DESAFIOS À COMUNIDADE IBERO-AMERICANA DE METADADOS EM REPOSITÓRIOS DIGITAIS PARA MAXIMIZAÇÃO DA INTEROPERABILIDADE

Ana Alice Baptista

INTRODUÇÃO

Acesso Aberto (AA) significa "acesso em linha permanente, imediato e gratuito ao texto integral de artigos de investigação por qualquer pessoa em qualquer ponto da web (HARNAD, 2007). Os repositórios digitais científicos são uma das formas mais comuns de implementar o AA; constituem o que normalmente se designa por via verde para o AA. Estes repositórios normalmente incorporam vários tipos de documentos científicos, mas os mais comuns são os artigos científicos e as teses e dissertações. A informação sobre estes documentos é armazenada em registros que se expõem ao exterior como registros de metadados em formato OAI-DC, que mais não é do que um formato XML criado no âmbito da *Open Archives Initiative* (OAI) para veicular metadados utilizando os famosos 15 termos do *Dublin Core Metadata Element Set* (DCMES) (DUBLIN CORE METADATA INITIATIVE, 2008). Estes registros podem depois ser coletados por outro software que obedeça ao protocolo *Open Archives Initiative – Protocol for Metadata Harvesting* (OAI-PMH) de modo a serem incorporados em bases de dados que suportam serviços de valor acrescentado à comunidade científica.

A utilização do formato OAI-DC para os registros de metadados garante interoperabilidade sintática. Contudo, qualquer breve análise desses registros revela de forma límpida os problemas que ainda subsistem em termos de qualidade dos metadados e, por isso, da interoperabilidade semântica. Pretende-se com este documento de posição apresentar à comunidade Ibero-Americana de metadados em repositórios digitais as soluções já há vários anos divisadas pelas comunidades internacionais de metadados e de web semântica. Estas passam pela aplicação rigorosa do quarto nível de interoperabilidade *Dublin Core* (DC), o que implica o desenvolvimento e a aplicação, em *Resource Description Framework* (RDF), de perfis de aplicação de metadados específicos. Desafia-se, assim, a comunidade Ibero-Americana de metadados em repositórios digitais a dar um passo em frente rumo à web semântica e aos Dados Abertos Ligados (DAL)[1] a fim de maximizar a interoperabilidade semântica entre os seus repositórios e entre estes e os do resto do mundo.

Este capítulo está dividido em cinco seções. Após esta introdução, apresenta-se um breve contexto sobre o paradigma *Linked Data*, após o que se disserta brevemente sobre a qualidade dos metadados. Por fim, apresenta-se a proposta a efetuar à comunidade Ibero-Americana de repositórios digitais, apresentando a posição do artigo e terminando com as considerações finais.

1 Tradução livre da expressão em inglês *Linked Open Data* (LOD).

LINKED DATA

> *The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF)* (WORLD WIDE WEB CONSORTIUM, 2013a).

A web semântica, também chamada de web de dados (BERNERS-LEE, 2009b), contrapõe-se à web de documentos, no sentido em que é inteligível por máquinas, ao contrário da segunda, inteligível por humanos. Informação inteligível por máquinas é associada a informação inteligível por humanos, possibilitando às máquinas a associação, relacionamento e processamento dessa mesma informação a um nível global e interoperável. A web semântica não é outra web, nem sequer uma nova versão da web atual: trata-se, antes, do **acoplamento** à web atual de um nível de significado que é processável por máquinas. a web atual não deixa de existir – apenas é estendida e enriquecida com significado.

Trata-se simplesmente de ter dados estruturados seguindo rigorosamente normas *de jure* e *de facto* específicas a domínios ou que os atravessam. É o caso das normas relativas aos termos DC ou à forma como estes termos e os valores com eles relacionados são codificados, conforme recomenda o *World Wide Web Consortium* (W3C).

Tim Berners-Lee (2009a) denomina de dados abertos cinco estrelas os que são veiculados em RDF e estão ligados, isto é, têm contexto acoplado (Figura 1). Ou seja, não basta que os dados estejam codificados em RDF e tenham identificadores; é necessário que seja possível que as máquinas os **entendam** sem intervenção humana adicional, é necessário dar-lhes contexto para que sejam semanticamente interoperáveis e, por isso, processáveis.

Também preocupada com as questões da interoperabilidade semântica a nível global, a *Dublin Core Metadata Initiative* (DCMI) lançou em 2009 uma recomendação intitulada *Interoperability Levels for Dublin Core Metadata* (NILSSON; BAKER; JOHNSTON, 2009). Esta recomendação, já traduzida para o português, apresenta quatro níveis de interoperablidade (Figura 2) e define como quarto nível, o mais alto, aquele em que exista um perfil de aplicação que empregue restrições aos registros (por exemplo, especifique como interpretar uma data ou quais os vocabulários controlados aceites para identificar o assunto ou o tipo do documento).

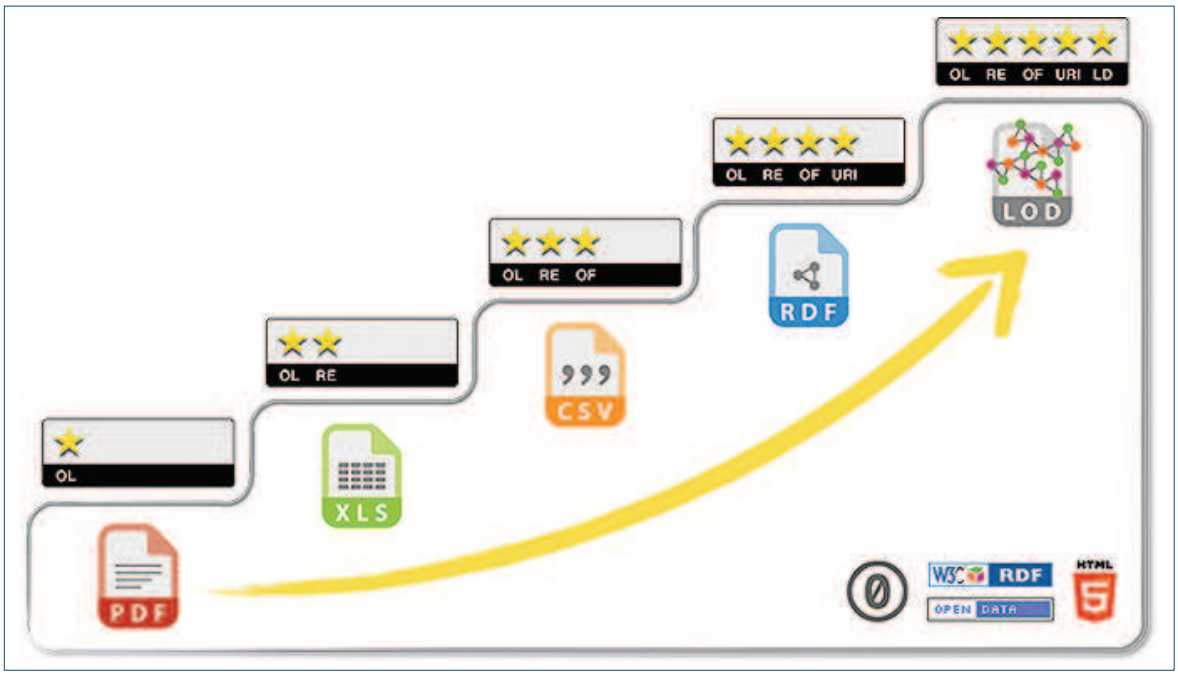

Figura 1 – Dados abertos cinco estrelas
Fonte: Hausenblas (2012).

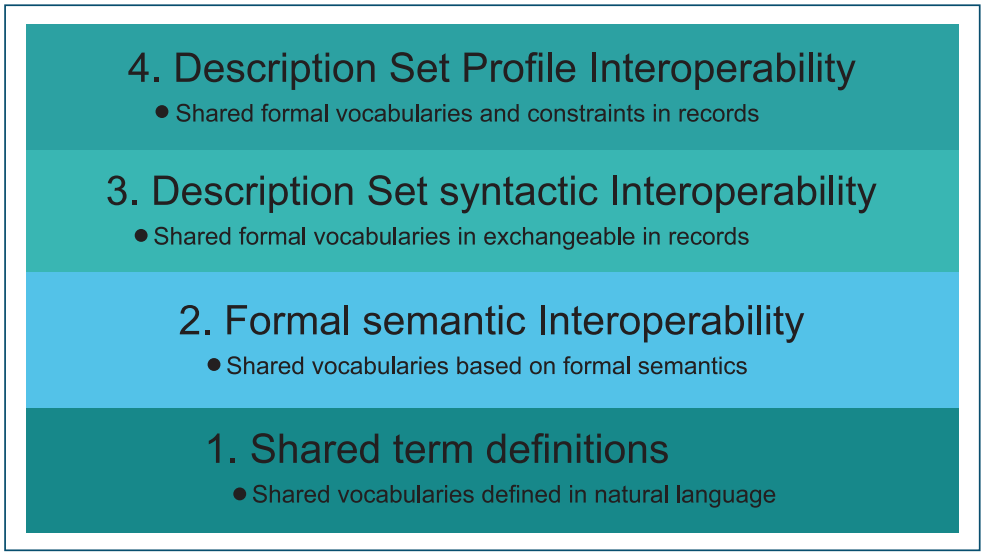

Figura 2 - Níveis de interoperabilidade *Dublin Core*
Fonte: Nilsson, Baker e Johnston (2009).

A definição de perfil de aplicação de metadados não é consensual na literatura. O Enquadramento de Singapura define-o como "um pacote de documentação" composto pelos seguintes componentes: requisitos funcionais, modelo de domínio, perfil de conjuntos de descrições, guias de utilização e guias de codificação (NILSSON; BAKER; JOHNSTON, 2008). Contudo, a definição que parece ter tido mais aceitação na comunidade é a de Heery e Patel (2000), que definem perfil da aplicação de metadados como "esquemas que consistem em elementos de dados provenientes de um ou mais *namespaces*, combinados pelos implemen-

tadores e otimizados para uma aplicação local em particular". Esta definição é relativamente antiga e não oferece aos perfis de aplicação de metadados o poder expressivo dos perfis de conjuntos de descrições do Enquadramento de Singapura, tal como definido por Coyle e Baker (2009) e por Nilsson (2008). No entanto, é bastante simples e clara, razão pela qual é a adotada neste capítulo.

São raros os repositórios digitais que têm definidos perfis de aplicação ou que sequer usam esquemas de metadados para além do DC simples, o OAI-DC (ANDRADE; BAPTISTA, 2015). Mesmo utilizando apenas OAI-DC, os repositórios não explicitam claramente como interpretar os valores associados às propriedades em cada um dos seus registros OAI-DC. Além disso, não há coerência entre registros de repositórios diferentes. É até possível que em alguns repositórios nem sequer haja coerência através dos seus próprios registros (aqui está um bom tema de investigação para um mestrando). A seção seguinte apresenta alguns exemplos e respetiva discussão sobre este assunto.

QUALIDADE DOS METADADOS

Uma consulta rápida aos registros OAI-DC dos repositórios digitais mostra claramente que há casos em que:

a) a semântica das propriedades não está a ser respeitada. Por exemplo, existem casos em que os valores *Open Access* ou *Restrict Access* aparecem associados à propriedade *dc:rights*, quando esta se refere aos direitos legais sobre o recurso e não ao tipo de acesso. Outro exemplo é a propriedade dc:identifier ser usada não para fornecer um identificador do recurso e sim outra informação como, por exemplo, a área disciplinar a que o recurso diz respeito;

b) não são utilizadas propriedades diferentes para veicular informação de natureza diferente. Um exemplo muito frequente é a repetição da propriedade *dc:date* com valores distintos, mas sem informação sobre o que cada uma das datas significa. Claramente, nestes casos será necessário utilizar propriedades diferentes. No caso particular do *dc:date*, pode-se utilizar refinamentos definidos pela própria DCMI, tais como *dct:created*, *dct:modified* ou *dct:issued*. O serviço *Linked Open Vocabularies* (LOV) permite a pesquisa de informação relativa a vários vocabulários abertos ligados, pelo que propriedades não existentes no DC Terms (DUBLIN CORE METADATA INITIATIVE, 2012), podem ser ali pesquisadas e identificadas;

c) os valores associados às propriedades não estão preparados para serem facilmente interpretados fora do âmbito estrito do repositório a que dizem respeito. Por exemplo, muitos repositórios utilizam os idiomas

e alfabetos oficiais dos países onde estão sediados para representar o assunto (*dc:subject*), o tipo de documento (*dc:type*), os direitos (*dc:rights*), entre outros. Alguns repositórios até usam texto para representar os valores associados a propriedades como *dc:relation* ou *dc:identifier* (para os quais as boas práticas recomendam a utilização de identificadores). Uma solução simples para este problema é usar, de forma complementar ou não, links para termos de vocabulários controlados codificados em linguagens apropriadas (SKOS, RDFS, OWL).

Outro problema comum à maior parte destes repositórios é o fato de todos os recursos serem descritos utilizando o mesmo conjunto de propriedades (as do OAI-DC). Na verdade, há propriedades que são específicas para determinados tipos de recursos. Mais do que isso, recursos do mesmo tipo podem necessitar de propriedades específicas de acordo com o seu domínio de conhecimento (ANDRADE; BAPTISTA, 2014).

Estes são alguns dos problemas facilmente identificáveis em breves análises aos registros de metadados dos repositórios digitais atuais. Os registros OAI-DC são facilmente convertíveis para registros RDF, mas só valeria a pena fazê-lo se esses registros RDF acrescentassem algo ao que atualmente dispomos nos registros OAI-DC. Para isso, esses registros RDF devem estar ligados entre si e ligados a outros registros RDF de natureza diferente (por exemplo, vocabulários controlados, ontologias, conjuntos de dados governamentais, conjuntos de dados estatísticos, dentre outros), ou seja, devem veicular dados cinco estrelas.

A Figura 3 mostra a *LOD cloud*, uma parte da web de dados que cumpre os princípios *Linked Data* (ABELE et al., 2017; BERNERS-LEE, 2009a), isto é, que são dados cinco estrelas (WORLD WIDE WEB CONSORTIUM, 2013). Na *LOD cloud* os dados estão classificados em diversas categorias, e a cada qual é atribuída uma cor. O gráfico é interativo e é possível obter mais informação sobre cada conjunto de dados, bem como verificar os tipos de ligações entre eles. Estes dados, mesmo tendo diferentes naturezas, estão no nível de interoperabilidade mais alta, o que implica que sejam passíveis de relacionamento se não automático, pelo menos com pouca intervenção humana. Quanto mais informação de contexto houver e quanto mais normalizada ela seja, menos intervenção humana se espera.

Um olhar mais aproximado sobre a *LOD cloud* mostra que as redes de repositórios digitais baseadas apenas no protocolo OAI-PMH não estão lá presentes. O OAI-PMH garante níveis de interoperabilidade sintática e uma interoperabilidade semântica básica mas, por si só, não é adequado quando se pretende atingir níveis de interoperabilidade semântica superiores. Passada já quase uma década e meia da criação dos primeiros repositórios digitais que implementam o protocolo OAI-PMH, é surpreendente, e até mesmo desanima-

dor, constatar que estão praticamente no mesmo ponto em termos de interoperabilidade semântica: permanecem em ilhas de interoperabilidade (BAPTISTA, 2010), isto é, são algo interoperáveis entre si, mas estão isolados do resto dos dados e aplicações na web.

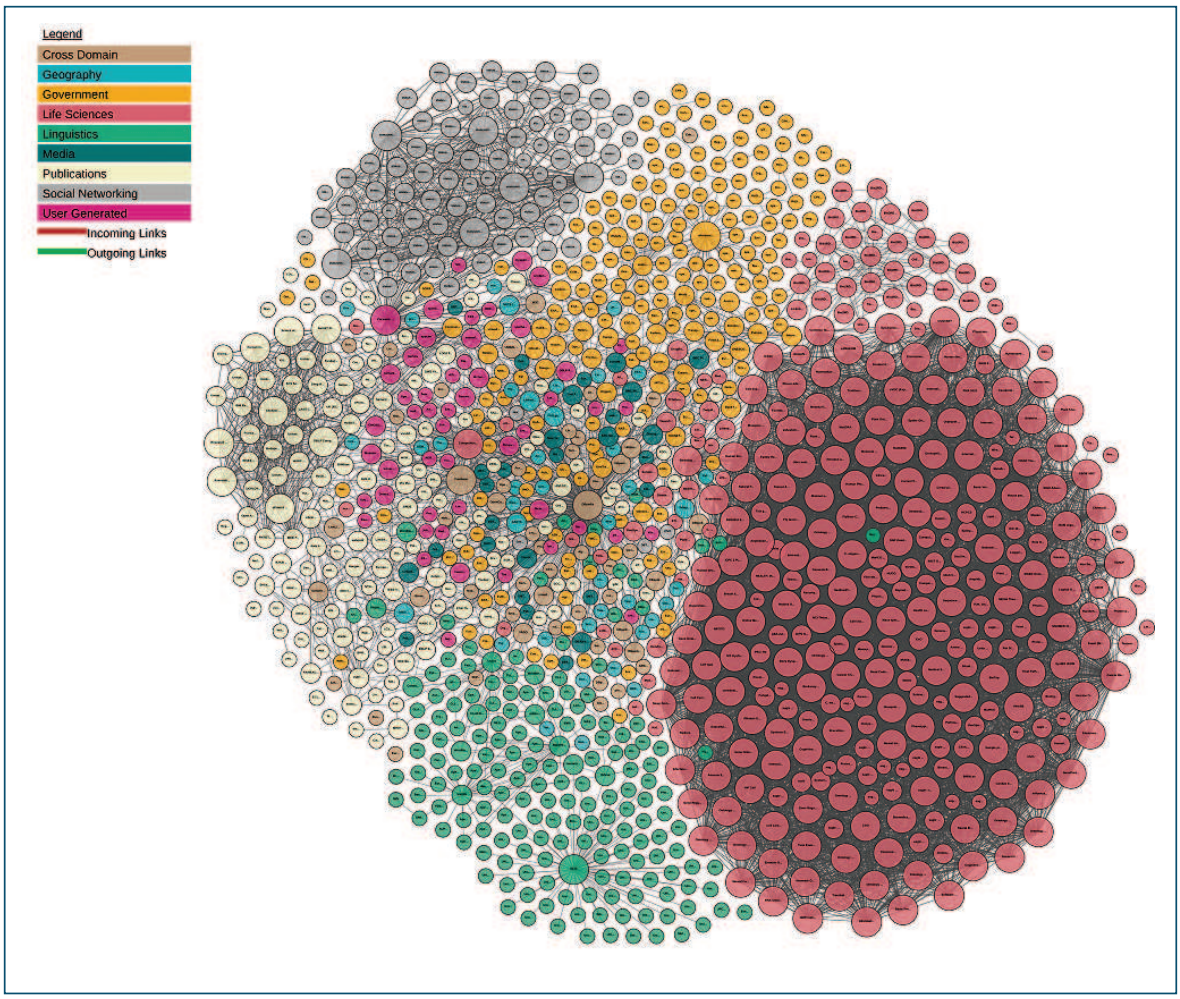

Figura 3 – *Linking open data cloud diagram* 2017
Fonte: Abele et al. (2017).

PROPOSTA

A posição defendida neste capítulo é, por isso, a seguinte: os repositórios digitais devem estar no nível de interoperabilidade quatro da DCMI e, para isso, devem ter perfis de aplicação rigorosamente definidos, os seus registros de metadados devem estar codificados em RDF e devem estar relacionados com outros dados na web (vocabulários, ontologias, outros dados) que lhes possam fornecer contexto (semântica).

Assim, propõe-se à comunidade Ibero-Americana de metadados em repositórios digitais que desenvolvam e implementem perfis de aplicação observando as seguintes sugestões:

a) determinar quais propriedades utilizar de acordo com o tipo e a natureza dos recursos. É de esperar que haja várias propriedades em comum que devem ser usadas em recursos de diferentes tipos e naturezas (como, por exemplo, as propriedades do DC Terms), mas também há propriedades que são específicas a determinados tipos de recursos. Por exemplo, supervisor, para teses e dissertações. Os implementadores poderão, ainda, querer incluir propriedades específicas relacionadas com a natureza do recurso. Por exemplo, determinados artigos no âmbito da Medicina poderiam ter associadas propriedades que permitissem informar sobre o diagnóstico ou sobre a terapêutica;

b) garantir que a semântica das propriedades é rigorosamente definida e aplicada. Se tiverem necessidade de utilizar propriedades com semântica diferente das que têm disponíveis, devem, em primeiro lugar, verificar se as propriedades já estão definidas em algum vocabulário/esquema e se são usadas pela comunidade (utilizando serviços específicos como, por exemplo, o LOV). Se as propriedades não estiverem definidas/declaradas em nenhum vocabulário/esquema, então elas podem ser criadas e disponibilizadas num vocabulário/esquema para que outros as possam também usar. Neste caso, devem estar relacionadas com outras propriedades que já existam, e que lhes fornecem, assim, contexto. Devem posteriormente ser incluídas em serviços como o LOV para que possam ser facilmente encontradas por outros que as queiram utilizar;

c) garantir que usam refinamentos de propriedades sempre que tal for imprescindível. Esta recomendação está relacionada com a anterior, mas refere-se especificamente às propriedades cuja semântica refina a de outras já existentes. Um caso frequente é o das datas. Existem vários tipos de datas como, por exemplo, *dct:modified*, *dct:valid*, ou *dct:created*. Todas têm uma semântica muito específica e, por vezes, oposta, apesar de todas se referirem a datas. Faz, por isso, sentido que os refinamentos sejam usados e não a propriedade mais geral, neste caso o dct:date. Antes de se criar qualquer propriedade nova, deve-se pensar muito bem, já que se trata de um equilíbrio difícil entre rigor semântico e interoperabilidade. A utilização de qualquer nova propriedade, ainda que semanticamente mais rigorosa, pode implicar perda de interoperabilidade;

d) utilizar esquemas de sintaxe para especificar como se devem interpretar os valores associados às propriedades. Um exemplo frequentemente utilizado é o da indicação da norma ISO 8601 para a representação das datas;

e) utilizar esquemas de vocabulário para relacionar com os valores de determinadas propriedades como, por exemplo, *dct:subject*, *dct:type* ou *dct:rights*. Se não existirem esquemas de vocabulário que possam usar,

os implementadores podem considerar criá-los e disponibilizá-los livremente à comunidade para uso e reuso. Os registros de metadados podem, assim, ter informação legível por humanos, com as etiquetas dos termos, e informação legível por máquinas com os links diretos para os termos.

f) utilizar, sempre que possível, informação legível por máquinas (*links*) complementada com informação legível por humanos (texto) para os valores associados às propriedades. Por exemplo, para os valores relacionados com a propriedade dct:publisher em vez de se utilizar apenas o nome de determinada editora, pode-se complementar essa informação com o link relativo a essa editora;

g) utilizar *links* persistentes, ou seja, que permaneçam no tempo.

Estas são as medidas principais a tomar para que se atinjam os níveis de interoperabilidade semântica necessários à incorporação na *LOD cloud* dos dados constantes nos registos OAI-PMH dos repositórios digitais. Para que estes dados passem realmente a fazer parte da LOD cloud, um outro passo será ainda necessário: a transformação dos registos OAI-PMH em RDF e a sua disponibilização aberta na rede como *Linked Open Data*.

CONSIDERAÇÕES FINAIS

Os repositórios digitais organizados em rede são sintaticamente interoperáveis entre si, mas atualmente ainda subsistem muitos problemas de interoperabilidade semântica. Estes devem-se à tecnologia utilizada mas também à qualidade dos metadados: nem dentro das suas próprias ilhas de interoperabilidade os repositórios digitais são semanticamente interoperáveis.

As soluções para estes problemas passam por começar a utilizar tecnologias da Web Semântica, como as baseadas em RDF e repensar a estrutura de metadados dos diversos repositórios digitais, utilizando e adaptando propriedades de diferentes esquemas adequadas às necessidades de cada repositório. Em poucas palavras, criar perfis de aplicação que estejam de acordo com as necessidades específicas dos repositórios.

A comunidade Ibero-Americana de metadados em repositórios digitais é uma comunidade muito dinâmica e empenhada, com vários casos de sucesso a nível mundial. Por isso, é uma comunidade que tem condições para liderar uma mudança no panorama das políticas de interoperabilidade dos repositórios digitais, para dar o passo em frente, evoluindo para o paradigma Linked Open Data de maneira calculada e sustentada. Este é o desafio que proponho e para o qual estou disposta a contribuir.

REFERÊNCIAS

ABELE, A. et al. **The linking open data cloud diagram**. 2017. Disponível em: <http://lod-cloud.net/>. Acesso em: 27 maio 2017.

BAPTISTA, A. A. A falar nos entendemos: a interoperabilidade entre repositórios digitais. In: GOMES, M. J.; ROSA, F. (Ed.). **Repositórios institucionais**: democratizando o acesso ao conhecimento. Salvador: EDUFBA, 2010. p. 71-90. Disponível em: <https://repositorio.ufba.br/ri/bitstream/ri/616/3/Repositorios%20institucionais.pdf>. Acesso em: 27 maio 2017.

ANDRADE, M. C.; BAPTISTA, A. A. Information needs of researchers in a bibliographic databases environment: a literature review. In: POLYDORATOU, P.; DOBREVA, M. (Ed.). **Let's put data to use**: digital scholarship for the next generation: proceedings of the 18th International Conference on Electronic Publishing. Tassaloniki: IOS Press, 2014. p. 30-38. Disponível em: <https://doi.org/10.3233/978-1-61499-409-1-30>. Acesso em: 31 maio 2017.

ANDRADE, M.; BAPTISTA, A. A. The use of application profiles and metadata schemas by digital repositories: results from a Survey. In: INTERNATIONAL CONFERENCE ON DUBLIN CORE AND METADATA APPLICATIONS, 2015, São Paulo. **Proceedings...** São Paulo: Dublin Core Metadata Initiative, 2015. p. 146–157. Disponível em: <http://dcevents.dublincore.org/IntConf/dc-2015/paper/view/362/373>. Acesso em: 27 maio 2017.

BERNERS-LEE, T. **Linked data**: design issues. 2009a. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 27 maio 2017.

BERNERS-LEE, T. **The next web**. 2009b. Disponível em: <http://www.ted.com/index.php/talks/tim_berners_lee_on_the_next_web.html>. Acesso em: 31 maio 2017. 1 vídeo (17min).

COYLE, K.; BAKER, T. **Guidelines for Dublin Core application profiles**. 2009. Disponível em: <http://dublincore.org /documents/2009/05/18/profile-guidelines/>. Acesso em: 31 maio 2017.

DUBLIN CORE METADATA INITIATIVE. **DCMI Metadata Terms**. 2012. Disponível em: <http://dublincore.org/documents/dcmi-terms/>. Acesso em: 31 maio 2017.

DUBLIN CORE METADATA INITIATIVE. **Dublin Core metadata element set, version 1.1**. 2008. Disponível em: <http://dublincore.org/documents/dces/>. Acesso em: 31 maio 2017.

HARNAD, S. **Green road to open access**: a leveraged transition. 2007. Disponível em: <http://users.ecs.soton.ac.uk/harnad/Temp/greenroad.html>. Acesso em: 31 maio 2017.

HAUSENBLAS, M. **5* open data**. 2012. Disponível em: <http://5stardata.info/en/>. Acesso em: 30 maio 2017.

HEERY, R.; PATEL, M. Application profiles: mixing and matching metadata schemas. **Ariadne**, v. 25, set. 2000. Disponível em: <http://www.ariadne.ac.uk/issue25/app-profiles/>. Acesso em: 31 maio 2017.

LINKED Open Vocabularies (LOV). Disponível em: <http://lov.okfn.org/dataset/lov/>. Acesso em: 30 maio 2017.

MILLER, P. **Transcript:** sir Tim Berners-Lee talks with Talis about the semantic web. 2008. Disponível em: <http://talis-podcasts.s3.amazonaws.com/twt20080207_TimBL.html>. Acesso em: 31 maio 2017.

NILSSON, M. **Description set profiles**: a constraint language for Dublin Core application profiles. 2008. Disponível em: <http://dublincore.org/documents/2008/03/31/dc-dsp/>. Acesso em: 31 maio 2017.

NILSSON, M.; BAKER, T.; JOHNSTON, P. **The Singapore Framework for Dublin Core application profiles**. 2008. Disponível em: <http://dublincore.org/documents/singapore-framework/>. Acesso em: 30 maio 2017.

NILSSON, M.; BAKER, T.; JOHNSTON, P. **Interoperability levels for Dublin Core metadata**. 2009. Disponível em: <http://dublincore.org/documents/interoperability-levels/>. Acesso em: 31 maio 2017.

WORLD WIDE WEB CONSORTIUM. **W3C semantic web activity**. 2013a. Disponível: <http://www.w3.org/2001/sw/>. Acesso em: 31 maio 2017.

WORLD WIDE WEB CONSORTIUM. **5 star linked data**. 2013b. Disponível em: <https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data>. Acesso em: 30 maio 2017.