

A Wikidata e os desafios da interoperabilidade na era dos dados abertos ligados na Web: uma breve reflexão

Wikidata and the challenges of interoperability in the age of linked open data in the web: a brief reflection

Linair Maria Campos

Universidade Federal Fluminense (UFF) - Brasil
linair@hotmail.com

Maria Luiza de Almeida Campos

Universidade Federal Fluminense (UFF) - Brasil
maria.almeida@pq.cnpq.br

Nilson Theobald Barbosa

Universidade Federal Fluminense (UFF) - Brasil
nilson@tbarbosa.org

Resumo

Hoje estamos vivendo um momento histórico onde o acesso à informação está mais difundido e móvel com a possibilidade de acesso a um imenso e diverso volume de dados em formato aberto, e em muitas das vezes interligados uns aos outros. As pessoas não querem apenas achar e acessar documentos, mas sim formular perguntas e obter respostas embasadas pelos dados disponíveis. Muitos são os avanços nesse sentido, e nesse contexto se destacam em especial a DBpedia e a Wikidata como fontes de dados centrais para o embrião de uma web semântica sendo construída de forma coletiva e democrática. Por outro lado, muitos ainda são os desafios a serem enfrentados, em particular no que tange à compatibilidade de vocabulários no acesso a essa cada vez maior massa de dados abertos ligados na

Abstract

Nowadays we are living a historic moment where access to information is widespread and mobile with the possibility of access to a huge and diverse volume of data in open formats, and often interconnected with each other. People not only want to find and access documents, but ask questions and get answers based on available data. There are many advances in this regard, and in this context DBpedia and Wikidata stand out in particular as central data sources for the embryo of a semantic web being built collectively and democratically. On the other hand, there are still many challenges to be faced, particularly regarding the compatibility of vocabularies in accessing this growing mass of linked open data. The aim of this paper is to discuss the possibility that Wikidata could be It is used as an

Web. O objetivo desse trabalho é discutir a possibilidade de a Wikidata poder ser utilizada como uma linguagem intermediária para compatibilidade de vocabulários e dados em um domínio específico, explorando estudos teóricos na área da Ciência da Informação para apoiar esse tipo de abordagem. A metodologia usada parte de uma revisão de literatura sobre o tema e possui caráter exploratório. Como resultado, apresentamos uma breve reflexão sobre o que foi apresentado, esperando contribuir para a compreensão dos desafios de interoperabilidade nas iniciativas de dados ligados abertos que envolvem a Wikidata e de que forma os aportes teóricos da Ciência da Informação podem contribuir.

intermediate language for the compatibility of vocabularies and data in a specific domain, exploring theoretical studies in the area of Information Science to support this type of approach. The methodology used is part of a literature review on the subject and is exploratory. As a result, we present a brief reflection on what has been presented, hoping to contribute to the understanding of the interoperability challenges in open linked data initiatives involving Wikidata and how information science theories can contribute.

Palavras-chave: wikidata, interoperabilidade semântica, linguagem intermediária, dados abertos ligados.

Keywords: *wikidata, semantic interoperability, intermediate language, linked open data.*

1. Introdução

No final do século XIX, Paul Otlet e La Fontaine conceberam um espaço onde a produção científica mundial poderia ser encontrada, com base em um esforço notável de coleta e organização da informação científica. Esse feito visionário abriu caminho para iniciativas que viriam surgir na área de organização do conhecimento e da informação, com foco no documento. Otlet, ainda, imaginou um mundo futuro onde a informação, em diversos meios, estaria disponível ao usuário em sua mesa, em uma tela que disponibilizaria não só textos, mas imagens, sons e vídeos oriundos de meios como televisão, microfilme e cinema (Rayward, 1991). Algum tempo depois, em 1945, Vannevar Bush imaginou esses recursos informacionais conectados, sendo considerado o precursor do hipertexto (Robredo, 2011).

Hoje, algumas décadas à frente, estamos presenciando a concretização dessas visões de futuro, e chegamos à era onde o acesso à informação está mais difundido e móvel, mas, além disso, vivemos a possibilidade de acesso a um imenso e diverso volume de dados (estruturados, semiestruturados e não estruturados), em formato aberto e em muitas das vezes interligados uns aos outros. Temos acesso a outro mundo de informação potencial, onde se insere um usuário cujo foco não é mais achar e acessar documentos, mas sim obter e fazer sentido desse enorme volume de dados, formulando perguntas para as quais espera respostas embasadas pelos dados disponíveis. As fronteiras para a aquisição do conhecimento se alargam, e somos desafiados a prover os meios tecnológicos e pensar os requisitos que as tecnologias devem atender. Precisamos também unir *expertises* diferenciadas e esforços, de modo a fazer as perguntas pertinentes, dar sentido e interpretar as respostas obtidas com base nesses dados. Muitos são os avanços nesse sentido, e nesse contexto se destacam em especial a DBpedia e a Wikidata como fontes de dados centrais para o embrião de uma web semântica sendo construída de forma coletiva e democrática, e cujas facilidades de acesso já hoje podemos perceber

como uma vasta e variada base de dados mundial, cujo conteúdo pode ser pesquisado pelo usuário comum, ainda que com algum pequeno grau de dificuldade (Burgstaller-Muehlbacher, 2016).

Por outro lado, muitos ainda são os desafios a serem enfrentados, em particular no que tange à compatibilidade de vocabulários no acesso a essa cada vez maior massa de dados abertos ligados na Web. O objetivo desse trabalho é discutir a possibilidade de a Wikidata poder ser utilizada como uma linguagem intermediária para compatibilidade de vocabulários e dados em um domínio específico, explorando estudos teóricos na área da Ciência da Informação para apoiar esse tipo de abordagem. A metodologia usada parte de uma revisão de literatura sobre o tema e possui caráter exploratório. Como resultado, apresentamos uma breve reflexão sobre o que foi apresentado, esperando contribuir para a compreensão dos desafios de interoperabilidade nas iniciativas de dados ligados abertos que envolvem a Wikidata e de que forma os aportes teóricos da Ciência da Informação podem contribuir.

Neste sentido, este trabalho pretende contribuir com pesquisas que visem minimizar os problemas de opacidade semântica já apontada por Pierre Levy em seus trabalhos, quando afirma que estamos em vias de constituir uma memória participativa comum ao conjunto da humanidade. Para Levy (2014), a limitação que temos hoje, no início do século XXI, para exploração desta memória imensa de dados são os problemas de entendimento do seu significado terminológico, de incompatibilidade dos sistemas de classificação e a diversidade linguística e cultural. Desta forma, a falta de modelos que possam ser tratáveis computacionalmente impede a automação da maior parte das operações cognitivas de análise, seleção, síntese e de interligação de informações potenciais, e assim “não sabemos ainda como transformar sistematicamente esse oceano de dados em conhecimento e ainda menos como transformar o meio digital em observatório reflexivo de nossas inteligências coletivas” (Lévy, 2014, p.23).

2. Dados interligados abertos e a web semântica

Bizer *et al.* (2009, p.1) definem dados interligados abertos (LOD) como “um conjunto de boas práticas para estruturar e publicar dados estruturados na web”.

Dados abertos são instrumentos para o avanço do conhecimento, na medida em que se constituem em fontes de dados úteis de livre acesso que têm sido publicadas por diferentes organizações de interesse público (Liu *et al.*, 2011). Quando esses dados são interligados em um contexto, podemos perceber de forma mais precisa o seu significado e a partir daí, obter conhecimento. Além disso, esses dados têm tido um crescimento constante e gradual, e juntamente com o apoio de governos de várias partes do mundo, são um indicativo da grande importância dos dados ligados abertos no contexto da informação da atualidade. Essa nuvem de dados, que são interligados na forma de um grafo, contém dados de diversos tipos, tais como: geográficos, governamentais, ciências biológicas, publicações, músicas, dentre outros. No centro do grafo encontra-se a DBpedia, a contrapartida semântica da Wikipedia.

As iniciativas de LOD utilizam padrões da web e são baseadas nos seguintes princípios: (1) uso de URIs (*uniform resource identifier*) como nomes para entidades; (2) uso de URIs via HTTP (*hypertext transfer protocol*), de modo que se possam buscar informações por esses nomes na web; (3) informações úteis associadas às URIs, usando padrões tais como RDF (*Resource Description Framework*) e SPARQL

(*Protocol and RDF Query Language*) (W3C, 2008); (4) inclusão de associações com outras URIs, de modo que se possam descobrir mais entidades (Bizer *et al.*, 2009).

URIs via HTTP na prática se constituem em um mecanismo para atribuir a cada entidade (concreta, abstrata, ou ainda um conceito qualquer) na web um identificador único, através do qual o recurso pode ser referenciado, ligado a outros recursos, ou se pode recuperar uma descrição do recurso que a URI representa. RDF é um formato padrão para representação de dados na web. Esse formato permite que se representem fatos através de triplas na forma de sujeito, predicado e objeto, que, por sua vez, representam entidades concretas ou abstratas do mundo real. A linguagem SPARQL permite buscas nesse conteúdo, distribuído em diferentes locais na web, de forma transparente, como se fosse uma única fonte de dados. Além disso, SPARQL também inclui um protocolo para criação de serviços de fornecimento de dados na web (*SPARQL endpoints*), os quais são acessíveis de forma usual através da web, e que aceitam pesquisas, sendo os resultados fornecidos em formatos padronizados tais como XML e RDF, o que facilita a sua interligação com outros dados na web (D'Aquin, 2012).

Cabe destacar que para que essa rede de informações com significado se estabeleça é necessário que pessoas façam um esforço extra na codificação de informações em representações passíveis de processamento automático. Com esse esforço computadores terão condições de processarem, interpretar e concatenarem dados. Nesse cenário se situa a importância de se planejar o modelo dessas informações, ou seja, quais os conceitos, as suas naturezas, características, e de que forma se relacionam uns com os outros.

Outro aspecto fundamental quando se trata de LOD é a possibilidade de identificar de forma única um recurso, de modo a estimular a sua referência de forma não ambígua e, nesse sentido, a Wikidata desempenha um papel importante.

3. Wikidata

A Wikidata é uma iniciativa com a chancela da Wikimedia Foundation (WMF), constituindo-se em uma “uma base de dados aberta que pode ser lida e editada tanto por pessoas como por máquinas.” (Wikidata, 2019). Ela fornece um ponto de ligação para a Wikipedia, Wikimedia Commons, outras wikis da comunidade Wikimedia e outras iniciativas ao redor do mundo (Wikidata, 2019), sendo seu conteúdo multilíngue. Como base de dados, pode-se entender um grafo baseado em um conjunto de ontologias e instâncias de suas classes, que espelham fatos sobre as entidades da ontologia (Färber *et al.*, 2015).

O conteúdo da Wikidata pode ser usado livremente (segue o contrato Creative Commons CC0 1.0), ou seja, é de domínio público, com a ressalva da citação da fonte, e pode ser ligado a outros dados da nuvem LOD.

O software na qual está hospedada a Wikidata permite a anotação semântica de dados em páginas wiki e sua exportação para RDF, sendo que a inserção de dados (seja de forma manual ou automatizada) é feita com supervisão dos próprios membros da comunidade da Wikidata (Martinelli, 2016), de modo que existe um processo colaborativo de curadoria para os dados inseridos.

Tendo em vista a Wikipedia como fonte de dados importados, em contraste com a DBpedia, elo central da nuvem LOD, a Wikidata possui uma qualidade maior e quantidade menor desses dados importados,

uma vez que possui um processo de curadoria manual desses dados com informação de proveniência, o que toma tempo (Ismayilov, 2016). A curadoria envolve não só os dados em si, mas ainda sua estrutura (Saorín et al., 2018).

O fato de a Wikidata fornecer identificadores únicos (URIs estáveis) para seus elementos de dados (conceitos e propriedades) é um importante aspecto para que seus dados sejam reutilizados em iniciativas de dados abertos ligados, sendo que nesse sentido, ainda, a Wikidata permite acesso para buscas por meio de SPARQL e de API específica da Mediawiki (Burgstaller-Muehlbacher et al., 2016). Quanto à sua estrutura, a Wikidata contém itens, propriedades, valores e afirmativas (*statements*).

Um **item** possui um identificador único começando pela letra Q e seguido de um número sequencial, um nome (ou *label*) e uma descrição textual (opcional). Diz respeito a uma entidade no mundo (seja um conceito geral ou particular). Por exemplo, o item Q3434562 diz respeito à obra “Five Laws of Library Science”, que é um conceito particular. Exemplo de um conceito geral seria Q5 (ser humano).

Uma **propriedade** é usada para associar um item a outro item ou a um valor e possui um identificador único começando pela letra P seguido de um número sequencial e também um nome (ou rótulo) e uma descrição textual (opcional). Por exemplo, a propriedade P170 é denominada “criador” e serve para indicar “maker of this creative work or other object (where no more specific property exists)”.

Propriedades podem interligar itens dentro da própria Wikidata, mas podem também interligar um item com uma fonte externa, como, por exemplo, o registro de autoridades da Library of Congress, caso em que a propriedade é denominada um **identificador** (*identifier*) (Wikidata, 2019). Por exemplo, a propriedade P244 (Library of Congress authority ID) é um identificador usado para ligar S. R. Ranganathan (Q457933) a um link externo na Library of Congress (<http://id.loc.gov/authorities/names/n50053919.html>), que contém a entrada padronizada para o nome de Ranganathan (Ranganathan, S. R. (Shiyali Ramamrita), 1892-1972).

Já as **afirmativas** (*Statements*) descrevem fatos sobre um Item na forma de uma tripla sujeito, predicado e objeto, onde o sujeito é o item, o predicado é a propriedade, e o objeto é outro item. Por exemplo, o item “Five Laws of Library Science” (Q3434562) consta na Wikidata como uma instância de (P31) uma Teoria (Q17737), cujo criador (P170) é S. R. Ranganathan (Q457933). Ou seja, aqui temos duas afirmativas: Q3434562 P31 Q17737 e Q3434562 P170 Q457933. Cabe destacar que não há na codificação dos itens nada que os diferencie como conceitos gerais ou particulares. Isso poderá ser feito por meio das propriedades (por exemplo, instância de, P31).

Na Wikidata é possível também fazer afirmativas específicas para dizer que o valor de uma propriedade é inexistente (*no value*) ou então é desconhecido (*unknown value*). Por exemplo, é possível afirmar que S. R. Ranganathan (Q457933) não era um compositor de algo (P86), ou afirmar que não sabemos a altura (P2048) de S. R. Ranganathan (Q457933). No primeiro caso, a afirmativa seria (Q457933) (P86) *no value*, e no segundo (Q457933) (P86) *unknown value*.

Outra possibilidade interessante é o uso de **qualificadores** (*qualifiers*), e ainda **classificação** (*rank*) e **referências** para uma afirmativa.

Qualificadores “permitem que as declarações sejam expandidas, anotadas ou contextualizadas além do que pode ser expresso em apenas um par simples de propriedade-valor” (Wikidata, 2019). Por

exemplo, podemos acrescentar qualificadores para a afirmativa que Ranganathan trabalhou na Universidade de Madras, limitando-a no tempo, conforme ilustrado na Figura 1.

Figura 1 – Exemplo de uso de um qualificador na Wikidata

Ranganathan (Q457933) (item) funcionário da corporação (P2828) (propriedade)
Universidade de Madras (Q1364464) (valor da propriedade)
Data inicial (P580) (qualificador) → 9 Jan 1924 (valor do qualificador)

Fonte: própria

Qualificadores (*qualifiers*) são informados por meio de uma funcionalidade específica (*add qualifier*) na interface de entrada de dados relacionada a uma afirmativa (*statement*).

Classificações (*ranks*), por outro lado, são usadas para indicar a associação de um grupo de valores diferentes para uma mesma propriedade associada a um item. Por exemplo, poderíamos usar *ranks* para indicar as diversas ocupações (P106) de Ranganathan, a saber, de acordo com a Wikidata: matemático (Q170790), bibliotecário (Q182436) e acadêmico (Q3400985). Como diversos valores podem existir associados a uma mesma propriedade, é possível ainda indicar, por meio do tipo do *rank*, se um desses valores é mais representativo, preciso, ou atual do que outro (*preferred rank*), ou se todos são semelhantes quanto a isso (*normal rank*). Pode-se ainda usar um *rank* para indicar afirmativas errôneas ou ultrapassadas (*deprecated rank*), como, por exemplo, a de que a terra é plana (Wikidata, 2019).

Por fim, referências podem ser utilizadas para remeter a fontes específicas de dados que possam corroborar a veracidade dos dados referenciados em uma afirmativa. Isso se dá pelo fato de que a Wikidata é uma base de dados secundária, ou seja, ela apenas fornece a informação de acordo com uma fonte (Wikidata, 2019). Nesse sentido, as afirmativas da Wikidata não são necessariamente verdadeiras, mas são fatos afirmados por fontes diversas, e que podem ter pontos de vista diferentes (Tanon et al., 2016).

Um item está descrito em uma página na Wikidata (cujo código faz parte da URL), onde outros links e propriedades estão presentes, fornecendo dados adicionais. Por exemplo, para o item S.R. Ranganathan (Q457933) está descrito na página <https://www.wikidata.org/wiki/Q457933>. Nessa página, encontramos dados tais como: que ele *nasceu em* (P569) 9 de agosto 1892 e *morreu em* (P582) 1947 e sua *área de trabalho* (P101) era ‘Matemático’ (Q170790). De maneira análoga, propriedades também estão descritas em páginas da Wikidata, porém a lei de formação da URL é diferente (embora inclua o código da propriedade). Por exemplo, <https://www.wikidata.org/wiki/Property:P170>.

Os itens na Wikidata podem opcionalmente ter “*aliases*”, ou seja, “nomes alternativos para itens que são colocados na coluna ‘Também conhecido como’ da tabela que aparece na parte superior de cada página de item da Wikidata” (Wikidata, 2019). Por exemplo, para o item S. R. Ranganathan (Q457933) existem *aliases* tais como “Shiyali Ramamrita Ranganathan” e “Ranganathan”. É possível fazer uma busca na Wikidata tanto por um *label* quanto por um *alias*.

A interface amigável da Wikidata facilita ainda a produção de dados na web semântica pelo cidadão comum (incluindo instituições), que não dispõe dos conhecimentos tecnológicos necessários para publicar dados abertos ligados de forma independente (Burgstaller-Muehlbacher et al., 2016). Cabe

destacar, ainda, o aspecto global da Wikidata, onde, possivelmente incentivado pela diversidade de idiomas, esta tende a trazer visões que contrastam com a anglo-americana, o que é bom para promover a diversificação do conhecimento e promover o ponto de vista de outras culturas (Piscopo, 2017).

Entretanto, o potencial da Wikidata ainda está longe de ser esgotado. Interligar dados entre fontes distintas não é tarefa trivial. Se por um lado a nuvem LOD permite um número enorme de conexões, estendendo o conhecimento possível, por outro lado, a incompatibilidade terminológica pode trazer desafios que precisam ser enfrentados, sob a pena de criarmos ligações espúrias ou com uma semântica que não é a desejada. As ontologias são elementos centrais na interligação de dados abertos, e no contexto da web não podemos esperar uma ontologia única, mesmo dentro de um mesmo domínio de conhecimento. É preciso lidar com o fato que várias ontologias vão existir e deverão ser compatibilizadas de alguma forma.

Estudos na área de compatibilização de vocabulários controlados como os tesouros têm sido conduzidos há décadas na área da Ciência da Informação. Esses estudos podem ser aplicados na compatibilização de ontologias, uma vez que as ontologias, assim como os tesouros, também são organizadas com base em uma estrutura taxonômica.

Neste sentido, no âmbito dos estudos informacionais, já desde o início dos anos 2000 se veem pesquisas nacionais nesta direção (Campos, 2005, 2007, 2009, 2018; Campos, Gomes e Campos, 2011, 2009; Rocha, Campos e Costa, 2017).

4. Compatibilidade de vocabulários e o papel das linguagens intermediárias

A compatibilidade de vocabulários pode ser definida de duas formas: (i) de acordo com Lancaster e Smith (1983), de forma *quantitativa*, através de uma medida que afere o nível de comunicação ou troca de dados entre dois sistemas; (ii) de acordo com Dahlberg, de forma *qualitativa*: “a qualidade de um sistema ordenado que permite que seus elementos possam ser usados juntos ou intercambiados com elementos de outro sistema ordenado” (Dahlberg, 1983, p. 5). Sendo que, por sistema ordenado, a autora entende “qualquer instrumento usado na organização, descrição e recuperação do conhecimento, composto por expressões verbais ou notacionais para conceitos e suas relações, dispostos de forma ordenada” (Dahlberg, 1983, p. 5). Seja expressa de forma quantitativa ou qualitativa, a noção subjacente à noção de compatibilidade entre vocabulários é a possibilidade de recuperar informação que pode coexistir ou ser conectada de forma coerente entre sistemas que utilizam esses vocabulários, permitindo um intercâmbio de informações entre esses sistemas.

Partindo dessas considerações, propomos que, no contexto desse trabalho, compatibilidade seja definida como a qualidade de um vocabulário de se articular com outro de temática afim, direta ou indiretamente (o relacionamento indireto pode se dar através de um terceiro vocabulário, para o qual os diferentes vocabulários se relacionam), seja para definir equivalências conceituais entre seus termos, estabelecendo relações de semelhança, seja para complementá-lo em seu escopo, estabelecendo relações de natureza ôntica, ou seja, relações espaço-temporais entre objetos (Gomes, Campos e Guimarães, 2010).

Algumas das estratégias de compatibilização de vocabulários privilegiam estudos teóricos da compatibilização de linguagens (Neville, 1970; Neville, 1972; Dahlberg, 1983), enquanto que outras têm um foco mais prático (Chen et al., 1997). Estas últimas têm-se voltado para buscar a compatibilização através de algoritmos de software que exploram, dentre outros, o tratamento computacional de padrões, tais como, identificação de nomes semelhantes, ou de inferências que permitem deduzir que termos são compatíveis devido ao contexto em que são utilizados.

Uma forma de compatibilizar linguagens é por meio de mapeamentos, onde se busca obter uma correspondência entre dois vocabulários, estabelecendo-se critérios de conversão de um vocabulário para o outro (Lancaster, 1986). Mapeamentos podem ser convenientes para o caso de haver apenas dois vocabulários a compatibilizar, porém no caso de se desejar incluir outros, o processo torna-se custoso uma vez que implicaria em mapear cada vocabulário para todos os outros em ambas as direções, ou seja, seria um mapeamento bidirecional de n para n , onde n é o número total de vocabulários. Uma alternativa, que vem resolver o problema da multiplicação de mapeamentos, é o uso de um vocabulário ou léxico intermediário, também chamado de linguagem de comutação.

Na abordagem do léxico intermediário, se temos quatro vocabulários a compatibilizar, são necessárias apenas quatro correspondências e não doze como seriam no mapeamento entre linguagens de indexação. Podemos considerar então que o léxico intermediário é um *vocabulário central* que atua como um mediador de mapeamentos entre n vocabulários com os quais queremos estabelecer compatibilidade. O mapeamento é feito entre cada vocabulário e o léxico. Assim, se quisermos saber qual o termo equivalente do vocabulário v_2 no vocabulário v_1 , basta ver qual o termo equivalente a v_2 no léxico intermediário e, a partir deste, qual o termo associado no vocabulário v_1 .

Dentro da linha de trabalho com léxicos intermediários, destaca-se o trabalho de Neville (1972). Sua proposta tem como objetivo examinar a possibilidade de delinear um método de aplicação geral, para converter as palavras-chave (*keywords*) de um sistema nos termos de outro sistema, a partir do estudo dos tipos de incompatibilidade que podem ocorrer dentro de uma mesma área temática¹. Neville parte do princípio que as incompatibilidades são de número limitado e seu tratamento deve ser feito com base nos tipos de incompatibilidade e não se tratando cada palavra-chave individualmente.

O método de Neville baseia-se no princípio que se devem compatibilizar os conceitos (os conteúdos conceituais dos termos, que estão expressos pelas definições) e não os termos somente. Para o autor, são os conceitos que são indexados, os termos são simplesmente rótulos, muitas vezes arbitrários, para os conceitos.

Neville considera ainda que dentro de uma mesma área temática, de modo geral, os vocabulários deveriam abarcar os mesmos conceitos, embora possam existir termos diferentes para denominar o mesmo conceito entre esses diferentes vocabulários. Partindo desse princípio, sua estratégia baseia-se em identificar os conceitos semelhantes e codificá-los de forma única em cada vocabulário. Essa codificação então permitiria que as palavras-chave de um vocabulário pudessem ser mapeadas para outros vocabulários, da mesma temática, que partilhassem desse esquema de codificação.

Para isso, Neville propõe uma abordagem de linguagem intermediária, que implementa essa codificação numérica de conceitos, e através da qual se torna possível o estabelecimento da

¹ A noção de reconciliação é aqui tomada no sentido de compatibilizar as palavras-chave de um tesouro de origem para as de um tesouro de destino, através de heurísticas, levando em conta a definição dos termos.

equivalência conceitual de termos de diferentes linguagens, denominada pelo autor como *reconciliação*, ou seja, “a possibilidade de integração e aproximação de sistemas que contemplam o mesmo tipo de literatura mas que adotam diferentes tesouros” (Neville, 1972).

Na proposta do autor, o estabelecimento das correspondências entre conceitos não necessariamente implica em correspondência de um para um. Pode haver casos, por exemplo, em que um conceito mais específico em um dos vocabulários seja coberto por um conceito mais amplo no outro vocabulário, ou ainda pode haver casos em que não haja correspondência alguma no outro vocabulário para um determinado conceito do vocabulário de origem. Na verdade, Neville propõe uma série de onze casos onde a reconciliação pode ser efetuada, a saber:

- 1) **Quando existe uma correspondência exata entre as palavras-chave:** as palavras-chave são idênticas e usadas da mesma forma em cada um dos tesouros. Formas plurais de nomes ou nomes em outro idioma são consideradas idênticas, desde que representem exatamente o mesmo conceito do tesouro de origem.
- 2) **Diferentes sinônimos são usados como palavras-chave para o mesmo conceito entre tesouros diferentes:** as palavras-chave são diferentes, mas expressam conceitos sinônimos. Neste caso, basta fazer uma equivalência simples.
- 3) **O tesouro de origem tem palavras-chave para um conceito que não existe no outro tesouro destino:** deve-se criar um termo não preferencial correspondente no tesouro de destino.
- 4) **A palavra-chave do tesouro de origem existe no tesouro de destino sob uma palavra-chave mais genérica:** a palavra-chave do tesouro de origem é muito específica para as necessidades do tesouro destino, mas considera-se útil incluí-la sob um termo mais genérico.
- 5) **O tesouro origem usa uma só palavra-chave para designar um conceito, enquanto que para o mesmo conceito o tesouro destino precisa usar duas ou mais palavras-chave em conjunto.** Os nomes dos termos do tesouro de destino combinados podem formar o mesmo nome do tesouro de origem ou não. O importante é que se informe que duas palavras-chave do tesouro de destino combinadas equivalem ao mesmo conceito da palavra-chave do tesouro de origem.
- 6) **O tesouro origem faz distinção entre homônimos, mas o tesouro destino não faz:** nesse caso, pode-se também considerar a inclusão de palavras-chave no tesouro de destino, de forma a tornar mais direta a reconciliação.
- 7) **Um tesouro usa palavras-chave separadas para distinguir um termo usado em sentidos diferentes (diferentes papéis), enquanto que o tesouro de destino não faz:** os diferentes papéis devem ser preservados no tesouro reconciliado.
- 8) **O tesouro de origem usa como palavra-chave um termo que por si só não representa um conceito claramente identificado, como, por exemplo, adjetivos:** Neville sugere que esses termos sejam desconsiderados no processo de reconciliação e em seu lugar deve-se usar termos que não são palavras-chave mas que combinados servem para fazer correspondência com os termos do tesouro destino.
- 9) **O tesouro de origem contém palavras-chave sinônimas:** as palavras-chave podem ser reconciliadas através da escolha, no tesouro reconciliado, de um termo preferido dentre os vários sinônimos do tesouro de origem.

- 10) **O tesouro de origem utiliza como palavras-chave termos cujo nome tem significado apenas para o uso no local de origem:** esses termos devem ser reconciliados para um termo mais genérico no tesouro de destino.
- 11) **Um tesouro usa um sistema de codificação arbitrário para alguns conceitos, como, por exemplo, a concatenação de um termo com outro que é nomeado como um radical alfanumérico (ex.: BEAMS + W4):** devem-se identificar no outro tesouro os termos que correspondem às possíveis combinações de radicais e se estabelecem as equivalências necessárias.

A reconciliação envolve também fazer certas adições em cada um dos tesouros, na maior parte dos casos como referências cruzadas, porém nenhuma palavra-chave é alterada, removida ou adicionada nos tesouros envolvidos, **como diretriz do processo de reconciliação**². Da mesma forma, as relações entre as palavras-chave de um tesouro não são afetadas. Embora, pela proposta do autor, não haja a necessidade de se alterar as *palavras-chave* dos vocabulários sendo compatibilizados, pode haver a necessidade de inclusão de *termos não preferenciais*, o que se por um lado tem o mérito de não afetar o processo de indexação, o qual utiliza palavras-chave, por outro lado tem o ônus de poder afetar o processo de atualização dos vocabulários envolvidos, ou seja, cada vez que se incluir ou alterar uma palavra-chave em um dos vocabulários, deve-se verificar se termos não preferenciais precisam ser atualizados ou incluídos.

Quando o processo de reconciliação de Neville é posto em prática a indexação de resumos com as palavras-chave dos tesouros possui as seguintes características: Cada organização vai continuar a usar as palavras-chave de seus tesouros e não as palavras-chave de outros tesouros participantes, exatamente como antes da reconciliação; Todas as palavras-chave agora são acompanhadas pelos seus códigos que foram gerados pelo processo de reconciliação; Cada organização participante pode agora interpretar as palavras-chave de outros tesouros participantes através da conversão dos códigos de terceiros para os seus códigos, através da aplicação da sua chave específica.

A abordagem de Neville tem um caráter teórico que privilegia um estudo detalhado dos diferentes tipos de correspondências entre termos dos vocabulários a compatibilizar, e pode ser usada como um conjunto de diretrizes para se pensar o uso de relações de equivalência, e em que cenários as mesmas devem ser estabelecidas. Apresenta, ainda, alguns casos não cobertos pelas relações do SKOS (Simple Knowledge Organization System) (<https://www.w3.org/2004/02/skos/>) como, por exemplo quando o tesouro origem usa uma só palavra-chave para designar um conceito, enquanto que no tesouro destino é preciso usar duas ou mais palavras-chave juntas.

Podemos aqui observar um paralelo das recomendações de Neville com determinados aspectos existentes na Wikidata, tais como os identificadores únicos (URIs) e as propriedades usadas para estabelecer as relações entre conceitos existentes em vocabulários compatíveis. Os identificadores únicos fornecem codificação numérica de conceitos e as propriedades estabelecem os tipos de correspondência entre os conceitos, em especial as que são inspiradas no SKOS, a saber: *exact match*, *close match*, *broad match*, *narrow match* ou *related match*. O que não observamos na Wikidata, até onde pudemos perceber, e que está presente em Neville, é a sistematização abrangente de situações (e não de casos individuais) de compatibilidade, com suas respectivas recomendações.

² Embora nada impeça, se for considerado útil, incluir no tesouro de destino uma nova palavra-chave, que por acaso existe no tesouro de origem. Entretanto, essa inclusão não é devida ao processo de reconciliação em si.


A proposta de uso de relações SKOS na Wikidata, conduzida por Neubert (2017), gerou uma série de discussões (o que é o procedimento normal nesse âmbito) que culminou com algumas recomendações breves de adoção dessas propriedades, como o seu uso associado a um domínio específico, como, por exemplo, o Tesouro para Economia STW (*STW Thesaurus for Economics*³). Entretanto, essas recomendações de uso estão voltadas para a propriedade em si, e não, como no caso de Neville, para as diversas e abrangentes possibilidades de compatibilização de vocabulários.


No caso das relações SKOS de modo geral, uma forma proposta (e colocada em prática, no âmbito do Tesouro de Economia) é por meio de qualificadores para uma afirmativa. Uma afirmativa estabelece a ligação entre um item da Wikidata e uma fonte externa, e o qualificador estabelece o tipo da ligação.

Por exemplo, o item *Overseas countries and territories* (Q1451600) pode ser ligado ao descritor externo 29738-2 (<http://zbw.eu/stw/version/latest/descriptor/29738-2/about>) por meio da propriedade específica para ligar itens ao Tesouro da Economia a fontes externas *STW Thesaurus for Economics ID* (P3911).

Um qualificador então é usado para especificar o tipo de relação que descreve essa ligação. No caso, isso é feito ao se associar a propriedade *mapping relation type* (P4390) ao item que descreve o tipo de relação em si, no caso, *close match* (Q39893184). A Figura 2 ilustra esse caso na Wikidata como um exemplo de uso da propriedade P4390.

Figura 2 – Exemplo de uso da propriedade *mapping relation type* (P4390)



Wikidata property example		Overseas countries and territories
		STW Thesaurus for Economics ID
		29738-2
		mapping relation type
		close match

Fonte: <https://www.wikidata.org/wiki/Property:P4390>

Um caso especial de uso das relações SKOS é a *exact match*, a qual pode também ser utilizada de forma direta, por meio de propriedade específica (P2888) e não por meio de qualificadores. Um exemplo de uso desse tipo de relação é: *soil* (Q36133) *exact match* (P2888) http://aims.fao.org/aos/agrovoc/c_7156. Recomenda-se o uso de *exact match* como propriedade quando se tem certeza que existe uma relação de transitividade entre os itens ligados. Nesse caso, sua semântica seria semelhante à da relação *sameAs* da linguagem OWL.

Entretanto, cabe destacar que embora essas relações sejam úteis para compatibilizar vocabulários como tesouros, para fins de indexação, elas nem sempre são adequadas para expressar determinadas relações aplicáveis a ontologias, do ponto de vista do uso de inferências. Por exemplo, quando se estabelece que um determinado conceito é mais amplo (*broad match*) do que outro, isso se aplica tanto a instâncias quanto a conceitos gerais. Esse tipo de compatibilização não será detalhado aqui, pois foge do escopo do presente trabalho.

³ Detalhes dessa discussão específica podem ser consultados na Wikidata (https://www.wikidata.org/wiki/Wikidata:Property_proposal/mapping_relation_type).

5. Conclusões

O crescente volume de dados digitais produzido em todas as áreas de atuação humana, notadamente a produção de dados científicos, requer atuação urgente dos pesquisadores e profissionais, especialmente das áreas da ciência da informação e da ciência da computação, para que estes dados possam efetivamente serem utilizados de forma inteligente na geração de conhecimento.

Nesse sentido, as iniciativas de construção de políticas e práticas para utilização de dados ligados abertos estão na ordem do dia para a criação de instrumentos que permitam a interligação semântica de dados entre diferentes repositórios e mesmo diferentes áreas do conhecimento.

Portanto, vemos as iniciativas para fortalecimento de estruturas como a DBpedia e Wikidata como essenciais, tanto como ferramentas de uso imediato quanto como paradigmas para construção deste processo de transformação da Web em um grande repositório semântico de dados abertos interligados. A utilização da Wikidata como uma linguagem intermediária que permita a interligação entre vocabulários heterogêneos, apoiada por pressupostos teóricos de compatibilidade de linguagens, como propomos neste trabalho, pode ser um caminho em rumo a este objetivo.

6. Referências Bibliográficas

- BIZER, C., HEATH, T., BERNERS-LEE, T. (2009). The story so far. *International Journal on Semantic Web and Information Systems*, v. 5, n. 3, p. 1-22.
- BURGSTALLER-MUEHLBACHER, S., WAAGMEESTER, A., MITRAKA, E., TURNER, J., PUTMAN, T., LEONG, J., SU, A. (2016). Wikidata as a semantic framework for the Gene Wiki initiative. *Database* (Oxford).
- CAMPOS, M. L. A. (2007). Integração de Ontologias: o domínio da bioinformática. *RECIIS. Revista Eletrônica de Comunicação, Informação & Inovação em Saúde*, v. 1, p. 117-121.
- CAMPOS, M. L. A. GOMES, H. E., CAMPOS, L. M. (2011). *Integração e compatibilização em ontologias*. In: Fabiano Couto Corrêa da Silva; Rodrigo de Sales. (Org.). *Cenários da Organização do Conhecimento*. 1.ed. Brasília: Thesaurus, v. 1, p. 169-200.
- CAMPOS, M. L. A. (2018). Compartilhamento de dados em ambiente de pesquisa: a interoperabilidade semântica em ambientes heterogêneos. *III Seminário do Grupo de Pesquisa MHTX, 2018, Belo Horizonte, ECI/UFMG*, p. 41-45.
- CAMPOS, M. L. A., CAMPOS, M. L. M., CAMPOS, L. M. (2009). Integração de Ontologias em Domínio interdisciplinar: experiência no campo da Biomedicina. *IX Congreso ISKO-España: Nuevas perspectivas para la difusión y organización del conocimiento, 2009, Valença*. p. 180-192.
- CAMPOS, M. L. A. (2009). Aspectos semânticos da compatibilização terminológica entre ontologias no campo da Bioinformática. *Anais do X Encontro Nacional de Pesquisa em Ciência da Informação – ENANCIB*.
- CAMPOS, M. L. A. (2005). A problemática da compatibilização terminológica e a integração de ontologias: o papel das definições conceituais. *VI Encontro Nacional de Pesquisa em Ciência da Informação - ENANCIB, Florianópolis*.

- CHEN, H., NG, T. D., MARTINEZ, J.; SCHATZ, B.R. (1997). A Concept Space Approach to addressing the vocabulary problem in scientific Information Retrieval: An experiment on the Worm Community System. *Journal of the American Society for Information Science*, v. 48, n.1, p. 17-31.
- DAHLBERG, I. (1983). Conceptual compatibility of ordering systems. *International Classification*, v. 10, n. 2, p. 5-8.
- D'AQUIN, M. (2012). Putting Linked Data to Use in a Large Higher-Education Organisation. *Interacting with Linked Data Workshop, Heraklion. Proceedings...Heraklion: ILD*, p. 9-21.
- FÄRBER, M., ELL, B., MENNE, C., RETTINGER, A. (2015) A Comparative Survey of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web*, p.1-5.
- GOMES, H. E., CAMPOS, M. L. A., GUIMARÃES, L. S. (2010). Organização da Informação e Terminologia: a abordagem onomasiológica. *Datagramazero*, v. 11, p. 03.
- ISMAYILOV, A., KONTOKOSTAS, D., AUER, S., LEHMANN, J., HELLMANN, S. (2016). Wikidata through the eyes of DBpedia. *Semantic Web*, v. 0, n. 0, p. 1–11.
- LANCASTER, F. W., SMITH, L. C. (1983). *Compatibility issues affecting information systems and services*. Paris: UNESCO.
- LANCASTER, F.W. (1986). *Compatibility and Convertibility. Vocabulary Control for Information Retrieval*. Information Resources Press, Arlington, VA, USA.
- LÉVY, P. (2014). *A esfera semântica. Tomo I: Computação, cognição, economia da informação*. Editora Annablume, Brasil.
- LIU, Q., BAI, Q.; DING, L.; PHO, H.; CHEN, Y.; KLOPPERS, C.FOX, P. (2011). *Linking Australian Government Data for Sustainability Science: A Case Study*. In: WOOD, D. *Linking Government Data*, New York: Springer.
- MARTINELLI, L. (2016). Wikidata: la soluzione wikimediana ai linked open data. *Aib studi*, v. 56, n. 1, p. 75-85.
- NEUBERT, J. (2017). Wikidata as a linking hub for knowledge organization systems? Integrating an authority mapping into Wikidata and learning lessons for KOS mappings. *Proceedings of the 17th European Networked Knowledge Organization Systems Workshop*, Greece.
- NEVILLE, H. H. (1970). Feasibility study of a scheme for reconciling thesauri covering a common subject. *Journal of Documentation*, v. 4, n. 26, p.313-36.
- NEVILLE, H. H. (1972). Thesaurus reconciliation. *Aslib Proceedings*, v.11, n.24, p. 620-626.
- PISCOPO, A., VOUGIOUKLIS, P., KAFFEE, L., PHETEAN, C., HARE, J., SIMPERL, E. (2017). What do Wikidata and Wikipedia Have in Common: An Analysis of their Use of External References. In *Proceedings of the 13th International Symposium on Open Collaboration*, ACM, New York, NY, USA.
- RAYWARD, W.B. (1991). The case of Paul Otlet, pioneer of information science, internationalist, visionary: reflections on biography. *Journal of Librarianship and Information Science*, v.23, n.3, p. 135–145.

ROBREDO, J. (2011). Do documento impresso à informação nas nuvens: reflexões. *Liinc em Revista*, n. 61, p. 19-42, 2011.

ROCHA, L. L., CAMPOS, M. L. A., COSTA, L. C. (2017). Diretrizes para a aplicação de ontologias na interligação de dados governamentais abertos brasileiros. *Tendências da Pesquisa Brasileira em Ciência da Informação*, v. 10, p. 1-28.

SAORÍN, T., PASTOR-SÁNCHEZ, J. (2018). Wikidata y DBpedia: viaje al centro de la web de datos. *Anuario ThinkEPI*, v. 12, p. 207-214.

TANON, T. P., VRANDEČIĆ, D., SCHAFFERT, S., STEINER, T., PINTSCHER, L. (2016). From Freebase to Wikidata: The Great Migration, *Proceedings of the 25th International Conference on World Wide Web*, Montréal, Québec, Canada.

WIKIDATA. (2019). Página web. Recuperado de:
<<https://www.wikidata.org/wiki/Wikidata:Introduction>>.