EIXO TEMÁTICO:
Organização e Representação da Informação e do Conhecimento

REALIZAÇÃO DE CONTROLE DE AUTORIDADE PARA PESQUISADORES EM *DATASETS* REFERIDOS COMO DE "PUBLICAÇÕES" NO *LINKED DATA*: UM MAPEAMENTO

AUTHORITY CONTROL FOR RESEARCHERS IN DATASETS REFERRED TO AS "PUBLICATIONS" IN THE LINKED DATA: A MAPPING

Antonio Victor Wolf Tadini - antoniovwt@gmail.com
José Eduardo Santarem Segundo - santarem@usp.br

Resumo: Este trabalho se volta para os *datasets* referidos como de "publicações" no *Linked Open Data*, que é concretização da Web Semântica, e para como é feito o controle de autoridade de pesquisadores nesse contexto. Como é elaborada a questão do controle de autoridade, especificamente para a identificação dos pesquisadores, nos *datasets* de publicação científica? E mais: existe um padrão que permita identificar, de modo unificado, os pesquisadores que são referidos nesses *datasets*? O trabalho objetiva obter uma análise criteriosa sobre como se resolve a referida questão. O método é uma análise exploratória dos recursos em RDF disponíveis no mais recente *Linked Open Data cloud diagram*, do qual foram selecionados os *datasets* enquadrados em "publicações". Neles, observou-se a ocorrência - ou ausência - de elementos que apontassem para soluções de controle de autoridade. Foi elaborada uma planilha para registro das observações e obtenção de dados estatísticos. Após as contabilizações, os *datasets* foram divididos em cinco categorias e a estatística mais significativa éa que isola a classe em que o controle é feito de modo ideal, em relação às outras duas em que cabe o controle: ela representa menos de 20% (ou 1/5), precisos 18,13%. Esse é o percentual de *datasets* que foram produzidos por entidades que, tendo a possibilidade ou a oportunidade - ou ainda, o dever - de fazerem o controle de autoridade, por meio dos denominados identificadores válidos, decidiram fazê-lo.

Palavras-chave: Web semântica. Controle de autoridade. Pesquisador. Publicação Científica.

Abstract: This work deals with datasets referred to as "publications" in the Linked Open Data, which is Semantic Web concretion, and with the control of researchers' authority in this context. We have looked into how the matter of authority control is tackled in the scientific publication datasets in the specific case of researchers' identification. We have also investigated whether there is a standard that allows researchers' unified identification in these datasets. In other words, this study aimed to analyze how this issue is approached in detail. The method employed herein consisted of an exploratory analysis of the resources in RDF available in the last Linked Open Data cloud diagram, during which the datasets framed in "publications" were selected. In these datasets, we observed the presence – or absence – of evidence pointing to authority control solutions. A spreadsheet was created to record observations and collect statistics. After analyses, the datasets were divided into five categories, and the most statistically significant category was the one that isolated the class in which the control was ideally performed, as compared to the other two categories where control was appropriate: this category represented less than 20% (or 1/5); more specifically 18,13%. This was the percentage of datasets produced by entities that due to possibility,

opportunity, or duty decided to conduct authority control by using the so-called valid identifiers.

Keywords: Semantic Web. Authority control. Researcher. Scientific Publication.

1 INTRODUÇÃO

Trata-se fundamentalmente de um estudo sobre Web Semântica e ambientes de publicação científica. Para que se entenda o que é Web Semântica, em um primeiro momento e de maneira simples, é possível olhar para esse nome e interpretar como uma rede (a Internet, largamente difundida) com significações nela injetadas. São atos de tratamento da informação, no âmbito da organização e representação da informação e do conhecimento.

Ainda de modo um tanto genérico, o que ocorre é que os também já bastante conhecidos *links* são aplicados ao máximo às informações que estão na nuvem - e não de maneira isolada e esporádica como anteriormente, - estabelecendo amarrações que fortalecem a rede. Mais precisamente, podemos pensar em vários tipos de *links* (RAMALHO, 2006, p.39), e é essa possibilidade de variação que permite que se fale em significados.

Este estudo parte da reunião de instrumentos teóricos que dizem respeito ao modo como se dá essa injeção de significações na Web para a constituição de estruturas; da menor possível até grandes estruturas de dados ligados.

A partir disso, é desenvolvida uma exploração acerca de uma questão bastante pontual, mas que muito importa para a efetivação de ambientes de publicação científica: a do controle de autoridade na identificação dos pesquisadores no *Linked Data*, que é uma concretização da Web Semântica.

> Authority control is the result of the process of maintaining consistency in the verbal form used to represent an access point and the further proccess of showing the relationships among names, works, and subjects. (TAYLOR; JOUDREY, 2009, p.249).

Em comum, o controle de autoridade e a Web Semântica têm a vocação para a organização da informação. Ambos se estabelecem conforme a lógica dos modelos de metadados: declarações, em que há recurso, propriedade e valor.

> The proccess of authority work [...] involves the proccess of documenting, in an authority record, the work done along with the decisions made. An authority record is a compilation of metadata about a person, a family, a corporate body, a place, a work, or a subject. (TAYLOR; JOUDREY, 2009, p.252).

A pesquisa se concentra na identificação de pesquisadores, logo os registros que aqui interessam são os relativos a pessoas. Sabe-se que hoje existem identificadores bastante poderosos, como o *Virtual International Authority File* (VIAF). Por meio de registros numéricos, ele conduz o processo de controle de autoridade congregando, principalmente, registros de cerca de trinta agências nacionais bibliográficas.

A análise de tecnologias e conceitos da Web Semântica, a que se propõe este trabalho, permite o contato com padrões e tendências mundiais no contexto do campo de informação e tecnologia, o qual se insere na Ciência da Informação. Ora, o imbricamento do *Linked Data* com o controle de autoridade é um exemplo perfeito disso.

De acordo com Santarem Segundo (2010, p. 16),

> os conceitos da Web Semântica, cunhada por Tim Berners-Lee e homologada pelo W3C, tem sido objeto de estudo das Ciências da Informacão e da Computacão e despertado interesse da comunidade, de um modo geral.

Disponibilizar dados dessa forma é sinônimo não só de arrojamento, mas de novas oportunidades, como as que dizem respeito à interoperabilidade e à própria possibilidade de buscas semânticas. Nesse contexto, passa a convir a adequação da organização dos recursos informacionais em ambientes digitais a essa tendência.

Ressalte-se que existe hoje uma quantidade razoável de dados estruturados abertos (*Linked Open Data*) - sendo boa parte representada por ambientes ditos de publicação.

Desse modo, entende-se que se impõe a necessidade do cuidado com o controle de autoridade dos pesquisadores representados nesses ambientes. Ele tanto pode estar sendo bem feito, de modo unificado e padronizado, quanto estar sendo completamente negligenciado.

Tal incerteza justifica uma tomada de conhecimento para que se pense em efetivação de ambientes de publicação científica no *Linked Data* e, destarte, desenvolvem-se os objetivos da pesquisa.

A pesquisa objetiva saber: como é trabalhada a questão do controle de autoridade, especificamente para a identificação dos pesquisadores, nos *datasets* de publicação científica? E mais: existe um padrão que permita identificar, de modo unificado, os pesquisadores que são referidos nesses *datasets*?

Desse modo, almeja-se a obtenção de uma análise criteriosa, com elementos

qualitativos e quantitativos, sobre como se resolve a questão do controle de autoridade nos *datasets* de publicação científica.

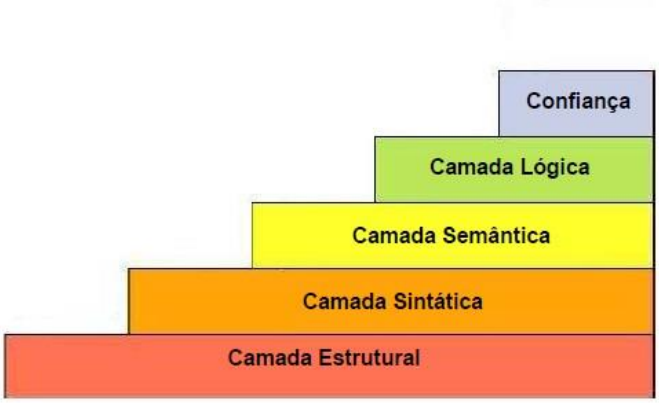
O objetivo específico é identificar, primordialmente entre os *datasets* que envolvem publicações científicas, a ocorrência de casos em que cabe controle com identificadores válidos (definidos mais à frente) e que, no entanto, isso não é feito, de modo que uma eventual adoção representasse uma oportunidade aproveitada; também se objetiva verificar a dimensão da realização do controle de autoridade pelos identificadores válidos. Ademais, considera-se que novas questões específicas podem se apresentar com a obtenção e análise dos resultados.

Neste trabalho, serão apresentados resultados parciais da investigação descrita acima, acompanhados do dimensionamento propiciado por eles e as decorrentes conclusões parciais.

2 WEB SEMÂNTICA

A Web Semântica, ou Web 3.0, é constituída por algumas camadas de padrões que se sobrepõem (RAMALHO, 2006, p.44). Pode-se organizá-las em um "espectro funcional" (RAMALHO, 2006, p.52), composto, em ordem, pelas seguintes camadas: estrutural, sintática, semântica, lógica e confiança.

FIGURA 1: "Espectro Funcional" da Web Semântica

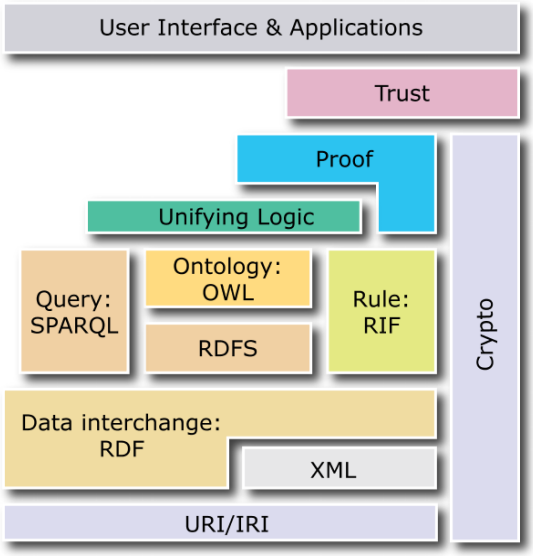


Fonte: RAMALHO, 2006, p. 52.

Existe um princípio na Web Semântica (KOIVUNEN; MILLER, 2001 apud RAMALHO, 2006, p.39) de que todo recurso físico ou abstrato é passível de ter um recurso informacional digital a ele equivalente, por meio de um URI que o identifique (SANTAREM SEGUNDO, 2010, p.73). Esse é o ponto de partida da sobreposição de

camadas, a camada estrutural.

FIGURA 2: Estrutura da Web Semântica (Layercake)



Fonte: http://www.w3.org/2007/03/layerCake.png (apud SANTAREM SEGUNDO, 2010, p.72).

Na constituição da Web Semântica, após a camada estrutural, a transição da camada sintática para a camada semântica se dá da seguinte maneira: a linguagem XML (*Extensible Markup Language*) é sobreposta pelo RDF (*Resource Description Framework*), que é um modelo de representação de recursos.

Uma estrutura RDF se concretiza através de declarações que descrevem recursos, são as chamadas triplas: sujeito-predicado-objeto ou recurso-propriedade-valor.

Por exemplo: Graciliano Ramos (sujeito/recurso) é autor de (predicado/propriedade) Vidas Secas (objeto/valor). Nessa declaração, “Graciliano Ramos” ocupa o *locus* do sujeito, mas nada impede que “Vidas Secas” seja sujeito em outra declaração: Vidas Secas (sujeito/recurso) tem como personagem (predicado/propriedade) Baleia (objeto/valor); e assim por diante. A estrutura de *links* flui.

Além disso, a sua conformidade com a teoria dos grafos (declarações são grafos) implica na possibilidade de aplicação de algoritmos de recuperação da informação que, por suas qualidades semânticas, são o principal caminho para a satisfação do usuário (SANTAREM SEGUNDO, 2010, p.51).

As ontologias, entre outras funções, inserem-se nesse contexto de

enriquecimento semântico para a recuperação da informação.

> Utilizar ontologias e suas relações é uma das maneiras de se construir uma relação entre termos dentro de um domínio, visto que elas possibilitam contextualizar dados, tornando mais eficiente a interpretação de documentos pelas ferramentas de recuperação da informação (SANTAREM SEGUNDO, 2010, p.100).

O acoplamento de ontologias (RAMALHO, 2006, p.41) é um enriquecimento da busca - potencialmente decisivo. Todas essas funcionalidades são capazes de operar desambiguações, reunir sinônimos, retornar informações relacionadas às que foram solicitadas, entre outras soluções.

Linked Data é o próximo assunto a ser tratado, mas já vale dizer: sua relação com o RDF pode ser evidenciada pela constatação de que o tamanho de um *dataset* pode ser apresentado pela sua quantidade de triplas, pois eles são constituídos dessas unidades que são as declarações em RDF.

3 LINKED DATA

O *Linked Data* é um formato de publicação de dados na Web, ou, mais especificamente,

> [...] um conjunto de melhores práticas para publicação e conexão de dados estruturados na Web, permitindo estabelecer links entre itens de diferentes fontes de dados para formar um único espaço de dados global (HEATH; BIZER, 2011 apud SANTAREM SEGUNDO, 2014, p.3868).

Considerando tais fontes de dados - os chamados *datasets* - na constituição da referida estrutura global, tem-se a DBpedia como a maior delas atualmente. Ela consiste em um esforço para a disponibilização dos dados contidos na Wikipédia de modo estruturado.

O *Linked Data* é entendido como concretização da Web Semântica. A navegação na DBpedia permite observar isso, tendo em vista o que Tim Berners-Lee (2006 apud SANTAREM SEGUNDO, 2014, p.3868) afirmou ser uma das aptidões do formato semântico: a de, a partir de um pouco de informação, obter-se muito mais, já que há um relacionamento estrutural previamente estabelecido.

Por exemplo, alguém está interessado em saber mais sobre Londrina; faz a busca; encontra na página de Londrina uma gama de *links* para cidades vizinhas; então, com simples cliques, enriquece sua pesquisa inicial com informações sobre as cidades vizinhas a Londrina.

Como dito anteriormente, é hoje notável a quantidade de informação que já está publicada no *Linked Data*, o que leva a crer que há coleções prontas de informações estruturadas, um enorme constructo à disposição. Por sinal, boa parte dele diz respeito a “publicações”, objeto desta pesquisa.

Mas vale o questionamento: até que ponto esse material está de fato disponível? Basicamente, são importantes, nesse sentido, as funcionalidades como as propiciadas pelo protocolo SPARQL (*Simple Protocol and RDF Query Language*), que permitem o aproveitamento do potencial semântico do *Linked Data* (SANTAREM SEGUNDO, 2014). A organização da informação ganha em sentido com operações de recuperação da informação.

4 METODOLOGIA

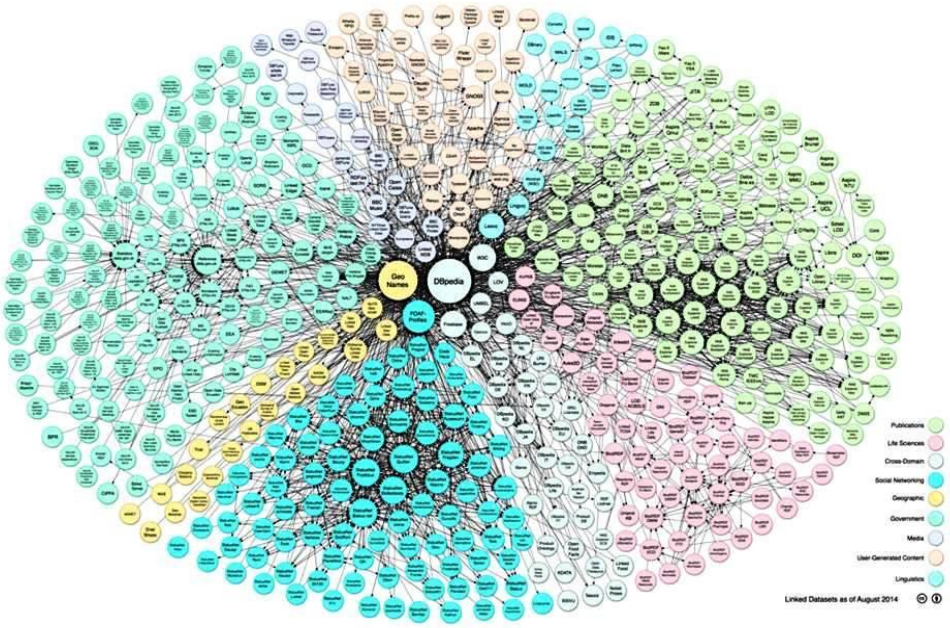
O método consiste no estudo exploratório dos recursos em RDF disponíveis no *Linked Open Data cloud diagram* (conforme atualizado em 30 de agosto de 2014), do qual foram selecionados os *datasets* enquadrados em “publicações”. São 133 *datasets*, que constituem o “objeto bruto” da exploração. Neles, observou-se a ocorrência - ou ausência - de elementos que apontassem para soluções de controle de autoridade.

Um instrumento basilar para essas observações é o documento modelo na Wikipédia intitulado *Authority Control*, que arrola, segundo dados extraídos do Wikidata, os identificadores aceitos para que o controle de autoridade seja válido, doravante referidos como identificadores válidos.

Foi elaborada uma planilha para registro das observações e obtenção de dados estatísticos para análise. Essas observações são concernentes à ocorrência ou ausência de identificadores válidos, bem como à necessidade de fazê-lo. Para cada *dataset*, um diagnóstico.

No diagrama, quando se clica no ícone de um *dataset*, tem-se acesso a uma página correspondente no *datahub* (<https://datahub.io/>). Nela, encontram-se documentos que ora exemplificam, ora constituem o *dataset* integralmente. As leituras precisaram, por vezes, ser realizadas na linguagem XML.

FIGURA 3 – *Linked Open Data clouddiagram* (atualizado em 30/08/2014)



Fonte: http://lod-cloud.net/

Com a planilha pronta, é possível identificar semelhanças e contrastes, para, então, estabelecer categorias. Feitas as devidas contabilizações, uma parcela importante da obtenção dos resultados ocorre nesse momento. Outra parcela é a que decorre da marcação quantitativa da ocorrência de cada identificador válido.

Dessa maneira, faz-se possível um primeiro dimensionamento - no material selecionado, - da realização do controle de autoridade, que integra os objetivos específicos da pesquisa.

5 RESULTADOS: APRESENTAÇÃO E DISCUSSÃO

Em uma primeira análise, caso a caso, percebeu-se, logo de início, que não se tratava exclusivamente de dados referentes a publicação científica, mas também a outros tipos de publicações.

Datasets como os denominados *British Museum Collection* e *STW Thesaurus for Economics*, por exemplo, apontaram nitidamente para então futuras nuances que seriam enfrentadas, entre as quais seria apenas uma a dos *datasets* que de alguma forma envolvem publicações científicas.

Mesmo sem uma classificação que fosse suficiente no que diz respeito à

consideração de tais nuances nesse momento, procedeu-se à análise sistemática dos documentos disponíveis de todos os *datasets*, com anotações para cada um.

À medida que se compreendiam as referidas nuances, os diagnósticos foram se tornando mais precisos e começaram a se repetir, apontando para algum tipo de classificação. Revisões e novas anotações foram necessárias.

Para exemplificar essa compreensão gradativa: os casos nos quais eram detectados identificadores válidos eram marcados. Outro exemplo: entendeu-se que em alguns deles não simplesmente é ausente o controle de autoridade, mas ele não é necessário, isto é, não cabe controle de autoridade. Ambos serão novamente abordados mais à frente.

Com o fim da observação caso a caso, foi possível estabelecer a seguinte classificação:

1. *datasets* aos quais não foi possível o acesso;
2. *datasets* em que não cabe controle;
3. *datasets* em que cabe controle, mas não há ou há de modo não uniforme;
4. *datasets* em que cabe controle, mas há apenas estruturas incipientes no sentido de identificação, de modo uniforme dentro do *dataset*;
5. *datasets* em que cabe controle e há controle homogêneo por meio dos identificadores válidos.

TABELA 1: Estatística de todos os *datasets* com classificação

	descrição da classe	nº. de *datasets*	percentual (%)
1	**sem acesso**	10,1	7,59
2	**não cabe controle**	33	24,81
3	**cabe, mas não há ou há de modo não uniforme**	23,55	17,71
4	**cabe e há estruturas incipientes para identificação**	50,05	37,63
5	**cabe e há controle**	16,3	12,26

Fonte: elaborado pelo autor

A presença de números não inteiros na coluna para os números de *datasets* se justifica pela divisão, que foi necessária, do *dataset* de nome DM2E, pois ele abrange dez conjuntos de dados com características distintas. A solução adotada foi atribuir um décimo (0,1) a cada um desses conjuntos, e distribuí-los nas categorias.

Junto da contabilização dos casos com presença de identificadores válidos,

que integram a classe 5, marcou-se quantitativamente a ocorrência de cada identificador, de modo que foram obtidos os seguintes resultados. Evidente que em parte dos casos há mais de um identificador.

TABELA 2: Estatística da ocorrência de identificadores

identificador	nº. de ocorrência(s)
VIAF	11,1
GND	6,6
ISNI	6
SUDOC	5
BNF	4
SELIBR	3
BIBSYS	1
BNE	1
KulturNav-id	1
LCAuth	1
NDL	1

Fonte: elaborado pelo autor.

Obtidos os resultados, isto é, contabilizadas as análises caso a caso, é interessante a realização de algumas operações para que se obtenham informações mais representativas e interpretações.

Inicialmente, despreza-se a classe 1 - pelo insucesso das tentativas de acesso, - a fim de serem observados os seguintes percentuais.

TABELA 3: Estatística com desprezo dos *datasets* sem acesso

	descrição da classe	nº. de *datasets*	percentual (%)
2	**não cabe controle**	33	26,85
3	**cabe, mas não há ou há de modo não uniforme**	23,55	19,16
4	**cabe e há estruturas incipientes para identificação**	50,05	40,72
5	**cabe e há controle**	16,3	13,26

Fonte: elaborado pelo autor.

Outra operação é conveniente para que siga a análise: os *datasets* em que não cabe controle de autoridade não podem ser considerados. Neles, o controle de autoridade não é uma questão, e o objetivo dessa pesquisa é observar como tal questão é resolvida, quando presente.

Mesmo que a essa altura da pesquisa não estejam precisados quais são os *datasets* que envolvem publicações científicas, sabe-se que, se um *dataset* tem essa característica, o controle de autoridade faz-se um problema. Desse modo, se não cabe controle, não há problema, logo justifica-se o desprezo. Não há risco de *datasets* de publicação científica serem excluídos com essa operação.

Vale ressaltar, todavia, a quantidade expressiva de *datasets* enquadrados em "publicações" nos quais não cabe controle de autoridade.

Novamente, se todo *dataset* que envolve publicações científicas demanda uma resolução para a questão do controle de autoridade, esses 33 *datasets* da classe 2 não envolvem publicações científicas diretamente.

Isso denota que pelo menos 26% dos *datasets* acessados enquadrados em "publicações" não envolvem publicações científicas ou isso se dá de modo superficial e indireto.

Voltando ao desprezo dessa classe 2, tal operação, então, justifica-se pela obtenção dos percentuais considerando como total os *datasets* em que cabe controle (classes 3, 4, 5), ou seja, aqueles em que existe a questão do controle de autoridade.

Nesses termos, finalmente foi possível a obtenção das seguintes estatísticas.

TABELA 4: Estatística apenas com *datasets* em que o controle de autoridade é uma questão

	descrição da classe	nº. de *datasets*	percentual (%)
3	**cabe, mas não há ou há de modo não uniforme**	23,55	26,20
4	**cabe e há estruturas incipientes para identificação**	50,05	55,67
5	**cabe e há controle**	16,3	18,13

Fonte: elaborado pelo autor

A estatística mais significativa é, desse modo, a que isola a classe 5 em relação às outras duas: percebe-se que ela representa menos de 20% (ou 1/5). Esse é o percentual de *datasets* que foram produzidos por entidades que, tendo a possibilidade ou a oportunidade - ou ainda, o dever - de fazerem o controle de autoridade, por meio de identificadores válidos, decidiram fazê-lo.

Outra análise pertinente é a que se depreende da superioridade do percentual da classe 4 sobre o da classe 3 (mais que o dobro). Ora, é de se pensar que existe uma mobilização que, de alguma forma, estabelece funcionalidades que contribuam,

mesmo que de modo incipiente, para a identificação padronizada dos pesquisadores; e ela é maior que o simples não cuidado quanto à questão.

Quanto à tabela 2 (Estatística de ocorrência de identificadores), destaca-se, em primeiro lugar, o VIAF (*Virtual International Authority File*); depois o GND, vinculado à Biblioteca Nacional da Alemanha; seguido pelo ISNI (*International Standard Name Identifier*), por identificadores franceses (SUDOC e BNF) e sueco (SELIBR).

Note-se que o ORCID, identificador cuja dinâmica é um pouco peculiar e que parece bastante viável, não é sequer uma vez utilizado; há apenas uma declaração, no *dataset* denominado VIVO Indiana University, de pretensão ao seu uso.

6 CONCLUSÕES PARCIAIS E PERSPECTIVAS

Este estudo até aqui é capaz de, por análise, demonstrar o tamanho da realização de controle de autoridade por identificadores válidos entre os *datasets* em que isso é pertinente, feito um recorte dos *datasets* disponíveis no *Linked Open Data* enquadrados como “publicações”. Analisando os casos em que cabe o controle, verificou-se que não há uma homogeneidade ou unicidade nos serviços utilizados para a resolução da questão do controle de autoridade.

Considera-se que a pesquisa deve prosseguir no sentido de precisar minuciosamente quais são os *datasets* que envolvem publicações científicas - e em que medida, - bem como, dado que todos eles estão entre os casos em que cabe controle de autoridade, definir quais são as intersecções com as três classes que foram definidas para esses casos. Novas estatísticas devem, então, ser produzidas para que os problemas de pesquisa continuem sendo atacados. O intuito é identificar a ocorrência de casos em que cabe controle com identificadores válidos, mas isso não é feito, de modo que uma eventual adoção representasse uma oportunidade aproveitada.

Além disso, ainda convém matizar a classe de *datasets* em que não cabe controle de autoridade (2), para que se entenda a natureza desses conjuntos de dados (estão enquadrados em “publicações”, mas não envolvem, pelo menos diretamente, publicações científicas). Boa parte deles são vocabulários controlados.

Não obstante, isso interessa para que seja pormenorizada a relação entre a classificação até então utilizada e a natureza de cada *dataset* do diagrama (ou de

categorias deles, quando possível).

REFERÊNCIAS

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.

_____. Web Semântica: Introdução a recuperação de dados usando Sparql. In: XV ENANCIB, 2014, Belo Horizonte/MG. **Além das nuvens: expandindo as fronteiras da Ciência da Informação**. Belo Horizonte/MG: ECI/UFMG, 2014. v. 1. p. 3863-3882.

TAYLOR, A. G.; JOUDREY, D. N. Metadata: access and authority control. In:_____. **The organization of information**. 3. ed. Westport, Conn.: Libraries Unlimited, 2009.

WIKIPEDIA. Module:Authority Control. Disponível em: <https://en.wikipedia.org/wiki/Module:Authority_control>. Acesso em: 7 mar. 2016.