em questão
E-ISSN 1808-5245

O catálogo da biblioteca e o *linked data*

Liliana Giusti Serra
Doutoranda; Universidade Estadual Paulista Julio de Mesquita Filho, São Paulo, SP, Brasil;
lgiustiserra@gmail.com

José Eduardo Santarém Segundo
Doutor; Universidade de São Paulo, Ribeirão Preto, SP, Brasil;
santarem@usp.br

Resumo: O objetivo do artigo é discorrer sobre as possibilidades de aplicação de conceitos da *web* semântica, por meio do *linked data*, no catálogo da biblioteca, favorecendo a troca de informações de forma dinâmica, legível por pessoas e máquinas, com dados estruturados e conectados. O método empregado é de pesquisa exploratória, realizada por meio de levantamento bibliográfico, com seleção de textos que discorrem sobre os catálogos, partindo dos objetivos enunciados por Cutter até a disponibilização dos mesmos na *Web*; e identificação da *web* semântica e *linked data*. Os resultados tomam como base a apresentação do formato *Machine-Readable Cataloging* (MARC) como elemento que iniciou a descrição em formato legível por máquinas e permitiu o intercâmbio de registros bibliográficos, mostrando-se, contudo, incapaz de satisfazer as necessidades descritivas atuais. A *web* semântica é apresentada, partindo dos componentes de metadados exemplificados por ferramentas e modelo de dados até o *linked data*. O *Resource Description Framework* é apresentado como um modelo de dados que permite a construção da estrutura necessária para a ligação de dados na *Web*, proporcionando flexibilidade na escolha dos elementos descritivos que são utilizados com recursos bibliográficos. Foram localizadas iniciativas de utilização de *linked data* em bibliotecas, como a estruturação de vocabulários e nomes de pessoas, identificando resultados em seu emprego. O texto conclui com a análise de ganhos oriundos do uso de *linked data* nos catálogos das bibliotecas e a necessidade de publicação de dados abertos e interoperáveis pelas instituições, ampliando as possibilidades de novas ligações de dados.

Palavras-chave: Catálogo. *Linked data*. *Web* semântica. Metadados. Estrutura de dados.

1 Introdução

As bibliotecas são formadas por coleções de fontes de informações armazenadas em variados suportes e formatos, cuja representação e disponibilização ao público é realizada por meio de catálogo. De acordo com Bermès (2013), os

catálogos eram considerados ferramentas bibliográficas que permitiam a recuperação da informação, além de contribuir com a gestão da coleção. O catálogo contém dados e conteúdos e não somente a representação bibliográfica, oferecendo outras possibilidades de localização da informação e utilização de produtos e serviços.

A necessidade de troca de registros entre instituições contribuiu com o desenvolvimento de padrões, formatos e protocolos, agregando maturidade a bibliotecários em relação à utilização de regras que favoreçam a interoperabilidade. Porém, embora utilizando e intercambiando dados, os catálogos ainda são identificados como silos de informação, restritos ao conteúdo presente na coleção de uma instituição.

A *web* semântica se propõe a agregar elementos descritivos ao conteúdo na *Web*, tornando os dados interligados, aumentando sua apropriação e reutilização, por meio de aplicação do *Linked Data* (LD).

Este artigo visa identificar as possibilidades de aplicação do LD nos catálogos das bibliotecas, favorecendo a troca de dados entre os registros presentes numa mesma coleção, no catálogo de outras instituições ou ainda na *Web*, criando relacionamentos e possibilidades de explorar recursos e documentos. Assim, espera-se identificar formas de aplicação do LD aos catálogos que, além de favorecer usuários ao prover descrições detalhadas de recursos bibliográficos, permite o reuso de dados por meio de uso de ontologias e compartilhamento de elementos descritivos.

2 Procedimentos metodológicos

Este texto é resultado de pesquisa bibliográfica realizada em literatura recolhida sobre catálogos de bibliotecas, *web* semântica e LD. Foram consultadas fontes como livros, artigos de periódicos e trabalhos apresentados em eventos, abrangendo o período de 2001 a 2016. A partir de breve histórico da evolução dos catálogos e identificação dos elementos e conceitos da *web* semântica e sua aplicação com LD, serão discutidas as possibilidades de aplicação nas bibliotecas, refletindo os resultados nos catálogos.

O delineamento dos conceitos pesquisados na literatura foi realizado com identificação de pontos relacionados aos catálogos, como o uso do formato *Machine-Readable Cataloging* (MARC) na representação de recursos bibliográficos, e os recursos da *web* semântica e sua aplicação realizada pelo LD, identificando as ferramentas empregadas – vocabulários controlados, ontologias, descrição de autoridades etc., - e o modelo de dados do *Resource Description Framework* (RDF). A partir deste arcabouço teórico, relatos de iniciativas de emprego do LD nos catálogos de bibliotecas foram analisados, com o intuito de identificar os benefícios das aplicações, com a publicação e reutilização de dados interoperáveis.

3 O catálogo

Em 1876, na obra *Rules for a dictionary catalog*, Charles A. Cutter estabeleceu os objetivos do catálogo como:

a) permitir a um usuário encontrar um livro por meio de autor, título e/ou assunto;
b) mostrar os livros que a biblioteca possui;
c) orientar o usuário na escolha do livro, embasado pela sua edição ou tipologia (MEY; SILVEIRA, 2009).

Observa-se desde o estabelecimento do conceito do catálogo que este é uma ferramenta para identificação de registros presentes em um acervo, por meio de descrição de elementos essenciais que permitam a identificação e localização da informação. O catálogo não se refere apenas à descrição das obras presentes em um acervo, mas à sua utilização pelo usuário, a partir do momento em que os dados podem ser consultados e diretamente acessados. Também não é mais limitado ao conjunto de registros existentes em um acervo, podendo ter seu alcance difundido para outras coleções ou ainda, na *Web*.

A representação não consiste somente no rigor técnico adotado na descrição, mas na forma como os dados são expostos aos usuários, contribuindo ou não com sua identificação e, consequente, uso. Portanto, a descrição não é limitada a elementos sintáticos – característica que marcou seus primórdios -,

mas também a presença de elementos semânticos, ampliando as opções de serviços que podem ser oferecidos aos usuários, além de proporcionar amplitude de respostas recebidas nas solicitações de pesquisa.

De acordo com Weibel (2008), a sintática estrutura e organiza os metadados de forma que eles sejam confiáveis e legíveis por máquinas. A semântica, por outro lado, provê significado aos dados, permitindo que as informações sejam identificadas e interpretadas. Na descrição bibliográfica, a sintática compreende a forma como os dados são distribuídos no registro, para que sejam reconhecidos e intercambiados. Tomando o formato MARC como exemplo, a sintática sinaliza onde inicia e termina um conjunto de dados, com elementos únicos ou repetitivos (campos/tags ou subcampos) e indicadores que norteiam o comportamento do campo, estruturando os registros para que sejam legíveis por máquinas. A semântica contextualiza sobre o que o dado discorre, identificando os valores dentro dos campos. Assim, a tag 100 indica que o dado armazenado contém a descrição da Entrada principal do recurso bibliográfico, enquanto a tag 650 armazena os dados dos assuntos atribuídos a este recurso.

Na época do catálogo impresso, as obras presentes no acervo eram descritas no formato de fichas em papel, medindo 7,5 x 12,5 cm, armazenadas em gavetas e agrupadas basicamente por ordem alfabética, pelo nome de autores, títulos e assuntos. A catalogação, por meio de regras internacionais de representação, orientou a adoção de padrões para favorecer o intercâmbio de registros entre bibliotecas, independentemente do idioma utilizado ou país da agência catalogadora. Desta forma, uma publicação que já havia sido catalogada pela Instituição A não precisaria ser descrita novamente pela Instituição B, visto que esta poderia aproveitar os dados representados pela Instituição A, incluindo-o em seu catálogo. Este movimento era decorrente de um livro ser lançado em tiragem de exemplares, podendo ser incorporado ao acervo de diversas bibliotecas. A partir do momento que uma obra já foi catalogada por uma instituição, seus dados podem ser reaproveitados para que outras bibliotecas, que também possuam um exemplar desta publicação, possam aproveitar os dados que já foram criados, agilizando a organização e tratamento da

informação e, consequentemente, a disponibilização do material para utilização do público.

Este processo permitiu à área da biblioteconomia o alcance de maturidade na representação de informação, uma vez que um item poderia estar presente em diversas instituições, estimulando a troca de registros. Esta situação não é observada em arquivos e museus, decorrente do fato dos itens presentes nestas instituições serem, na maioria das vezes, registros únicos, sem a existência de exemplares, não justificando a estipulação de um padrão descritivo visando intercâmbio. Como a importação de registros visa agilizar a inclusão e disponibilização de um item no catálogo, aproveitando registros que foram criados por outras instituições, e como as coleções de arquivos e museus usualmente não possuem o mesmo item que possa ser importado, a troca de registros não se faz necessária com a mesma intensidade que em bibliotecas.

Este fato acarreta em dificuldades de compartilhamento de dados entre bibliotecas, museus e arquivos, visto que cada segmento adota padrão descritivo especifico às suas necessidades, resultando em interoperabilidade de poucos elementos, usualmente pelo formato *Dublin core*, que, com seus poucos elementos descritivos, possui características flexíveis que suportam metadados essenciais e comuns às três áreas.

Os padrões para descrição bibliográfica foram fixados em regras como o *Anglo American Cataloguing Rules*, segunda edição (AACR2) e o *International Standard Bibliographic Description* (ISBD), porém o intercâmbio dos dados foi favorecido com estipulação do formatos e protocolos.

Com as fichas em papel, o intercâmbio era proporcionado por meio de catálogos impressos que eram distribuídos entre as bibliotecas, representando os registros existentes nos acervos. Este produto facilitava a identificação de elementos descritivos como ponto de acesso, dados de imprenta, assuntos etc., dispensando análise detalhada da publicação, uma vez que os elementos essenciais para sua descrição já haviam sido registrados de forma padronizada. Às instituições era facultado incluir as fichas catalográficas recebidas de outras bibliotecas em seus catálogos, eliminando a etapa de criação da ficha manuscrita ou datilografada. Esta situação, entretanto, não se mostrava totalmente

satisfatória, visto que uma obra é descrita pela instituição de acordo com seu escopo e características do público usuário, com necessidade de ajustes das fichas recebidas, acarretando trabalho adicional nos registros importados. Assim, a descrição feita pela Instituição A pode não atender plenamente a política descritiva adotada pela Instituição B, sem contar o emprego de elementos auxiliares distintos na elaboração do registro como vocabulários controlados, catálogo de autoridades, tabela de classificação etc.

A *Library of Congress* desenvolveu o formato MARC na década de 1960. Destaca-se que ele não foi desenvolvido com o propósito primeiro de intercâmbio de registros, mas para facilitar a rotina bibliotecária de desdobramento de fichas catalográficas, onde o bibliotecário criaria a ficha matriz e as demais fichas (autores, assuntos, título, série, classificação etc.) seriam geradas de forma automática (TENNANT, 2002). O formato MARC foi pensado para atender às necessidades descritivas de catálogos em papel, transferindo aos computadores as informações contidas nas fichas analógicas, sem, contudo, explorar recursos tecnológicos. Esta situação não é coerente com o cenário atual.

De acordo com Taylor (1999[1] apud THOMALE, 2010), um dos objetivos do formato MARC consistia em armazenar os metadados para que fossem compartilhados entre bibliotecas e permitir a visualização dos registros, baseados em regras originais que existiam antes da tecnologia de computadores. Ao receber um registro bibliográfico em formato que pudesse ser alterado, com inclusão, edição e exclusão de *tags* pela biblioteca diretamente no computador, o intercâmbio de registro encontrou cenário favorável para importação massiva, alterando metadados quando necessário, uma vez que ao dado importado era possível realizar ajustes ou aplicar instrumentos adotados pela instituição.

O MARC é um esquema de metadados robusto que tem auxiliado as bibliotecas há mais de quarenta anos, permitindo descrição detalhada e armazenamento de dados bibliográficos, de autoridades etc., onde os registros podem ser criados, compartilhados e validados (MITCHELL, 2013).

> Catálogo é um meio de comunicação, que veicula mensagens sobre os registros do conhecimento, de um ou vários acervos, reais ou

> ciberespaciais, apresentando-as com sintaxe e semântica próprias e reunindo os registros do conhecimento por semelhanças, para os usuários desses acervos. O catálogo explicita, por meio de mensagens, os atributos das entidades e os relacionamentos entre elas. (MEY; SILVEIRA, 2009, p. 12).

Com a disponibilização do *Online Public Access Catalogue* (OPAC) na *Web*, o alcance dos catálogos foi ampliado, permitindo tanto a importação de registros diretamente dos sítios das bibliotecas, como a exploração de recursos que automatizam os processos de busca, recuperação e reutilização de metadados. Para Bermès (2013), este é o cenário observado na ocasião da publicação de seu livro, com registros descritos de forma padronizada e estruturada e que podem ser acessados à distância tanto por usuários como por agentes computacionais.

Apesar de protocolos como o z39.50 e o *Open Archives Initiative - Protocol for Metadata Harvesting* (OAI-PMH) permitirem o intercâmbio de registros, os catálogos das bibliotecas são ilhas de informação, sem interligação de dados, exigindo a elaboração de pesquisas nos sítios das instituições ou em projetos de bibliotecas digitais, sem compartilhamento de elementos descritivos. A implementação de LD como aplicação prática da *web* semântica possui elementos que podem alterar esta situação.

4 *Web* semântica e *linked data*

Os metadados permitem a criação, organização, descrição, identificação e acesso aos recursos de informação. Eles são estabelecidos por padrões e boas práticas de forma a permitir qualidade, consistência e interoperabilidade. Segundo Gilliland (2008), os metadados são compostos por quatro componentes: conteúdo de dados, estrutura, semântica e formato de intercâmbio.

O conteúdo de dados é descrito por meio de regras orientando o formato e a sintaxe dos metadados que preencherão os elementos descritivos e é representado pelas normativas de catalogação como o AACR2, ISBD e *Resource Description and Access* (RDA, Recurso, Descrição e Acesso, tradução nossa) etc.

A estrutura é composta pelos esquemas de representação com suas regras, indicando ordem dos elementos, campos obrigatórios, campos repetitivos, hierarquia etc. Os campos do MARC, *Dublin core*, *Encoded Archival Description* (EAD), *Web Ontology Language* (OWL), *Simple Knowledge Organization System* (SKOS) etc. são exemplos de estruturas.

A semântica é identificada pelos instrumentos que são utilizados para preencher com nomes e termos os elementos de conteúdo. São representados pelos vocabulários controlados, tesauros, ontologias, autoridades etc., que usualmente são utilizados para preenchimento de campos (*tags*) que acolhem dados sobre a temática do recurso bibliográfico.

O formato de intercâmbio é caracterizado pela sintaxe, com padrões de metadados legíveis por máquinas, muitas vezes sendo um padrão da estrutura de dados. Neste conjunto estão o MARC XML, *Metadata Encoding & Transmission Standard* (METS), *Metadata Objetc Description Schema* (MODS), *Dublin core* XML *Schema*, *JavaScript Object Notation* (JSON) etc.

De acordo com Mitchell (2013), os esquemas de metadados são definidos em cinco blocos:

a) modelo de dados;
b) regras de conteúdo;
c) esquemas e vocabulários;
d) serialização;
e) intercâmbio de dados.

O modelo de dados é a forma onde o relacionamento entre recursos e seus metadados são documentados. Os códigos de catalogação e demais normativas para descrição de dados constituem as regras de conteúdo, orientando como um dado deve ser criado e representado. Vocabulários e demais esquemas de representação de informação permitem a descrição e armazenamento dos metadados, favorecendo a reutilização dos elementos. A serialização de dados está centrada nos formatos e padrões adotados, orientando a forma como os dados serão gravados e armazenados em sistemas. A estrutura utilizada para descrição permitirá que os dados sejam intercambiados, com o compartilhamento de registros.

Observa-se que as visões de Gilliland (2008) e Mitchell (2013) são semelhantes, apesar de utilizarem com algumas diferenças nas descrições. Comparativamente, Mitchell (2013) reúne estrutura e semântica como um único bloco de metadados, enquanto Gilliland (2008) aborda estes elementos em separado. Também se destaca que Mitchell (2013) agrega em seu estudo o modelo de dados, componente inexistente em Gilliland (2008).

Para Gilliland (2008), os metadados somente permitirão aos usuários a descoberta e o acesso aos objetos de informação se:

a) conferirem autenticidade e completude do conteúdo;
b) definirem e documentarem o contexto do conteúdo;
c) permitirem identificação e exploração das relações existentes entre os objetos de informação;
d) proporcionarem pontos de acesso para comunidade de usuários variada e ampla;
e) oferecerem algumas informações que poderiam ser obtidas com bibliotecários no processo de referência tradicional.

> O papel da estrutura vem crescendo de acordo com as capacidades de processamento por máquinas, tornando-se cada vez mais poderosas e sofisticadas. Comunidades de informação estão cientes que quanto mais estruturado for o objeto de informação, mais a estrutura poderá ser explorada para pesquisa, manipulação e ser inter-relacionada com outros objetos de informação. (GILLILAND, 2008, p. 6, tradução nossa).

O conceito da *web* semântica propõe a criação de estrutura de conteúdo para sítios da *Web*, criando ecossistemas que permitam a agentes computacionais e pessoas a realização de tarefas. Desta forma, os sítios deixam de ser páginas isoladas e fora de contexto na *Web*, proporcionando interligação entre eles. Não se trata de uma *Web* separada da atual, mas uma extensão desta, agregando contexto de forma a ser legível por pessoas e máquinas (BERNERS-LEE; HENDLER; LASSILA, 2001).

> A *web* semântica é um conjunto de tecnologias que está sendo desenvolvidas pela W3C a fim de conseguir uma visão: a visão da *Web* de dados, uma *Web* onde os dados estruturados que estão atualmente isolados em bases de dados, possam ser expressados de

> uma maneira que tornaria possível sua interpretação e processamento por máquinas em novas aplicações e serviços. Para fazer isso, os dados precisam ser compartilhados em uma rede global e interligados por identificadores únicos e confiáveis. (BERMÈS, 2013, p. 118, tradução nossa).

A *World Wide Web Consortium* (W3C) é uma comunidade internacional que desenvolve padrões abertos para garantir o crescimento da *Web* em longo prazo. Liderada por Tim Bernes-Lee e Jeffrey Jaffe, estabelece princípios sobre a construção de padrões que garantam o acesso da *Web* para todas as pessoas, independentemente de *hardware*, *software*, infraestrutura, idioma, cultura, localização geográfica ou condições físicas e mentais (WORLD WIDE WEB CONSORTIUM, 2016).

A *web* semântica promove os padrões como o modelo de dados *Resource Description Framework* (RDF), ferramentas para descrição semântica e lógica dos dados como o RDF *Schema*, OWL e o SPARQL[2], linguagem para o protocolo RDF, que nortearão uma *Web* mais inteligente e eficiente (BERMÈS, 2013).

Uma informação contextualizada deve ser descrita por meio de modelo de dados, que é identificado como um conjunto de classes e propriedades. O RDF é um modelo de dados que se propõe a representar e transmitir metadados de forma que sejam legíveis por máquinas e favoreçam a integração entre sistemas de informação na *Web*. Neste modelo, os pedaços de informação são expressos em três elementos: sujeito, predicado e objeto, também identificado por triplas. Os sujeitos e predicados são representados por identificadores (*Uniform Resource Identifier* - URIs), garantindo que um recurso seja único. O predicado ou relacionamento também é identificado como propriedade e pode ter comportamentos diferentes, de acordo com a expressão criada. Assim, uma propriedade pode ser unidirecional ou reversa, relacionando um item a outro, permitindo a criação de inferências. Um objeto pode ser representado por uma URI ou ainda por uma tripla ou texto. A composição de diversas triplas forma um grafo (BERMÈS, 2013).

O *Linked Open Data* (LOD) é resultado dos estudos do grupo *Semantic Web Education and Outreach*, fundado em 2006 pela W3C, com o propósito de

estimular a aplicação dos conceitos de *web* semântica, tornando a *Web* global e interligada. O LD também é conhecido como a *Web* de dados e suas ações são centradas em criar boas práticas e guias, estimulando que organizações disponibilizem gratuitamente conteúdo em triplas RDF com *links* na *Web* com URIs.

Para Berners-Lee, Hendler e Lassila (2001), o *Linked Open Data* possui dois conceitos onde *linked* define a capacidade de um dado publicado na *Web* se conectar facilmente com informações relacionadas e, a partir destas ligações, este dado (*data*) poder ser acessado por computadores e pessoas. O segundo conceito estabelece que os dados que serão relacionados (*linkados*) e reutilizados devem ser abertos e livres de restrições de direitos autorais. O mesmo deve ocorrer com vocabulários, ontologias, esquemas etc., com as instituições disponibilizando instrumentos abertos, auxiliando na construção da *web* semântica (MITCHELL, 2013).

Em 2006, Berners-Lee publicou quatro princípios que norteiam o LD:

a) a utilização de URIs para nomear itens;
b) emprego de URIs HTTP para que os nomes possam ser consultados por pessoas;
c) ao consultar uma URI fornecer informação RDF útil;
d) incluir declaração RDF com *links* para outras URIs, aumentando as possibilidades de descobertas de itens relacionados (SANTAREM SEGUNDO, 2015).

Estes princípios são orientações na construção de estrutura para LD, porém não são obrigatórios. Não empregá-los não impede a realização de ligação entre os dados, porém oportunidades podem ser perdidas (BERMÈS, 2013). O RDF possui grande flexibilidade na escolha dos elementos descritivos que serão empregados para recursos bibliográficos. A escolha da URI como identificador é um elemento do RDF, buscando obter persistência dos recursos.

A aplicação de ontologias e vocabulários controlados agregam eficiência na recuperação dos recursos ao permitir que relacionamentos sejam criados com contexto, auxiliando na interpretação dos dados tanto por pessoas como por agentes computacionais. Ao incluir elementos com estrutura semântica de

conceitos e representações, a recuperação deixa de ser unicamente sintática, realizada a partir dos metadados descritivos, e passa a ser contextualizada, com relação organizada entre termos pertencentes a um domínio. A aplicação de vocabulários e ontologias reconhecidos por comunidades e aceitas internacionalmente favorece a publicação e recuperação de dados na *Web* em ambientes semânticos (SANTAREM SEGUNDO, 2015). Ambientes semânticos são identificados como locais onde os dados estão estruturados de forma que sejam processados por máquinas e que permitam a ligação dos elementos descritivos, favorecendo a geração, o reaproveitamento e a relação com outros grupos de dados. Ainda segundo o autor, conjuntos de dados publicados por uma instituição constituem um *dataset*[3]. Para Bermès (2013), uma das regras do LD é permitir a ligação de um *dataset* a outro, agregando recursos de fontes variadas. Observando esta questão, compreende-se que as aplicações de LD irão se expandir na medida que a quantidade de *datasets* disponibilizados e interligados forem ampliados, permitindo que dados de informações diversas possam ser interligados, contribuindo para a criação de rede contextualizada de informações. Entretanto, não é necessário que um *dataset* identifique a forma e o modelo em que um outro *dataset* foi estruturado.

> Isto significa que vários *datasets* são conectados por *links*, mas a forma e o modelo de um *dataset* local não precisa ter o conhecimento da forma e do modelo do outro. Isto é possível porque cada recurso representado por uma URI é independente e cada tripla é auto descritiva, independente das declarações feitas sobre um mesmo recurso, ou seja, um recurso pode ser tanto um objeto e um sujeito dentro de um mesmo contexto ou um grafo RDF. (BERMÈS, 2013, p. 126, tradução nossa).

5 *Linked data* e os catálogos

A criação e manutenção de metadados é custosa às instituições. Apesar de existirem regras que norteiam a descrição, a catalogação não é única, visto que deve ser realizada com aderência à política descritiva da instituição e ao público ao qual ela atende. Quanto mais especializado for o acervo e o usuário, maior detalhamento será demandado na descrição, tornando o processo moroso e caro. Catalogações exaustivas costumam despertar o sentimento identificado por

Bermès (2013) como "*database hugging*", com profissionais e instituições coibindo que metadados sejam copiados e distribuídos livremente. Este entendimento é o oposto do pregado pelo LD, onde o compartilhamento de dados é o objetivo principal, visando interligar os sítios e tornar as informações interoperáveis e reutilizáveis.

O formato MARC, apesar de ajudar no intercâmbio de registros há muitos anos, mostra-se na atualidade uma ferramenta limitada. A descrição de objetos de informação com a aplicação da norma RDA, por sua vez fruto dos conceitos do *Functional Requirements for Bibliographic Records* (FRBR), favorece o processamento dos dados por máquinas, em decorrência de seu modelo entidade-relacionamento. Em virtude da impossibilidade do MARC em acomodar a criação de relacionamentos entre registros, a *Library of Congress* anunciou a iniciativa de um novo padrão de estrutura dos dados, identificado como *Bibliographic Framework Transition Initiative* (BIBFRAME), lançado em 2011 (COYLE, 2012). Ao se identificar que o formato MARC, por ter sido desenvolvido em uma época antes da *Web*, não é mais capaz de representar toda a quantidade de informações de interesse das bibliotecas e facilitar a troca de dados entre instituições e comunidades relacionadas, faz-se necessário sua substituição (MARCUM, 2011). "O novo projeto de estrutura bibliográfica será focado no ambiente da *Web*, princípios e mecanismos do *Linked data* e no RDF como modelo de dados." (MARCUM, 2011).

Segundo Coyle (2012), a *Library of Congress* começou a estruturar sua listagem de cabeçalhos de assunto – *Library of Congress Subject Headings* (LCSH) – com padrões semânticos, permitindo compartilhamento da terminologia da instituição. Mitchell (2016) informa que a *National Library of Medicine* (NLM) também está desenvolvendo e testando aplicações em LD e no modelo RDF, incluindo o vocabulário *Medical Subject Headings* (MeSH).

Tanto o LCSH quanto o MeSH são iniciativas de controle e compartilhamento de terminologias. Em relação às autoridades, movimentos são observados com a disponibilização de nomes de pessoas no *Virtual International Authority File* (VIAF), que conta com a colaboração de bibliotecas nacionais de diversos países. Mitchell (2016) cita também as plataformas

ORCID, *ResearcherID*, *Scopus Author ID* etc., que, no âmbito acadêmico, estão publicando informações nos padrões do LD para contribuir com a desambiguação e disponibilização de dados de autores em grafos, além de permitir identificadores persistentes para este conjunto de pessoas.

ORCID é uma organização sem fins lucrativos que busca criar uma rede de pessoas que participam em pesquisas e inovação para que sejam identificados de forma única e conectados a suas afiliações. Atende uma comunidade formada por indivíduos, instituições de pesquisas e acadêmicas, repositórios de dados, bibliotecas, editores, escritórios de patentes, provedores de serviços etc. (ORCID, 2016). O *ResearchID* é uma solução para eliminar a ambiguidade entre autores na comunidade de pesquisa acadêmica. Cada membro recebe um identificador único, permitindo aos pesquisadores administrarem suas listas de publicações, citações e fator de impacto. Os dados são integrados com a *Web of Science* e compatível com o ORCID (RESEARCHERID, 2015). *Scopus* é o maior banco de dados de citações e resumos de literatura revisada por pares. A ferramenta possui um identificador único para autores, permitindo que identificações diferentes utilizadas por um autor ou alterações de afiliação sejam agrupadas, reunindo a produção. Possui integração com o ORCID.

Bermès (2013) pontua que o LD pode ser empregado por bibliotecas que possuem *datasets* relevantes e cujos dados podem ser reutilizados. Neste espectro, são destacadas instituições que possuem expressivos acervos com obras raras, coleções únicas, ou ainda apurado controle de autoridades, tesauros, vocabulários controlados e demais instrumentos desta natureza. Além destas instituições, as bibliotecas também podem se beneficiar do LD com a utilização de fontes heterogêneas distribuídos em *datasets* na *Web* como o *DBpedia*[4] ou *Geonames*[5], permitindo que dados de contextualização sejam oferecidos aos usuários diretamente nos catálogos. A interoperabilidade é funcionalidade básica nestes dois casos, afinal a biblioteca pode tanto dispor seus dados para serem utilizados na *Web*, como usufruir de *datasets* publicados.

Em uma pesquisa realizada em 2014, visando identificar casos de uso em LD por bibliotecas, observou-se que as principais motivações eram:

a) enriquecer a descrição do registro bibliográfico;

b) relacionar registros harmonizando diversas fontes, como se fosse uma fonte de referência;
c) automatizar o controle de autoridades;
d) enriquecer a aplicação utilizada pela instituição;
e) publicar dados de forma ampla, apresentando casos de uso potenciais (MITCHELL, 2016).

O sítio *Linked Data for Libraries* (LD4L) é um projeto colaborativo encabeçado pelas universidades *Cornell, Harvard* e *Stanford* que visa desenvolver estudos para facilitar a transição dos fluxos de trabalhos na produção de metadados à comunidade LD, contribuindo com estudos sobre o BIBFRAME, criação de ontologias etc. (LINKED DATA FOR LIBRARIES, 2016). As universidades *Columbia, Princeton* e a *Library of Congress* também participam do projeto.

6 Considerações finais

As bibliotecas têm como prática o desenvolvimento de trabalhos de padronização da representação descritiva, com esforços canalizados nos catálogos e na construção de listas de autoridades e terminologias. O uso de instrumentos como vocabulários controlados e tesauros é uma prática recorrente na Biblioteconomia. Agregar estrutura a estes instrumentos representa um avanço no tratamento descritivo, permitindo que dados sejam preenchidos com contexto e que sejam interoperáveis, favorecendo a reutilização de recursos.

A utilização de tecnologia de LD nos catálogos deve ser estimulada, principalmente em instituições que possuam acervos raros, notáveis ou acurado tratamento de autoridades, tanto de pessoas como assuntos. Nestas situações recomenda-se a disponibilização dos instrumentos e registros existentes, de forma a permitir que outras instituições possam vincular seus metadados a eles, aumentando as chances de novas ligações. Compartilhar registros, com possibilidade de interoperabilidade, contribui para que as bibliotecas enriqueçam

seus acervos, proporcionando informação rica e variada aos usuários, mas também favorece que novas relações entre registros sejam criadas.

Com o LD a cooperação entre bibliotecas deixa de ser a de expor dados disponíveis para importação, no estilo copia e cola como é feito hoje, mas de interligar dados de forma dinâmica, proporcionando a complementação e significado no resultado de busca. Isto se refletirá em catálogos enriquecidos, com informações adicionais, favorecendo a seleção da fonte desejada pelo usuário. Ao selecionar uma autoridade Pessoa em uma fonte aberta e com estrutura de dados interligados como o VIAF por exemplo, informações adicionais como datas, local de nascimento e/ou atividade, obras publicadas, temas de interesse, relacionamentos com outros autores etc., são incluídos ao registro de forma dinâmica, minimizando o tempo de descrição necessário, agregando contexto à autoridade.

Por mais que os OPACs estejam publicados e disponíveis para consulta na *Web*, não é recorrentemente identificada interligação entre os catálogos e demais sítios na *Web* ou ainda com outros catálogos, a não ser quando se trata de um *link* para acesso a um conteúdo descrito, como numa *tag* MARC 856|u, por exemplo. Esta *tag* é empregada para armazenar uma URL vinculada ao registro bibliográfico, na forma de um *link*. Assim, é possível representar que o conteúdo descrito ou uma fonte complementar podem ser consultados por meio da *Web*. Com o LD, um dado na *Web* pode remeter ao catálogo da Instituição A que, por sua vez, possui dados interligados com o catálogo da Instituição B, que utiliza a *DBPedia* para aumentar a contextualização de seus metadados e o ORCID na descrição de autoridades Pessoa. Assim, novas relações vão sendo construídas na medida em que são ampliadas as ligações entre termos e *datasets,* proporcionando dados agregados de contexto.

Ao utilizar instrumentos descritivos abertos para realizar a descrição de registros, o processo de catalogação é otimizado, com a aplicação de metadados padronizados para pessoas, instituições, locais, idiomas, assuntos etc., proporcionando agilidade na inclusão de novos registros no catálogo e, consequentemente, eficácia na disponibilização de recursos de informação. Isto

também torna o processo de catalogação menos oneroso, demandando menor quantidade de horas empenhadas nesta atividade.

Um ponto que pode ser desafiador aos bibliotecários será a identificação dos instrumentos que serão adotados para favorecer a interligação de dados. Devem ser selecionadas as ontologias adequadas, assim como os catálogos de autoridades, cuja seleção deve ser orientada pelo escopo da instituição e de seu público usuário.

Com a criação de vínculos entre registros outras possibilidades de visualização dos dados passam a ser possíveis, contribuindo para que inferências entre os registros sejam identificadas ou criadas. Assim, pode-se, por exemplo, identificar os participantes de um movimento artístico específico somente por meio dos vínculos existentes entre eles; ou ainda, contextualizar dados por meio de metadados cronológicos, geográficos etc., identificando pessoas ou temáticas em um determinado período ou espaço. Estas possibilidades mostram-se desafiadoras, tanto na construção de instrumentos para compartilhamento, como a elaboração de catálogos dinâmicos, cuja descoberta leva para além do acervo da instituição.

Referências

BERMÈS, E. Enabling your catalogue for the semantic web. In: CHAMBERS, Sally (Ed.). **Catalogue 2.0**: the future of library catalogue. Chicago: Neal-Schuman, 2013. p. 117-142.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. **The semantic web.** Scientific American**,** New York, p. 24-30, May 2001. Disponível em: <https://www.scientificamerican.com/article/the-semantic-web/>. Acesso em: 27 jul. 2016.

COYLE, K. **Linked data tools**: connecting on the web. Chicago: ALA TechSource, 2012. (Library Technology Reports, v. 48, n. 4).

DBPEDIA. **About DBPedia.** 2016. Disponível em: <http://wiki.dbpedia.org/about>. Acesso em: 2 nov. 2016.

GEONAMES. **About GeoNames**. Disponível em: <http://www.geonames.org/about.html>. Acesso em: 2 nov. 2016.

GILLILAND, A. J. Setting the stage. In: BACA, Murtha (Ed.). **Introduction to metadata**. 2. ed. Los Angeles: J. Paul Getty Trust, 2008. p. 1-19.

LINKED DATA FOR LIBRARIES. Gateway. 2016. Disponível em: <https://www.ld4l.org/>. Acesso em: 11 ago. 2016.

MARCUM, D. **A bibliographic framework for the digital age**. 2011. Disponível em: <https://www.loc.gov/bibframe/news/framework-103111.html>. Acesso em: 11 ago. 2016.

MEY, E. S. A.; SILVEIRA, N. C. **Catalogação no plural**. Brasília: Briquet de Lemos, 2009.

MITCHELL, E. T. **Library linked data**: research and adoption. Chicago: ALA TechSource, 2013. (Library Technology Reports, v. 49, n. 5).

MITCHELL, E. T. **Library linked data**: early activity and development. Chicago: ALA TechSource, 2016. (Library Technology Reports, v. 52, n. 1).

ORCID. **What is ORCID**. 2016. Disponível em: <http://orcid.org/about/what-is-orcid/>. Acesso em: 2 nov. 2016.

RESEARCHERID. **What is ResearcherID?** 2015. Disponível em: <http://www.researcherid.com>. Acesso em: 2 nov. 2016.

SANTAREM SEGUNDO, J. E. Web semântica, dados ligados e dados abertos: uma visão dos desafios do Brasil frente às iniciativas internacionais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16., 2015, João Pessoa. **Anais...** João Pessoa: UFPB, 2015. p. 219-239. Disponível em: <http://inseer.ibict.br/ancib/index.php/tpbci/article/view/207/272>. Acesso em: 31 jul. 2016.

TENNANT, R. MARC must die. **Library Journal**, New York, p. 26-27, 15 out. 2002. Disponível em: <http://soiscompsfall2007.pbworks.com/f/marc+must+die.pdf >. Acesso em: 31 ago. 2016.

THOMALE, J. Interpreting MARC: where's the bibliographic data? **Code4Lib**, [S.l.], n. 11, set. 2010. Disponível em: <http://journal.code4lib.org/articles/3832>. Acesso em: 31 ago. 2016.

WEIBEL, S. **Metadata:** semantics; structure; syntax. 2008. Disponível em: <http://weibel-lines.typepad.com/weibelines/2008/02/metadata-semant.html>. Acesso em: 31 out. 2016.

WIKIPEDIA. **Conjunto de dados.** 2016. Disponível em: <https://pt.wikipedia.org/wiki/Conjunto_de_dados>. Acesso em: 2 nov. 2016.

WIKIPEDIA. **SPARQL.** 2016. Disponível em: <https://en.wikipedia.org/wiki/SPARQL>. Acesso em: 24 out. 2016.

WORLD WIDE WEB CONSORTIUM. **W3C mission**. 2016. Disponível em: <https://www.w3.org/Consortium/mission>. Acesso em: 2 nov. 2016.

The library catalogue and the linked data

Abstract: The article aims to discuss the possibilities of applying semantic web concepts in the library catalogue, through the application of linked data, favoring the exchange of dynamic information, readable by people and machines. The method adopted is an exploratory research through literature, with a selection of texts that discuss the catalogues, semantic web and linked data. The results focus on RDF, presented as a data model that allows the construction of the structure required for Web data connection, providing flexibility in the choice of descriptive elements that are used with library resources. Initiatives of linked data use were identified in libraries, acknowledging results in their employment. The text concludes with the identification of gains from the use of linked data in library catalogues and the need for the data publication be open and interoperable among institutions, expanding the possibilities of new data connections.

Keywords: Catalogue. Linked data. Semantic web. Metadata. Data structure.

Recebido em: 19/08/2016

Aceito em: 21/11/2016

[1] TAYLOR, A. G. **The Organization of Information**. Englewood, CO: Libraries Unlimited, 1999.

[2] Linguagem de consulta semântica para banco de dados capaz de recuperar e manipular dados armazenados em formato RDF. É o padrão utilizado pelo RDF *Data Access Working Group* (DAWG) do Consórcio *World Wide Web* (W3C), reconhecida como uma das tecnologias chave da *web* semântica (WIKIPEDIA, 2016).

[3] *Datasets* ou "conjunto de dados" são coleções de dados onde, para cada elemento, são indicadas características (WIKIPEDIA, 2016).

[4] *DBPedia* é uma comunidade que permite extrair informação estruturada da *Wikipedia*. Esta iniciativa almeja facilitar a extração de dados presentes no *Wikipedia* para que seja utilizado e visualizado em outros pontos da *Web*, além de permitir navegação, ligação e melhorias na própria *Wikipedia* (DBPEDIA, 2016).

[5] *Geonames* é uma ontologia, disponibilizada de forma gratuita, que contém mais de dez milhões de nomes de lugares (GEONAMES, 2016).