CARACTERIZAÇÃO DO SEGMENTO DE PUBLICAÇÕES NO LINKING OPEN DATA, UM ESTUDO EXPLORATÓRIO[1]

Characterization of the publications segment in the *open data*: an exploratory study

Antonio Victor Wolf Tadini
Universidade de São Paulo (USP). Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto.
antoniovwt@gmail.com

Caio Saraiva Coneglian
Universidade Estadual Paulista (UNESP). Câmpus de Marília. Faculdade de Filosofia e Ciências.
caio.coneglian@gmail.com

José Eduardo Santarem Segundo
Universidade de São Paulo (USP). Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto.
santarem@usp.br

RESUMO: O Linking Open Data contempla uma gama de dados diversa e seus datasets são divididos em variadas categorias. Dentre essas categorias, a de publicações se destaca por conter informações de diversas áreas do conhecimento, com potencial para ser utilizada por pesquisadores para identificarem o impacto de publicações e como fonte para localizar e relacionar publicações científicas. No entanto, a definição desta categoria requer estudos devido ao caráter vasto que pode abordar, faltando uma compreensão mais aprofundada do significado de seus dados no âmbito do Linking Open Data. Desta forma, este trabalho tem como objetivo caracterizar a categoria denominada de publicações, comparando o entendimento sobre essa categoria àquilo que se pôde verificar pelo exame da natureza dos dados nela publicados, utilizando uma metodologia exploratória e aplicada. Por meio desse trabalho, foi possível dividir em tipos os datasets publicados, apontando deficiências na definição dos tipos apresentada pelos criadores do projeto Linking Open Data. Desta forma, identificou-se que a categoria de publicações possui uma amplitude muito grande, compreendendo inclusive sistemas de representação de conhecimento, que deveriam estar contidos em uma nova categoria, visto que se diferenciam na sua origem da categoria "publicações".

PALAVRAS-CHAVE: Linked Data. Web Semântica. Web de Dados. Representação do Conhecimento. Classificação.

1 Pesquisa oriunda do Programa Institucional de Bolsas de Iniciação Científica do CNPq.

ABSTRACT: The Linking Open Data comprehends a varied range of data and its datasets are divided into several categories. Among these categories, the one related to publications stands out for keeping information from several knowledge areas, with potential to be used by researchers as a tool to identify the impact of publications and as source to locate and relate scientific publications. However, this category definition requires studies due to the wide spectrum of types it can approach as it lacks a deeper understanding of the meaning of its data, in the range of the Linking Open Data. Therefore, this work aims to characterize the category called "publications", comparing its understanding to what we could verify from the exam of the nature of the data published in it, through the use of an exploratory and applied methodology. Through this work, we could divide the datasets published into types, showing deficiencies in the type definitions presented by the designers of the Linking Open Data Project. Actually, it was identified that the "publications" category comprehends a wide range including knowledge representation systems which should be contained in a new category, as they differ in their origin from the "publications" category.

KEYWORDS: Linked Data. Web Semântica. Web de Dados. Representação do Conhecimento. Classificação.

1 Introdução

Na virada do século, a publicação massiva de documentos na World Wide Web e a navegação hipertextual entre eles por meio de browsers já eram uma realidade sedimentada (BRIN; PAGE, 1998 apud BIZER; HEATH; BERNERS-LEE, 2009). Um dos motivos que possibilitou esse cenário está na possibilidade de interligar um recurso informacional a outro, visto que isso possibilitou que os dados estivessem interconectados, conduzindo a uma expansão nas relações e na quantidade de dados dentro da Web.

Com a popularização da Web e o consequente aumento no volume de dados ali contidos, a dificuldade de localizar os conteúdos e de ter mecanismos computacionais capazes de compreender o significado das relações existentes tornou a navegação e a recuperação de informações um processo bastante custoso.

Diante desse cenário, a Web Semântica, proposta em 2001, buscava ser uma forma de organizar e de tornar mais expressivos os conteúdos contidos na Web. Em síntese, um dos princípios da Web Semântica era possibilitar que os links fossem dotados de expressividade, isto é, a natureza da relação não seria necessariamente implícita, mas seria dotada de significado.

Essa característica foi fundamental para que a compreensão da natureza

dos links e, por consequência, a navegação no hipertexto pudessem ser operadas não apenas por humanos, mas também de modo automático (RAMALHO, 2006). Shadbolt, Hall e Berners-Lee (2006) entendem que o RDF proporciona a representação minimalista do conhecimento da Web. É assim que, da Web de Documentos, em que os recursos são interligados sem um significado explícito, nasce a Web de Dados, em que as ligações expressam semântica (BIZER; HEATH; BERNERS-LEE, 2006).

Neste contexto surge o Linked Data, em que diversos conjuntos de dados são interligados utilizando os princípios da Web Semântica e da Web de Dados. Os ideais do Linked Data deram origem ao Projeto Linking Open Data, que reúne centenas de datasets interligados, seguindo normativas pré-definidas.

O Linking Open Data divide os conjuntos de dados nas doravante denominadas categorias, como, por exemplo, as de: saúde, redes sociais, mídia e publicações, trazendo contribuições a distintas áreas do conhecimento. No entanto, uma categoria em específico, a de publicações, a princípio apresenta papel significativo para todas as categorias, porém falta uma compreensão mais aprofundada do significado dos dados desta categoria no âmbito do Linking Open Data.

Desta forma, este trabalho tem como objetivo caracterizar a categoria do Projeto Linking Open Data (LOD) denominada publicações, comparando o entendimento do projeto sobre essa categoria àquilo que se pôde verificar pelo exame da natureza dos dados nele publicados.

Para desenvolver esta pesquisa, utilizou-se uma metodologia quantitativa, de caráter exploratório, uma vez que buscaram-se na literatura subsídios teóricos para discutir e compreender a temática, e aplicado, pois analisou e explorou os datasets da categoria de publicações.

2 Web Semântica, Linked Data e Linking Open Data

Com a popularização e disseminação de dados dentro da Web, a dificuldade de encontrar os dados nesse ambiente aumentou significativamente, especialmente devido à pouca expressividade das ligações e dos documentos para os agentes computacionais. Neste cenário, a Web Semântica foi proposta como uma extensão da Web, visando dar significado às informações e contribuir, assim, para um melhor

uso da Web pelas máquinas, e consequentemente para as pessoas. (BERNERS-LEE; HENDLER; LASSILA, 2001).

Desde 2001, quando a Web Semântica foi proposta, vem ocorrendo um processo de maturação, que pode ser visto principalmente pelo desenvolvimento de uma série de tecnologias, que visam tornar concreta e implementável a Web Semântica.

Uma das tecnologias com mais influência na promoção da Web Semântica é o Resource Description Framework (RDF). O RDF é um modelo de dados que permite a representação do conhecimento em forma de rede, sendo uma tecnologia fundamental para a Web Semântica (FERREIRA; SANTOS, 2013). A figura 1 expressa a base do RDF, que relaciona um recurso, chamado de sujeito, por meio de uma propriedade, ou predicado, a um valor, também referido como objeto.

Figura 1: Tripla RDF, com as duas maneiras de denominar seus elementos.

Fonte: Elaborado pelos autores

Além do RDF, tecnologias como o eXtensible Markup Language (XML), utilizado como sintaxe para a representação dos dados, a Web Ontology Language (OWL), linguagem recomendada para a construção de ontologias, e o Uniform Resource Identifier (URI), utilizado para a descrição única de recursos na Web, foram essenciais para tornar a proposta da Web Semântica real.

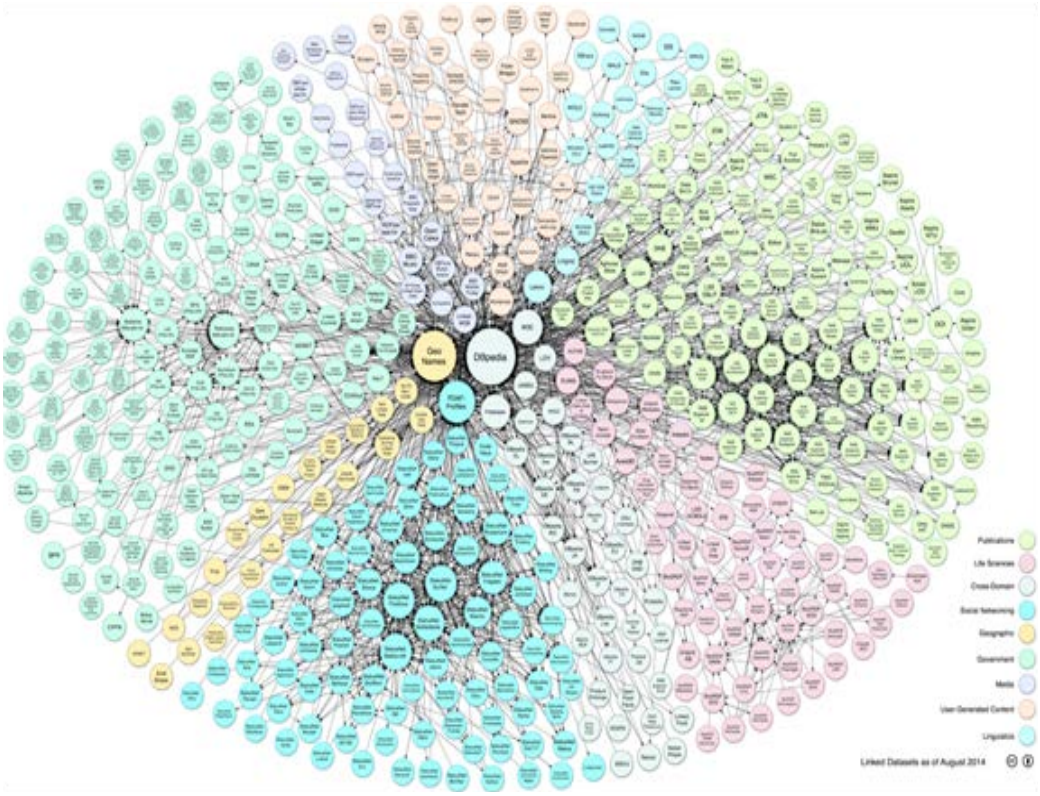
Com a estabilização dessas tecnologias, a Web Semântica iniciou um processo de materialização, centrada principalmente em criar aplicações de seus princípios na representação, na organização e na recuperação da informação.

O principal expoente dessa materialização é o Linked Data, que se refere a um "conjunto de melhores práticas" (HEATH; BIZER, 2011) que devem ser aplicadas na constituição da Web de Dados. Seu objetivo é criar, a partir da interligação de diferentes fontes, um único espaço de dados global. Vale destacar que o uso do RDF é uma das práticas que embasam o Linked Data.

O principal representante da aplicação do Linked Data é o Projeto Linking

Open Data (LOD). Ele foi originado pelo Open Data Movement, que objetiva viabilizar a publicação de grandes volumes de dados, os chamados datasets, com ligações entre si e de modo aberto. Os dados podem ser acessados a partir do Linking Open Data cloud diagram, que é renovado de tempos em tempos. A divisão desse diagrama, como demonstrado na figura 2, é representada pelas denominadas categorias, sendo cada uma delas vinculada a uma cor distinta. No caso, a categoria de "publicações" está vinculada à cor verde (à direita).

Figura 2: Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/.



Fonte: THE LINKING..., 2014.

Conforme apresentado na figura 2, existe um grande número de categorias representadas pelo Linking Open Data, demonstrando a importância e a presença desta iniciativa em distintas áreas do conhecimento. A publicação dos diagramas iniciou no ano de 2007, porém foi a partir de março de 2009 que a classificação em categorias foi feita pela primeira vez.

Para compreender a distribuição dos datasets entre as categorias do Linking Open Data, bem como averiguar a situação quanto à rastreabilidade desses datasets, apresenta-se a tabela 1, contemplando em números absolutos e percentuais como se distribuem entre as categorias os datasets do Linking Open Data, demonstrando os dados dos números totais de datasets, e dentre estes quantos são rastreáveis e

quantos não são.

Vale destacar que a análise publicada por Schmachtenberg, Bizer e Paulheim (2014b) considerou os datasets dentro do LOD como sendo rastreáveis ou não-rastreáveis. Os datasets rastreáveis são aqueles que permitem que robôs automáticos, chamados de crawlers, consigam explorar os datasets de forma automatizada.

Tabela 1: Distribuição de datasets do Linking Open Data

Tópico	Total		Rastreável		Não Rastreável	
	Datasets	%	Datasets	%	Datasets	%
Rede social	520	47,66	520	51,28	0	0
Governo	199	18,24	183	18,05	16	20,77
Publicações	138	12,65	96	9,47	42	54,54
Ciências da Vida	85	7,79	83	8,19	2	2,59
Conteúdo gerado pelo usuário	51-	4,67	48	4,73	3	3,89
Domínio-cruzado	47	4,31	41	4,04	6	7,79
Geográficos	27	2,47	21	2,07	6	7,79
Mídia	24	2,2	22	2,17	2	2,59
Total	1091	100	1014	100	77	100

Fonte: Adaptado de SCHMACHTENBERG; BIZER; PAULHEIM, 2014a; 2014b.

A partir dos dados apresentados na tabela 1, direcionando na perspectiva da categoria de publicações, objetivo deste trabalho, é possível verificar que essa categoria apresenta um conjunto considerável de datasets, acima dos dez por cento. No entanto, ao analisar a quantidade de datasets que não são rastreáveis, a categoria de publicações passa a contemplar o principal conjunto, correspondendo a número superior a 50%; mais de 30% dos datasets dessa categoria foram considerados não-rastreáveis.

Esta alta porcentagem de datasets não-rastreáveis da categoria de publicações aponta que uma parte considerável dos datasets de publicações não estão permitindo a rastreabilidade de seus dados, o que diminui um pouco as possibilidades em que os dados de publicações podem ser utilizados.

A partir dos apontamentos teóricos de Web Semântica, Linked Data e Linking Open Data, e as considerações sobre os datasets da categoria de publicações, é possível ter um panorama inicial de como essa temática está inserida dentro das propostas citadas. Visando aprofundar a compreensão desta categoria, na próxima seção apresentam-se mais características dela.

3 Resultados e Discussões: a categoria "publicações" no LOD

A categoria de publicações aparece como uma seara abordada pelo Linking Open Data que desperta bastante interesse pela comunidade acadêmica, por, em tese, reunir os conteúdos gerados a partir de pesquisa científica, bem como informações a esse respeito. Dentro de um contexto em que diversos pesquisadores estudam e disponibilizam os seus dados de pesquisa abertamente, o Linking Open Data pode se tornar um ambiente indispensável para a publicação e divulgação desses dados.

Desta forma, para compreender mais especificamente a categoria de publicações do Linking Open Data, recorreu-se primeiramente à DBpedia (plataforma que é produto do esforço em disponibilizar de modo estruturado, aberto e semântico os conteúdos da Wikipedia), que define "publicação", dando parâmetros de como se interpreta esse termo na comunidade do Linked Data.

> Publicar é tornar um conteúdo disponível para o público em geral. Embora usos específicos do termo possam variar entre os países, ele é usualmente aplicado a texto, imagens, ou outro conteúdo audiovisual em qualquer mídia tradicional, incluindo papel (jornais, revistas, catálogos, etc.). A palavra publicação significa o ato de publicar, e também se refere a quaisquer cópias publicadas (ABOUT…, 2017, tradução nossa).

Para que seja possível entender especificamente a concepção do Linking Open Data para a categoria "publicações", convém que se faça referência a Schmachtenberg, Bizer e Paulheim (2014a), que se propõem a delimitar e subdividir cada categoria, bem como citar alguns datasets proeminentes para cada uma delas. Para a categoria "publicações", objeto deste estudo, os autores cumprem essa tarefa por meio do excerto abaixo:

> A categoria publicações envolve datasets de bibliotecas, informação sobre publicações científicas e sobre conferências científicas, listas de leitura de universidades, e bases de dados de citação. Os mais conhecidos datasets incluem o dataset da Biblioteca Nacional da Alemanha, o dataset L3S DBLP, e o dataset Open Library (SCHMACHTENBERG; BIZER; PAULHEIM, 2014a, p.4, tradução nossa).

A partir da explanação feita pelos autores, chega-se ao questionamento se os quatro tipos de datasets apresentados são de fato taxativos, ou meramente exemplificativos. Ou seja, todos os datasets classificados na categoria analisada se enquadram necessariamente em algum desses quatro tipos, ou trata-se apenas de uma listagem de exemplos sem compromisso com a exaustividade? Além disso, questiona-se ainda se poderia um dataset contemplar mais de um dos tipos

apresentados. Não se pode esquecer que todas essas possibilidades são transpassadas pelo que foi alertado por Meusel et al. (2015) quanto aos equívocos e inconsistências, por eles verificados, na classificação das categorias, o que atribuíram ao fato de ela ter sido realizada manualmente, e não de forma automática.

Adicionalmente, ao analisar o tipo "informação sobre publicações científicas e sobre conferências científicas", pode-se indagar se seria esta uma espécie de classe destinada, inclusive, à miscelânea. Em caso afirmativo, questiona-se ainda: o tipo supracitado foi inserido a partir (no momento) da divisão das categorias, ou foi somente um elemento de expressão de linguagem, colocado a posteriori, referente a todo o restante da categoria de publicações que não se enquadra em algum dos outros três tipos?

Com base nas questões levantadas, buscou-se explorar os dados da categoria de publicações dentro do Linking Open Data, visando fornecer um panorama mais completo e embasado sobre a categoria em questão.

Desta forma, busca-se contrapor aos conceitos apresentados uma caracterização verificada, realizada em duas etapas: (I) análise dos dados estatísticos levantados pelo rastreamento do LOD descrito em Schmachtenberg, Bizer e Paulheim (2014b); e (II) investigação dos datasets de publicações a partir de resultados de estudos anteriores, além de novos dados obtidos pelos autores, que não foram publicados.

Primeiramente, analisaram-se os dados estatísticos do LOD, em que se têm algumas informações que caracterizam a categoria de publicações, presentes na Tabela 2.

Tabela 2: Estatísticas levantadas pelo rastreamento do LOD para a categoria "publicações"

		uso (%)
Predicados* mais usados	1. owl:sameAs	32,20%
	2. dct:language	25,42%
	3. rdfs:seeAlso	23,73%
Vocabulários** mais usados	1. dct	81,73%
	2. foaf	69,23%
	3. bibo	41,34%
Métodos alternativos de acesso	SPARQL	9,62%
	Dump	3,85%
Informações de licença	string "licen", dc:/dct:rights, etc.	3,85%
* ver Fig.1. ** exceto rdf, rdfs e owl		

Fonte: Adaptado de SCHMACHTENBERG; BIZER; PAULHEIM, 2014b.

É possível observar que, quanto aos dados referentes ao diagrama (rastreado) de 2014, os links dos datasets da categoria "publicações" expressam, mais frequentemente: a ideia de sinônimo, a especificação da língua do recurso, e a ideia de remissiva (ou relação sugerida). Também nota-se que o Dublin Core é largamente utilizado, assim como nas outras categorias, o que confirma a aderência do dcterms ao modelo RDF; em seguida, têm destaque um vocabulário para conexão de pessoas (foaf) e outro para descrição de citações e referências bibliográficas (bibo).

Na adoção do SPARQL, a categoria de publicações tem posição mediana entre as outras categorias, ao passo que destacam-se governo e ciências da vida. Quanto a informações de licença, publicações apresenta um uso muito baixo - em relação a governo (29,57%), - mas equiparado a ciências da vida e mídia, por exemplo. Maiores comparações são possíveis por meio da fonte da Tabela 2.

Esses resultados são expressivos, mas em conjunto não possibilitam vislumbrar implicações mais importantes do que as que já foram descritas. Esta pesquisa os apresenta mais como informações adicionais do que como instrumento de análise.

Assim, realizou-se uma investigação mais aprofundada nos datasets de publicação. Nesta fase, cada dataset da categoria "publicações" foi classificado, para fins estatísticos, em um tipo e, eventualmente, em uma subdivisão. Neste processo, o rol de quatro tipos citado anteriormente foi pouco expressivo e pouco significativo para a classificação dos datasets.

Desta forma, para solucionar essa problemática, realizou-se uma reformulação dos tipos contemplados pela categoria "publicações", bem como algumas subdivisões existentes, com base nas características dos datasets existentes e nas estatísticas apresentadas. O quadro 1 apresenta esses tipos, com comentários sobre as características deles.

Quadro 1: Tipos e subtipos da categoria "publicações"

Tipo	Subtipo	Exemplos	Comentário
Informação sobre publicações científicas e/ou sobre conferências científicas	Informação sobre publicações científicas	BibBase RKB Explorer OAI Theses.fr TWC IEEE vis	Metadados referentes a recursos informacionais que sejam publicações científicas, conforme a definição da DBpedia para publicação somada à presença do caráter científico dos documentos representados.
	Informação sobre conferências científicas	Colinda Multimedia Lab University Ghent	Dados referentes a eventos (ex: data e localização), de caráter científico, bem como os eventuais metadados referentes aos anais, ou seja, também existe aqui informação sobre publicações científicas, eventualmente.

Sistema de representação do conhecimento	Tesauro	Agrovoc Skos STW Thesaurus for Economics	Linguagem documentária pós-controlada e, caracteristicamente, pós-coordenada. Compõe-se de hierarquia e associações. Serve-se à indexação e à recuperação.
	Lista de cabeçalhos de assuntos	LCSH Lista de Encabezamientos de Materia	Linguagem documentária pós-controlada e pré-coordenada. Serve-se à indexação e à recuperação.
	Sistema de classificação	JITA MSC	Linguagem documentária pré-controlada. Classificação em sentido estrito.
	Outros sistemas (sem tipo especificado)	Ariadne O'Reilly Semantic Quran Worldcat(FAST)	É um tipo residual, para os datasets que são sistemas de representação do conhecimento, embora não se enquadrem nos outros três tipos. Não foi necessário, aqui, abrir novos subtipos.
Listas de leitura de universidade	Não Contém	Aspire Brunel Aspire Plymouth Nottingham Trent Resource Lists	O nome do tipo é autoexplicativo. São as bibliografias referentes a disciplinas oferecidas em universidades. Aqui se destaca quantitativamente a iniciativa Aspire.
Dataset de biblioteca	Dataset de biblioteca nacional	Bluk BNB Data Bnf.fr Datos Bne.es DNB	Catálogos de e informações sobre bibliotecas ou recursos informacionais que compõem seu acervo - sobretudo livros. Dá-se foco à unidade de informação, e não ao caráter científico que possa estar envolvido (apesar de isso ser considerado). Contudo, nota-se que é pertinente a presença desse tipo, e mesmo dos datasets da subdivisão "outros", na categoria "publicações". Os registros do catálogo de uma biblioteca, por exemplo, podem ser publicados e entendidos como publicação segundo a definição da DBpedia (ABOUT…, 2017).
	Dataset de biblioteca digital	DM2E Gutenberg Libris Verrijktkoninkrijk	
	Outros	B3kat Iobid Organizations Iobid Resources Sudoc.fr	
Ficheiro de autoridade	Não Contém	Idref.fr Radatana (BIBSYS) Viaf	Conjuntos de registros de pontos de acesso preferenciais, codificados numericamente, para a identificação de autores de publicações.
Não-classificado	Não Contém	Dutch Ships and Sailors Swedish Open Cultural Heritage	A inclusão desse tipo foi uma opção, para que não fossem obtidas classificações de difícil sustentação.

Fonte: Elaborado pelos autores.

Os tipos apresentados no quadro 1, inserem três novas classificações aos tipos inicialmente propostos por Schmachtenberg, Bizer e Paulheim (2014a): "sistema de representação do conhecimento", "ficheiro de autoridade" e "não-classificado". Este último foi uma opção, para que não fossem obtidas classificações insustentáveis; até mesmo pelos riscos alertados por Meusel et al. (2015). Além disso, o tipo base de dados de citação não foi contemplado, sendo excluído da análise.

Partindo da classificação de tipos desenvolvida, realizou-se a análise de cada um dos datasets pertencentes à categoria de publicações, identificando o tipo e a possível subdivisão a que tal dataset pertence. Vale observar que, se um dataset analisado apresentasse características de dois tipos ou subtipos, este era dividido ao meio. O resultado obtido está expresso na Tabela 3.

Tabela 3: Distribuição dos datasets da categoria de "publicações" conforme os tipos

Tipos	Datasets	%	Subdivisões		
			Tipos	Datasets	%
Informação sobre publicações científicas e/ ou sobre conferências científicas	56	45,53	Informação sobre publicações científicas	53	43,09
			Informação sobre conferências científicas	3	2,44
Não-classificado	19	15,45			
Sistema de representação do conhecimento	17,5	14,23	Tesauro	7	5,69
			Outros sistemas (sem tipo especificado)	6,5	5,28
			Lista de cabeçalhos de assuntos	2	1,63
			Sistema de classificação	2	1,63
Listas de leitura de universidade	14	11,38			

Dataset de biblioteca	12,5	10,16	Outros	6	4,88
			Dataset de biblioteca nacional	3,5	2,85
			Dataset de biblioteca digital	3	2,44
Ficheiro de autoridade	4	3,25			
Total	123	100			

Fonte: Elaborado pelos autores.

A distribuição dos datasets expressa na Tabela 3 permite identificar que o subtipo "informação sobre publicações científicas" se destaca como o mais numeroso: mais de 40% dos datasets analisados. Acredita-se que isso está menos relacionado com o risco de ter sido usado como depósito residual dos inclassificados, haja vista que se tomou cuidado com isso na análise, e mais relacionado ao fato de esse subtipo ser tido, perceptivelmente, como o mais característico à categoria de "publicações", como se tal categoria tivesse sido criada para ele. Esse número também é explicado por esse tipo comportar a grande maioria dos datasets publicados através da aplicação RKB Explorer[2], de modo semelhante ao que ocorre no tipo "listas de leitura de universidades", com a solução Aspire[3] (Talis Aspire Reading Lists).

É necessário destacar a presença notável de sistemas de representação do conhecimento como datasets na categoria "publicações", que levaram à constituição de um novo tipo de dataset. Tal medida pareceu bem mais apropriada do que somá-las à possível miscelânea do tipo "informação sobre publicações científicas e/ou sobre conferências científicas", problemática que será discutida mais adiante. Representam quase 15% dos datasets, número que se considera alto para um tipo aparentemente inusitado.

Vale notar que os datasets que foram incluídos no tipo "sistema de representação do conhecimento" se identificam com espécies de linguagens documentárias tradicionalmente compreendidas e aplicadas no âmbito da Ciência da Informação (CINTRA et al., 2002; CURRÁS, 2010), com exceção dos que se aproximaram disso, mas que foram por opção descritos como "outros sistemas (sem tipo especificado)", pois não houve necessidade dessa pormenorização neste momento; três espécies foram identificadas, e correspondem às demais subdivisões. Como exemplo da classificação realizada, apresenta-se no quadro 2 o resultado obtido com os três datasets citados por Schmachtenberg, Bizer e Paulheim (2014a):

2 "RKB Explorer is a Semantic Web application that is able to present unified views of a significant number of heterogeneous data sources regarding a given domain" (GLASER; MILLARD, 2007, p.1).

3 Talis Aspire Reading Lists é uma solução da Talis Information Systems, organização britânica que provê soluções e consultorias relativas a gerenciamento de bibliotecas e informação, tendo relação estreita com o W3C; é assim que o site institucional apresenta a referida solução atualmente: "Improve your student learning experiences, and support your teaching and learning strategies. They can also create major workflow efficiencies across your institution. We make it simple to create and manage resource lists that integrate fully with your current systems. At the same time, we provide powerful library back office functionality".

Biblioteca Nacional da Alemanha, L3S DBLP e Open Library, analisando os tipos de cada um dos datasets.

Quadro 2: Classificação dos datasets: Biblioteca Nacional da Alemanha, L3S DBLP e Open Library

Dataset	Tipo	Subdivisão
Biblioteca Nacional da Alemanha	Dataset de biblioteca	Dataset de biblioteca nacional
	Informação sobre publicações científicas e/ou sobre conferências científicas	Informação sobre publicações científicas
L3S DBLP	Informação sobre publicações científicas e/ou sobre conferências científicas	Informação sobre publicações científicas
Open Library	Dataset de biblioteca	Outros

Fonte: Elaborado pelos autores.

Os exemplos apresentados indicam a abrangência na classificação dos tipos da categoria publicações, pois os exemplos de datasets dados por Schmachtenberg, Bizer e Paulheim (2014a) estão inseridos em diversos tipos, inclusive pertencendo a subdivisões dos tipos em que foram classificados.

Visando aprofundar as análises dos resultados alcançados, na próxima seção é realizada a discussão de como a categoria publicações está posta frente ao contexto do Linking Open Data, partindo do elevado número de datasets tratando de representação do conhecimento.

3.1 A presença notável de sistemas de representação do conhecimento como datasets na categoria "publicações"

Ao mesmo tempo em que a descrição postulada por Schmachtenberg, Bizer e Paulheim (2014a) consiste em uma análise retrospectiva dos dados que já estavam publicados e foram analisados a posteriori, ela também comporta o caráter de expressar o que se esperava dessa categoria quando ela foi concebida, a exemplo da ausência de bases de dados de citação, mesmo que estejam expressamente manifestadas na referida descrição.

Como foi explicado anteriormente, a elevada quantidade de datasets no tipo "informação sobre publicações científicas" pode ser reflexo dessa expectativa. Nesse cenário, acredita-se, ainda, que a natureza dos dados que por excelência deveriam compor a categoria "publicações" - na expectativa de quem o concebeu - seria a de dados que consistissem eles mesmos em publicação científica em formato semântico. Essa hipótese lança luz sobre uma outra questão aqui problematizada: qual a razão de datasets que são sistemas de representação do conhecimento - conforme o

entendimento ancorado na Ciência da Informação - estarem classificados como "publicações"? Em resposta a isso, o que se tem são hipóteses, todas verossímeis, e que podem até mesmo ter composto a polissemia ou as múltiplas funções da palavra "publicação", de modo a princípio conveniente, diante da dificuldade no momento de classificar dados tão diversos como os presentes no LOD.

Assim, pode-se justificar que esses sistemas são, eles mesmos, as publicações. Seria o conhecimento disposto em dados, e não em documentos. Essa visão é um exemplo do que apresenta Marcondes (2011), que afirma se estar, nessas situações, diante de autênticas publicações digitais semânticas; o conhecimento, pela via da terminologia e, em seguida, do vocabulário controlado, tem sua representação desenvolvida de modo a ser ela mesma uma verdadeira base de conhecimento, ou seja, os dados são conhecimento.

Fato é que nenhum outro tipo dos elencados por Schmachtenberg, Bizer e Paulheim (2014a) carrega consigo essa proposta, de modo que, se a intenção for encontrar conhecimento propriamente dito nos dados publicados na categoria "publicações", o tipo "sistema de representação do conhecimento" é o mais indicado. Os outros tipos mais orbitam ao redor das "publicações" do que consistem em, justificando-se pela pertinência ao tema (ideia que se representa muito bem pela palavra "sobre") e pelo seu caráter utilitário. Assim, os autores teriam se esquecido de destinar um tipo especificamente para esses datasets.

É assim que se chega à segunda hipótese: existe o caráter utilitário desses sistemas de representação. Os datasets são vocabulários controlados que apóiam as publicações, estão junto delas, e é possível pensar que, nessa medida, compõem o tema "publicações". Sua presença nessa categoria se justifica praticamente pela mesma via que os datasets tipificados como "informação sobre publicações científicas", e seria uma saída considerar que Schmachtenberg, Bizer e Paulheim (2014a) teriam alocado esses sistemas no referido tipo. O mesmo se aplicaria ao novo tipo "ficheiro de autoridade".

Por fim, a última hipótese se sustenta sobre o momento de classificação do LOD em categorias, operação pertinente à representação temática. Os sistemas de representação do conhecimento verificados, então, teriam restado sem classe. Meramente por terem sido publicados, eles são incorporados à categoria "publicações", o que nem faria tanto sentido se se pensar que o LOD inteiro foi publicado, de modo

aberto e conectado. Seria mais pelo raciocínio expresso no ornitorrinco de Umberto Eco (LARA, 2001), basicamente um objeto de classificação problemática. E, claro, esse tipo de precariedade é endossada pelos equívocos alertados por Meusel et al. (2015).

Foi a fundamentação dessas hipóteses que levou à opção por tratar esses datasets como um caso especial e, portanto, como um novo tipo, não contemplado pela lista de Schmachtenberg, Bizer e Paulheim (2014a).

No intuito de aprofundar, agora, a contextualização da categoria "publicações" perante as demais, pode-se comparar a categoria de publicações com a de ciências da vida: esta última parece ser uma parte da primeira, e, no entanto, dado o nível elevado de desenvolvimento do tema e sua relevância, dela se desmembrou. No diagrama do LOD aqui analisado, publicado em 2014, "publicações" tem mais datasets. Contudo, basta uma rápida análise do diagrama do LOD publicado no início de 2017 para notar que a categoria de "ciências da vida" se desenvolveu fortemente, tanto em número de datasets quanto no tamanho deles, passando a ser a maior entre as categorias.

4 Considerações Finais

A quantidade de dados publicados no contexto do Linked Data vem aumentando consideravelmente, o que é comprovado pelo aumento visível no total de datasets que compõem o projeto do Linking Open Data. Essa expansão acompanha uma necessidade de estudar e compreender o conteúdo que está sendo publicado, para que assim, seja possível avaliar o impacto real desses conjuntos de dados na Web como um todo e principalmente tornar possível o consumo desses dados de forma mais adequada, atingindo o real objetivo da disponibilização de dados no padrão Linked Data.

Um possível impacto da disponibilização desses datasets tange aos dados de publicações científicas, que podem auxiliar em pesquisas bibliométricas, na localização de referências, na identificação de linhas de pesquisa, entre outros. Neste contexto, disponibilizar e utilizar dados de publicações científicas estruturados em formatos compatíveis com a Web Semântica pode contribuir significativamente para diversas áreas do conhecimento, colaborando com a Ciência como um todo. A partir destas considerações, este trabalho buscou compreender como os dados

do Linking Open Data pertencentes à categoria classificada como publicações estão organizados, bem como os tipos de dados que estão publicados nesta categoria.

Desta forma, uma primeira constatação realizada foi que não há uma perfeita identificação dos tipos de dados da categoria de publicações, ao analisar a natureza dos datasets. Outro aspecto é que, enquanto as denominações "listas de leitura de universidade" e "informação sobre conferências científicas" são precisas e correspondentes ao que foi verificado (embora sejam poucos os seus datasets), as denominações "dataset de biblioteca" e "informação sobre publicações científicas" são vagas, ainda que possam ser adequadas.

Além disso, a descrição de Schmachtenberg, Bizer e Paulheim (2014a) sobre esta categoria não manifesta explicitamente se a conceituação é meramente exemplificativa, o que implica na expectativa, de quem a interpreta, em ter todas as nuances da categoria contempladas. Contudo, foi com dificuldade que se aplicaram os tipos por eles elencados para qualificar cada dataset de "publicações", resultando em algumas subdivisões residuais e na abertura de novos tipos.

O caso especial dos sistemas de representação do conhecimento ou foi alocado discretamente no tipo "informação sobre publicações científicas e sobre conferências científicas", o que parece forçoso, ou foi negligenciado a despeito de, nessa hipótese, serem os datasets por excelência desta categoria. É bem verdade que nem todos esses datasets podem ser considerados ciência em forma de dados, na medida em que alguns apenas cumprem a funcionalidade imediata de um vocabulário controlado. O que importa é que esse material merecia um tratamento que o especificasse. Vislumbra-se que é possível esmiuçar e refletir mais sobre os datasets desse caso.

A dificuldade na compreensão do que é a categoria "publicações", ou seja, de quais são os sentidos em jogo no gesto classificatório que o coloca entre as demais categorias do LOD, bem como no gesto classificatório que lhe é interno, investigado com profundidade neste estudo, demonstram as sutilezas polissêmicas da palavra "publicação". É possível que seja essa indefinição um motivo para a relativa estagnação verificável no diagrama publicado em 2017. Restam ainda imprecisos o que se esperava - ou se espera - dessa categoria, bem como o que ela aceita.

REFERÊNCIAS

ABOUT: Publication. In: DBPEDIA. 2017. Disponível em: <http://dbpedia.org/page/Publication>. Acesso em: 9 maio 2017.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, v. 284, n. 5, p. 34 - 43, May 2001. Disponível em: <https://www.scientificamerican.com/article/the-semantic-web/>. Acesso em: 13 maio 2017.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked Data: the story so far. International Journal on Semantic Web and Information Systems, v. 5, n. 3, p. 1-22, 2009. Disponível em: <http://eprints.soton.ac.uk/271285/1/bizer-heath-berners-lee-ijswis-linked-data.pdf>. Acesso em: 19 abr. 2017.

CINTRA, A. M. M. et al. Para entender as linguagens documentárias. 2. ed. rev. e ampl. São Paulo: Polis, 2002.

CURRÁS, E. Ontologias, taxonomia e tesauros em teoria de sistemas e sistemática. Tradução: Jaime Robredo. Brasília: Thesaurus, 2010.

FERREIRA, J. A.; SANTOS, P. L. V. A. C. O modelo de dados Resource Description Framework (RDF) e o seu papel na descrição de recursos. Informação & Sociedade, João Pessoa, v. 23, n. 2, p. 13-23, maio/ago. 2013. Disponível em: <http://www.ies.ufpb.br/ojs/index.php/ies/article/view/15436/9681>. Acesso em: 17 jan. 2016.

GLASER, H.; MILLARD, I. RKB Explorer: Application and Infrastructure. CEUR Workshop Proceedings, 2007, [S.l.]: CEUR-WS.org, 2007. Disponível em: <http://ceur-ws.org/Vol-295/paper13.pdf>. Acesso em: 7 jun. 2017.

HEATH, T.; BIZER, C. Linked Data: Evolving the Web into a Global Data Space. ISOTANI, S.; BITTENCOURT, I. I. Dados Abertos Conectados. São Paulo: Novatec, 2015. Disponível em: <http://ceweb.br/livros/dados-abertos-conectados/>. Acesso em: 16 jan. 2017.

LARA, M. L. G. O unicórnio (o rinoceronte, o ornitorrinco...), a análise documentária e a linguagem documentária. DataGramaZero, v. 2, n. 6, p. 0-0, 2001. Disponível em: <http://www.brapci.ufpr.br/brapci/v/a/1251>. Acesso em: 17 Maio 2017.

MARCONDES, C. H. O papel das relações semânticas na organização e representação do conhecimento em ambientes digitais. In: SILVA, F. C. C. da; SALES, R. de (Orgs.) Cenários da organização do conhecimento: linguagens documentárias em cena. Brasília: Thesaurus, 2011. p. 129-168.

MEUSEL, R. et al. Towards Automatic Topical Classification of LOD Datasets. CEUR workshop proceedings, Florence, v. 1409, May 2015. Disponível em: <http://

ceur-ws.org/Vol-1409/paper-03.pdf>. Acesso em: 21 fev. 2017.

MOREIRO GONZÁLEZ, J. A. Linguagens documentárias e vocabulários semânticos para a web: elementos conceituais. Salvador: EDUFBA, 2011.

RAMALHO, R. A. S. Web Semântica: aspectos interdisciplinares da gestão de recursos informacionais no âmbito da Ciência da Informação. 2006. 120f. Dissertação (Mestrado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2006. Disponível em: <http://hdl.handle.net/11449/93709>. Acesso em: 7 mar. 2016.

SANTAREM SEGUNDO, J. E. Representação iterativa: um modelo para repositórios digitais. 2010. 244f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010. Disponível em: <http://hdl.handle.net/11449/103346>. Acesso em: 7 mar. 2016.

SCHMACHTENBERG, M.; BIZER, C.; PAULHEIM, H. Adoption of the linked data best practices in different topical domains. In: MIKA, P. et al. (Eds.). The Semantic Web – ISWC 2014: 13th International Semantic Web Conference, proceedings part 1. Heidelberg: Springer, 2014. p. 245-260.

_____; _____; _____. State of the LOD Cloud 2014. 2014. Disponível em: <http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/>. Acesso em: 21 jan. 2017.

SHADBOLT, N.; HALL, W.; BERNERS-LEE, T. The Semantic Web Revisited. IEEE Intelligent Systems Journal. v. 21, n. 3, p. 96-101, jan./fev. 2006. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1637364>. Acesso em: 10 maio 2017.

TADINI, A. V. W.; SANTAREM SEGUNDO, J. E. Realização de controle de autoridade para pesquisadores em datasets referidos como de "publicações" no Linked Data: um mapeamento. In: SEMINÁRIO EM CIÊNCIA DA INFORMAÇÃO (SECIN), 6., 2016, Londrina. Fenômenos Emergentes na Ciência da Informação. Londrina: UEL, 2016. v. 6. p. 1189-1201. Disponível em: <http://www.uel.br/eventos/cinf/index.php/secin2016/secin2016/paper/viewFile/353/209>. Acesso em: 8 fev. 2017.

THE LINKING Open Data cloud diagram. Disponível em: <lod-cloud.net>. Acesso em: 21 jan. 2017.

_____. Disponível em: <lod-cloud.net>. Acesso em: 17 maio 2017.