Fabrico/Ciência: um Ambiente *Linked Data* para o Mapeamento da Ciência

Rafael Port da Rocha

RESUMO

Este trabalho investiga contribuições de Arquivos Abertos, Web 2.0 e *Linked Data* para o mapeamento da Ciência. Apresenta o ambiente Fabrico/Ciência, que combina características de Arquivos Abertos, Web 2.0 e *Linked Data*. Descreve e analisa as características de Fabrico/Ciência em função dos passos que envolvem o processo de analise do mapeamento da ciência e das vantagens obtidas pela exploração de recursos de Arquivos Abertos, Web 2.0 e *Linked Data*. Conclui que a ferramenta apresenta um ambiente para pesquisas no mapeamento da Ciência, em especial no que diz respeito à colheita, representação, integração e interligação de metadados; produção e preparação coletiva de dados e exploração de redes e similaridades. Identifica como pesquisa o desenvolvimento de uma ontologia, para representar, no ambiente, recursos bibliográficos, de citação e de redes.

PALAVRAS-CHAVE: Mapeamento da Ciência. *Linked Data*. Web Semântica. Arquivos Abertos. Web 2.0.

1 Introdução

Na *web*, os Arquivos Abertos e a Iniciativa dos Arquivos Abertos visam, respectivamente, tornar textos de resultados de pesquisas visíveis e usáveis por usuários da Internet, e promover padrões de interoperabilidade que visam a difusão dos Arquivos Abertos. Estas ações garantem maior visibilidade aos trabalhos de pesquisa e viabilizam a disseminação de conteúdos que até então eram de circulação restrita, como teses, dissertações, 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 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. Pode proporcionar a socialização e o compartilhamento de esforços e produtos, e provocar um alargamento das fronteiras destes estudos, estimulando a adesão de outros colaboradores, como pesquisadores interessados em identificar os rumos, as relações, e os canais ligados à 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; LASSILA, 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 *Resource Description Framework* (RDF)[1].

Tendo a Web Semântica como plataforma, surge, sob a denominação Linked Data, uma iniciativa para publicar conjuntos de dados já existentes na forma de triplas RDF. *Linked Data* é o termo usado para descrever recomendações de melhores práticas para expor, compartilhar e conectar pedaços de dados e conhecimento

[1] Especificada em <http://www.w3.org/RDF/>

na Web Semântica, usando *Uniform Resource Identifier* (URI)[2] e 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* configuram uma plataforma promissora para publicação de dados de estudos bibliométricos, à medida dispõem de uma estrutura simples para representar e integrar 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 al, 2011).

Este artigo investiga benefícios ao mapeamento da Ciência trazidos pela exploração de Arquivos Abertos, Web 2.0, Web Semântica e *Linked Data*. Sua estrutura reflete o método de investigação. Inicialmente define mapeamento da Ciência e caracteriza suas ferramentas de acordo com as principais etapas do processo. A seguir, investiga como ferramentas e recursos utilizados em Arquivos Abertos, Web 2.0, Web Semântica e *Linked Data* podem trazer benefícios para o mapeamento da Ciência. Então, apresenta e analisa os benefícios para o mapeamento da Ciência da ferramenta Fabrico/Ciência, que é um ambiente que investiga o uso de Arquivos Abertos, Web 2.0, Web Semântica e Linked Data como ferramental para o mapeamento da Ciência.

2 Ferramentas para o mapeamento da Ciência

Mapeamento da Ciência é um tópico de pesquisa da bibliometria que "busca encontrar representações de conexões intelectuais em um sistema dinâmico e mutante do conhecimento científico" (SMALL, 1997, p.275, tradução nossa). Cobo et al (2011) destacam como técnicas mais populares de análise bibliográfica: coautoria, cotermos, cocitação e acoplamento bibliográfico. A coautoria é uma técnica para analisar autores (e suas instituições e países) relacionados por produzirem obras em conjunto. Cotermos permite investigar conceitos associados aos campos de pesquisa, através da análise de relações estabelecidas pela coocorrência nos documentos de palavras-chaves ou de termos extraídos de título, resumo ou do próprio documento. Cocitação é a técnica que investiga autores, documentos e revistas relacionados por citarem os mesmos itens, e acoplamento bibliográfico compreende no agrupamento de documentos por citarem o(s) mesmo(s) item(s). Para a análise do mapeamento da Ciência,

[2] Identificador usado para as páginas e recursos da web

Cobo et al (2011) apresentam a seguinte sequência de atividades, que são incorporadas em ferramentas de mapeamento da Ciência:

a) **busca de dados**: busca de dados em fontes bibliográficas;
b) **pré-processamento dos dados**: implica em gerenciar erros ortográficos e dados faltantes, reduzir de dados, fatiar dados no tempo e pré-processar redes para reduzir nós;
c) **normalização**: extraídas as redes de relacionamento entre as unidades de análise selecionadas, medidas de similaridades entre os dados são derivadas, através do uso de diversas medidas de similaridade, formando agrupamentos;
d) **mapeamento**: 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.;
e) **métodos de análise**: técnicas como análise de redes (como análise temporal) são usadas para extrair conhecimento a partir da rede;
f) **visualização**: técnicas de visualização de redes são empregadas para possibilitar uma melhor compreensão dos resultados;
g) **interpretação**: em que o analista interpreta os resultados e os mapas utilizando-se de suas experiências e conhecimentos.

tenda-se imagem antiga como aquela produzida anteriorm

3 Arquivos Abertos e o mapeamento da Ciência

Os Arquivos Abertos, aliados aos padrões de interoperabilidade de informação promovidos pela Iniciativa dos Arquivos Abertos e a *softwares* livres desenvolvidos segundo estes padrões, proporcionaram uma ampliação dos horizontes da comunicação científica, oportunizando a publicação de baixo custo e o amplo acesso a conteúdos que anteriormente eram de circulação limitada, como revistas editadas por unidades acadêmicas, teses e dissertações, artigos não avaliados por pares (*pre-prints*) e relatórios de pesquisa.

Para promover a integração entre Arquivos Abertos, a Iniciativa dos Arquivos Abertos desenvolveu o padrão *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH)[3]. Este padrão estabeleceu uma arquitetura de interoperabilidade composta por provedores de dados, que são os repositórios de arquivos abertos, 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 que permite que provedores de serviços colham descrições (metadados) de itens armazenados em provedores

[3] Especificado em: <http://www.openarchives.org/pmh/>

de dados. Sob a arquitetura estabelecida por OAI-PMH surge um sistema global de repositórios distribuídos e interoperáveis. Este sistema global proporciona novo modelo desagregado para publicação acadêmica, em que provedores de serviços integram repositórios nacionais ou temáticos, como os portais brasileiro Biblioteca Digital Brasileira de Teses e Dissertações[4] (BDTD), britânico *British Library EthOS* (EThOS)[5] e europeu *Europe E-theses Portal* (DART)[6] de teses e dissertações, o diretório de revistas eletrônicas *Directory of Open Access Journals* (DOAJ)[7], e o portal europeu de repositórios acadêmicos *Digital Repository Infrastructure Version for European Research* (DRIVER)[8].

O modelo global proporcionado pela Iniciativa dos Arquivos Abertos traz contribuições ao mapeamento da Ciência, em especial no que diz respeito às atividades de busca e pré-processamento de dados. Entretanto, a efetividade dessas contribuições implica na observância de critérios como quantidade, diversidade e expressividade dos metadados, assim como padronização e facilidades para obtenção dos metadados dados, e credibilidade das informações.

Os Arquivos Abertos dispõem de informações bibliográficas em grande quantidade e diversidade (1812 provedores de dados registrados na Iniciativa dos Arquivos Abertos)[9], incluindo documentos que passam por avaliações (como artigos, teses, dissertações). Também há preocupação em disponibilizar informações sobre os repositórios, a fim de orientar seus usuários sobre seus conteúdos e políticas. Nesse sentido, a especificação OAI-PMH inclui comandos para identificação dos repositórios e das coleções destes; e a Parceria SHERPA desenvolve um diretório de repositórios que disponibiliza informações sobre a qualidade das bases de dados nele incluídas (OpenDOAR)[11] e uma base de dados sobre políticas de direito autoral e de arquivamento, e de editores de repositórios (ROMEO)[12].

A especificação OAI-PMH dispõe sobre uma arquitetura padronizada para a representação e colheita dos metadados, fato que traz facilidades para a obtenção e manipulação desses dados. Essa especificação estipula que repositórios devem obrigatoriamente disponibilizar seus matadados, para colheita, no padrão *Dublin Core* Simplificado. *Dublin Core* Simplificado[13] representa metadados que permitem a investigação de autoria, coautoria, cotermos, e dos tipos dos documentos. Entretanto, não contempla a representação de metadados para análise de citação e não envolve a descrição de características de cada autor (como instituição, formação, país), embora o elemento colaborador (dc:*contributor*) permita a representação de instituições que colaboraram com o trabalho, como executoras ou patrocinadoras, por exemplo. *Dublin Core* Simplificado também não estipula

[4] Acesso: <http://bdtd.ibict.br/>

[5] Acesso: <http://ethos.bl.uk>

[6] Acesso: <http://www.dart-europe.eu/basic-search.php>

[7] Acesso: <http://www.doaj.org/>

[8] Acesso: <http://www.driver-repository.eu/>

[9] OAI *Register*. Acesso: <http://www.openarchives.org/Register/BrowseSites>

[10] Parceria SHERPA. Acesso: <http://www.sherpa.ac.uk/>

[11] *Directory of Open Access Repositories*. Acesso: <http://www.opendoar.org/>

[12] SHERPA RoMEO. Acesso: <http://www.sherpa.ac.uk/romeo>

[13] *Dublin Core Metadata Element* Set, Version 1.1. Especificado em: <http://dublincore.org/documents/dces/>

regras para codificação de valores, como nomes de autores e tipos dos documentos.

Entretanto, além de *Dublin Core* Simplificado, muitas revistas eletrônicas e repositórios de teses e dissertações dispõem seus metadados através de esquemas mais expressivos. Para a descrição das teses e dissertações, o modelo internacionalmente recomendado, e normalmente utilizado, é o perfil de aplicação de *Dublin Core* para teses e dissertações *Interoperability Metadata Standard for Electronic Theses and Dissertations* (ETD-DS)[14], que qualifica e estende *Dublin Core* com informações sobre orientador, banca, programa, grau do título concedido, entre outras. As revistas eletrônicas que utilizam do *software* livre *Open Journal System*[15]/ Serviço Eletrônico de Editoração de Revistas (SEER)[16], além do *Dublin Core* Simplificado, disponibilizam metadados no formato *National Library of Medicine* / NISO - *Journal Article Tag Suite* (NLM-JATS)[17], que permite a descrição de autores discriminados pela instituição.

O padrão OAI-PMH define uma arquitetura de interoperabilidade que proporciona o desenvolvimento de provedores de serviços qualificados tendo como fonte os dados colhidos dos arquivos abertos. Entretanto, os provedores de serviços atuais limitam-se basicamente a prover a busca em múltiplos repositórios. Ferramentas para o mapeamento da ciência poderiam ser desenvolvidas na forma de 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 coautoria e cotermos). A disponibilização desses serviços permitiria que pesquisadores, interessados em identificar rumos, contextos e conexões de suas pesquisas, pudessem também explorar mapas da Ciência.

4 Mapeamento da Ciência e a Web 2.0

A Web 2.0 ou Social surge como uma nova versão da web em que seus usuários mudam de meros consumidores de informações e serviços para parceiros na sua construção. A web passa a ser a plataforma, assumindo serviços que anteriormente eram obtidos através da instalação de softwares nas estações de trabalho, e fornecendo novos serviços que exploram a sua estrutura em rede. Os sistemas desenvolvidos para essa plataforma possuem uma arquitetura, chamada de Arquitetura de Participação, pois são projetados para atender a contribuições de usuários (O'REILLY, 2007).

[14] Especificado em <http://www.ndltd.org/standards/metadata>

[15] Disponível em: <http://pkp.sfu.ca/?q=ojs>

[16] Disponível em: <http://seer.ibict.br/>

[17] Especificado em: <http://dtd.nlm.nih.gov>

> A estrutura tecnológica se expande de maneira conjunta com as interações sociais dos sujeitos que utilizam a Internet. Cada vez que uma pessoa cria um novo link, a rede se completa e, portanto, enriquece. A ideia de uma arquitetura de participação se baseia no princípio de que as novas tecnologias potencializam o intercâmbio e a colaboração ente os usuários (COBO ROMANI; PARDO KUKLINSKI, 2007, p.47, tradução nossa).

Na Web 2.0 a tecnologia tira proveito da Inteligência Coletiva, isto é de uma "inteligência distribuída por toda parte, incessantemente valorizada, coordenada em tempo real, que resulta em uma mobilização efetiva das competências (LÉVY, 1998, p.28)

> [...] se as tecnologias se orientam a serem mediadoras entre as inteligências dos indivíduos e da sociedade, estas poderiam realmente ver potencializadas suas capacidades criativas. Desta perspectiva, a sociedade pode ser entendida como um sistema que alcança um nível superior de inteligência coletiva que transcende, no tempo e espaço, as inteligências individuais que a constituem. (COBO ROMANI; PARDO KUKLINSKI, 2007, p.47, tradução nossa).

A Web 2.0 também está ligada ao fenômeno da cauda longa, em que a economia e a cultura passam de um modelo baseado em hits (produtos da tendência dominante) e avançam em direção a uma grande quantidade de nichos (ANDERSON, 2006). Isso é viabilizado devido à democratização das ferramentas de produção (facilidade em produzir livros, vídeos, ou periódicos eletrônicos) e de distribuição (rede viabiliza acesso aos nichos) e da ligação entre a oferta e a procura (negócios se deslocam para os nichos) (ANDERSON, 2006). Arquivos Abertos contribuem para o alargamento da cauda longa nas publicações científicas.

Focados na arquitetura de participação, em tirar proveito da inteligência coletiva, e em explorar também os nichos, surgiram na Web 2.0 ambientes de redes sociais, como Facebook, de escrita coletiva, como Wikipedia, e de catalogação social, como Filmow[18], que são espécies de redes sociais na qual seus membros desenvolvem coletivamente um catálogo sobre determinados tipos de recursos, como filmes, livros e músicas. Também surgiram ambientes que exploram a classificação social (folksonomias), isto é, que permitem que os usuários de um recurso possam classificá-lo de acordo com usas impressões, tarefa que até então era restrita aos autores dos recursos (que indicam palavras chaves) e aos indexadores das bases de dados que armazenam o recurso. Em folksonomias, termos atribuídos livremente por usuários são chamados etiquetas.

A Web 2.0 voltada para o desenvolvimento da ciência e o estudo dos seus reflexos está sendo tratada sob a etiqueta Science 2.0. Em Science 2.0 surgem ambientes que exploram wikis, catalogação social e folksonomias. *Open Wetware*[19] é um exemplo ambiente wiki para a documentação de projetos de pesquisa; PLOS ONE[20] é uma revista eletrônica (com revisão por pares) em que usuários comentam e ranqueiam artigos; e Bibsonomy[21],

[18] Filmow. Acesso em <http://filmow.com/>

[19] Acesso: <http://openwetware.org/>

[20] Acesso: <http://www.plosone.org/>

[21] Acesso: <http://www.bibsonomy.org/>

CiteULike[22] e Connotea[23] são ambientes catalogação social de artigos e outros documentos acadêmicos e científicos, e que também utilizam folksonomias .

Ambientes de folksonomias operam sobre um grafo tripartido formado por etiquetas, usuários (produtores de etiquetas) e recursos (documentos da web etiquetados). Estes ambientes permitem a análise de redes baseada em co-ocorrências, como redes de co-documentos, formadas por documentos indexados por mesmas etiquetas ou mesmos usuários, redes de cousuários, formadas por usuários que indexam via mesmas etiquetas ou por mesmos documentos, e coetiquetas, formadas por etiquetas atribuídas a mesmos documentos (PETERS, 2009). Muitos destes instrumentos são similares aos utilizados no mapeamento da ciência, como coautoria, cotermos, cocitação, mas, na Web 2.0, estes instrumentos manifestam-se com características próprias, como a exploração de redes de cotermos através da navegação por nuvens de termos relacionados, além de incluírem nessa exploração informações produzidas pelos usuários das informações científicas (como etiquetas e comentários).

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. Utilizam técnicas de análise de redes para investigar questões como a colaboração, assim como aplicam leis e princípios bibliométricos, que investigam questões como impacto, relevância, ineditismo, vida média, obsolescência, elitismo, etc. Já ferramentas da Web 2.0, como folksonomias, voltam-se à exploração social do espaço do problema, e à descoberta inesperada de coisas neste espaço, isto é, serendipidade (MATHES, 2004). Uma aproximação entre Web 2.0 e o mapeamento da ciência leva a inclusão e a análise de novos tipos de informações (como etiquetas e avaliações de usuários), a socialização da análise dos dados, e a inclusão de novos públicos, como pesquisadores que buscam rumos, relações e nichos de suas pesquisas. Nessa perspectiva a Web 2.0 contribui ao mapeamento da Ciência através de ambientes como Bibsonomy, *CiteULike* e Connotea.

A Web 2.0 também pode ser explorada pelo mapeamento da Ciência em ações coletivas de preparação e pré-processamento de dados, em que usuários criam catálogos coletivos e atuam na limpeza dos dados, identificando inconsistências, redundâncias, etc. Por exemplo, para auxiliar no processo de desambiguação de nomes no processo de identificação única de autores, o repositório de auto arquivamento arXiv[24] desenvolveu um método em que os autores indicam, nos seus perfis, seus identificadores no Facebook (WARNER, 2010). Dessa forma, a correlação entre identificadores de pessoas no arXiv e Facebook é estabelecida, e a

[22] Acesso: <http://www.citeulike.org/>

[23] Acesso: <http://www.connotea.org/>

[24] Acesso: <http://arxiv.org/>

lista de publicações do autor no arXiv é publicada em sua página no Facebook.

5 Mapeamento da Ciência e *Linked Data*

Assim como a web pode ser vista como uma grande rede distribuída de hiperdocumentos que são identificados e interligados através de suas URIs, *Linked Data* é vista como a web dos dados, em que dados estão armazenados em bases de dados que estão distribuídas através da mesma rede que suporta os hiperdocumentos, e são identificados e interligados via URIs. De forma semelhante aos hiperdocumentos, URIs são usadas para estabelecer a ligação entre dados, isto é, um determinado dado, quando referencia um outro dado, utiliza a URI deste como referência.

Hoje há produção e publicação crescente de dados bibliográficos no cenário *Linked Data*. Várias iniciativas estão em andamento como: a publicação de artigos, conferências, organizações e pessoas da área da Web Semântica (*Semantic Web Dog Food*)[25]; a publicação da base de dados bibliográfica da Ciência da Computação, *Computers Science Bibliography* DBPL (*Faceted* DBLP)[26]; dos ativos de Ciência produzidos pela Universidade de Muenster, *Linked Open Data Universit of Muenster* (LODUM)[27]; dos registros de autoridades, instituições e assuntos da biblioteca alemã (HANNEMANN; KETT, 2010); da produção acadêmica da Universidade de Economia de Praga (HLADKA et al, 2012); e o projeto *Virtual International Authority File* (VIAF)[28], coordenado pelas bibliotecas nacionais dos Estados Unidos, França e Alemanha, com a adesão de várias outra bibliotecas, que experimentam publicar seus nomes de autoridades. Uma grande expectativa ocorre no projeto *Open Citations*[29]. Suas pretensões são ambiciosas, como podemos observar em seu escopo e propósito, divulgados em sua página:

> O Projeto Open Citations é de âmbito mundial, pretende mudar a face da publicação e da comunicação científica. Especificamente, visa tornar possível a publicação de informação bibliográfica e de citação em RDF, e fazer com que links de citação sejam tão fáceis de percorrer como os links da Web. (UNIVERSITY OF OXFORD, 2012, tradução nossa).

A dimensão atual da base de dados de Open Citations já é considerável:

> Open Citations é um banco de dados de citações da literatura biomédica, colhidas a partir das listas de referência de todos os artigos de acesso aberto da PubMed Central que fazem referência a aproximadamente 20% dos artigos da PubMed Central (cerca de 3,4 milhões de artigos), incluindo todos os artigos altamente citados em cada campo da biomédica. Todos os dados estão disponíveis gratuitamente para download e reutilização. (UNIVERSITY OF OXFORD, 2012, tradução nossa).

[25] Acesso: <http://data.semanticweb.org/>

[26] Acesso: <http://dblp.l3s.de/d2r/>

[27] Acesso: <http://data.uni-muenster.de/>

[28] Acesso: <http://www.oclc.org/research/activities/viaf/>

[29] Acesso: <http://opencitations.net/>

A publicação de dados via *Linked Data* também ocorre em outras áreas. Sob a denominação *Linked Science*, são investigadas 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; BAGLATZI; KESSLER, 2011). *Linked Open Government Data* designa esforços em publicar como *Linked Data* dados abertos governamentais. Iniciativas pioneiras são encabeçadas pelos governos da Inglaterra e Estados Unidos (DING; PERSISTERAS; HAUSENBLAS, 2012). Essas fontes de dados abrem novos horizontes aos estudos de mapeamento da ciência, uma vez que dados tradicionais de mapeamento da ciência podem ser cruzados com outros tipos de dados, como ativos de pesquisa, financiamentos de instituições de fomento, avaliações sobre instituições de pesquisa e programas de pós-graduação, dados demográficos, etc.

Várias ontologias foram desenvolvidas para descrever recursos acadêmicos e de pesquisa, como *Semantic Web for Research Communities* (SWR)[30], utilizada para publicações da comunidade da Web Semântica (*Semantic Web Dog Food*); *Linked Science Core* (LSC)[31], ligada a *Linked Science*, em que pesquisas relacionam pesquisadores, temas, datas, locais, publicações; *Bibliographic Ontology* (BIBO)[32], que especifica tipos de recursos bibliográficos (como artigo, livro, apresentação, periódico, evento); *DataCite Ontology* (*DataCite*)[33], que especifica vários tipos de identificadores de recursos; *The Bibliographic Reference Ontology* (BiRO)[34], ontologia para referências bibliográficas, estruturada de acordo com o modelo funcional para registros bibliográficos desenvolvido pela Federação Internacional de Associações e Instituições Bibliotecárias (*Functional Requirements for Bibliographic Records*)[35]; e *Citation Typing Ontology* (CiTO)[36], que caracteriza a natureza das citações e é usada no projeto *Open Citations*.

Linked Data adota a Web Semântica como plataforma de representação de informação. Bases de dados neste ambiente são compostas por triplas representadas na linguagem RDF, e o significado destas triplas é formalmente especificado através de ontologias expressas na linguagem *Web Ontology Language* 2 (OWL)[37]. A plataforma da Web Semântica proporciona vantagens para a integração e interligação de dados. O uso de URI como identificador evita redundância uma vez que um dado externo pode ser referenciado, ao invés de ser localmente replicado. A integração da informação é facilitada, pois ocorre no nível semântico e não no nível estrutural, visto que RDF e OWL são instrumentos para adicionar semântica aos recursos, sem fazer suposição sobre a estrutura desses recursos.

[30] Especificada em: <http://ontoware.org/swrc/>

[31] Especificada em: <http://linkedscience.org/lsc/ns/>

[32] Especificada em: <bibliontology.com>

[33] Especificada em: <http://www.essepuntato.it/lode/http://purl.org/spar/datacite>

[34] Especificada em: <http://www.essepuntato.it/lode/http://purl.org/spar/biro>

[35] Especificada em: <http://archive.ifla.org/VII/s13/frbr/frbr.htm>

[36] Especificada em: <http://www.essepuntato.it/lode/http://purl.org/spar/cito>

[37] Especificada em: <http://www.w3.org/TR/2009/REC-owl2-syntax-20091027/>

Mesur (*Studying science from large-scale usage data*)[38] é um projeto de mapeamento da ciência que abrange uma amostra significativa de grandes editoras e instituições acadêmicas do mundo, integrando metadados de 50 milhões de documentos (RODRIGUEZ; BOLLEN;VAN DE SOMPEL, 2007), grande parte colhida via protocolo OAI-PMH. Este projeto adotou a arquitetura da Web Semântica, representado os dados obtidos em RDF e de acordo com uma ontologia que abrange bibliografia, citação e aspectos de uso da comunidade acadêmica. Este projeto comprova a eficiência da tecnologia da Web Semântica como solução para integrar uma quantidade gigantesca de metadados provenientes de várias fontes.

[38] Acesso: <http://mesur.informatics.indiana.edu/>

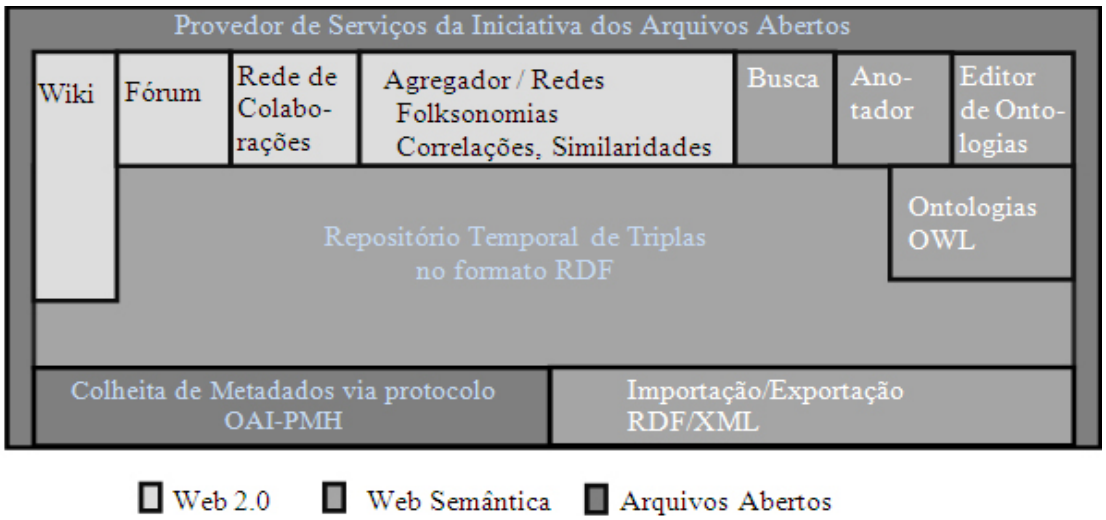
6 O Ambiente Fabrico/Ciência

Fabrico/Ciência[39] é um ambiente web para explorar as fronteiras entre Arquivos Abertos, Web 2.0, Web Semântica e *Linked Data*, no campo da Ciência. Da Web 2.0, adota padrão de projeto de *software* de O'Reilly (2007) e ferramentas de escrita colaborativa (wiki), classificação social (folksonomias), comunicação (fórum) e relacionamento (rede social cujo foco é o que está sendo construído coletivamente). A arquitetura do ambiente é apresentada na figura 1, com seus componentes discriminados quanto suas origens: Arquivos Abertos, *Linked Data*/Web Semântica e Web 2.0.

A produção e a representação da informação ocorrem segundo a arquitetura da Web Semântica, isto é, os metadados são representados através de triplas RDF e de acordo com Ontologias (figura 1). A fim de permitir exploração de dados que mudam com o tempo e registrar seus autores, a estrutura de uma tripla RDF é estendida com informações sobre período de validade da informação e seus produtores. O ambiente também permite: a representação e a visualização de múltiplas ontologias, o desenvolvimento de ontologias, a produção manual de metadados e a busca. As ontologias e os metadados são produzidos, respectivamente, via anotador e editor de ontologias (figura 1). O anotador permite que usuários anotem recursos da web de acordo com restrições estabelecida por ontologias. O editor de ontologias é uma configuração do anotador, em que os objetos anotados são os artefatos da ontologia que está sendo desenvolvida e as restrições são impostas pela ontologia que especifica OWL.

[39] Acesso: <http://www.ufrgs.br/fabrico/ciencia>

Figura 1 – Arquitetura do Fabrico/Ciência



Fonte: Autor

Sob a ótica da Web 2.0, o anotador é um ambiente de catalogação coletiva, pois permite que usuários descrevam coletivamente recursos da web (de acordo com ontologias). O anotador mantém o registro histórico das atualizações e dos atualizadores, e viabilizando uma rede social de colaboração entre os usuários (figura 1), cujos objetos de ligação são as anotações feitas. Também possibilita que usuários anotem as ontologias armazenadas, estabelecendo correlações entre ontologias. Discussões envolvendo cada artefato desenvolvido pelo editor de ontologias ou pelo anotador podem ser exploradas através da criação de fóruns (figura 1) ou páginas wiki (figura 1). Os fóruns também armazenam as mensagens como triplas RDF e a semântica das mensagens segue uma ontologia baseada no modelo de colaboração *Issue Based Information System* (IBIS)[40].

Em ambientes de folksonomias, usuários atribuem livremente etiquetas a recursos, e operações de agregação são usadas para explorar redes de co-ocorrências, como codocumentos, cousuários e coetiquetas, assim como ligações entre objetos dessas três redes (como usuários de etiqueta). No ambiente Fabrico, folksonomias estão associadas ao anotador, para explorar os valores anotados pelos usuários, incluindo co-ocorrência e similaridade. Fabrico/Ciência apresenta a frequência de ocorrência dos valores de uma propriedade através de listas ou través de nuvens em que os valores são ordenados alfabeticamente, sendo que o tamanho da letra de cada valor indica a sua frequência. Explora relações de co-ocorrencia e de similaridade, indicando valores relacionados por co-ocorrerem em um mesmo recurso ou valores relacionados por serem similares. A similaridade entre esses valores é calculada pelos coeficientes de similaridade de Jaccard[41], Dice[42] e Cosine[43], tendo como fonte a co-ocorrência. As operações disponíveis em Fabrico/Ciência para analisar propriedades e seus valores são

[40] Descrito em: GRANT, D. Issue-Based Information System (IBIS). In. OLSEN, S. **Group planning and problem-solving methods in engineering management**, New York: Wiley-Interscience, 1982. p. 203-246

[41] Fórmula disponível em: <http://en.wikipedia.org/wiki/Jaccard_index>

[42] Fórmula disponível em: <http://en.wikipedia.org/wiki/Dice%27s_coefficient>

[43] Fórmula disponível em: <http://en.wikipedia.org/wiki/Cosine_similarity>

apresentadas no quadro 1.

A navegação nas redes de valores relacionados por co-ocorrencia ou por medidas de similaridade é feita de forma semelhante a utilizadas em ambiente de folksonomias. Cada valor é representado através de uma página. Cada página de valor apresenta, na forma de lista ou nuvem de frequência, valores relacionados pela co-ocorrência ou por similaridade. A cada valor apresentado na lista e na nuvem está associado um hiperlink que leva à página que o descreve. 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 (ou lista), os assuntos a ele relacionados (cotermos), 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) coautores, similaridade com coautores e assuntos relacionados.

Sob o ponto de vista da arquitetura OAI-PMH, o ambiente é um provedor de serviços (figura 1), pois colhe de dados bibliográficos de provedores de dados (repositórios, bases de dados de teses e dissertações, revistas eletrônicas), oferecendo aos seus usuários serviços qualificados que envolvem ferramentas da Web 2.0, como folksonomias, fóruns e redes sociais, e da Web Semântica, como busca, anotação semântica, integração e ligação com dados de outras bases de dados do tipo *Linked Data*.

Quadro 1 – Análise de valores no Fabrico/Ciência – Módulo Agregador

Análise	Exibição	Exemplo
Frequência dos valores de uma propriedade	Nuvem e lista	Autores que mais publicam. Assuntos mais frequentes
Total de recursos descritos pela propriedade Totais de valores da propriedade Média de valores por recurso	Valores	Total de autores ou de recursos. Média de assuntos ou autores por recurso
Co-propriedades: Outras propriedades que co-ocorrem com a propriedade analisada	Nuvem e lista	Correlações entre propriedades
Co-ocorrência: Outros valores da propriedade que co-ocorrem com o valor analisado. Valores de outra propriedade que co-ocorrem com o valor analisado	Nuvem e lista	Cotermos. Coautores. Autores de um assunto. Assuntos de um autor.
Similaridade: Outros valores da propriedade que são similares ao valor analisado, tendo como base para o cálculo da similaridade as co-corrências de cada valor. Valores de propriedade de recursos que são similares pelo valor analisado, tendo como base para o cálculo da similaridade as co-corrências de cada valor	Nuvem e lista	Grau de similaridade entre assuntos pela co-ocorrência. Grau de similaridade entre autores pela coautoria. Autores que têm artigos de assuntos similares. Assuntos que possuem artigos de autores similares

Fonte: Autor

Fabrico/Ciência aplica e adapta o ambiente Fabrico (ROCHA,2010), que é um anotador semântico de recursos para Web2.0. Considerando as funcionalidades para ferramentas de mapeamento da ciência levantadas por Cobo et ali (2011), Fabrico/Ciência possui as características destacadas a seguir.

Busca de dados. Permite a representação de qualquer ontologia desenvolvida para a Web Semântica (em OWL); 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), da exploração de índices de similaridade, 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 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 fóruns de discussão e textos wiki, junto a qualquer propriedade.

Normalização. Permite a exploração de redes de co-ocorrência de valores para qualquer propriedade da ontologia, e de redes formadas por valores relacionados por medidas de similaridade tendo como base a co-ocorrência.

Mapeamento, Análise e Visualização. A ferramenta está focada em oferecer recursos para exploração dos dados nos moldes da Web 2.0 (*serendipty*). Através dela, autores, assuntos e redes de coautoria e cotermos são apresentadas de forma semelhante à folksonomias (assim como qualquer outra propriedade da ontologia e suas co-ocorrências e similaridades). Por estar voltada para a exploração de dados no cenário da Web 2.0, métodos complexos de mapeamento e análise não são contemplados, pois tais tipos de necessidades podem ser atendidos por ferramentas específicas mediante exportação/importação dos dados.

7 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 (Arquivos Abertos 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).

Conclui-se que o Fabrico/Ciência apresenta predicados para dar apoio a diversas pesquisas na área do mapeamento da ciência, com destaque para investigações que envolvem fontes de dados (colheita, representação/ontologias, agregação de dados de usuários, relacionamento com outros tipos de dados disponibilizados via *Linked Data*), pré-processamento (integração baseada em ontologias, interligação com *Linked Data*, preparação coletiva de dados) e normalização de dados (redes de similaridades).

Com relação à atividade de mapeamento, não dispões de métodos para construção de mapas e reduções de redes. Entretanto, a partir da análise realizada por este trabalho, uma nova atividade de pesquisa é desencadeada que envolve desenvolver uma ontologia para representar redes, que seja alinhada com ontologias recentes e populares para representar recursos bibliográficos (BIBO e BiRO) e citação (*DataCite* e CiTO), e observando os resultados obtidos pelo projeto Mesur no desenvolvimento de sua ontologia, que combina redes, citação e recursos bibliográficos. O desenvolvimento dessa ontologia viabilizará o novas pesquisas que contemplam atividades de mapeamento, visualização e métodos de análise.

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 coautoria e cotermos. Disso, temos como primeiro resultado, o uso da ferramenta para analisar a qualidade dos metadados colhidos da revista Em Questão, da Fabico/UFRGS (BETANCOURT ; ROCHA, 2012).

FABRICO/CIÊNCIA: a *Linked Data* environment for Science mapping

ABSTRACT

This work investigates contributions of Open Archives, Web 2.0 and Linked Data for Science mapping. It describes the Fabrico/Ciência, a web application which combines resources of Open Archives, Web 2.0 and Linked Data, and analyzes Fabrico/Ciência according to activities in the science mapping analysis workflow, and its benefits for science mapping of exploring features of Open Archives, Web 2.0 and Linked Data. It concludes that Fabrico/Ciência is appropriate for science mapping, especially with respect to the production and preparation of collective data, the operation of networks and similarity measures, and metadata harvesting, representation, integration and interconnection. It points as a next-step research, the development of an ontology for representing networks, bibliographic and citation resources in Fabrico/Ciência.

KEYWORDS: Science mapping. Linked Data. Semantic Web. Open Archives. Web 2.0.

Referências

ANDERSON, C. **A Cauda Longa**: a nova dinâmica de marketing e vendas. Rio de Janeiro: Elsevier, 2006

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. **Scientific American Magazine**, v. 284, n.5, p. 34-43. 2001.

BETANCOURT, S.; ROCHA, R. Metadados de qualidade e visibilidade na comunicação científica. In: SIMPÓSIO BRASILEIRO DE COMUNICAÇÃO CIENTÍFICA, 3, 2012, Florianópolis. [**Anais...**] Florianópolis, 2012.

COBO, M. J. et al. Science mapping software tools: review, analysis, and cooperative study among tools. **Journal of the American Society for Information Science and Technology**, New York, v. 62, n.7, p.1382-1402, 2011.

COBO ROMANÍ, C.; PARDO KUKLINSKI, H. **Planeta Web 2.0**: inteligencia colectiva o medios fast food. Barcelona, México:Grup de Recerca d'Interaccions Digitals, Universitat de Vic.Flacso. 2007.

DING, L.; PERISTERAS, V.; HAUSENBLAS, M. Linked Open Government Data. **IEEE Intelligent Systems**, Los Alamitos, Califórnia, v. 27, v. 3, p. 11-15, 2012.

HANNEMANN, J.; KETT, J. Linked Data for libraries. In: WORLD LIBRARY AND INFORMATION CONGRESS-MEETING 149, 76, 2010, Gothenburg. **Proceedings…**, 2010. Disponível em: <http://conference.ifla.org/past/ifla76/149-hannemann-en.pdf>. Acesso em : 6 nov. 2012.

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

KAUPPINEN, T.; BAGLATZI, A. KESSLER, C. **Linked Science**: interconnecting scientific assets. 2011. Disponível em: <http://linkedscience.org/wp-content/uploads/2012/02/linked-science-bookchapter-revised-2011-11-16.pdf>. Acesso em: 6 nov 2012.

KELLER, M. et al. **Linked data for libraries, museums, and archives: survey and workshop**. Washington, D.C: Commission on Preservation and Access, 2011. Disponível em: <http://www.clir.org/pubs/abstract/reports/pub152>. Acesso em: 6 nov. 2012.

LÉVY, P. **A Inteligência coletiva**: por uma antropologia do ciberespaço. São Paulo: Loyola, 1998.

MATHES, A. **Folksonomies**: cooperative classification and communication through shared metadata. Urbana : Univ. of Illinois, 2004. Disponível em: <academic/computer-mediated-communication/ folksonomies.html>. Acesso em 30 ago. 2012.

O'REILLY, T. What is Web 2.0: design patterns and business models for the next generation of software. **Journal of Digital Economics**, n. 65, p. 17-37, 2007. Disponível em: <http://mpra.ub.uni-muenchen.de/4578/>. Acesso em: 6 nov. 2012.

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...** Florianópolis: Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina. 2010.

RODRIGUEZ, M.; BOLLEN, J.; VAN DE SOMPEL, H. A Practical ontology for the large-scale modeling of scholarly artifacts and their usage. In JOINT CONFERENCE ON DIGITAL LIBRARIES, Vancouver, June 2007. **Proceedings…** New York: ACM, 2007. Disponível em: <http://arxiv.org/abs/0708.1150>. Acesso em: 6 nov 2012.

SMALL, H. Update on Science mapping: creating large document paces. **Scientometrics**, Oxford e Budapeste, v.38, n.2, 1997.

UNIVERSITY OF OXFORD. Image Bioinformatics Research Group. **About the JISC OpenCitations Project**. 2012. Disponível em: <http://opencitations.net/about/>. Acesso em 29 set. 2012

WARNER, S. Author identifiers in scholarly repositories. **Journal of Digital Information**, Austin, Texas, v. 11, n 1, 2010. Disponível em <http://arxiv.org/abs/1003.1345>. Acesso em: 6 nov 2012.

Rafael Port da Rocha
Doutor em Ciência da Computação pela Universidade Federal do Rio Grande do Sul (UFRGS).
Professor vinculado ao Departamento de Ciência da Informação da Universidade Federal do Rio Grande do Sul (UFRGS).
E-mail: r2ocha@yahoo.com.br

Recebido em: 30/09/2012

Aceito em: 09/11/2012