

FABRICO/CIÊNCIA: Um Ambiente *Linked Data* para o Mapeamento da Ciência

Eixo temático: Métodos, Técnicas e Ferramentas para Estudos Bibliométricos e Cientométricos

Modalidade: Apresentação oral

1 INTRODUÇÃO

Com a Web, surgiram os Arquivos Abertos (AA) e a Iniciativa dos Arquivos Abertos (IAA) que atuam, respectivamente, no sentido de tornar textos de resultados de pesquisas visíveis e usáveis por usuários da Internet, e promover padrões de interoperabilidade que visam a disseminação dos AAs. Estas ações garantem maior visibilidade aos trabalhos de pesquisa e viabilizam a disseminação de conteúdos que até então eram de circulação restrita, como teses e dissertações e relatórios de pesquisa. Para estudos bibliométricos, oportunizam fontes de acesso livre e conteúdos até então de circulação restrita.

A Web também evoluiu para um cenário em que seus usuários passaram de meros consumidores de informação para colaboradores, que produzem e agregam valores às informações e operam através de redes (como redes de relacionamento). Neste cenário, chamado de Web 2.0, ambientes passam a permitir que usuários adicionem valores aos conteúdos, na forma de comentários, avaliações, recomendações, relações, denúncias e classificação social (chamada de *folksonomia*, em que assuntos são livremente atribuídos pelos usuários). Para os estudos bibliométricos, a Web 2.0 pode contribuir com um cenário de análise, produção e preparação coletiva de dados, que já está familiarizado em operar redes. Pode proporcionar a socialização e o compartilhamento de esforços e produtos, e provocar um alargamento das fronteiras destes estudos, oportunizando a inclusão de novos tipos de colaboradores, como pesquisadores interessados em identificar os rumos, as relações, e os canais ligados às suas pesquisas.

A Web Semântica apresenta-se como uma nova Web que visa dar significado aos recursos da Web atual, permitindo que estes sejam compreendidos tanto por pessoas quanto por sistemas computacionais (Berners-Lee, Hendler e Lassina, 2001). A plataforma da Web Semântica estende a da web atual através de uma camada semântica, na qual o significado dos

recursos é descrito através de metadados, que são produzidos de acordo com ontologias. A Web Semântica traz facilidades para o compartilhamento de informação uma vez que a interoperabilidade de informação ocorre em uma camada acima dos aspectos estruturais dos recursos, isto é, na camada semântica. A interoperabilidade e a integração da informação são facilitadas, pois os metadados são representados em estruturas simples: triplas expressas na linguagem RDF (Resource Description Framework).

Tendo como plataforma a Web Semântica, surge, sob a denominação Linked Data, uma iniciativa para publicar conjuntos de dados já existentes (hoje armazenados em bases de dados relacionais e documentos XML) na forma de triplas RDF. Com o objetivo de permitir a ligação entre dados publicados por diversas organizações, Linked Data estabelece como princípio usar URIs para identificar unicamente as entidades representadas nesses conjuntos de dados. Web Semântica e Linked Data aprenam-se como uma plataforma promissora para publicação de dados usados em estudos bibliométricos, à medida dispõem uma estrutura simples para a representação de dados e para integrar dados, trazendo, por exemplo, facilidades para as atividades preparatórias de estudos bibliométricos ligados ao pré-processamento de dados. O uso de URI para identificar entidades como autores, instituições e artigos é um instrumento valioso para a identificação de descrições repetidas destas entidades (tanto redundantes como complementares), principalmente quando entidades provenientes de várias fontes são integradas. O controle de autoridades via Linked Data é mais imediato, internacionalizado, transparente e controlável (Keller et ali, 2011).

Este artigo apresenta o Fabrico Ciência, que é um ambiente que investiga o uso de Arquivos Abertos, Web 2.0, Web Semântica e Linked Data como ferramental para estudos bibliográficos. A seção 2 discute benefícios de IAA, Web 2.0, Web Semântica e Linked Data para ferramentas de mapeamento da ciência, a seção 3 apresenta como Fabrico Ciência explora estes benefícios. O estado atual, resultados já obtidos e próximas etapas são discutidos na seção 4.

2 MAPEAMENTO DA CIÊNCIA NO CENÁRIO IAA, WEB 2.0 E LINKED DATA

Mapeamento da Ciência é um tópico de pesquisa da bibliometria que busca encontrar representações de conexões intelectuais em no sistema dinâmico e mutante do conhecimento científico (Small, 1997 apud Cobo et ali, 2011). Para a análise do mapeamento da ciência, Cobo et ali (2011) apresentam a seguinte sequência de atividades, que são incorporadas em ferramentas de mapeamento da ciência: a **busca de dados** em fontes bibliográficas; o **pré-processamento** dos **dados**, que implica em gerenciar erros ortográficos e dados faltantes,

reduzir de dados, fatiar dados no tempo e pré-processar redes para reduzir nodos; **a extração das redes**, que abordam co-autoria, co-termos, co-citação, acoplamento bibliográfico, tendo como unidades autores, documentos, país, instituição; **a normalização**, em que similaridades entre os dados são detectadas, através do uso de diversas medidas de similaridade, formando agrupamentos (*clusters*); **o mapeamento**, em que algoritmos são aplicados sobre a rede para a construção de mapas, envolvendo reduções de dimensão, formação de sub-redes (via algoritmos de agrupamento), extração de espinha dorsal de rede, etc.; **métodos de análise**, em que técnicas como análise de redes (como análise temporal) são usadas para extrair conhecimento a partir da rede; e **visualização**, em que técnicas de visualização de redes são empregadas para possibilitar uma melhor compreensão dos resultados; e **interpretação**, em que o analista interpreta os resultados e os mapas utilizando-se de suas experiências e conhecimentos.

Para promover a integração entre repositórios de Arquivos Abertos, a IAA desenvolveu o padrão OAI-PMH¹. Este padrão estabeleceu uma arquitetura de interoperabilidade composta por provedores de dados, que são os repositórios de arquivos abertos, e provedores de serviços, que são serviços de terceiros que utilizam as informações de provedores de dados para oferecer serviços de alto nível (como busca em múltiplos repositórios), e por um protocolo de comunicação que permite que provedores de serviços colham descrições (metadados) de itens armazenados em provedores de dados. Ferramentas para o mapeamento da ciência poderiam ser desenvolvidas como provedores de serviços, capazes de colher dados de provedores de dados, como revistas eletrônicas, bases de dados de teses e dissertações e repositórios temáticos. Estas ferramentas poderiam prover serviços de alto nível, incluindo mecanismos de extração, análise e visualização de redes (como redes de co-autoria e co-termos). A disponibilização desses serviços permitiria que pesquisadores, interessados em identificar rumos, contextos e conexões de suas pesquisas, pudessem também explorar mapas da ciência.

Na Web 2.0 foram desenvolvidas ferramentas analisam dados produzidos colaborativamente, que agregam, agrupam e permitem a visualização desses dados. Por exemplo, ambientes de classificação social (*folksonomias*) operam sobre um grafo tripartido formado por etiquetas (assuntos atribuídos livremente por usuários), usuários (produtores de

¹ Open Archives Initiative Protocol for Metadata Harvesting -<http://www.openarchives.org/pmh/>

etiquetas) e recursos (documentos da web etiquetados). Estes ambientes de classificação social permitem a análise redes de co-relações, como redes de co-documentos, formadas por documentos indexados por mesmas etiquetas ou mesmos usuários, co-usuários, formadas por usuários que indexam via mesmas etiquetas ou por mesmos documentos, e co-etiquetas, formadas por etiquetas atribuídas a mesmos documentos (Peters 2009). Muitos destes instrumentos são similares aos utilizados nos estudos bibliométricos, como co-autoria, co-terminos, co-citação, mas, na Web 2.0, estes instrumentos apresentam-se com características próprias, como a exploração de redes de co-terminos através da navegação por nuvens de termos relacionados. As ferramentas de mapeamento da ciência são voltadas para estudos em profundidade, por pequenos grupos, para públicos e propósitos específicos, instaladas em estações de trabalho de redes locais. Já ferramentas como *folksonomias* voltam-se, à exploração do espaço do problema pelo povo, e à descoberta inesperada de coisas neste espaço, isto é, *serendipity* (Mathes, 2004). A disponibilização de ferramentas para estudos bibliométricos na plataforma Web 2.0 pode trazer benefícios, à medida que viabiliza a inclusão e a análise de novos tipos de informações (como etiquetas e avaliações de usuários), a adoção de ações coletivas de preparação e pré-processamento de dados, a socialização da análise dos dados, e a inclusão de novos públicos.

Hoje há produção e publicação crescente de dados bibliográficos no cenário Web Semântica/Linkd Data, que demonstram sua importância para estudos de mapeamento da ciência. Várias ontologias são desenvolvidas para descrever recursos acadêmicos e de pesquisa, como SWRC², BIBO³, CiTO⁴, SWAP⁵, SWC⁶, LSC⁷. Várias iniciativas em disponibilizar dados bibliográficos como Linked Data estão em andamento, como: a publicação de artigos, conferências, organizações e pessoas da área da Web Semântica (Semantic Web Dog Food⁸); a publicação da base de dados bibliográfica da Ciência da Computação DBPL (Faceted DBLP⁹); dos ativos de ciência produzidos pela Universidade de Muenster (LODUM¹⁰); dos registros de autoridades, instituições e assuntos da biblioteca

² Semantic Web for Research Communities: ontoware.org/swrc/

³ Bibliographic Ontology: bibliontology.com

⁴ Citation Typing Ontology: imageweb.zoo.ox.ac.uk/pub/2009/citobase/Citation_typing_using_CiTO.doc

⁵ Scholarly Works Appl. Profile: www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile

⁶ Semantic Web Conference: data.semanticweb.org/ns/swc/swc_2009-05-09.html

⁷ Linked Science Core <http://linkedscience.org/lsc/ns/>

⁸ Semantic Web Dog Food: <http://data.semanticweb.org/>

⁹ DBLP: <http://dblp.uni-trier.de/db/>. - Facet DBLP: <http://dblp.l3s.de/d2r/>

¹⁰ Linked Open Data Universit of Muenster: <http://data.uni-muenster.de/>

alemã (Hannemann, 2010); da produção acadêmica da Universidade de Economia de Praga (Hladka et al., 2012); e o projeto VIAF¹¹, coordenado pelas bibliotecas nacionais dos Estados Unidos, França e Alemanha, com a adesão de várias outras bibliotecas, que experimentam publicar seus nomes de autoridades. Sob a denominação Linked Science, é investigada a publicação e a interconexão semântica de dados bibliográficos e de ativos de ciência (trabalhos, processos, modelos, dados, métodos e métricas de avaliação), proporcionando maior transparência dos resultados obtidos (Kauppinen, Baglatzo, Kessler, 2012).

3 O AMBIENTE FABRICO/CIÊNCIA

Fabrico Ciência¹² é um ambiente que visa explorar as fronteiras entre Arquivos Abertos, Web 2.0, Web Semântica e Linked Data, no campo da Ciência. Da Web 2.0, adota o modelo de software (O'Reilly, 2007), e mecanismos de colaboração e de exploração de redes (*folksonomias*). A produção e a representação da informação ocorrem de acordo com a arquitetura da Web Semântica (triplas RDF); e as informações podem ser obtidas/ligadas tanto de repositórios Linked-Data, como colhidas de Arquivos Abertos via OAI-PMH³.

Considerando as funcionalidades para ferramentas de mapeamento da ciência levantadas por Cobo et al. (2011), Fabrico Ciência possui as seguintes características. **Busca:** permite a representação de qualquer ontologia desenvolvida para a Web Semântica (em OWL¹³); a produção e edição de triplas RDF de acordo com ontologias; a junção de ontologias e seus conjuntos de dados (como ontologias/dados de publicações e entidades); a importação/ligação de Linked Data; e a colheita de dados de arquivos abertos. **Pré-processamento:** permite a identificação de dados errados e repetidos através da agregação de valores de propriedades (exibidos em nuvens), de relatórios sobre valores de propriedades (totais, médias) e de mecanismo de busca; oferece recursos colaborativos da Web 2.0 para editar valores de propriedades, que registram as alterações e os usuários que as realizaram, mostram quais usuários alteraram um determinado valor, quais os valores foram alterados por um usuário, quais as últimas alterações, e exploram uma rede de co-colaboração em torno de usuários que alteram mesmos valores, agrupando-os via medidas de similaridades com base nas alterações. Para incentivar a discussão e registrar o conhecimento adquirido, dispõe de

¹¹ Virtual International Authority File- <http://www.oclc.org/research/activities/viaf/>

¹² Fabrico/Ciência: <http://www.ufrgs.br/fabrico/ciencia.php>

¹³ OWL: Web Ontology Language. <http://www.w3.org/TR/owl-features/>

fóruns de discussão e textos wiki, junto a qualquer propriedade. **Extração de Redes e Normalização:** permite a exploração de redes de co-ocorrência de valores para qualquer propriedade da ontologia, medidas de similaridade entre valores de uma propriedade tendo como base a co-ocorrência, e a co-ocorrência entre valores de duas propriedades distintas (ex.; assunto e autor). **Mapeamento, Análise e Visualização:** a ferramenta está focada em oferecer recursos para exploração dos dados nos moldes da Web 2.0 (*serendipity*). Através dela, autores, assuntos e redes de co-autoria e co-termos são apresentadas de forma semelhante à *folksonomias* (assim como qualquer outra propriedade da ontologia e suas co-relações). Por exemplo, os assuntos são apresentados de forma agregada através de uma nuvem, em que o tamanho das letras representa a sua frequência. Cada assunto dessa nuvem é ligado a uma página em que são apresentados, na forma de nuvem (onde o tamanho da letra representa a frequência), os assuntos a ele relacionados (co-termos), medidas de similaridade entre o assunto e seus assuntos relacionados, e os autores de itens de esse assunto. Cada autor dessa nuvem está ligado a uma página, que apresenta (em nuvem de frequência) co-autores, similaridade com co-autores e assuntos relacionados. Por estar voltada para a exploração de dados no cenário Web 2.0, métodos complexos de mapeamento e análise não são contemplados, pois tais tipos de necessidades podem ser atendidas por ferramentas específicas mediante exportação/importação dos dados. Fabrico Ciência aplica e adapta o ambiente Fabrico (Rocha, 2010), que é um anotador semântico de recursos da web, para Web2.0.

4 CONSIDERAÇÕES FINAIS

Através do Fabrico Ciência, são exploradas as interfaces entre Web 2.0, Web Semântica, Linked Data e Arquivos Abertos como ferramental para o Mapeamento da Ciência, em especial no que diz respeito à disponibilização de informação (AA e Linked Data), à integração /interoperabilidade de informação (Linked Data e Web Semântica), à preparação de dados (Web 2.0) e à exploração de dados (Web 2.0). No momento, o Fabrico Ciência é experimentado através da colheita de metadados de revistas eletrônicas de acesso livre (via OAI-PMH), para prover um ambiente Web 2.0 que permita análises e descobertas dos rumos das pesquisas (e das revistas), através exploração de artigos, autores, assuntos e de redes de co-autoria e co-termos. Disso, temos com primeiro resultado, o uso da ferramenta para analisar a qualidade dos metadados fornecidos (Betancourt e Rocha, 2012). Parte-se agora para explorar a integração da informação sob a ótica da Web Semântica/Linked Data,

envolvendo uso de ontologias e coleções Linked Data existentes sobre dados bibliográficos e autoridades.

REFERÊNCIAS

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. **Scientific American Magazine**, May 2001.

BETANCOURT, S. ROCHA, R. Metadados de qualidade e visibilidade na comunicação científica. In Simpósio Brasileiro de Comunicação Científica, 3, 2012, Florianópolis. 2012.

COBO, M.J. et ali. Science Mapping Software Tools: Review, Analysis, and Cooperative Study Among Tools. **Journal of the American Society for Information Science and Technology**, v. 62, n.7, 2011

HANNEMANN, J.; KETT, J. Linked Data for Libraries. In World Library and Information Congress, 76, 2010, Gothenburg. **Proceedings...** IFLA, August 2010,

HLADKA, J. MYNARZ, J.; SKLENAK, V. Experience with transformation of bibliographic data into linked data. **Journal of Systems Integration**, v.3, n.1, 2012

KAUPPINEN, T.; BAGLATZI, A. KESSLER, C. Linked Science: Interconnecting Scientific Assets. In CRITCHLOW. T.; Van DAM, K **Data Intensive Science**. CRC Press, USA, 2012.

KELLER, M. et ali. Linked Data for Libraries, Museums, and Archives: Survey and Workshop Report. **CLIR Reports** n. 152, October 2011.

MATHES, A. Folksonomies - cooperative Classification and Communication through shared metadata. Computer Mediated Communication – LIS590CMC, Urbana : Univ. of Illinois, 2004. Disponível em <<http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>>.

O'REILLY, T. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. **International Journal of Digital Economics**, No. 65, March 2007.

PETERS, I. **Folksonomies: Indexing and Retrieval in the Web 2.0**. Berlin:Gruyter, 2009.

ROCHA, R . Desenvolvimento de Ontologias apoiado pela anotação semântica de textos. In. Seminário de Pesquisa em Ontologias no Brasil,3, Florianópolis, 2010. **Anais...** 2010.

SMALL, H. Update on science mapping: Creating large document paces. **Scientometrics**, v.38, n.2, 1997