conceito como um gráfico RDF, possibilitando descrever formalmente os termos e relacionamentos existentes em um tesauro [...]

O SPARQL é uma linguagem computacional que possibilita realizar consultas semânticas em um determinado *dataset* (ARARAKI, SIMIONATO, SANTOS, 2017). Essa linguagem é aquela que auxilia nas consultas em dados iguais aos da *Web* Semântica, que são os dados estruturados, por meio do padrão RDF. Contudo, como ressaltado por

DuCharme (2013) não se limita, somente, a consultar dados armazenados em formato RDF, sendo esse um de seus aspectos mais significativos.

De acordo com Segaran, Evans e Taylor (2009), URI, RDF e SPARQL é o que se necessita para o início da construção de aplicativos que considerem o aspecto semântico, visto que o URI garante que se esteja falando da mesma coisa que outra pessoa, o RDF é o padrão para a representação e compartilhamento de dados semânticos e o SPARQL propicia a consulta a esses dados.

Esses recursos e linguagens da *Web* Semântica são o que permite conectar dados. Dados Conectados são: "[...] um conjunto de boas práticas para publicar e conectar conjuntos de dados estruturados na *Web*, com o intuito de criar uma '*Web* de Dados'." (BIZER *et al*., 2006 apud ISOTANI; BITTENCOURT, 2015, p.31). A *Web* de Dados apresenta os seguintes padrões: "mecanismo de identificação e acesso universal: URIs e HTTP; modelo padrão para representação de dados: RDF; linguagem de consulta para acesso aos dados: SPARQL." (ISOTANI; BITTENCOURT, 2015).

O XML E o RDF, tem como principal objetivo "[...] criar uma rede de informações a partir de dados distribuídos." (SANTARÉM SEGUNDO, 2015, p. 6), codificando as informações em estruturas triplas (sujeito – verbo – objeto), fazendo declarações para especificar um item particular, pessoas, páginas da *Web*, etc.; suas propriedades, profissões, graus parentescos, etc; e seus valores, proporcionando distinções sobre as informações pesquisadas na rede. Tais itens são identificados por URIs, permitindo a criação e acesso por qualquer usuário. (BERNERS-LEE; HENDLER; LASSILA, 2001). XML é usado, porque o RDF estrutura documentos da W*eb* tendo por base essa linguagem (BROOKS, 2008). Portanto, pode-se inferir que é por meio do XML que os recursos são descritos.

Assim sendo, percebe-se que a *Web* Semântica: "[...] visa facilitar a obtenção, classificação e organização de informações, de forma estruturada, de modo que os dados estejam disponíveis tanto para as pessoas quanto para as máquinas." (PINHEIRO, 2009, p. 27), além de "[...] diminuir ou eliminar as dificuldades relacionadas com o acesso à informação." (PINHEIRO, 2009, p. 27). Por causa disso, pode-se afirmar que a *Web* Semântica se configura como uma solução tecnológica para atender da melhor forma possível as necessidades de informação que cada usuário apresente. Uma vez que a estruturação dos dados levará ao alcance da informação de modo mais satisfatório.

Antes de começar qualquer trabalho é requerido avaliar qual o tipo de dado que se tem disponível. Para isso, utiliza-se o princípio de "5 estrelas", proposto por Tim Berners-Lee; quanto mais estrelas os dados tiverem, mais abertos são e por consequência melhores

conectados. Porém só são considerados dados abertos conectados aqueles que possuem ao menos quatro estrelas. Para ter uma estrela basta que os dados estejam disponíveis na *Web* com licença aberta. Como exemplo de dados com uma estrela, Isotani e Bittencourt (2015) exemplificam com um arquivo em formato *pdf* que disponibiliza informações sobre a temperatura de Galway, na Irlanda. Já possui uma estrela, porque foi publicado em formato aberto, mas não permite a manipulação e reutilização facilmente.

Para passar a ter duas estrelas, além de estar sob licença aberta, deve apresentar um formato estruturado. O exemplo apresentado pelos autores é a disponibilização em uma planilha *Excel* em XLS. Esse formato permite a utilização de dados mais facilmente por máquinas e pessoas. O problema é que os dados continuam 'trancados' em um documento (no exemplo anterior em *pdf*, agora em *excel*), além de estar disponível em um formato proprietário da Microsoft. (ISOTANI; BITTENCOURT, 2015).

Consegue-se três estrelas, quando se usa um formato aberto e não proprietário. Como exemplo, há o formato *csv* (*Comma-Separated Values*); a manipulação torna-se simples e não fica restrita ao *software* proprietário. O *csv* armazena os dados de modo tabular, sendo que as colunas separam-se por vírgulas e os registros por novas linhas. No entanto, esse formato não é capaz de representar detalhes importantes, e sua manipulação não é flexível. (ISOTANI; BITTENCOURT, 2015).

"A quarta estrela está relacionada com a utilização de URIs para identificar os recursos na *Web*." (ISOTANI; BITTENCOURT, 2015, p. 50). Seu uso é ainda mais simples do que em *csv*. A utilização do URI permite o compartilhamento de dados que são publicados. O formato para representação desses dados é o RDF, como exemplo os autores disponibilizam uma página em *HyperText Markup Language* (HTML) e explicam que o RDF permite a serialização e disponibilização de diferentes formas, sendo que o RDFa, concede atributos às *tags* HTML, possibilitando a disponibilização em HTML com atributos de um documento RDF.

Isotani e Bittencourt (2015, p. 53) apresentam o que é possível fazer com a disponibilização do URI:

> a) conectar e combinar estes dados com outros dados;
> b) reusar estes dados em outros contextos;
> c) melhorar a busca e a compreensão dos dados apresentados;
> d) possibilitar inferência através de dados parciais;
> e) permitir navegação entre documentos; entre outros.

Para alcançar cinco estrelas basta conectar um primeiro *dataset* com um segundo *dataset*. No exemplo disponibilizado pelos autores, os dados sobre a temperatura de Galway são conectados aos dados que contém informações sobre a cidade (Galway) ou sobre o país (Irlanda). Tudo isso permite o acesso a outros dados de forma automática, possibilitando às máquinas fazer uma descrição da cidade, ou apresentação de fotos turísticas, ou fornecer um mapa com a localização de Galway, na Irlanda. (ISOTANI; BITTENCOURT, 2015).

Percebe-se que para conseguir cinco estrelas é preciso seguir uma série de recomendações que farão com que a estruturação e o significado dos dados estejam ao alcance tanto das pessoas, quanto das máquinas.

Quando se pensa em representar um conhecimento, três problemas tornam-se visíveis: descrição, representação e interpretação. Na descrição deve-se ter prudência ao determinar um vocabulário para representação de um conceito. A segunda dificuldade é a representação dos dados de forma que seja mais expressivo quanto possível, diminuindo a ambiguidade. Na interpretação é preciso ser muito restrito para que as interpretações ocorram de forma correta. (ISOTANI; BITTENCOURT, 2015). Ou seja, é preciso considerar aspectos que um bibliotecário que realiza catalogação tem contato. Quando irá catalogar precisa descrever de uma forma padrão, e além disso, pensar em um modo de representar o recurso informacional, fazendo com que o usuário tenha a mesma interpretação que ele, para conseguir selecionar o recurso que lhe será útil.

Na *Web* para transpor o obstáculo da interpretação, faz-se uso da linguagem *Web Ontolgy Language* (OWL), que mostra tanto para as pessoas, quanto para as máquinas como os conceitos estão relacionados. A OWL "[...] é considerada a linguagem de ontologias da *Web* e é bastante utilizada para o desenvolvimento de aplicações baseadas na *Web* Semântica." (ISOTANI; BITTENCOURT, 2015, p. 108), ou seja, a OWL é a linguagem que expressa as ontologias (ROCHA, 2012). Dessa forma, a representação na *Web* ocorre por meio da OWL e da ontologia. As ontologias disponibilizam "[...] uma estrutura conceitual comum sobre a qual podemos desenvolver bases de conhecimento compartilháveis e reutilizáveis. [...] facilitam a interoperabilidade e a fusão das informações [...]" (ISOTANI; BITTENCOURT, 2015, p. 96).

Isotani e Bittencourt (2015) apontam as fases do ciclo de vida de uma ontologia:

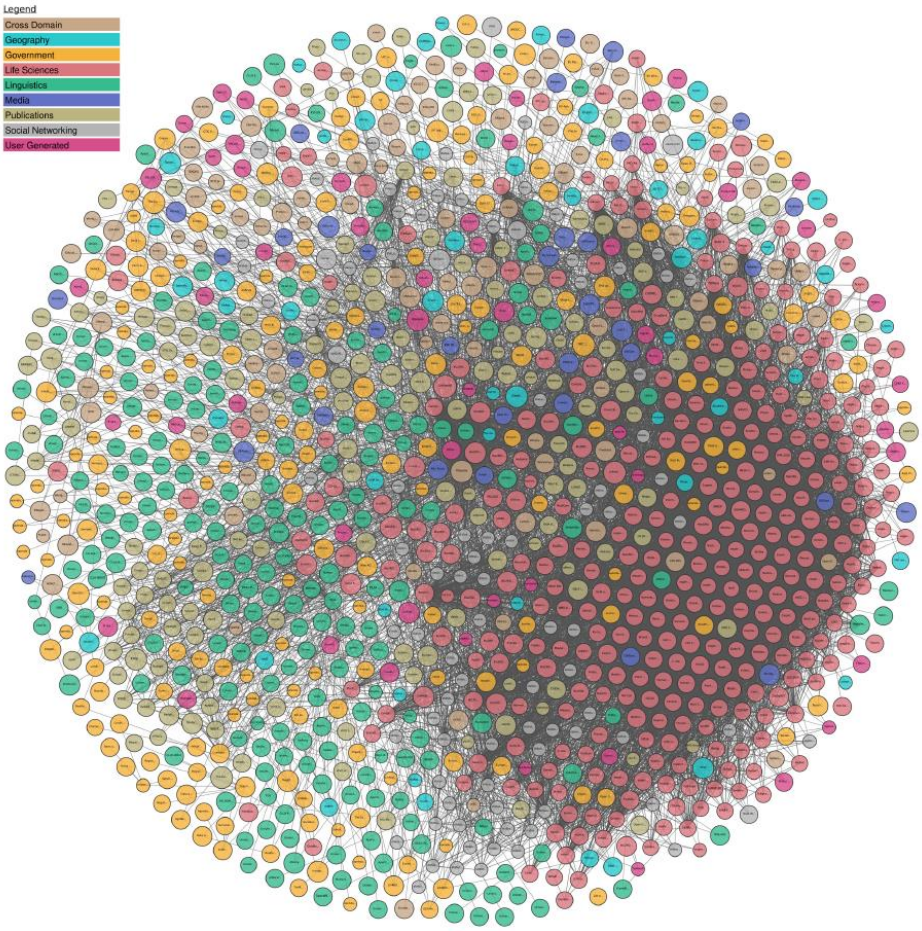
- "Especificação": é mostrada a utilidade da ontologia e a que usuários pode ser útil;
- "Conceitualização": organização do conhecimento de domínio da ontologia;

- “Formalização”: estabelecimento de conceitos, relações e axiomas (verdades que são consideradas absolutas);
- “Implementação”: colocar em prática a ontologia, por intermédio de uma linguagem;
- “Manutenção”: depois de implementar, ela deve ser mantida.

Portanto, é preciso aprender a trabalhar com essas tecnologias para conectar dados, pois: “[...] tanto o mercado quanto a academia exigem base de dados conectados, altamente compartilháveis, que permitam a interoperabilidade e a possibilidade de lidar com o acúmulo de conhecimento (isto é, novos dados conectados) disponível na *Web*.” (ISOTANI; BITTENCOURT, 2015, p. 99-100). Nesse viés, o *Linked Open Data* vai além do *Linked Data*, preocupando-se, além da ‘interoperabilidade técnica’, com a ‘interoperabilidade legal’ dos dados. (ARAKAKI, 2016, p. 28). É possível afirmar que os dados abertos conectados são o caminho da Ciência, por isso é preciso aprender a lidar com eles, para oferecer ao usuário a melhor forma de recuperar conteúdos relevantes, que possibilitarão a construção do conhecimento, que posteriormente virá a ser divulgado.

Como já mencionado, a iniciativa consolidada denominada como *Linking Open Data* apresenta *datasets* que foram publicados no formato de dados ligados, pelos contribuintes para o projeto de conexão da comunidade de dados abertos e outras pessoas e organizações, visualizado pela figura 3.

Figura 3 – Nuvem do *Linking Open Data*



Fonte: Abele e McCrae (2020).

A figura 3 é a última atualização da nuvem de *datasets*, realizada em julho de 2020. A coleta para contribuição à nuvem está baseada em metadados coletados e organizados pelos colaboradores do projeto, depositados no *DataHub*.

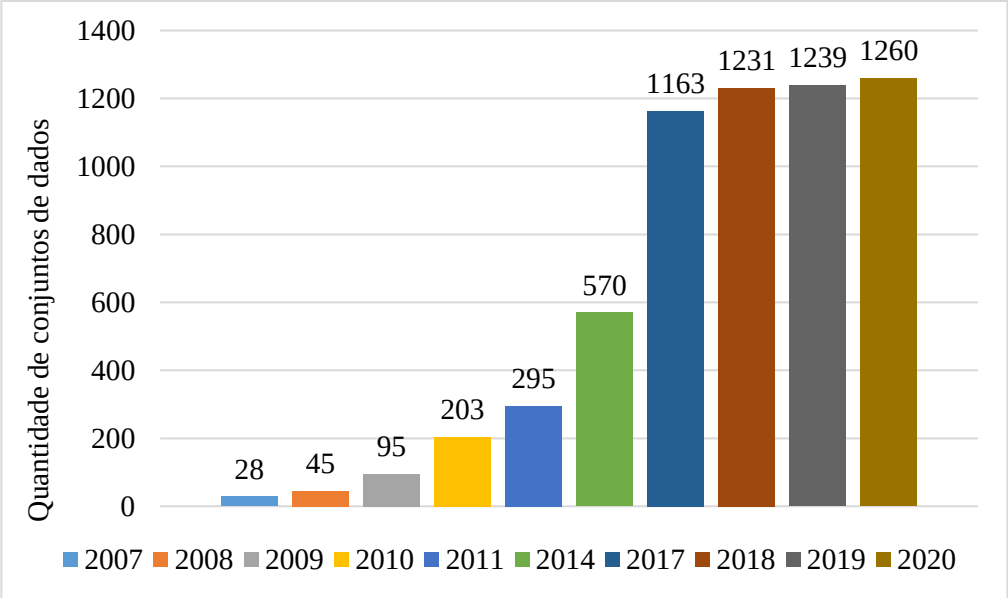
4 INICIATIVA *LINKING OPEN DATA* (LOD): DADOS LINGUÍSTICOS

Diante de tudo o que foi exposto, é notável que o *Linking Open Data* (LOD) consiste em uma aplicação de todos os princípios *Linked Open Data*. Di Noia *et al*. (2012) destaca que os dados que fazem parte da nuvem LOD são de qualidade e podem ser usados como fonte principal de informação. O desafio consiste no fato de explorar toda essa quantidade de dados. É perceptível que "[...] os dados publicados não estão apenas disponíveis para visualização e acesso, mas estão estruturados de modo a permitir sua efetiva manipulação, uso e reuso." (NININ, 2018, p. 43).

O LOD foi fundado em 2007 (BIZER, 2008) e desde esse ano sua manutenção é feita pela W3C. Ele apresenta uma série de *datasets* distribuídos por categorias: "[...] dados de publicação; ciências da vida; domínio geral (*cross-domain*); dados geográficos; dados governamentais; mídia; dados de uso geral; dados de redes sociais e linguística." (SANTARÉM SEGUNDO; SIMIONATO, 2016, p. 7).

Os *datasets* dessa iniciativa passam por um constante crescimento, como pode ser visto no gráfico abaixo:

Gráfico 1 – Crescimento dos *datasets* do LOD por ano



Fonte: elaborado pela autora, dados retirados de Abele e McCrae (2020).

Desses 1260 *datasets*, 226 são dados linguísticos que estão distribuídos em sete categorias: *Corpora* (Corpus); *Lexicons and Dictionaries* (Léxicos e Dicionários); *Terminologies, Thesauri and Knowledge Bases* (Terminologias, Tesauros e Bases de Conhecimento); *Linguistic Resource Metadata* (Metadados de Recursos Linguísticos);

Linguistic Data Categories (Categorias de Dados Linguísticos); *Typological Databases* (bases de dados tipológicas); *Other* (Outros).

Interessante notar que há uma categoria que abarca os tesauros, pois apresentam aspectos em comum. O tesauro, conforme Dodebei (2002) tem a função de recuperar a informação, ele surgiu da necessidade de manipulação de uma variedade de documentos, na década de 1940.

4.1 Mapeamento dos dados linguísticos

Na categoria Corpus, há 73 *datasets*, descritos no quadro abaixo:

Quadro 1 – Mapeamento dos dados linguísticos: *Corpora*

Datasets	**Descrição**	**Nº de triplas**	***Websites***
Atlante Sintattico d'Italia (ASIt)	Atlas sintático da Itália que busca mostrar as diferenças gramaticais, ou seja, as variedades linguísticas do idioma. Reúne-se neste conjunto mais de 54.000 sentenças e mais de 240 dialetos diferentes.	420.000	http://purl.org/asit/alld
Brown Corpus in RDF/NIF	Apresenta 1.014.312 palavras retiradas de 500 documentos, incluindo jornais, livros e documentos do governo.	14.335.131	https://old.datahub.io/dataset/brown-corpus-in-rdf-nif
DBpedia abstract corpus	Compreende resumos da *Wikipedia* em seis idiomas (holandês, inglês, francês, alemão, italiano e espanhol)	743.532.157	http://downloads.dbpedia.org/2015-04/ext/nlp/abstracts/
KORE 50 NIF NER Corpus	Composto por 50 frases de diferentes áreas (música, celebridades, negócios). Tem como objetivo mostrar a dificuldade em tirar a ambiguidade, quando se refere a entidades, contendo uma considerável quantidade de nomes próprios de pessoas. Para solucionar o problema da ambiguidade é preciso recorrer ao contexto.	1.410	https://old.datahub.io/dataset/kore-50-nif-ner-corpus
Manually Annotated	Traz em torno de 500.000 palavras de dados tanto	1.000.000	http://www.anc.org/MASC/Home.html

Sub-Corpus (MASC) of the Open American National Corpus	escritos como falados do inglês americano contemporâneo. Essas palavras são retiradas do *Open American National Corpus*, além de apresentar diferentes anotações.		
News-100 NIF NER Corpus	Formado por 100 artigos de notícias na língua alemã, publicados em 2010, contendo a palavra "Golf". Considerando que essa palavra pode significar três coisas: uma localização geográfica, um esporte e um modelo de carro.	12.289	https://old.datahub.io/dataset/news-100-nif-ner-corpus
Ontologies of Linguistic Annotations (OLiA)	Esse conjunto concede uma taxonomia para as categorias de dados, modelos para esquemas de anotações, além dos relacionamentos que se estabelecem entre os dados.	43.775	http://purl.org/olia
Reuters-128 NIF NER Corpus	Abrange 128 artigos de notícias econômicas, escritos em língua inglesa.	6.967	https://old.datahub.io/dataset/reuters-128-nif-ner-corpus
RSS-500 NIF NER CORPUS	Consiste em sentenças, escolhidas aleatoriamente de uma compilação de dados de todos os principais jornais do mundo.	10.038	https://old.datahub.io/dataset/rss-500-nif-ner-corpus
Universal Dependencies Treebank Croatian[4]	Descrição simplificada das relações gramaticais da língua croata.	139.023	https://github.com/UniversalDependencies/UD_Croatian
Universal Dependencies Treebank Indonesian	Descrição simplificada das relações gramaticais do indonésio.	121.923	https://github.com/UniversalDependencies/UD_Indonesian
Universal Dependencies Treebank Slovak	Descrição simplificada das relações gramaticais do eslovaco.	106.043	https://github.com/UniversalDependencies/UD_Slovak
Universal Dependencies Treebank	Descrição simplificada das relações gramaticais do grego antigo.	206.966	https://github.com/UniversalDependencies/UD_Ancient_Greek-PROIEL

[4] Todos os *datasets* do Quadro 1 a partir desse integram o projeto denominado *Universal Dependencies*, que tem como objetivo o desenvolvimento de um corpus em vários idiomas, por meio de uma descrição simplificada das relações gramaticais, permitindo uma análise comparativa entre línguas.

Ancient_Greek -PROIEL			
Universal Dependencies Treebank Japanese	Descrição simplificada das relações gramaticais da língua japonesa.	92.033	https://github.com/UniversalDependencies/UD_Japanese
Universal Dependencies Treebank Latin	Descrição simplificada das relações gramaticais do latim.	47.303	https://github.com/UniversalDependencies/UD_Latin
Universal Dependencies Treebank Danish	Descrição simplificada das relações gramaticais do dinamarquês.	100.733	https://github.com/UniversalDependencies/UD_Danish
Universal Dependencies Treebank German	Descrição simplificada das relações gramaticais da língua alemã.	293.088	https://github.com/UniversalDependencies/UD_German
Universal Dependencies Treebank Turkish	Descrição simplificada das relações gramaticais da língua turca.	56.418	https://github.com/UniversalDependencies/UD_Turkish
Universal Dependencies Treebank Gothic	Descrição simplificada das relações gramaticais da língua gótica.	56.128	https://github.com/UniversalDependencies/UD_Gothic
Universal Dependencies Treebank Galician	Descrição simplificada das relações gramaticais do galego.	138.852	https://github.com/UniversalDependencies/UD_Galician
Universal Dependencies Treebank Finnish-FTB	Descrição simplificada das relações gramaticais do finlandês.	159.314	https://github.com/UniversalDependencies/UD_Finnish-FTB
Universal Dependencies Treebank Portuguese-Bosque	Descrição simplificada das relações gramaticais da língua portuguesa.	227.653	https://github.com/UniversalDependencies/UD_Portuguese-Bosque
Universal Dependencies Treebank Vietnamese	Descrição simplificada das relações gramaticais da língua vietnamita.	43.754	https://github.com/UniversalDependencies/UD_Vietnamese
Universal Dependencies Treebank Chinese	Descrição simplificada das relações gramaticais da língua chinesa.	123.283	https://github.com/UniversalDependencies/UD_Chinese
Universal Dependencies Treebank Portuguese-BR	Descrição simplificada das relações gramaticais da língua portuguesa do Brasil.	298.323	https://github.com/UniversalDependencies/UD_Portuguese-BR

Universal Dependencies Treebank Portuguese	Descrição simplificada das relações gramaticais da língua portuguesa.	209.977	https://github.com/UniversalDependencies/UD_Portuguese
Universal Dependencies Treebank Russian-SynTagRus	Descrição simplificada das relações gramaticais da língua russa.	1.068.483	https://github.com/UniversalDependencies/UD_Russian-SynTagRus
Universal Dependencies Treebank French	Descrição simplificada das relações gramaticais da língua francesa.	391.107	https://github.com/UniversalDependencies/UD_French
Universal Dependencies Treebank Japanese-KTC	Descrição simplificada das relações gramaticais da língua japonesa.	267.631	https://github.com/UniversalDependencies/UD_Japanese-KTC
Universal Dependencies Treebank Bulgarian	Descrição simplificada das relações gramaticais da língua búlgara.	156.319	https://github.com/UniversalDependencies/UD_Bulgarian
Universal Dependencies Treebank Sanskrit	Descrição simplificada das relações gramaticais do sânscrito.	1.031	https://github.com/UniversalDependencies/UD_Sanskrit
Universal Dependencies Treebank Spanish	Descrição simplificada das relações gramaticais do espanhol.	423.346	https://github.com/UniversalDependencies/UD_Spanish
Universal Dependencies Treebank Old_Church_Slavonic	Descrição simplificada das relações gramaticais de dados de uma igreja antiga eslava.	57.507	https://github.com/UniversalDependencies/UD_Old_Church_Slavonic
Universal Dependencies Treebank Swedish	Descrição simplificada das relações gramaticais da língua sueca.	96.819	https://github.com/UniversalDependencies/UD_Swedish
Universal Dependencies Treebank Dutch	Descrição simplificada das relações gramaticais da língua holandesa.	209.411	https://github.com/UniversalDependencies/UD_Dutch
Universal Dependencies Treebank Irish	Descrição simplificada das relações gramaticais da língua irlandesa.	23.686	https://github.com/UniversalDependencies/UD_Irish
Universal Dependencies Treebank Catalan	Descrição simplificada das relações gramaticais do catalão.	530.766	https://github.com/UniversalDependencies/UD_Catalan

Universal Dependencies Treebank Swedish-LinES	Descrição simplificada das relações gramaticais do sueco.	79.812	https://github.com/UniversalDependencies/UD_Swedish-LinES
Universal Dependencies Treebank Swedish_Sign_Language	Descrição simplificada das relações gramaticais da língua sueca de sinais.	672	https://github.com/UniversalDependencies/UD_Swedish_Sign_Language
Universal Dependencies Treebank Italian	Descrição simplificada das relações gramaticais da língua italiana.	272.913	https://github.com/UniversalDependencies/UD_Italian
Universal Dependencies Treebank Russian	Descrição simplificada das relações gramaticais da língua russa.	99.389	https://github.com/UniversalDependencies/UD_Russian
Universal Dependencies Treebank Norwegian	Descrição simplificada das relações gramaticais do norueguês.	310.222	https://github.com/UniversalDependencies/UD_Norwegian-Bokmaal
Universal Dependencies Treebank Greek	Descrição simplificada das relações gramaticais do grego.	59.156	https://github.com/UniversalDependencies/UD_Greek
Universal Dependencies Treebank Hindi	Descrição simplificada das relações gramaticais de uma língua do norte da Índia.	351.704	https://github.com/UniversalDependencies/UD_Hindi
Universal Dependencies Treebank Spanish-AnCora	Descrição simplificada das relações gramaticais do espanhol.	547.681	https://github.com/UniversalDependencies/UD_Spanish-AnCora
Universal Dependencies Treebank Uyghur	Descrição simplificada das relações gramaticais da língua uigur, específica de um povo que habita a Ásia Central.	6.442	https://github.com/UniversalDependencies/UD_Uyghur
Universal Dependencies Treebank Ancient_Greek	Descrição simplificada das relações gramaticais do grego antigo.	244.993	https://github.com/UniversalDependencies/UD_Ancient_Greek
Universal Dependencies Treebank *Czech*	Descrição simplificada das relações gramaticais do tcheco.	1.503.732	https://github.com/UniversalDependencies/UD_Czech

Universal Dependencies Treebank Slovenian-SST	Descrição simplificada das relações gramaticais da língua eslovena.	29.488	https://github.com/UniversalDependencies/UD_Slovenian-SST
Universal Dependencies Treebank Polish	Descrição simplificada das relações gramaticais da língua polonesa.	83.571	https://github.com/UniversalDependencies/UD_Polish
Universal Dependencies Treebank Hebrew	Descrição simplificada das relações gramaticais da língua hebraica.	115.535	https://github.com/UniversalDependencies/UD_Hebrew
Universal Dependencies Treebank Persian	Descrição simplificada das relações gramaticais da língua persa.	151.625	https://github.com/UniversalDependencies/UD_Persian
Universal Dependencies Treebank English-ESL	Descrição simplificada das relações gramaticais da língua inglesa.	97.681	https://github.com/UniversalDependencies/UD_English-ESL
Universal Dependencies Treebank Latvian	Descrição simplificada das relações gramaticais da língua letã.	20.798	https://github.com/UniversalDependencies/UD_Latvian
Universal Dependencies Treebank Czech-CAC	Descrição simplificada das relações gramaticais da língua tcheca.	493.306	https://github.com/UniversalDependencies/UD_Czech-CAC
Universal Dependencies Treebank Czech-CLTT	Descrição simplificada das relações gramaticais da língua tcheca.	35.084	https://github.com/UniversalDependencies/UD_Czech-CLTT
Universal Dependencies Treebank Romanian	Descrição simplificada das relações gramaticais da língua romena.	218.516	https://github.com/UniversalDependencies/UD_Romanian
Universal Dependencies Treebank Finnish	Descrição simplificada das relações gramaticais do finlandês.	181.022	https://github.com/UniversalDependencies/UD_Finnish
Universal Dependencies Treebank Tamil	Descrição simplificada das relações gramaticais da língua tâmil.	8.635	https://github.com/UniversalDependencies/UD_Tamil
Universal Dependencies Treebank Basque	Descrição simplificada das relações gramaticais da língua basca.	121.433	https://github.com/UniversalDependencies/UD_Basque

Universal Dependencies Treebank Dutch-LassySmall	Descrição simplificada das relações gramaticais da língua holandesa.	98.107	https://github.com/UniversalDependencies/UD_Dutch-LassySmall
Universal Dependencies Treebank Latin-PROIEL	Descrição simplificada das relações gramaticais do latim.	165.201	https://github.com/UniversalDependencies/UD_Latin-PROIEL
Universal Dependencies Treebank English	Descrição simplificada das relações gramaticais da língua inglesa.	254.830	https://github.com/UniversalDependencies/UD_English
Universal Dependencies Treebank Coptic	Descrição simplificada das relações gramaticais da língua cóptica.	5.220	https://github.com/UniversalDependencies/UD_Coptic
Universal Dependencies Treebank Kazakh	Descrição simplificada das relações gramaticais da língua cazaque.	6.023	https://github.com/UniversalDependencies/UD_Kazakh
Universal Dependencies Treebank Galician-TreeGal	Descrição simplificada das relações gramaticais do galego.	24.219	https://github.com/UniversalDependencies/UD_Galician-TreeGal
Universal Dependencies Treebank Hungarian	Descrição simplificada das relações gramaticais da língua húngara.	42.032	https://github.com/UniversalDependencies/UD_Hungarian
Universal Dependencies Treebank Estonian	Descrição simplificada das relações gramaticais da língua estoniana.	234.351	https://github.com/UniversalDependencies/UD_Estonian
Universal Dependencies Treebank Ukrainian	Descrição simplificada das relações gramaticais da língua ucraniana.	1.676	https://github.com/UniversalDependencies/UD_Ukrainian
Universal Dependencies Treebank Arabic	Descrição simplificada das relações gramaticais da língua árabe.	242.056	https://github.com/UniversalDependencies/UD_Arabic
Universal Dependencies Treebank Latin-ITTB	Descrição simplificada das relações gramaticais do latim.	291.295	https://github.com/UniversalDependencies/UD_Latin-ITTB
Universal Dependencies Treebank Slovenian	Descrição simplificada das relações gramaticais da língua eslovena.	140.418	https://github.com/UniversalDependencies/UD_Slovenian

Universal Dependencies Treebank English-LinES	Descrição simplificada das relações gramaticais da língua inglesa.	82.821	https://github.com/UniversalDependencies/UD_English-LinES

Fonte: elaborado pela autora.

Nessa categoria, de um modo geral, há dados de atlas, de jornais, de resumos, de fala e escritos, ontologias, além de dados que possibilitam a análise comparativa entre línguas.

A categoria Léxicos e Dicionários contém 72 *datasets*, descritos no quadro que se segue:

Quadro 2 – Mapeamento dos dados linguísticos: *Lexicons and Dictionaries*

Datasets	**Descrição**	**Nº de triplas**	***Websites***
Apertium RDF	Agrupamento de todos os dicionários bilíngues do RDF *Apertium*.	8.842.516	http://linguistic.linkeddata.es/apertium/
Apertium RDF ES-GL	Dicionário bilíngue (espanhol – galego).	206.284	https://old.datahub.io/dataset/apertium-rdf-es-gl
Apertium RDF EO-ES	Dicionário bilíngue (esperanto – espanhol).	390.198	https://old.datahub.io/dataset/apertium-rdf-eo-es
Apertium RDF ES-RO	Dicionário bilíngue (espanhol – romeno).	400.366	https://old.datahub.io/dataset/apertium-rdf-es-ro
Apertium RDF EN-ES	Dicionário bilíngue (inglês – espanhol).	576.322	https://old.datahub.io/pt_BR/dataset/apertium-rdf-en-es
Apertium RDF EO-FR	Dicionário bilíngue (esperanto – francês)	726.281	https://old.datahub.io/pt_PT/dataset/apertium-rdf-eo-fr
Apertium RDF EN-CA	Dicionário bilíngue (inglês – catalão).	759.601	https://old.datahub.io/pt_PT/dataset/apertium-rdf-en-ca
Apertium RDF EN-GL	Dicionário bilíngue (inglês – galego).	425.117	https://old.datahub.io/dataset/apertium-rdf-en-gl
Apertium RDF EO-CA	Dicionário bilíngue (esperanto – catalão).	426.301	https://old.datahub.io/dataset/apertium-rdf-eo-ca
Apertium RDF OC-CA	Dicionário bilíngue (occitano – catalão).	346.346	https://old.datahub.io/dataset/apertium-rdf-oc-ca
Apertium RDF ES-PT	Dicionário bilíngue (espanhol – português).	279.245	https://old.datahub.io/pt_PT/dataset/apertium-rdf-es-pt
Apertium RDF EO-EN	Dicionário bilíngue (esperanto – inglês).	617.772	https://old.datahub.io/pt_PT/dataset/apertium-rdf-eo-en
Apertium RDF OC-ES	Dicionário bilíngue (occitano – espanhol).	317.162	https://old.datahub.io/dataset/apertium-rdf-oc-es

Apertium RDF PT-CA	Dicionário bilíngue (português – catalão).	163.149	https://old.datahub.io/pt_PT/dataset/apertium-rdf-pt-ca
Apertium RDF ES-CA	Dicionário bilíngue (espanhol – catalão).	730.501	https://old.datahub.io/ca/dataset/apertium-rdf-es-ca
Apertium RDF EU-EN	Dicionário bilíngue (basco – inglês).	265.466	https://old.datahub.io/pt_PT/dataset/apertium-rdf-eu-en
Apertium RDF FR-CA	Dicionário bilíngue (francês – catalão).	152.002	https://old.datahub.io/dataset/apertium-rdf-fr-ca
Apertium RDF CA-IT	Dicionário bilíngue (catalão – italiano).	180.851	https://old.datahub.io/pt_BR/dataset/apertium-rdf-ca-it
Apertium RDF ES-AST	Dicionário bilíngue (espanhol – asturiano).	825.540	https://old.datahub.io/dataset/apertium-rdf-es-ast
Apertium RDF ES-AN	Dicionário bilíngue (espanhol – aragonês).	71.997	https://old.datahub.io/pt_PT/dataset/apertium-rdf-es-an
Apertium RDF FR-ES	Dicionário bilíngue (francês – espanhol).	495.614	https://old.datahub.io/dataset/apertium-rdf-fr-es
Apertium RDF PT-GL	Dicionário bilíngue (português – galego).	234.065	https://old.datahub.io/dataset/apertium-rdf-pt-gl
Apertium RDF EU-ES	Dicionário bilíngue (basco – espanhol).	262.336	https://old.datahub.io/pt_PT/dataset/apertium-rdf-eu-es
Basque EuroWordNet-lemon lexicon (3.0)	Apresenta o léxico da língua basco.	1.215.583	https://old.datahub.io/dataset/basque-eurowordnet-lemon-lexicon-3-0
Catalan EuroWordNet-lemon lexicon (3.0)	Mostra o léxico catalão.	1.841.180	https://old.datahub.io/dataset/catalan-eurowordnet-lemon-lexicon-3-0
CopyrightTermBank	Apresenta dados referentes à terminologia sobre direitos autorais e outros conceitos relacionados.	11.068	https://old.datahub.io/dataset/copyrighttermbank
EuroSentiment	Compreende léxicos que se referem a sentimentos.	975.566	https://old.datahub.io/dataset/eurosentiment
EMN	Traz a terminologia da *European Migration Network* (EMN) em RDF.	117.115	https://old.datahub.io/dataset/emn
Galician EuroWordNet-lemon lexicon (3.0)	Exibe o léxico galego.	734.979	https://old.datahub.io/dataset/galician-eurowordnet-lemon-lexicon-3-0
IATE RDF	Os dados foram convertidos do TBX.	8.081.142	https://old.datahub.io/dataset/iate-rdf

Lexicon of Syntactic and Semantic Framework	Contém em sua ontologia mais de 70.000 entradas lexicais da língua croata.	70.366	http://www.ss-framework.com
Linked Old Germanic Dictionaries	Apresenta recursos lexicais (como listas de palavras, dicionários etimológicos) de língua alemã em diferentes momentos históricos.	349.021	https://old.datahub.io/dataset/germlex
Open Bantu isiXhosa Lexicon	Traz dados lexicais, e morfológicos além de traduções em inglês que são conectados ao *WordNet RDF*.	0	https://github.com/MMoOn-Project/OpenBantu/tree/master/xho
Open Multilingual Wordnet	Disponibiliza documentos em 20 idiomas (albanês, árabe, dinamarquês, inglês, persa, finlandês, francês, hebraico, italiano, japonês, basco, catalão, galego, espanhol, indonésio, malaio, norueguês, português e tailandês), com o objetivo de facilitar o uso de *wordnets* em vários idiomas.	10.145.232	https://old.datahub.io/dataset/xwn
MultiWordNet WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical.	468.044	http://compling.hss.ntu.edu.sg/omw/
Croatian WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical da língua croata.	289.146	http://compling.hss.ntu.edu.sg/omw/
Chinese WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical da língua chinesa.	45.753	http://compling.hss.ntu.edu.sg/omw/
Albanet WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical do albanês.	61.243	http://compling.hss.ntu.edu.sg/omw/

Swedish WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical do sueco.	50.073	http://compling.hss.ntu.edu.sg/omw/
Romanian WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical do romeno.	516.915	http://compling.hss.ntu.edu.sg/omw/
WOLF WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical do francês.	606.250	http://compling.hss.ntu.edu.sg/omw/
DanNet WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical do dinamarquês.	40.183	http://compling.hss.ntu.edu.sg/omw/
OpenWN-PT WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical do português.	496.697	http://compling.hss.ntu.edu.sg/omw/
Japanese WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical do japonês.	1.007.643	http://compling.hss.ntu.edu.sg/omw/
FinnWordNet WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical do finlandês.	1.221.878	http://compling.hss.ntu.edu.sg/omw/
Persian WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical da língua persa.	182.224	http://compling.hss.ntu.edu.sg/omw/
Multilingual Central Repository (as part of Open Multilingual WordNet)	Repositório central multilíngue de bancos de dados lexicais.	198.465	http://compling.hss.ntu.edu.sg/omw/
Slovak WordNet (as	Banco de dados lexical do eslovaco.	280.726	http://compling.hss.ntu.edu.sg/omw/

part of Open Multilingual WordNet)			
Greek WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical da língua grega.	163.659	http://compling.hss.ntu.edu.sg/omw/
Thai WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical do tailandês.	703.957	http://compling.hss.ntu.edu.sg/omw/
sloWNet WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical do esloveno.	414.557	http://compling.hss.ntu.edu.sg/omw/
ItalWordnet WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical do italiano.	468.044	http://compling.hss.ntu.edu.sg/omw/
IceWordNet WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical do islandês.	106.288	http://compling.hss.ntu.edu.sg/omw/
Norwegian WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical da língua norueguesa.	31.467	http://compling.hss.ntu.edu.sg/omw/
plWordNet WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical do polonês.	384.425	http://compling.hss.ntu.edu.sg/omw/
Chinese WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical do chinês.	677.855	http://compling.hss.ntu.edu.sg/omw/
Arabic WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical do árabe.	202.006	http://compling.hss.ntu.edu.sg/omw/

Princeton WordNet (as part of Open Multilingual WordNet)	Banco de dados que apresenta compilações de diferentes léxicos.	1.403.855	http://compling.hss.ntu.edu.sg/omw/
Open WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical.	399.121	http://compling.hss.ntu.edu.sg/omw/
Wordnet WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical.	508.860	http://compling.hss.ntu.edu.sg/omw/
BulTreeBank WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical do búlgaro.	60.494	http://compling.hss.ntu.edu.sg/omw/
Hebrew WordNet (as part of Open Multilingual WordNet)	Banco de dados lexical do hebreu.	47.512	http://compling.hss.ntu.edu.sg/omw/
English WordNet	Banco de dados da língua inglesa.	0	https://en-word.net/
Wordnet-Wikipedia/DBpedia instance mapping	Apresenta um mapeamento das instâncias do *WordNet* para a *Wikipedia*.	7.702	https://github.com/jmccrae/wn-wiki-instances
Global WordNet Association Interlingual Index	Traz um índice global do *WordNet*.	352.995	https://github.com/globalwordnet/ili
PanLex	Banco de dados léxico que documenta traduções entre lexemas de vários idiomas.	786.205.297	http://panlex.org
Parole/Simple 'lexinfo' Ontology & lexicons	Informação morfológica, sintática e semântica. Na língua espanhola há 7572 entradas e no catalão 20545 entradas.	10.279	http://gilmere.upf.edu/corpus_data/ParoleSimpleOntology/ParoleOntology.html
PDEV-Lemon	Dicionário que mostra a relação dos verbos com substantivos e outras palavras.	233.372	https://old.datahub.io/pt_BR/dataset/pdev-lemon

SALDO-RDF	Trata-se de um *Swedish Associative Thesaurus*, apresentando um léxico em RDF.	0	https://old.datahub.io/pt_BR/dataset/saldo-rdf
SIMPLE	Dados do léxico italiano *SIMPLE*, em formato RDF.	372.294	https://old.datahub.io/pt_BR/dataset/simple
WordNet-RDF	Uma versão em RDF do *WordNet* de Princeton, que traz dados do léxico para o inglês.	8.903.374	https://old.datahub.io/pt_BR/dataset/wordnet-rdf
xLiD-Lexica	Contém dados extraídos da *Wikipedia* em julho de 2013 nas seguintes línguas: inglês, alemão, espanhol, catalão, esloveno e chinês.	300.000.000	https://old.datahub.io/pt_BR/dataset/xlid-lexica

Fonte: elaborado pela autora.

Pode-se dizer que nessa categoria encontram-se dados lexicais, dicionários, traduções e ontologias. Um *dataset* dessa categoria que se considerou importante destacar é o *Linked Old Germanic Dictionaries*, que apresenta a língua alemã em diferentes momentos históricos, podendo atender às necessidades de pesquisadores que trabalham, especificamente, na área de Linguística Histórica.

A categoria denominada Terminologias, Tesauros e Bases de Conhecimento abarca 15 *datasets*, descritos no quadro subsequente:

Quadro 3 – Mapeamento dos dados linguísticos: *Terminologies, Thesauri and Knowledge Bases*

Datasets	**Descrição**	**Nº de triplas**	***Websites***
associations	Coleção das associações e mapeamento para entidades *DBpedia*, compreende 780.000 associações humanas do *Edinburgh Associative Thesaurus*, e mapeamento de 790 associações às entidades *DBpedia*.	1.680.000	https://w3id.org/associations
Bibliography of Linguistic Literature (BLL) Thesaurus	Apresenta um vocabulário bilíngue.	0	http://data.linguistik.de/bll
EARTh	Tesauro de Referências de Aplicações Ambientais (*Environmental*	133.315	http://thesaurus.iia.cnr.it/index.php/vocabularies/earth

	Applications Reference Thesaurus).		
Gemeenschappelijke Thesaurus Audiovisuele Archieven? Common Thesaurus Audiovisual Archives	Utilizado para descrever programas de TV, contendo as seguintes categorias: assuntos; pessoas mencionadas; entidades nomeadas (nomes de empresas, bandas de música, etc); localizações; gêneros; fabricantes e apresentadores.	992.797	http://data.beeldengeluid.nl/gtaa/GTAA
GEneral Multilingual Environmental Thesaurus	Tesauro em mais de vinte idiomas com termos que se relacionam ao ambiente, além de dados ambientais. É publicado pela *European Environment Agency*.	20.229.105	http://www.eionet.europa.eu/gemet/
Geological Survey of Austria (GBA) – Thesaurus	Tesauro bilíngue, composto por quatro categorias: litologia, escala de tempo geológica, unidades geológicas, unidades tectônicas e classificação.	40.912	http://resource.geolba.ac.at/
Linked Clean Energy Data (reegle.info)	Exibe dados que se referem à energia limpa. As informações são de quatro tipos: perfis políticos e regulamentos dos países, perfis organizacionais, documentos de resultados do projeto, e um tesauro sobre energias renováveis, eficiência energética e alterações climáticas para reutilização pública.	334.049	http://data.reegle.info/
Open Data Thesaurus	Traz conceitos e entidades disponíveis em inglês e alemão.	8.000	http://vocabulary.semantic-web.at/PoolParty/wiki/OpenData
Social Semantic Web Thesaurus	Apresenta informações sobre pessoas, organizações, aplicações e tecnologias, etc.	20.000	http://vocabulary.semantic-web.at/PoolParty/wiki/semweb

STW Thesaurus for Economics	Fornece cerca de dezoito mil termos de entrada.	112.000	http://zbw.eu/stw
Pleiades	Dicionário geográfico para estudos do mundo antigo, mantido pelo *Institute for the Study of the Ancient World*, apresenta locais e nomes antigos.	2.600.000	http://pleiades.stoa.org/
TheSoz Thesaurus for the Social Sciences (GESIS)	Tesauro de Ciências Sociais com cerca de doze mil entradas.	438.901	http://lod.gesis.org/thesoz/
ThIST	Tesauro italiano de Ciências da terra, tem como objetivo apoiar a gestão de recursos ambientais. Seu conteúdo é disponibilizado pelo *Istituto Superiore per la Protezione e la Ricerca Ambientale*.	396.885	http://purl.org/NET/ThISTWebPage
UMTHES	Tesauro mantido pela *Federal Environment Agency*, da Alemanha, que traz a tradução para o inglês da maioria dos termos.	300.000	http://data.uba.de/umt/
WordNet 3.0 (VU Amsterdam)	*Dataset* para o qual foi feita a conversão do pacote Princeton.	4.573.749	http://semanticweb.cs.vu.nl/lod/wn30/

Fonte: elaborado pela autora.

Nota-se que nessa categoria há muitos dados relacionados ao meio ambiente, como energia limpa, ciências da terra, entre outros.

Na categoria Metadados de Recursos Linguísticos há 2 *datasets*, descritos no quadro abaixo:

Quadro 4 – Mapeamento dos dados linguísticos: *Linguistic Resource Metadata*

Datasets	**Descrição**	**Nº de triplas**	***Websites***
ISOcat-metadata	Foi tomado um subconjunto do conjunto de termos *ISOcat* (plataforma colaborativa para manter um conjunto de categorias de dados), e o reconstruiu por meio de uma estrutura de árvore. Isso permite aos usuários ter uma visão hierárquica das terminologias linguísticas.	830	https://old.datahub.io/dataset/isocat-metadata

Linguistic Metadata (LIME) vocabulary	Vocabulário de metadados linguísticos para descrição do ativo lexical de ontologias e *datasets* RDF.	158	https://old.datahub.io/dataset/lime

Fonte: elaborado pela autora.

Nessa categoria há um *dataset* muito interessante que permite a visualização hierárquica das terminologias linguísticas. Isso é significativo, pois aqueles que não conhecem a área, poderão ter uma visão geral do que ela engloba.

A quinta categoria denominada Categorias de Dados Linguísticos contém 14 *datasets* descritos no quadro que se segue:

Quadro 5 – Mapeamento dos dados linguísticos: *Linguistic Data Categories*

Datasets	**Descrição**	**Nº de triplas**	***Websites***
Automated Similarity Judgment Program lexical data	Há uma coleta de 40 palavras de 5500 idiomas, sendo elas representadas de um modo simplificado da fonética.	5.000.000	cldbstest.eva.mpg.de/asjp
DBpedia in Spanish	Dados de ontologia da versão 2014.	169.101.647	http://es.dbpedia.org
DBpedia Spotlight NIF NER Corpus	Apresenta sessenta frases em linguagem natural, retiradas de dez artigos diferentes do *New York Times*, trazendo a anotação de 249 entidades *DBpedia*.	3.425	https://old.datahub.io/dataset/dbpedia-spotlight-nif-ner-corpus
GeoWordNet	Construído a partir da integração completa do *WordNet* e *GeoNames*, além da parte italiana do *MultiWordNet*. É um *dataset* público que compreende 3.698.238 entidades, 3.698.237 de relações entre entidades, 334 conceitos, 182 relações entre conceitos, 3.698.238 relações entre instâncias e conceitos e 13.562 nomes de entidades em inglês e italiano.	53.390.969	http://geowordnet.semanticmatching.org/
Glottolog	Oferece informações sobre literatura descritiva para todas as línguas do mundo, além de uma classificação de idioma, e bases de conhecimento para nomes, códigos e locais.	10.000.000	http://glottolog.org

Intercontinental Dictionary Series	Traz 1.200 palavras em 200 línguas.	2.000.000	http://lingweb.eva.mpg.de/ids/
IWN	Corresponde ao *ItalWordNet*, que foi criado em Pisa. Composto por instâncias únicas.	515.816	https://old.datahub.io/dataset/iwn
lexinfo	Apresenta a ontologia de categorias lexicais.	4.374	http://lexinfo.net/
MASC-BN-NIF	Corpus em inglês de diferentes gêneros textuais, tanto escrito quanto falado, além de anotações semânticas.	5.620.229	https://old.datahub.io/dataset/masc-bn-nif
OLiA Discourse	Extensão da *OLiA* que se refere às características do discurso.	4.175	https://old.datahub.io/dataset/olia-discourse
SALDOM-RDF	Léxico sueco morfológico em formato RDF.	8.349.115	https://old.datahub.io/dataset/saldom-rdf
SweFN-RDF (Swedish FrameNet)	Léxico semântico em RDF.	339.385	https://old.datahub.io/dataset/swefn-rdf
Wikilinks RDF/NIF	Contém mais de 40 milhões de menções de mais de 3 milhões de entidades. As menções são mostradas em forma de links que remetem para as páginas da *Wikipedia*.	533.016.300	https://old.datahub.io/dataset/wikilinks-rdf-nif
World Loanword Database	Fornece vocabulários de 41 idiomas de todo o mundo, trazendo informações sobre o empréstimo de cada palavra, permitindo encontrar palavras emprestadas, palavras de origem e idiomas doadores em cada um dos 41 idiomas, além de facilitar a comparação das palavras emprestadas entre os idiomas.	1.000.000	http://wold.livingsources.org/

Fonte: elaborado pela autora.

Essa categoria, de um modo geral, constitui-se de dados de fonética, discurso, de fala e escritos, lexicais, e os empréstimos entre as línguas.

A penúltima categoria, Bases de dados tipológicas, traz apenas um *dataset* descrito no quadro subsequente:

Quadro 6 – Mapeamento dos dados linguísticos: *Typological Databases*

Datasets	**Descrição**	**Nº de triplas**	***Websites***
Phonetics Information Base and Lexicon (PHOIBLE)	Traz dados de um inventário fonológico, que mostra informações linguísticas e não linguísticas.	133.076	http://phoible.org

Fonte: elaborado pela autora.

Esse *dataset* fornece uma descrição detalhada dos fonemas (sons) da língua, considerando sua função. Os dados são relevantes, pois conhecendo os fonemas de uma dada língua é possível pronunciá-la melhor, além de estudar as dificuldades de um aluno de língua estrangeira, quanto aos fonemas que diferem de sua língua materna.

Na categoria Outros há 49 *datasets*, descritos no quadro abaixo:

Quadro 7 – Mapeamento dos dados linguísticos: *Other*

Datasets	**Descrição**	**Nº de triplas**	***Websites***
ALPINO RDF Treebank	Dados em RDF que foram exportados do *ALPINO Dutch Treebank*, que se compõe de frases holandesas.	1.799.901	https://old.datahub.io/pt_BR/dataset/alpino-rdf
BabelNet	Dicionário enciclopédico multilíngue, que apresenta uma ontologia responsável por conectar os conceitos em uma grande rede de relações semânticas, composta de 13.801.844 milhões de nós.	1.971.744.856	https://old.datahub.io/pt_BR/dataset/babelnet
Chat Game corpus	Traz dados que são resultado de um jogo de organização de objetos, por meio de uma configuração mediada por computador.	15.750	https://old.datahub.io/pt_BR/dataset/chat-game-corpus
CLLD[5]*-afbo*	Mostra os dados de uma pesquisa mundial de empréstimo de afixos.	12.453	http://afbo.info
CLLD-APICS	Atlas que mostra as estruturas linguísticas de *pidgin* e crioulo.	463.009	http://apics-online.info
CLLD-EWAVE	Atlas Mundial Eletrônico das Variedades de Inglês.	320.130	http://ewave-atlas.org
CLLD-GLOTTOLOG	Banco de dados bibliográfico de línguas menos conhecidas.	8.322.839	http://glottolog.org
CLLD-PHOIBLE	Apresenta os dados do *PHOIBLE* (repositório de fonologia interlinguística).	1.664.872	http://phoible.org

[5] CLLD significa *Cross-Linguistic Linked Data*.

CLLD-SAILS	Traz dados das línguas indígenas da América do Sul.	564.738	http://sails.clld.org
CLLD-WALS	Publica dados do *World Atlas of Language Structures*. Expõe dados fonológicos, gramaticais e lexicais, coletados de materiais descritivos.	1.478.985	http://wals.info
CLLD-WOLD	Fornece vocabulários de 41 idiomas com informações referentes ao empréstimo de palavras, sendo que *WOLD* é a sigla de *World Loanword Database*.	1.764.648	http://wold.clld.org
Cornetto1.2	Banco de dados do léxico holandês que possui semelhanças com o *WordNet*, mas apresenta mais relações semânticas.	792.747	http://www2.let.vu.nl/oz/cltl/cornetto/
Dbnary	*Dataset* extraído do *wiktionary* em várias línguas (búlgaro, holandês, inglês, finlandês, francês, alemão, grego, indonésio, italiano, japonês, latim, lituano, malgaxe, português, russo, servo-croata, sueco e turco).	191.801.778	http://kaiko.getalp.org/about-dbnary
Dbpedia	Extrai informações da *Wikipedia* e as estrutura para que sejam disponibilizadas na *Web*.	9.500.000.000	http://dbpedia.org/
DBpedia in Dutch	Extrai informações estruturadas da *Wikipedia* em holandês.	79.674.010	http://nl.dbpedia.org
De-gaap-ontology-lexicon	Ontologia do léxico financeiro alemão-inglês, que contém 728 frases anotadas bilíngues.	246.522	https://old.datahub.io/pt_BR/dataset/de-gaap-ontology-lexicon
FAO geopolitical ontology	Ontologia geopolítica que faz o gerenciamento de nomes em vários idiomas (inglês, francês, espanhol, árabe, chinês, russo e italiano); mapeamento de sistemas de codificação padrão (ONU, ISO, FAOSTAT, AGROVOC, DBPedia, etc); e fornece relações entre territórios (fronteiras terrestres, membros de grupos, etc.); além de rastrear mudanças históricas. Para a *Food and*	22.495	http://www.fao.org/countryprofiles/geoinfo.asp?lang=en

	Agriculture Organization of the United Nations (FAO) estes dados constituem-se como uma rede de conhecimento para erradicação da fome.		
FiESTA	Disponibiliza um formato genérico para anotações linguísticas e comportamentais.	245	https://old.datahub.io/pt_BR/dataset/fiesta
FrameBase schema	Base de conhecimento aberta e vinculada para representar de maneira uniforme uma ampla gama de conhecimento, abordando a heterogeneidade semântica entre várias fontes de conhecimento estruturado.	500.000	http://www.framebase.org/
Framester	Traz dados que representam o resultado do projeto *Framester* (centro de recursos abertos ligados à linguística).	32.105.685	https://w3id.org/framester
gemet-annotated	Conceitos traduzidos em trinta idiomas.	231.297	https://old.datahub.io/pt_BR/dataset/gemet-annotated
General Ontology of Linguistic Description	Ontologia para a linguística descritiva.	3.000	https://old.datahub.io/pt_BR/dataset/gold
Greek Wordnet	Traz uma base de dados lexical da língua grega.	356.595	https://old.datahub.io/pt_BR/dataset/greek-wordnet
Ietflang	Mapeamento das *tags* de idioma IETF (código de idioma abreviado definido pela *Internet Engineering Task Force* – IETF).	100.000	https://old.datahub.io/pt_BR/dataset/ietflang
ISOcat	Contém uma estrutura para definir categorias conforme as normas ISO / IEC 11179, e apresenta descrições linguísticas, como definições de categoria de dados, instruções de domínios de valor associados e exemplos.	25.000	http://www.isocat.org
JRC-Names-MLODE	Exibe grandes listas de nomes e suas variantes ortográficas.	1.458.828	https://ec.europa.eu/jrc/en/language-technologies/jrc-names
Language Name Authority List	Vocabulário controlado das línguas oficiais da União Europeia com seu respectivo código.	423.313	https://op.europa.eu/en/web/eu-vocabularies/at-dataset/-

			/resource/dataset/language
lemonUby	Os dados que permitem descrever de modo lexical as entidades ontológicas.	32.916.476	https://old.datahub.io/pt_BR/dataset/lemonuby
LemonWiktionary	Dados de um dicionário eletrônico.	5.000.000	https://old.datahub.io/pt_BR/dataset/lemonwiktionary
Lexvo	Traz informações sobre idiomas, palavras, caracteres e outras entidades relacionadas à linguagem do ser humano.	353.815	http://www.lexvo.org
lingvoj? Languages of the World (Multilingual RDF Descriptions)	Apresenta dados de diferentes idiomas como forma de preservação das línguas.	22.442	http://www.lingvoj.org/
Linked hypernyms	Anexa artigos de entidade na *Wikipedia* em inglês, alemão e holandês.	73.000.000	https://old.datahub.io/pt_BR/dataset/linked-hypernyms
LODAC BDLS	Dicionário de Ciências da Vida, que se constitui de dois tipos de dados: um traz a relação entre nomes científicos e nomes comuns de espécies; e o outro estabelece uma relação entre termos científicos japoneses e ingleses.	5.764.962	http://lod.ac/wiki/LODAC_BDLS
MExiCo	Modelo de dados para coleções de dados contendo anotações linguísticas e interações multimodais.	305	https://old.datahub.io/pt_BR/dataset/mexico
MLSA - A Multi-layered Reference Corpus for German Sentiment Analysis	Traz dados que correspondem às anotações de frases, deixando mais claro seu grau de polaridade negativo ou positivo.	21.000	http://iggsa.sentimental.li/index.php/downloads/
Multext-East	Dados multilíngues (búlgaro, croata, tcheco, inglês, estoniano, húngaro, lituano, macedônio, persa, polonês, romeno, russo, sérvio, eslovaco, esloveno, ucraniano) utilizados para pesquisa e desenvolvimento de engenharia de linguagem.	127.651	http://nl.ijs.si/ME/V4/

Muninn World War I	Projeto de pesquisa acadêmica multidisciplinar e multinacional que busca realizar investigações em milhões de registros referentes à Primeira Guerra Mundial em arquivos de todo o mundo.	38.927.740	rdf.muninn-project.org
OLAC Metadata	Metadados de recursos linguísticos.	5.143.342	http://www.language-archives.org/archive_records.php
Ontos News Portal	Extrai fatos (pessoas, organizações, relações entre eles) que são mesclados para uma enorme rede de informações, além de referenciar o artigo de onde veio o fato.	2.677.965	http://news.ontos.com
Polymath Virtual Library (Authority data) - Fundación Ignacio Larramendi	*Dataset* de autoridade das pessoas (espanhol, latino-americano, brasileiro e português) que possuem conhecimento em diversos assuntos.	0	http://www.larramendi.es/i18n/estaticos/contenido.cmd?pagina=estaticos/bibliotecaIL
Predicate Model For Ontologies (PreMOn)	Fornece uma ontologia OWL para modelar classes semânticas.	32.611.819	http://premon.fbk.eu
SentimentWortschatz	Dados em alemão para análise de sentimento, para isso são listadas palavras com polaridade positiva e negativa. Além de adjetivos e advérbios que expressam os pontos positivos e negativos de maneira explícita, também há substantivos e verbos que expressam a polaridade, porém de modo implícito.	30.339	http://asv.informatik.uni-leipzig.de/download/sentiws.html
SLI Galnet, the Galician wordnet, at version 3.0.26, in RDF format	Dados lexicais multilíngues (galego, catalão, basco, espanhol, português e inglês) conectados.	22.478.245	http://sli.uvigo.gal/download/SLI_Galnet/SLI_Galnet_RDF_3.0.26.tar.gz
TDS	Apresenta bancos de dados de linguística usados para pesquisa em tipos de língua e linguística.	3.135	https://old.datahub.io/pt_BR/dataset/tds

wiktionary.dbpedia.org	Dados de linguagem, parte do discurso, sentidos, definições, sinônimos, taxonomias e traduções.	0	http://wiktionary.dbpedia.org
WordNet (RKBExplorer)	Repositório semântico cujos dados têm origem no *WordNet*.	2.727.001	http://wordnet.rkbexplorer.com
WordNet 2.0 (W3C)	Conversão padrão de *Princeton WordNet* para RDF / OWL, descrevendo como foi a conversão além de fornecer exemplos de como ele pode ser consultado para uso em aplicativos da *Web* Semântica.	710.000	http://www.w3.org/TR/wordnet-rdf
zhishi.lemon	Léxico de dados conectados em chinês com traduções para o espanhol e o inglês.	7.036.338	https://old.datahub.io/pt_BR/dataset/zhishi-lemon
Zhishi.me	Dados da língua chinesa, que foram extraídos de três enciclopédias: *Chinese Wikipedia*, *Hudong Baik*e e *Baidu Baike*.	125.000.000	http://zhishi.me

Fonte: elaborado pela autora.

Observa-se que nessa última categoria há um pouco de cada tipo de dados, como frases, dicionário enciclopédico, dados fonológicos, lexicais e ortográficos, ontologias, línguas indígenas, etc. Considerou-se o *dataset CLLD-SAILS*, pertencente a esta categoria, muito importante, pois ao trazer dados das línguas indígenas da América do Sul faz com que se tenha um registro, permitindo que não se perca a língua, que é uma forma de manifestação cultural de um povo.

Tendo sido finalizado o mapeamento dos dados linguísticos é possível afirmar que os dados disponíveis, são extremamente ricos e úteis e podem ser aproveitados para diferentes finalidades.

Como forma de colaborar com *Iniciativa Linking Open Data* considerou-se que a divisão das categorias está coerente com os dados por elas abarcados, com exceção da última categoria denominada “Outros”. Afirma-se isso, pois nesta categoria há *datasets* que podem ser realocados para outras categorias. Na categoria em questão, aparecem dados de frases, atlas das línguas *pidgin* e crioulo, e dados de línguas indígenas. Esses tipos de dados poderiam ser acrescentados à categoria Corpus, pois assemelham-se aos que lá se encontram. Em Outros é possível encontrar dados de dicionários, enciclopédias, vocabulários, léxicos, variantes ortográficas, entre outros. Neste caso, eles poderiam ir para a categoria Léxicos e Dicionários, porque ao considerar suas características, nota-se que se encaixam neste grupo.

Também há dados que tratam de empréstimos de afixos, ontologias, fonológicos e da Primeira Guerra Mundial. De acordo com o que foi observado, eles poderiam integrar a quinta categoria, que traz a Categoria de dados linguísticos. Além disso, traz dados de um banco de dados próprio da Linguística. Aqui foi pensado em agregá-los à categoria Base de Dados Tipológicas.

Ademais, percebeu-se que como forma de direcionar o usuário, algumas categorias poderiam dividir-se em subcategorias. A primeira, Corpus, seria subdividida em atlas sintático, jornais, *Wikipedia*, ambiguidade, dados de fala, ontologias, comparação entre línguas; Léxicos e Dicionários teria a seguinte subdivisão dicionários, enciclopédias, léxico, vocabulário, traduções; Terminologia, tesauros e bases de conhecimento seria subdividida nas áreas que tem tesauros e/ou terminologias, como ciências biológicas, ciências sociais, ciências humanas; e a quinta categoria, que é Categoria de dados linguísticos, seria subdividida em fonética e fonologia, ontologias, discurso, léxico, morfologia e semântica.

4.2 Tecnologias empregadas nos dados linguísticos

Após o mapeamento dos dados, foram verificadas quais tecnologias são utilizadas pelos *datasets* linguísticos que foram descritos. Para este levantamento considerou-se as tecnologias que poderiam ser empregadas para realizar a descrição dos dados.

Nos quadros que se seguem (8, 9, 10, 11, 12, 13 e 14) os dados linguísticos foram considerados dentro de suas categorias. Foram verificados os usos das tecnologias URI, XML, RDF, RDF-S, OWL e SKOS. Quando os *datasets* apresentam determinada tecnologia encontra-se um "X", e se não faz uso dela o espaço não é preenchido, mostrando a ausência de determinada tecnologia.

Quadro 8 – Tecnologias empregadas nos *datasets* linguísticos: *Corpora*

Datasets	**URI**	**XML**	**RDF**	**RDF-S**	**OWL**	**SKOS**
Atlante Sintattico d'Italia (ASIt)	X	X	X	X	X	
Brown Corpus in RDF/NIF	X	X	X	X	X	X
DBpedia abstract corpus	X	X	X	X	X	
KORE 50 NIF NER Corpus	X	X	X	X	X	
Manually Annotated Sub-Corpus (MASC) of the Open American National Corpus	X	X	X	X		
News-100 NIF NER Corpus	X	X	X	X	X	
Ontologies of Linguistic Annotations (OLiA)	X	X	X	X	X	X
Reuters-128 NIF NER Corpus	X	X	X	X	X	
RSS-500 NIF NER CORPUS	X	X	X	X	X	

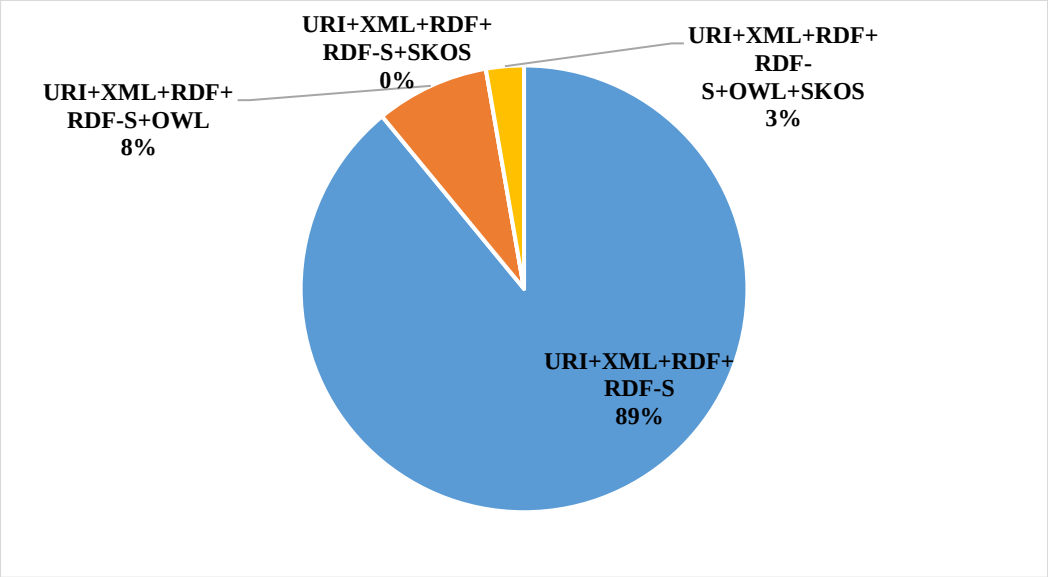
Universal Dependencies Treebank Ancient_Greek	X	X	X	X		
Universal Dependencies Treebank Ancient_Greek-PROIEL	X	X	X	X		
Universal Dependencies Treebank Arabic	X	X	X	X		
Universal Dependencies Treebank Basque	X	X	X	X		
Universal Dependencies Treebank Bulgarian	X	X	X	X		
Universal Dependencies Treebank Catalan	X	X	X	X		
Universal Dependencies Treebank Chinese	X	X	X	X		
Universal Dependencies Treebank Coptic	X	X	X	X		
Universal Dependencies Treebank Croatian	X	X	X	X		
Universal Dependencies Treebank Czech	X	X	X	X		
Universal Dependencies Treebank Czech-CAC	X	X	X	X		
Universal Dependencies Treebank Czech-CLTT	X	X	X	X		
Universal Dependencies Treebank Danish	X	X	X	X		
Universal Dependencies Treebank Dutch	X	X	X	X		
Universal Dependencies Treebank Dutch-LassySmall	X	X	X	X		
Universal Dependencies Treebank English	X	X	X	X		
Universal Dependencies Treebank English-ESL	X	X	X	X		
Universal Dependencies Treebank English-LinES	X	X	X	X		
Universal Dependencies Treebank Estonian	X	X	X	X		
Universal Dependencies Treebank Finnish	X	X	X	X		
Universal Dependencies Treebank Finnish-FTB	X	X	X	X		
Universal Dependencies Treebank French	X	X	X	X		
Universal Dependencies Treebank Galician	X	X	X	X		
Universal Dependencies Treebank Galician-TreeGal	X	X	X	X		
Universal Dependencies Treebank German	X	X	X	X		
Universal Dependencies Treebank Gothic	X	X	X	X		
Universal Dependencies Treebank Greek	X	X	X	X		
Universal Dependencies Treebank Hebrew	X	X	X	X		
Universal Dependencies Treebank Hindi	X	X	X	X		
Universal Dependencies Treebank Hungarian	X	X	X	X		
Universal Dependencies Treebank Indonesian	X	X	X	X		
Universal Dependencies Treebank Irish	X	X	X	X		
Universal Dependencies Treebank Italian	X	X	X	X		
Universal Dependencies Treebank Japanese	X	X	X	X		

Universal Dependencies Treebank Japanese-KTC	X	X	X	X		
Universal Dependencies Treebank Kazakh	X	X	X	X		
Universal Dependencies Treebank Latin	X	X	X	X		
Universal Dependencies Treebank Latin-ITTB	X	X	X	X		
Universal Dependencies Treebank Latin-PROIEL	X	X	X	X		
Universal Dependencies Treebank Latvian	X	X	X	X		
Universal Dependencies Treebank Norwegian	X	X	X	X		
Universal Dependencies Treebank Old_Church_Slavonic	X	X	X	X		
Universal Dependencies Treebank Persian	X	X	X	X		
Universal Dependencies Treebank Polish	X	X	X	X		
Universal Dependencies Treebank Portuguese	X	X	X	X		
Universal Dependencies Treebank Portuguese-Bosque	X	X	X	X		
Universal Dependencies Treebank Portuguese-BR	X	X	X	X		
Universal Dependencies Treebank Romanian	X	X	X	X		
Universal Dependencies Treebank Russian	X	X	X	X		
Universal Dependencies Treebank Russian-SynTagRus	X	X	X	X		
Universal Dependencies Treebank Sanskrit	X	X	X	X		
Universal Dependencies Treebank Slovak	X	X	X	X		
Universal Dependencies Treebank Slovenian	X	X	X	X		
Universal Dependencies Treebank Slovenian-SST	X	X	X	X		
Universal Dependencies Treebank Spanish	X	X	X	X		
Universal Dependencies Treebank Spanish-AnCora	X	X	X	X		
Universal Dependencies Treebank Swedish	X	X	X	X		
Universal Dependencies Treebank Swedish_Sign_Language	X	X	X	X		
Universal Dependencies Treebank Swedish-LinES	X	X	X	X		
Universal Dependencies Treebank Tamil	X	X	X	X		
Universal Dependencies Treebank Turkish	X	X	X	X		
Universal Dependencies Treebank Ukrainian	X	X	X	X		
Universal Dependencies Treebank Uyghur	X	X	X	X		
Universal Dependencies Treebank Vietnamese	X	X	X	X		

Fonte: elaborado pela autora.

Quase todos os *datasets* dessa categoria fazem uso, apenas, das quatro tecnologias básicas – URI, XML, RDF e RDF-S – (Gráfico 2). A explicação para isso pode ser o fato de que esta categoria abarca *datasets* que apresentam frases, sentenças, palavras, resumos, artigos ou notícias. Não sendo tão necessária a construção de ontologias, mas apenas o estabelecimento de conexões entre os componentes do corpus.

Gráfico 2 – Tecnologias empregadas em *Corpora*



Fonte: elaborado pela autora.

Quadro 9 – Tecnologias empregadas nos *datasets* linguísticos: *Lexicons and Dictionaries*

Datasets	**URI**	**XML**	**RDF**	**RDF-S**	**OWL**	**SKOS**
Albanet WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
Apertium RDF	X	X	X	X	X	
Apertium RDF CA-IT	X	X	X	X	X	
Apertium RDF EN-CA	X	X	X	X	X	
Apertium RDF EN-ES	X	X	X	X	X	
Apertium RDF EN-GL	X	X	X	X	X	
Apertium RDF EO-CA	X	X	X	X	X	
Apertium RDF EO-EM	X	X	X	X	X	
Apertium RDF EO-ES	X	X	X	X	X	
Apertium RDF EO-FR	X	X	X	X	X	
Apertium RDF ES-NA	X	X	X	X	X	
Apertium RDF ES-AST	X	X	X	X	X	
Apertium RDF ES-CA	X	X	X	X	X	
Apertium RDF ES-GL	X	X	X	X	X	
Apertium RDF ES-PT	X	X	X	X	X	
Apertium RDF ES-RO	X	X	X	X	X	

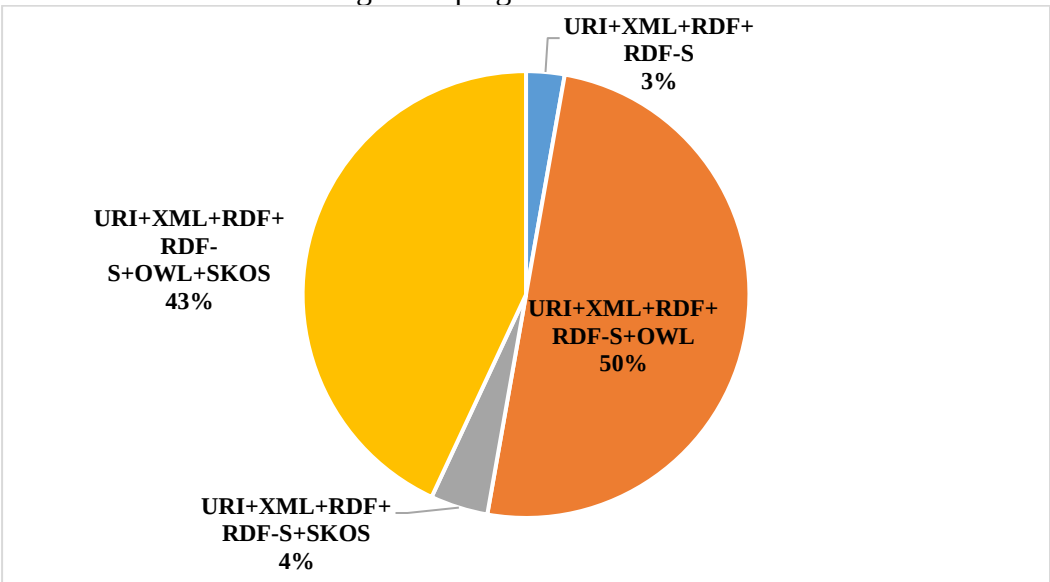
Apertium RDF EU-EM	X	X	X	X	X	
Apertium RDF EU-ES	X	X	X	X	X	
Apertium RDF FR-CA	X	X	X	X	X	
Apertium RDF FR-ES	X	X	X	X	X	
Apertium RDF OC-CA	X	X	X	X	X	
Apertium RDF OC-ES	X	X	X	X	X	
Apertium RDF PT-CA	X	X	X	X	X	
Apertium RDF PT-GL	X	X	X	X	X	
Arabic WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
Basque EuroWordNet-lemon lexicon (3.0)	X	X	X	X	X	
BulTreeBank WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
Catalan EuroWordNet-lemon lexicon (3.0)	X	X	X	X	X	X
Chinese WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
Chinese Taiwan WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
CopyrightTermBank	X	X	X	X		X
Croatian WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
DanNet WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
English WordNet	X	X	X	X	X	X
European Migration Network (EMN)	X	X	X	X		X
EuroSentiment	X	X	X	X		
FinnWordNet WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
Galician EuroWordNet-lemon lexicon (3.0)	X	X	X	X	X	X
Global WordNet Association Interlingual Index	X	X	X	X	X	X
Greek WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
Hebrew WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
IATE RDF	X	X	X	X	X	X
IceWordNet WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
ItalWordnet WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
Japanese WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
Lexicon of Syntactic and Semantic Framework	X	X	X	X	X	
Linked Old Germanic Dictionaries	X	X	X	X	X	
Multilingual Central Repository (as part of Open Multilingual WordNet)	X	X	X	X	X	X

MultiWordNet WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
Norwegian WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
Open Bantu isiXhosa Lexicon	X	X	X	X	X	
Open Multilingual Wordnet	X	X	X	X	X	X
Open WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
OpenWN-PT WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
PanLex	X	X	X	X	X	
Parole/Simple 'lexinfo' Ontology & lexicons	X	X	X	X	X	
PDEV-Lemon	X	X	X	X	X	
Persian WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
plWordNet WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
Princeton WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
Romanian WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
SALDO-RDF	X	X	X	X	X	
SIMPLE	X	X	X	X	X	
Slovak WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
sloWNet WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
Swedish WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
Thai WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
WOLF WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
Wordnet-Wikipedia/DBpedia instance mapping	X	X	X	X		X
Wordnet WordNet (as part of Open Multilingual WordNet)	X	X	X	X	X	X
WordNet-RDF	X	X	X	X	X	X
xLiD-Lexica	X	X	X	X		

Fonte: elaborado pela autora.

A maior parte dos *datasets* dessa categoria, utiliza, conjuntamente, URI, XML, RDF, RDF-S e OWL e/ou SKOS (Gráfico 3). Por estarem presentes neste grupo dicionários (bilíngues e etimológicos), léxicos, terminologias, traduções, entre outros, justifica-se a necessidade de tecnologias como OWL e SKOS para a representação de conceitos.

Gráfico 3 – Tecnologias empregadas em *Lexicons and Dictionaries*



Fonte: elaborado pela autora.

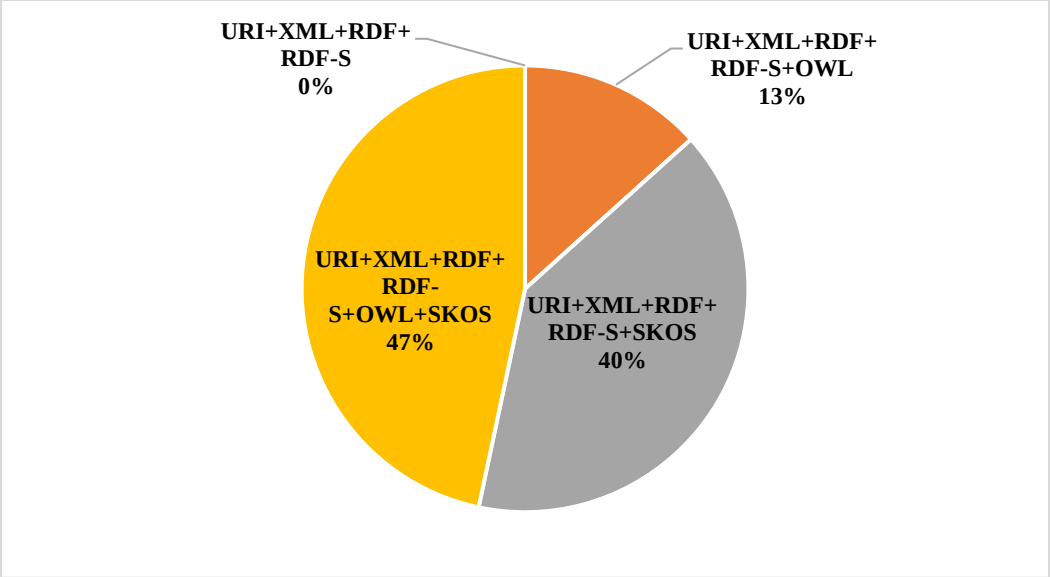
Quadro 10 – Tecnologias empregadas nos *datasets* linguísticos: *Terminologies, Thesauri and Knowledge Bases*

Datasets	**URI**	**XML**	**RDF**	**RDF-S**	**OWL**	**SKOS**
Associations	X	X	X	X	X	
Bibliography of Linguistic Literature (BLL) Thesaurus	X	X	X	X	X	X
EARTh	X	X	X	X		X
Gemeenschappelijke Thesaurus Audiovisuele Archieven ? Common Thesaurus Audiovisual Archives	X	X	X	X		X
GEneral Multilingual Environmental Thesaurus	X	X	X	X		X
Geological Survey of Austria (GBA) – Thesaurus	X	X	X	X		X
Linked Clean Energy Data (reegle.info)	X	X	X	X		X
Open Data Thesaurus	X	X	X	X	X	X
Pleiades	X	X	X	X	X	X
Social Semantic Web Thesaurus	X	X	X	X	X	X
STW Thesaurus for Economics	X	X	X	X	X	X
TheSoz Thesaurus for the Social Sciences (GESIS)	X	X	X	X	X	X
ThIST	X	X	X	X		X
UMTHES	X	X	X	X	X	X
WordNet 3.0 (VU Amsterdam)	X	X	X	X	X	

Fonte: elaborado pela autora.

Nenhum *dataset* dessa categoria apresenta apenas URI, XML, RDF e RDF-S; 47% utiliza essas quatro tecnologias junto com OWL e SKOS; 40% faz uso das quatro tecnologias mais o SKOS; 13% emprega OWL junto com as outras quatro tecnologias (Gráfico 4). O uso predominante do SKOS pode ser explicado pelo fato de que a maioria dos *datasets* refere-se a tesauros e o SKOS é empregado para representar tesauros.

Gráfico 4 – Tecnologias empregadas em *Terminologies, Thesauri and Knowledge Bases*



Fonte: elaborado pela autora.

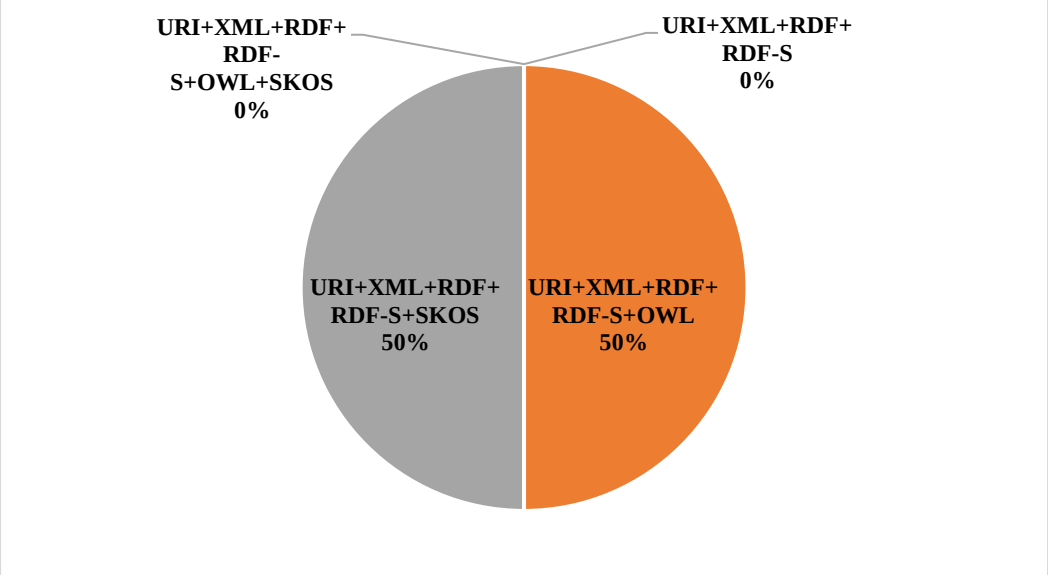
Quadro 11 – Tecnologias empregadas nos *datasets* linguísticos: *Linguistic Resource Metadata*

Datasets	**URI**	**XML**	**RDF**	**RDF-S**	**OWL**	**SKOS**
ISOcat-metadata	X	X	X	X	X	
Linguistic Metadata (LIME) vocabular	X	X	X	X		X

Fonte: elaborado pela autora.

Essa categoria é composta por dois *datasets* que apresentam metadados linguísticos. Um utiliza URI, XML, RDF, RDF-S e SKOS e o outro utiliza URI, XML, RDF, RDF-S e OWL (Gráfico 5).

Gráfico 5 – Tecnologias empregadas em *Linguistic Resource Metadata*



Fonte: elaborado pela autora.

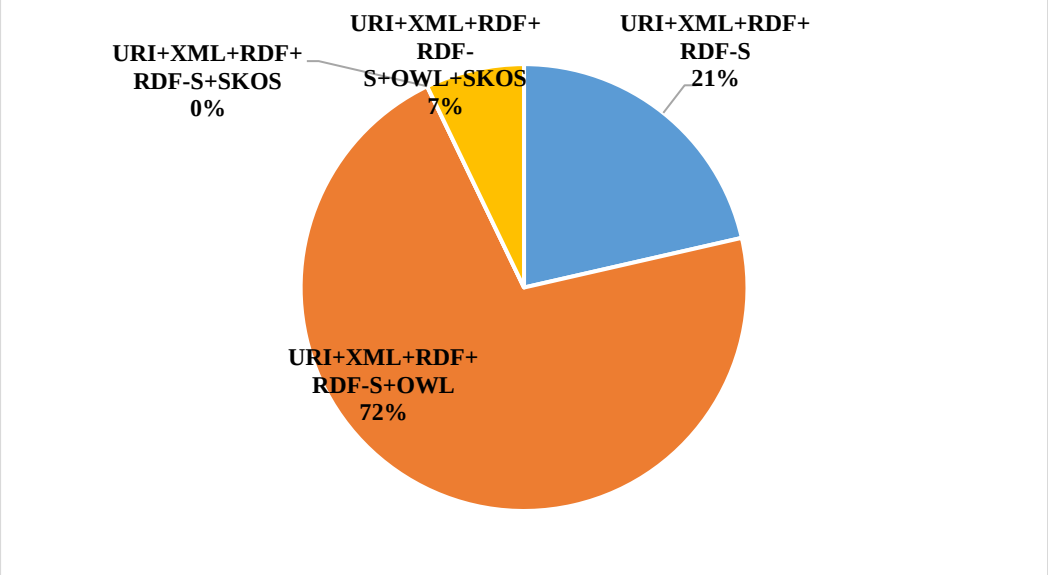
Quadro 12 – Tecnologias empregadas nos *datasets* linguísticos: *Linguistic Data Categories*

Datasets	**URI**	**XML**	**RDF**	**RDF-S**	**OWL**	**SKOS**
Automated Similarity Judgment Program lexical data	X	X	X	X	X	
DBpedia in Spanish	X	X	X	X	X	
DBpedia Spotlight NIF NER Corpus	X	X	X	X	X	
GeoWordNet	X	X	X	X		
Glottolog	X	X	X	X	X	X
Intercontinental Dictionary Series	X	X	X	X		
IWN	X	X	X	X	X	
Lexinfo	X	X	X	X	X	
MASC-BN-NIF	X	X	X	X	X	
OLiA Discourse	X	X	X	X	X	
SALDOM-RDF	X	X	X	X	X	
Swedish FrameNet RDF (SweFN-RDF)	X	X	X	X	X	
Wikilinks RDF/NIF	X	X	X	X	X	
World Loanword Database	X	X	X	X		

Fonte: elaborado pela autora.

A maioria dos datasets dessa categoria faz uso de URI, XML, RDF, RDF-S e OWL (Gráfico 6). O uso de OWL junto às demais tecnologias ocorre, porque neste grupo há ontologias, relações entre entidades e conceitos, etc.

Gráfico 6 – Tecnologias empregadas em *Linguistic Data Categories*



Fonte: elaborado pela autora.

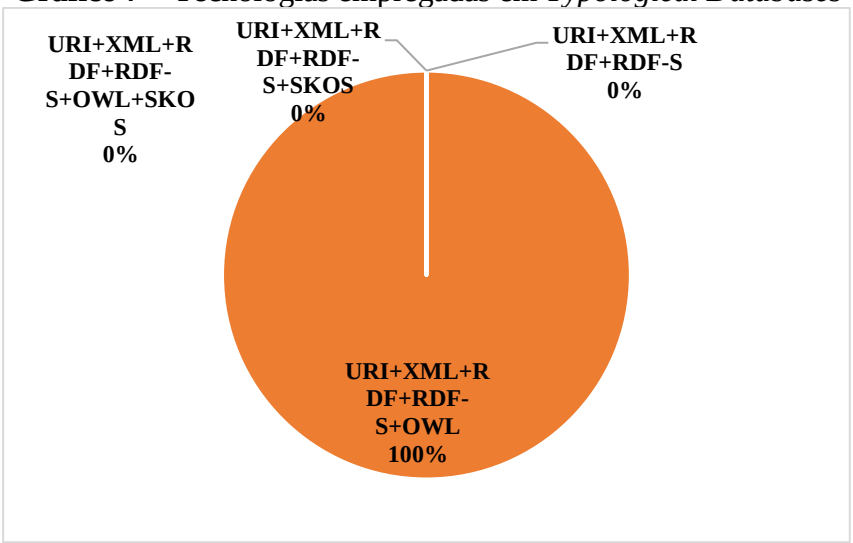
Quadro 13 – Tecnologias empregadas nos *datasets* linguísticos: *Typological Databases*

Datasets	**URI**	**XML**	**RDF**	**RDF-S**	**OWL**	**SKOS**
Phonetics Information Base and Lexicon (PHOIBLE)	X	X	X	X	X	

Fonte: elaborado pela autora.

Essa categoria é constituída, somente, por um *dataset*, que se trata de um repositório de dados de inventário fonológico em diferentes idiomas, além disso inclui informações genealógicas e geográficas. São empregadas cinco tecnologias: URI, XML, RDF, RDF-S e OWL que justifica-se por ser um inventário acrescido de outras informações (Gráfico 7).

Gráfico 7 – Tecnologias empregadas em *Typological Databases*



Fonte: elaborado pela autora.

Quadro 14 – Tecnologias empregadas nos *datasets* linguísticos: *Other*

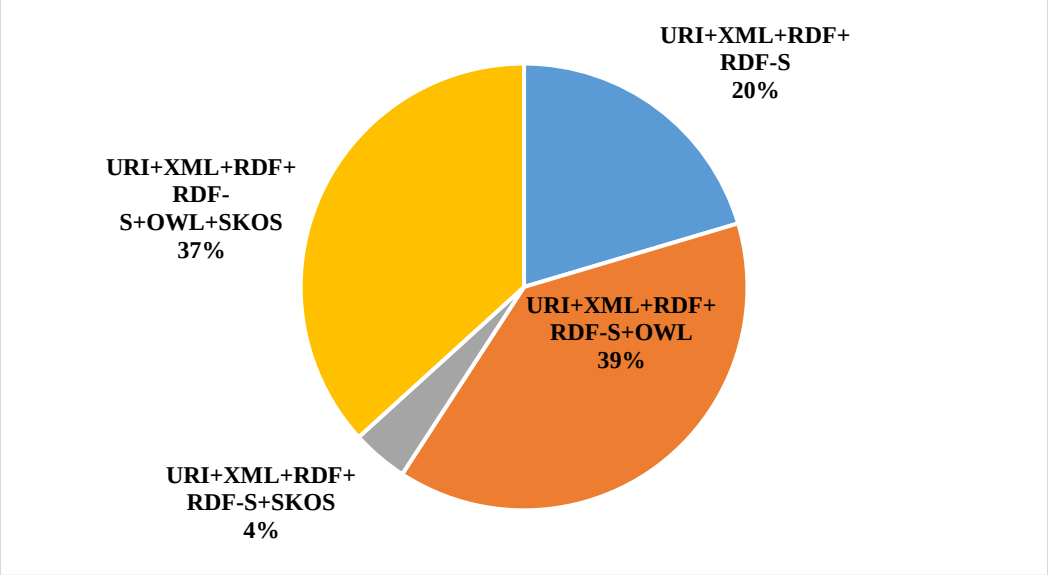
Datasets	**URI**	**XML**	**RDF**	**RDF-S**	**OWL**	**SKOS**
ALPINO RDF Treebank	X	X	X	X		
BabelNet	X	X	X	X	X	
Chat Game corpus	X	X	X	X		
CLLD-afbo	X	X	X	X	X	X
CLLD-APICS	X	X	X	X	X	X
CLLD-EWAVE	X	X	X	X	X	X
CLLD-GLOTTOLOG	X	X	X	X	X	X
CLLD-PHOIBLE	X	X	X	X	X	X
CLLD-SAILS	X	X	X	X	X	X
CLLD-WALS	X	X	X	X	X	X
CLLD-WOLD	X	X	X	X	X	X
Cornetto1.2	X	X	X	X	X	
Dbnary	X	X	X	X	X	
DBpedia	X	X	X	X	X	X
DBpedia in Dutch	X	X	X	X	X	
de-gaap-ontology-lexicon	X	X	X	X		
FAO geopolitical ontology	X	X	X	X	X	
FiESTA	X	X	X	X	X	X
FrameBase schema	X	X	X	X	X	
Framester	X	X	X	X	X	X
gemet-annotated	X	X	X	X		X
General Ontology of Linguistic Description	X	X	X	X	X	
Greek Wordnet	X	X	X	X	X	
Ietflang	X	X	X	X		
ISOcat	X	X	X	X	X	X
JRC-Names-MLODE	X	X	X	X	X	X
Language Name Authority List	X	X	X	X	X	X
lemonUby	X	X	X	X	X	
LemonWiktionary	X	X	X	X	X	
Lexvo	X	X	X	X	X	X
lingvoj ? Languages of the World (Multilingual RDF Descriptions)	X	X	X	X	X	X
linked hypernyms	X	X	X	X	X	
LODAC BDLS	X	X	X	X		X
MExiCo	X	X	X	X	X	
MLSA - A Multi-layered Reference Corpus for German Sentiment Analysis	X	X	X	X		
Multext-East	X	X	X	X		
Muninn World War I	X	X	X	X	X	
OLAC Metadata	X	X	X	X		
Ontos News Portal	X	X	X	X		

Polymath Virtual Library (Authority data) - Fundación Ignacio Larramendi	X	X	X	X	X	X
PreMOn	X	X	X	X	X	X
SentimentWortschatz	X	X	X	X		
SLI Galnet, the Galician wordnet, at version 3.0.26, in RDF format	X	X	X	X	X	
TDS	X	X	X	X	X	
wiktionary.dbpedia.org	X	X	X	X	X	
WordNet (RKBExplorer)	X	X	X	X		
WordNet 2.0 (W3C)	X	X	X	X	X	
zhishi.lemon	X	X	X	X	X	
Zhishi.me	X	X	X	X	X	

Fonte: elaborado pela autora.

O maior número dos *datasets* dessa categoria usa as tecnologias URI, XML, RDF e RDF-S junto com OWL (Gráfico 8). Isso acontece, pois há vários dados de dicionários e ontologias.

Gráfico 8 – Tecnologias empregadas em *Other*



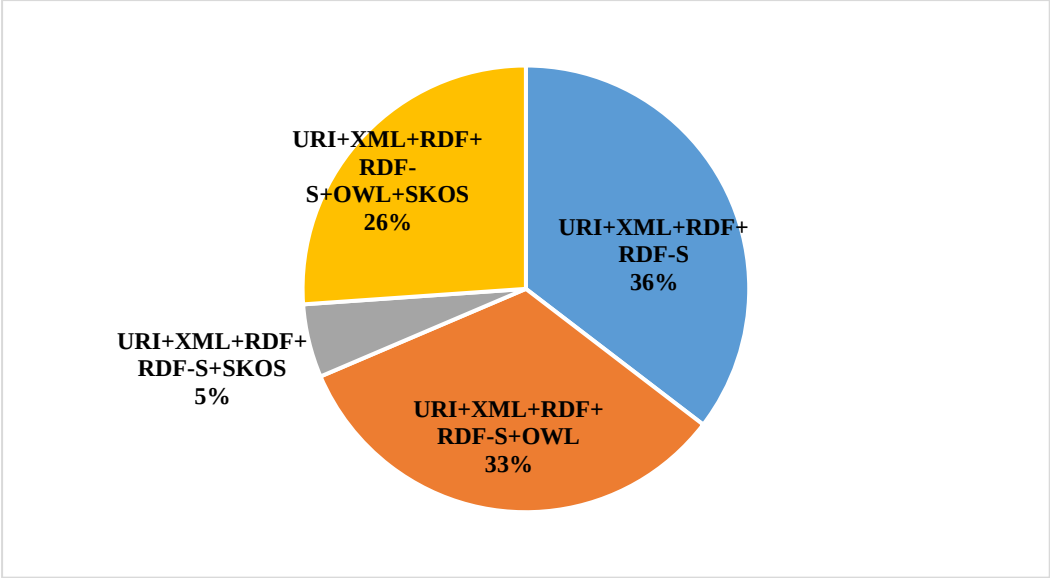
Fonte: elaborado pela autora.

De um modo geral, quando se observa os *datasets* linguísticos como um todo, é notável que todos os *datasets* fazem uso de quatro tecnologias: URI, XML, RDF e RDF-S. A utilização do URI possibilita uma recuperação unívoca dos *datasets*, promovendo a unicidade; o XML como uma linguagem de marcação, proporciona a criação de *tags* próprias para descrição dos conteúdos da *Web* e dá a base para o RDF; o RDF é de extrema

importância, pois viabiliza a constituição de uma rede informacional por meio dos dados; e o RDF-S traz um vocabulário básico além de propiciar a construção de novos vocabulários.

Essas quatro tecnologias aliadas à linguagem OWL são empregadas em 33% dos *datasets* (Gráfico 9). Seu uso pode ser explicado pelo fato de que essa linguagem possibilita a construção de ontologias, ou seja, por meio dela é possível dizer à máquina o que significa cada dado, concedendo, assim, o caráter semântico. Em 5% dos *datasets* (Gráfico 9) faz-se uso das quatro tecnologias citadas anteriormente juntamente com o SKOS, que assim como OWL é utilizado para construir ontologias, mas de um modo mais simples. Os *datasets* que utilizam todas as tecnologias (URI, XML, RDF, RDF-S, OWL e SKOS) ao mesmo tempo representam 26% dos *datasets* (Gráfico 9).

Gráfico 9 – Tecnologias empregadas nos *datasets* linguísticos



Fonte: elaborado pela autora.

Isso indica que os *datasets* linguísticos são um excelente exemplo de aplicabilidade das tecnologias semânticas. Visto que ao observar o emprego das tecnologias em cada categoria, é perceptível que na maioria dos *datasets* analisados, com exceção daqueles de *Corpora*, utiliza-se URI, XML, RDF, RDF-S, OWL e/ou SKOS.

5 CONSIDERAÇÕES FINAIS

Percebe-se que a *Web* constitui-se como um ambiente repleto de informações. Para que todo o conteúdo disponível seja utilizado da melhor maneira, precisa-se estruturar os dados aplicando tecnologias que são recomendadas pelo *World Wide Web Consortium* (W3C). Para isso, há o desenvolvimento do *Linked Data* que se constitui como princípios para ligação de dados.

Como forma de complementar a teoria de um ponto de vista aplicado encontra-se a iniciativa *Linking Open Data* (LOD). Tim Berners-Lee propôs um sistema de classificação baseado em estrelas para o LOD: uma estrela – dados disponíveis na *Web* com licença aberta; duas estrelas – dados disponíveis na *Web* com licença aberta e formato estruturado; três estrelas – dados disponíveis na *Web* com licença aberta e formato estruturado não proprietário; quatro estrelas – dados disponíveis na *Web* com licença aberta, formato estruturado não proprietário e utilização de URIs para identificar os recursos; cinco estrelas – dados disponíveis na *Web* com licença aberta, formato estruturado não proprietário, utilização de URIs para identificar os recursos e conectados com outros *datasets*.

Os metadados constituem-se como um importante componente nesse universo da *Web* Semântica, pois são responsáveis por propiciar a descrição e manutenção dos recursos digitais, além de garantir a troca de dados, permitindo, assim, seu reaproveitamento.

A iniciativa *Linking Open Data* (LOD) apresenta 1260 *datasets* de diversas áreas (domínio geral, geografia, governo, ciências da vida, linguística, mídia, publicações, redes sociais, uso geral) conectados entre si. Esses dados são de qualidade e podem ser usados como fonte principal de informação. Destaca-se o crescimento contínuo no número de *datasets*: em 2007 havia apenas 28 e em pouco mais de dez anos (2020) esse número aumentou quarenta e cinco vezes. Isso demonstra o sucesso de tal iniciativa.

Desses 1260 *datasets*, 226 são dados linguísticos. Eles estão distribuídos em sete categorias: Corpus; Léxicos e Dicionários; Terminologias, Tesauros e Bases de Conhecimento; Metadados de Recursos Linguísticos; Categorias de Dados Linguísticos; Bases de dados tipológicas; Outros. Com exceção dessa classificação, por meio da qual se poderia supor quais tipos de dados poderiam ser encontrados, não havia sido feito um levantamento que demonstrasse de modo mais específico os tipos de dados presentes nessa categoria.

Desse modo, esse trabalho teve como objetivo geral mapear os *datasets* linguísticos localizados na iniciativa *Linking Open Data*. Partindo desse objetivo foram traçados três

objetivos específicos. Atendendo ao primeiro deles, os *datasets* linguísticos foram descritos, demonstrando que tipos de dados são encontrados, quantos *datasets* há em cada uma das sete categorias e o número de triplas RDF. Isso permitiu verificar que os dados linguísticos disponibilizados são extremamente úteis. Por exemplo, na categoria Corpus encontram-se dados de atlas, de jornais, de resumos, de fala e escritos que podem constituir um objeto de pesquisa da Linguística, além de outros dados que possibilitam a análise comparativa entre línguas; em Léxicos e Dicionários há um *dataset* denominado *Linked Old Germanic Dictionaries* que apresenta a língua alemã em diferentes momentos históricos, podendo ser utilizado por pesquisadores que trabalham com Linguística Histórica; na categoria Terminologia, Tesauros e Bases de Conhecimentos há muitos dados relacionados ao meio ambiente, como energia limpa, ciências da terra, entre outros, que podem servir como vocabulário controlado para indexação de documentos relacionados a esses temas; em Metadados de Recursos Linguísticos há um *dataset* que permite a visualização hierárquica das terminologias linguísticas, fornecendo uma visão geral do que é englobado pela área; na Categoria de Dados Linguísticos há dados de fonética, discurso, de fala e escritos, lexicais, e os empréstimos entre as línguas, o que propicia uma série de estudos, como análise de discurso, uma análise de como se dá o empréstimo entre línguas, etc.; em Bases de Dados Tipológicas há um *dataset* que mostra dados de um inventário fonológico, que é uma descrição detalhada dos fonemas (sons) da língua considerando sua função, podem ser utilizados por distintas pessoas, pois ao conhecer os fonemas de uma dada língua é possível pronunciá-la melhor, além de possibilitar o entendimento das dificuldades de um aluno de língua estrangeira quanto aos fonemas que diferem de sua língua materna; na última categoria, Outros, destaca-se o *dataset CLLD-SAILS*, que ao trazer dados das línguas indígenas da América do Sul faz com que se tenha um registro, permitindo que não se perca a língua, preservando, assim, uma das formas de manifestação cultural de um povo.

Por meio dessas informações, cumpriu-se o segundo objetivo ao propor subcategorias, contribuindo assim para sua organização: Categoria Corpus – subcategorias: Atlas sintático, Jornais, Wikipedia, Ambiguidade, Dados de fala, Ontologias, Comparação entre línguas; Categoria Léxicos e Dicionários – subcategorias: Dicionários, Enciclopédias, Léxico, Vocabulário, Traduções; Categoria Terminologia, Tesauros e Bases de Conhecimento – subcategorias: Ciências Biológicas, Ciências Sociais, Ciências Humanas; Categoria de dados linguísticos – subcategorias: Fonética e Fonologia, Ontologias, Discurso, Léxico, Morfologia e Semântica.

Identificando as tecnologias utilizadas pelos *datasets* linguísticos descritos, terceiro objetivo, foi perceptível que elas são utilizadas de acordo com o que exige cada categoria. A maioria deles utiliza URI, XML, RDF, RDF-S e OWL e/ou SKOS. A exceção é a categoria Corpus, na qual grande parte dos *datasets* não utiliza OWL e/ou SKOS, pois não precisam construir ontologias considerando que há dados de frases, sentenças, palavras, resumos, artigos ou notícias. Também notou-se um maior uso do SKOS na categoria Terminologias, Tesauros e Bases de Conhecimento, justificado por apresentar tesauros e o SKOS é uma linguagem mais simples que OWL frequentemente utilizada para construção de tesauros na *Web*.

Conclui-se que o movimento *Linking Open Data* é de extrema relevância, para que seja possível desenvolver uma sociedade, visto que com os dados abertos todos terão acesso às informações que levam à construção do conhecimento. Além disso, ao estarem conectados, a estruturação do conhecimento dar-se-á de modo mais dinâmico, pois uma informação direcionará à outra. A semântica dos dados permitirá que o alcance da informação ocorra de modo mais rápido, uma vez que a máquina "entenderá" o que o usuário está buscando, e por isso lhe retornará informações pertinentes.

A iniciativa *Linking Open Data* cumpre de modo satisfatório o seu papel, que é mostrar como é viável a conexão de dados abertos, por meio das tecnologias prescritas, para que se estabeleça uma padronização, levando a interoperabilidade. Além de servir de referência para os que tem alguma dúvida acerca da aplicação das tecnologias propostas. Quanto aos dados linguísticos de tal iniciativa, ficou perceptível que são de extrema relevância e empregam as tecnologias adequadamente. Para tornar a categoria de dados linguísticos melhor do que se apresenta atualmente, aconselha-se criar as subcategorias sugeridas, pois possibilitarão uma navegação mais adequada para o usuário.

Quanto às categorias dos *datasets* linguísticos que foram descritos, nota-se que a categoria *Terminologies, Thesauri and Knowledge Bases* (Terminologias, Tesauros e Bases de Conhecimento) é importante para a Biblioteconomia e Ciência da Informação, visto que nesse grupo há instrumentos para representação de recursos informacionais de modo a possibilitar o controle do vocabulário e consequentemente uma recuperação eficiente. Considera-se que os *datasets* desse grupo deveriam ser alvo de um estudo mais aprofundado, visto que os tesauros são um instrumento tradicional e nessa iniciativa foram adaptados para a *Web*, utilizando as tecnologias semânticas.

REFERÊNCIAS

ABELE, A., MCCRAE, J. **The Linked Open Data Cloud**. Disponível em: http://lod-cloud.net/. Acesso em: 07 set. 2020.

ALVES, R. C. V. **Metadados como elementos do processo de catalogação**. 2010. 132 f. Tese (Doutorado em Ciência da Informação) – Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, 2010. Disponível em: https://repositorio.unesp.br/bitstream/handle/11449/103361/alves_rcv_dr_mar.pdf?sequence=1&isAllowed=y. Acesso em: 15 out. 2018.

ARAKAKI, F. A. ***Linked data***: ligação de dados bibliográficos. 2016. 144 f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista "Júlio de Mesquita Filho", Marília, 2016. Disponível em: https://www.marilia.unesp.br/Home/Pos-Graduacao/CienciadaInformacao/Dissertacoes/arakaki_fa_me.pdf. Acesso em: 18 mar. 2018.

ARAKAKI, F. A.; SIMIONATO, A. C.; SANTOS, P. L. V. A. da C. Catalogação e tecnologia: interseções com a web semântica. **Informação@Profissões**, Londrina, v. 6, n. 2, p. 03 – 19, jul./dez. 2017. Disponível em: http://www.uel.br/revistas/uel/index.php/infoprof/article/view/32003/23612. Acesso em: 01 ago. 2018.

ARAÚJO, L de R.; SOUZA, J. F. de. Aumentando a transparência do governo por meio da transformação de dados governamentais abertos em dados ligados. **Revista Eletrônica de Sistemas de Informação**, Curitiba, v. 10, n. 1, artigo 7, p. 1 - 15, jan. - jun. 2011. Disponível em: http://www.periodicosibepes.org.br/index.php/reinfo/article/view/880/pdf. Acesso em: 20 jul. 2018.

BEAL, A. **Gestão estratégica da informação**: como transformar a informação e a tecnologia da informação em fatores de crescimento e auto desempenho nas organizações. 4. ed. São Paulo: Atlas, 2004.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic *Web*: A new of *Web* content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American**, p. 35-43, 2001. Disponível em: http://www-sop.inria.fr/acacia/fabien/lecture/licence_travaux_etude2002/TheSemantic*Web*/. Acesso em: 13 out. 2017.

BIZER, C. *et al*. Linked Data on the Web (LDOW2008). In: International World Wide Web Conference, 17, 2008, Beijing. **Anais**… Beijing, 2008. Disponível em: http://www2008.org/papers/pdf/p1265-bizer.pdf. Acesso em: 13 ago. 2018.

BROOKS, T. A. Watch this: LOD - linking open data. **Information Research**, Borås, v. 13, n. 4, dez. 2008. Disponível em: http://InformationR.net/ir/13-4/TB0812.html>. Acesso em: 20 jul. 2018.

DI NOIA, T. *et al*. Linked open data to support content-based recommender systems. In: **I SEMANTICS 2012**, 8th Int. Conf. on Semantic Systems, Graz, sept. 5-7 2012. Disponível em:

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.912.7248&rep=rep1&type=pdf. Acesso em: 10 set. 2018.

DODEBEI, V. L. D. **Tesauro**: linguagem de representação da memória documentária. Rio de Janeiro: Interciência, 2002.

DUCHARME, B. **Learning SPARQL**: querying and updating with SPARQL 1.1. 2. nd. Sebastopol: O'Reilly, 2013.

FERMOSO-GARCÍA, A. M. *et al*. Sistema de modelado semántico para catalogación, clasificación, consulta y publicación en abierto de información bibliográfica. **El profesional de la información**, Barcelona, v. 27, n. 2, p. 410-418, 2018. Disponível em: http://www.elprofesionaldelainformacion.com/contenidos/2018/mar/20.pdf. Acesso em: 10 ago. 2018.

GIACOMELLI, K.; SOBRAL, A. O Curso de Linguística Geral e a constituição do campo científico de estudo da linguagem. **Revista Prolíngua**, João Pessoa, v. 11, n. 2, out./dez. 2016. Disponível em: https://periodicos.ufpb.br/index.php/prolingua/article/view/32224. Acesso em: 12 dez. 2020.

GIL, A. C. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 2008.

GONÇALVES, A. de O.; JACYNTHO, M. D. de A. Um método para publicação semântica Linked Data de bases de dados convencionais e um estudo de caso real de artigos acadêmicos. **Transinformação**, Campinas, v. 32, e180051, 2020. Disponível em: http://dx.doi.org/10.1590/1678-9865202032e180051. Acesso em: 12 out. 2020.

ISOTANI, S; BITTENCOURT, I. I. **Dados abertos conectados**. São Paulo: Novatec, 2015.

LAUFER, C. **Guia de Web Semântica**. São Paulo: Governo do Estado de São Paulo. 2015.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to information retrieval**. Cambridge: Cambridge University Press, 2008.

MOLLICA, M. C.; GONZALEZ, M. (Org.). **Linguística e Ciência da Informação**: diálogos possíveis. Curitiba: Appris, 2011.

NININ, D. M. **Linked open data em coleções de patrimônio cultural**: aspectos da representação da informação para humanidades digitais. 2018. 104 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal de São Carlos, 2018. Disponível em: https://repositorio.ufscar.br/bitstream/handle/ufscar/10538/NININ_Debora_2018.pdf?sequence=4&isAllowed=y. Acesso em: 15 out. 2018.

NOWACK, B. **The Semantic Web Technology Stack (not a piece of cake...)**. BNODE. 2009. Disponível em: http://bnode.de/media/2009/07/08/semantic_web_technology_stack.png. Acesso em: 7 abr. 2018.

OLIVEIRA, D. A. As questões éticas da democratização da informação. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**, João Pessoa, v. 8, n. 2, p. 1-11, 2013. Disponível em: https://revistas.ancib.org/index.php/tpbci/article/view/293/293. Acesso em: 12 dez. 2020.

OPEN DEFINITION, 2017. Disponível em: http://opendefinition.org/. Acesso em: 13 out. 2017.

PINHEIRO, J. M. dos S. *Web* Semântica: uma rede de conceitos. **Cadernos UniFOA**, Volta Redonda, ano IV, n.9, p.23-27, abr. 2009. Disponível em: http://*Web*.unifoa.edu.br/cadernos/edicao/09/23.pdf. Acesso em: 13 out. 2017.

RAMALHO, R. A. S.; VIDOTTI, S. A. B. G.; FUJITA, M. S. L. Web semântica: uma investigação sob o olhar da Ciência da Informação. **DataGramaZero**, [s. l.], v. 8, n. 6, dez. 2007.

ROCHA, R. P. de. Fabrico/Ciência: um ambiente Linked Data para o Mapeamento da Ciência. **Em Questão**, Porto Alegre, v. 18, Edição Especial, p. 281 - 297, dez. 2012. Disponível em: http://www.seer.ufrgs.br/EmQuestao/article/view/33279/0. Acesso em: 27 jan. 2018.

ROZSA, V.; DUTRA, M. L.; NHACUONGUE, J. A. Linked open data no contexto acadêmico: identificação e análise de vocabulários utilizados na academia e na pesquisa científica. **Brazilian Journal of Information Science**: Research Trends, Marília, v. 11, n. 3, p. 34 - 52, 2017. Disponível em: http://www2.marilia.unesp.br/revistas/index.php/bjis/article/view/6780/4651. Acesso em: 20 jul. 2018.

SANTARÉM SEGUNDO, J. E. *Web* semântica, dados ligados e dados abertos: uma visão dos desafios do brasil frente às iniciativas internacionais. *In*: XVI ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 2015, João Pessoa. **Anais** [...] João Pessoa: Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação, 2015. p. 1 – 20. Disponível em: http://www.brapci.inf.br/index.php/article/download/43838. Acesso em: 27 jan. 2018.

SANTARÉM SEGUNDO, J. E.; CONEGLIAN, C. S. Web semântica e ontologias: um estudo sobre construção de axiomas e uso de inferências. **Informação e Informação**, Londrina, v. 21, n. 2, p. 217 – 244, maio/ago. 2016. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/26417/20131. Acesso em: 27 jul. 2018.

SANTARÉM SEGUNDO, J. E.; SIMIONATO, A. C. Uma abordagem sobre a estrutura do geonames e suas contribuições para o *linking open data*. *In*: XVII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 2016, Salvador. **Anais** […] Salvador: Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação, 2016. p. 1 – 19. Disponível em: http://www.brapci.inf.br/index.php/article/view/0000021802/2c4d4ee17d51b7a3b5c0b1f875649f6f. Acesso em: 05 ago. 2018.

SEGARAN, T.; EVANS, C.; TAYLOR, J. **Programming the Semantic Web**: build flexible applications with graph data. Sebastopol: O'Reilly, 2009.

SERRA, L. G.; SANTARÉM SEGUNDO, J. E. O catálogo da biblioteca e o linked data. **Em Questão**, Porto Alegre, Online First, p. 1 -19, 2017. Disponível em: http://www.brapci.inf.br/index.php/article/download/50279. Acesso em: 20 jul. 2018.

SOUZA, R. R.; ALVARENGA, L. A *Web* Semântica e suas contribuições para a ciência da informação. **Ciência da Informação**, Brasília, v. 33, n. 1, p. 132-141, jan./abril 2004. Disponível em: http://www.scielo.br/pdf/ci/v33n1/v33n1a16.pdf. Acesso em: 27 jan. 2018.

TADINI, A. V. W.; CONEGLIAN, C. S.; SANTARÉM SEGUNDO, J. E. Caracterização do segmento de publicações no linking open data, um estudo exploratório. **Revista Conhecimento em Ação**, Rio de Janeiro, v. 2, n. 2, jul/dez. 2017. Disponível em: https://revistas.ufrj.br/index.php/rca/article/view/11699/9739. Acesso em: 01 ago. 2018.

WHAT IS OPEN? **OPEN KNOWLEDGE**, 2018. Disponível em: https://okfn.org/opendata/. Acesso em: 15 jan. 2018.

WOOD, D. *et al*. **Linked Data**: structured data on the Web. Shelter Island: Manning Publications, 2014.

WORLD WIDE WEB CONSORTIUM, 2015. Disponível em: http://www.w3c.br/Padroes/WebSemantica. Acesso em: 18 mar. 2018.