

DOS SILOS DE DADOS À WEB DOS DADOS: BIBLIOTECAS E O LINKED DATA

FROM DATA SILOS TO WEB OF DATA: LIBRARIES AND LINKED DATA

Liliana Giusti Serra^a

José Eduardo Santarém Segundo^b

RESUMO

Objetivo: mostrar como os catálogos atualmente são opacos, sem elementos semânticos, e como podem ser enriquecidos com *datasets* presentes na *Web* para que passem a ser semânticos e proporcionem outras possibilidades de descoberta aos usuários. **Metodologia:** foi utilizada a técnica de análise conceitual, buscando constatar a partir de estudos o quanto o catálogo é de fato um silo de dados e como esta situação pode ser alterada por meio do *linked data*. **Resultados:** por meio de exemplos de registros bibliográficos foi possível incluir relacionamentos entre os dados presentes no catálogo, assim como de páginas da *Web*, proporcionando contexto aos usuários e enriquecendo as informações recuperadas em pesquisas realizadas no catálogo. **Conclusões:** os catálogos das bibliotecas podem ser fontes confiáveis de dados e contribuir com a *Web* dos dados, além de proporcionarem mais possibilidades de descobertas aos usuários.

Descritores: Catálogo. Web Semântica. Linked Data. Silos de Dados.

1 INTRODUÇÃO

As bibliotecas constroem seus catálogos reunindo dados de diversas naturezas, tanto bibliográficos, como de autoridades. Esta atividade tem como objetivo aumentar a qualidade dos dados e, conseqüentemente, as possibilidades de descobertas de recursos aos usuários, consolidando informações sobre autores e registros. Bibliotecas trabalham em seus catálogos com profissionais

^a Doutora em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita Filho (UNESP). Profissional da informação dos sistemas SophiA Biblioteca e SophiA Acervo. lgiustiserra@gmail.com

^b Doutor em Ciência da Informação pela Universidade Estadual Paulista (UNESP); Docente no Departamento de Educação, Informação e Comunicação da Universidade de São Paulo (USP); Docente do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista (UNESP). E-mail: santarem@usp.br.O

dedicados ao detalhamento de descrições e conjuntos de metadados. Quanto mais trabalhados os metadados, melhores as chances de recuperação e a credibilidade da agência catalogadora. Entretanto, segundo Bermès (2013), nos dias de hoje o foco não está mais nos catálogos, mas nos dados.

Antes da disponibilização de catálogos na *Web*, o acesso aos dados ficava muitas vezes restrito ao espaço físico da biblioteca ou dependente de convênios ou acordos de cooperação para troca ou aquisição de fichas catalográficas ou acesso a catálogos coletivos. Catálogos bem detalhados possuem alto custo, afinal o trabalho de profissionais experientes e capacitados para realizar a representação é oneroso. Enquanto grandes bibliotecas conseguem formar e manter um corpo técnico qualificado, esta não é a realidade de todas as instituições.

Cada biblioteca trabalha no seu catálogo, investindo na representação dos recursos presentes em seus acervos. A adoção de padrões internacionais para a representação proporciona a possibilidade de interoperabilidade de registros entre bibliotecas, em movimento cooperativo para intercâmbio de dados. Entretanto, recursos iguais podem ter usos diferentes de acordo com a comunidade atendida, o que acarreta diferentes representações de um mesmo recurso. Ao copiar os metadados de uma biblioteca quase sempre é necessário realizar ajustes para adequar a descrição ao público usuário.

Ao copiar ou produzir registros para os seus catálogos, a biblioteca está criando um silo de dados, onde os metadados não possuem conexão com os demais dados presentes no catálogo, com os catálogos de outras bibliotecas ou com a *Web*.

Analisando com mais apuro um catálogo, observamos que os registros possuem dados que podem fornecer mais alternativas de recuperação, entretanto essas possibilidades são reduzidas às autoridades que são compartilhadas entre os registros (autores, assuntos etc.), sem promover relacionamento entre os próprios itens presentes na coleção. Mesmo assim, ao compartilhar autoridades não é garantido que serão criadas relações entre os registros além das que foram estabelecidas sintaticamente pelas autoridades, exigindo que os vínculos entre os registros de uma mesma coleção sejam investigados pelos usuários, de acordo com seus conhecimentos prévios. Além disso, os registros que não compartilham autoridades não possuem vínculo com

os demais registros presentes em um catálogo, formando silos dentro do silo.

A *Web* semântica (WS) pode promover a alteração desta situação, transformando um dado opaco em um nó que pode ser conectado com outro conjunto de dados dentro de uma mesma coleção e desta para a *Web*, aumentando as possibilidades de descobertas, favorecendo a serendipidade.

O conceito da *Web* semântica foi apresentado em 2001 por Berners-Lee, Hendler e Lassila e discorre sobre a expansão da *Web* tradicional, sintática (dos documentos) à *Web* dos dados, proporcionando a troca e o reuso de dados, agregando compreensão para máquinas e seres humanos. A proposta não era moldar a *Web* em uma estrutura rígida ou exigir a adoção de padrões para construí-la, mas mantê-la aberta e descentralizada, porém agregando significado ao conteúdo, construída com boas práticas para o modelo e a publicação de dados, interligados e interoperáveis com ontologias (BERNERS-LEE; HENDLER; LASSILA, 2001).

O *linked data* (LD) é uma aplicação da WS e consiste em criar *links* entre dados provenientes de diferentes fontes na *Web*. Este vínculo pode ser construído a partir de dados publicados de forma que sejam legíveis por máquinas e que seus significados e relacionamentos sejam definidos explicitamente. Para Singer (2009) é fundamental que as bibliotecas façam parte da *Web* dos dados e o LD pode ser um caminho para alcançar este objetivo. De acordo com Byrne e Goddard (2010) o LD pode ser compreendido por dois aspectos: o uso do modelo *Resource Description Framework* (RDF) para publicação de dados na *Web*; e o uso de *Uniform Resource Identifiers* (URI) para explicitar a relação entre as entidades vinculadas. O LD pode promover essa transformação nos catálogos das bibliotecas, o que pode vir a alterar consideravelmente a forma como os recursos são descritos e o alcance da catalogação.

2 PROCEDIMENTOS METODOLÓGICOS

Este artigo é fruto de uma pesquisa bibliográfica realizada em textos obtidos em artigos de periódicos, livros, trabalhos apresentados em eventos e em documentos dispostos na *Web* que abrangem o período de 2001 a 2018. A técnica adotada nesta pesquisa foi de análise conceitual.

Análise conceitual é uma técnica que trata os conceitos como

classes de objetos, eventos, propriedade ou relacionamentos. A técnica envolve definir com precisão o significado de um determinado conceito identificando e especificando as condições sob as quais qualquer entidade ou fenômeno é (ou poderia ser) classificado sob um conceito em questão (FURNER, 2004, p. 233, destaques do autor, tradução nossa)¹.

Por meio da análise conceitual é possível levantar questionamentos sobre um tema e, a partir destes questionamentos, expandir o entendimento de um conceito. Esta técnica favorece a obtenção de compreensão de um determinado conceito por seus usuários na forma de seu emprego, assim como a crença de que a compreensão obtida por estes usuários venha a ser considerada um pré-requisito para propor o desenvolvimento de outros conhecimentos.

Foram analisadas as possibilidades de aplicação de conceitos da WS e do LD nos catálogos das bibliotecas, favorecendo o enriquecimento dos dados e proporcionando ganhos semânticos nas descrições existentes, primeiramente dentre os registros existentes dentro de um catálogo, depois deste com outros catálogos de outras bibliotecas e, posteriormente, com a *Web*.

3 SILOS DE DADOS

De acordo com Coyle (2010), o catálogo da biblioteca possui diversas funções, mas para o usuário a principal é ser o local de descoberta dos registros presentes em uma coleção. “O resultado definitivo da inclusão de HTTP URIs expandirá o impacto potencial que a coleção de uma biblioteca local pode ter no amplo mundo da informação” (SHIEH; REESE, 2015, p. 209, tradução nossa)².

Para Hyland (2010), dentre os benefícios do LD estão a descentralização e exposição de silos com dados legados, permitindo o desenvolvimento de outras aplicações, melhorias na descoberta de recursos e reutilização de dados. Esta estrutura altera a configuração de silos de informação que estão distribuídos na *Web* (DENAUX *et al*, 2017; WU *et al*, 2017; BERNÈS, 2013; ALLEMANG, 2010;

¹ *Conceptual analysis is a technique that treats concepts as classes of objects, events, properties, or relationships. The technique involves precisely defining the meaning of a given concept by identifying and specifying the conditions under which any entity or phenomenon is (or could be) classified under the concept in question (FURNER, 2004, p. 233).*

² *The ultimate result of the embedded http URIs will expand the potential impact that a local library collection may have on the wider information world (SHIEH; REESE, 2015, p. 209).*

HYLAND, 2010), permitindo a interligação de dados que, como consequência, apresentarão mais e novos resultados sobre um termo pesquisado.

O relatório final do W3C *Library Linked Data Incubator Group* (W3C, 2011) alcançou alguns entendimentos como:

- 1) O termo Biblioteca refere-se ao conjunto de instituições do âmbito cultural e histórico, incluindo bibliotecas, museus e arquivos, independentemente se suas coleções são pequenas ou grandes, se são formadas por recursos analógicos ou digitais, ou sob custódia de instituições públicas ou privadas;
- 2) Dados de biblioteca (*library data*) se referem a qualquer tipo de informação digital produzida e sob curadoria de bibliotecas. Foram consideradas três categorias de dados de biblioteca: conjunto de dados (*datasets*), conjuntos de valores (*element sets*) e vocabulário de classes e propriedades (*value vocabularies*).
- 3) São considerados dados em LD aqueles publicados de acordo com os princípios definidos para facilitar o vínculo entre *datasets*, ou seja, utilizando URIs como identificadores únicos de recursos, expressos sob o modelo RDF, especificando as relações existentes entre as coisas;
- 4) São considerados dados abertos aqueles que não possuem restrição legal para uso, reuso ou distribuição. O LD não precisa necessariamente ser aberto, porém ao usar dados desta natureza, potencializam-se as possibilidades de uso.

O relatório final constatou que os dados de bibliotecas não se apresentam integrados com os recursos da *Web*, conservando a estrutura de silos de dados. Muitos dos padrões adotados por bibliotecas (MARC 21, ISO 2709, protocolo z39.50 etc.) são utilizados somente por esta comunidade, além do fato que muitos dados de recursos bibliográficos serem representados em linguagem natural ou por meio de códigos, armazenados em subcampos ou posições no MARC 21.

Restringir o acesso a metadados de bibliotecas é um comportamento identificado por Bermès (2013) como “*database hugging*” (abraçar banco de dados, tradução nossa), que consiste em inibir a exposição pública de dados, com consumo restrito somente à instituição catalogadora. Ao optar por este caminho, as bibliotecas mantêm a postura de silos de dados, com dados

encerrados em uma única fonte, sem vínculo com *datasets* externos que podem contribuir para o entendimento de contextos e o enriquecimento de dados, restringindo as possibilidades de uso e descobertas.

Para Malmsten (2009) é preferível publicar dados a não os publicar caso não estejam perfeitos. Este é o entendimento que o “ótimo” é inimigo do “bom”. É preferível publicar um “bom” *dataset* ao invés de adiar sua publicação até que ele seja considerado “ótimo”, o que pode não ser alcançado nunca, até porque é relativo estabelecer o que é “bom” e o que é “ótimo”. É preferível ter alguma informação a não ter nenhuma. “A *Web* semântica é para derrubar silos, não para construir novos³” (BYRNE; GODDARD, 2010, não paginado, tradução nossa).

Ao fazer relações de itens e autoridades de seu próprio acervo, a biblioteca continua autocentrada, porém não é mais um silo de informações, pois seus dados estão descritos com a estrutura da *WS*. Além de vincular elementos do acervo, esta aplicação pode dar visibilidade à biblioteca, uma vez que os dados da coleção podem ser recuperados por outras fontes.

[...] isso parece um pouco de desserviço aos nossos usuários e somos incapazes de guiá-los facilmente entre vários formatos, edições e representações de uma determinada obra, sem sequer abordar a noção de “trabalhos relacionados” (por exemplo *West Side Story*). Isso é especialmente trágico, dada o vasto esforço e recursos necessários para criar registros e autoridades normalizados que o *software* não usa. A informação está lá, mas a tecnologia está nos enfraquecendo (SINGER, 2009, p. 117, tradução nossa)⁴.

O Projeto *InterParty* foi lançado em 2003 pela *European Commission* (EC) em parceria com a *EDItEUR*, *British Library*, *Royal Swedish Library*, *International Federation of Library Associations* (IFLA), *Library of Congress* (LOC), *Online Computer Library Center* (OCLC), *Digital Object Identifiers* (DOI) *Foundation*, entre outros, e tinha como objetivo a construção de uma rede para distribuição de metadados comuns aos envolvidos na indústria do livro, buscando otimizar o tratamento e descrição de dados

³ *The semantic web is about breaking down silos, not building better ones* (BYRNE; GODDARD, 2010, não paginado).

⁴ *And yet, it seems a bit of a disservice to our users that we are unable to easily guide them to the various formats, editions, and representations of a given work, without even addressing the notion of “related works” (e.g., West Side Story). This is especially tragic, given the vast amount of effort and resources that go into creating normalized records and authorities that the software does not use. The information is there, but the technology is failing us* (SINGER, 2009, p. 117).

de publicações impressas (MACEWAN, 2004). Neste projeto foi mencionado a perspectiva histórica de bibliotecas em criar silos independentes de dados.

Também podemos apreciar, de uma perspectiva de biblioteca, o caminho histórico que levou a "silos de dados" independentes. No passado, a necessidade de tal nível de interconectividade não era aparente nem fácil de alcançar. Hoje, o crescimento da *Web* destacou a necessidade de metadados que podem "viajar" através desses padrões e barreiras de sistemas (MACEWAN, 2004, p. 432, tradução nossa)⁵

Segundo DeWeese e Segal (2014), o LD provê benefícios às bibliotecas ao aumentar as possibilidades de interoperabilidade com o reuso de conjuntos de dados e vocabulários, a inclusão de URIs aos metadados e o desenvolvimento de serviços, como *Applicance Program Interface* (API) para publicação dos dados.

A estrutura utilizada para bibliotecas disponibilizarem seus dados em projetos de WS consiste nos padrões de triplas e modelo de dados RDF, com aplicação de vocabulários ou ontologias para proporcionar a compreensão semântica dos dados, identificando conceitos, objetos e suas relações de forma singular. É possível criar triplas entre os dados presentes no catálogo ou ainda desses com outras fontes, ampliando a disseminação dos dados no acervo com fontes internas e externas.

A Figura 1 ilustra triplas que podem ser criadas dentro do contexto de uma biblioteca. O autor utilizado foi o escritor Monteiro Lobato e a obra analisada é "Reinações de Narizinho", que foi publicada pela editora Brasiliense.

Figura 1 - Representação de recursos, propriedades e valores entre biblioteca e usuários



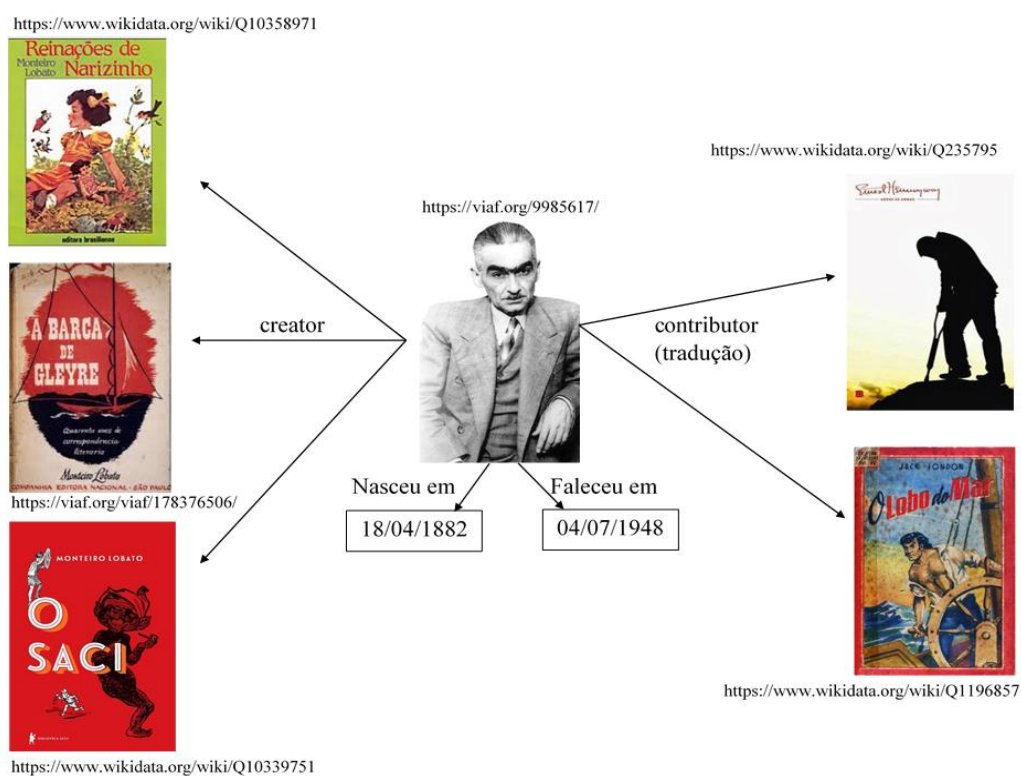
Fonte: Elaborado pelos autores a partir de W3C (2014)

⁵ *We can also appreciate from a library perspective the historical path which has led to independent "data silos." In the past the need for such a level of inter-connectivity was neither apparent nor easy to achieve. Today the growth of the Web has highlighted the need for metadata which can "travel" across these standards and systems barriers* (MACEWAN, 2004, p. 432)

Na Figura 1 as URIs foram retiradas do *Virtual International Authority File* (VIAF) para identificar o autor Monteiro Lobato, e da *Wikidata* para a obra “Reinações de Narizinho” e a editora Brasiliense.

O livro faz parte do acervo de uma biblioteca que, evidentemente, possui outras obras. Cada recurso possui sua representação, formando grafos relacionando dados. A seguir serão feitos alguns recortes no acervo, apresentando grafos simples de obras de quatro autores. O primeiro autor é Monteiro Lobato (Figura 2). Nele são representadas três obras em que Lobato é o autor (“Reinações de Narizinho”, “A barca de Gleyre” e “O Saci”), e duas onde ele é tradutor (*contributor*): “Adeus às armas”, de Ernest Hemingway, e “O lobo do mar”, de Jack London.

Figura 2 - Recorte do acervo – Monteiro Lobato



Fonte: elaborado pelos autores.

O segundo recorte ilustra as obras de Machado de Assis presentes no acervo desta biblioteca (Figura 3). O autor é o criador de todos os recursos ilustrados (“Dom Casmurro”, “Quincas Borba”, “O alienista” e “Memórias póstumas de Brás Cubas”).

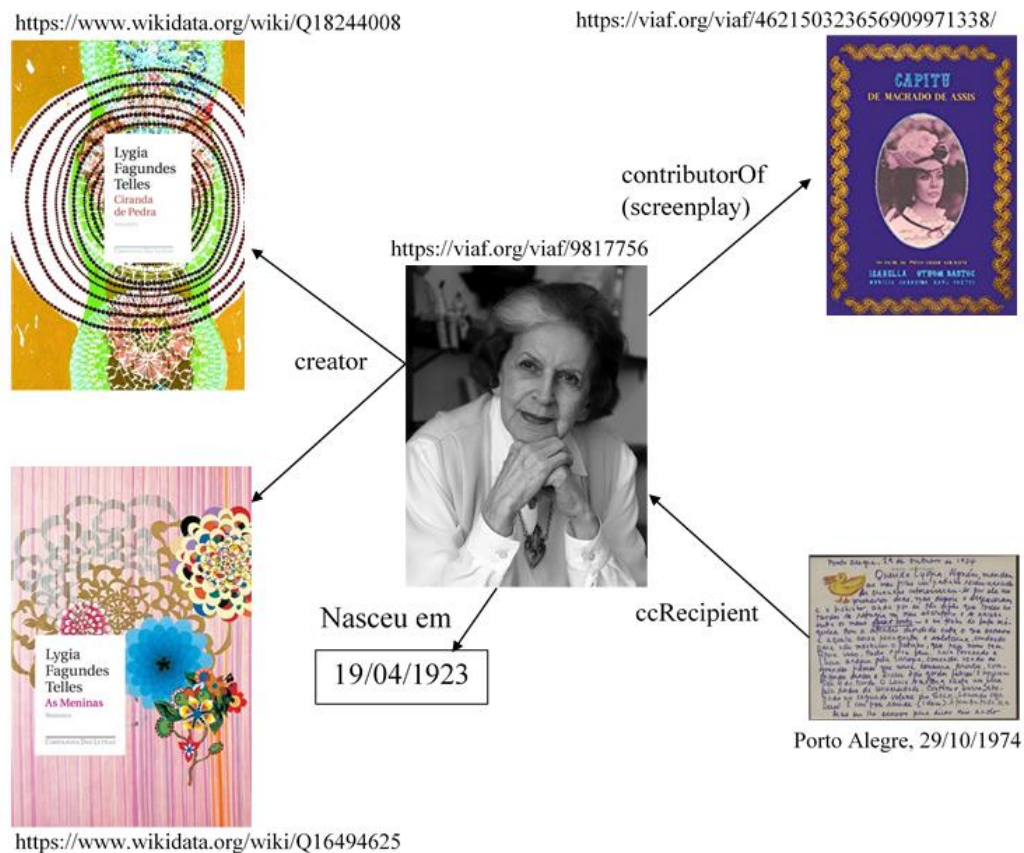
Figura 3 - Recorte do acervo – Machado de Assis



Fonte: elaborado pelos autores.

O terceiro grafo ilustra as obras de Lygia Fagundes Telles constantes nesta biblioteca (Figura 4). Aqui estão descritas duas obras da escritora (“Ciranda de pedra” e “As meninas”) e o filme “Capitu”, onde ela fez o roteiro. A biblioteca também conta com um manuscrito: uma carta endereçada à escritora.

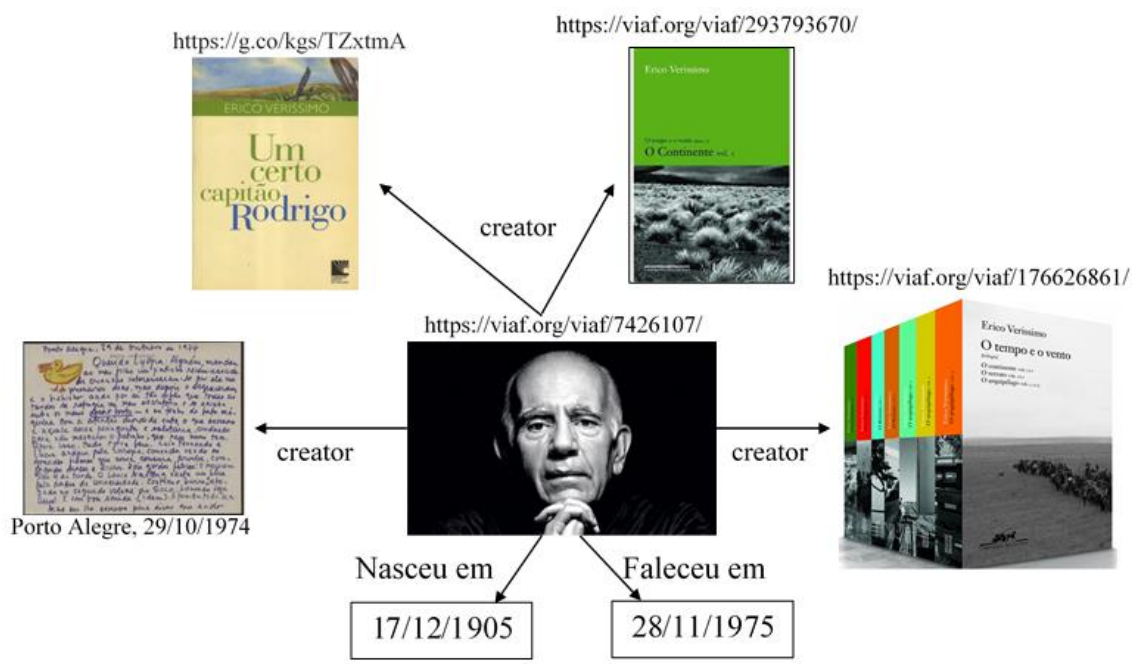
Figura 4 - Recorte do acervo – Lygia Fagundes Telles



Fonte: elaborado pelos autores.

O quarto grafo apresenta as publicações de Erico Veríssimo que fazem parte da coleção da biblioteca (Figura 5). Aqui estão representadas três obras de autoria do escritor (“Um certo capitão Rodrigo”, “O continente” e “O tempo e o vento”) e um manuscrito (carta).

Figura 5 - Recorte do acervo – Erico Veríssimo



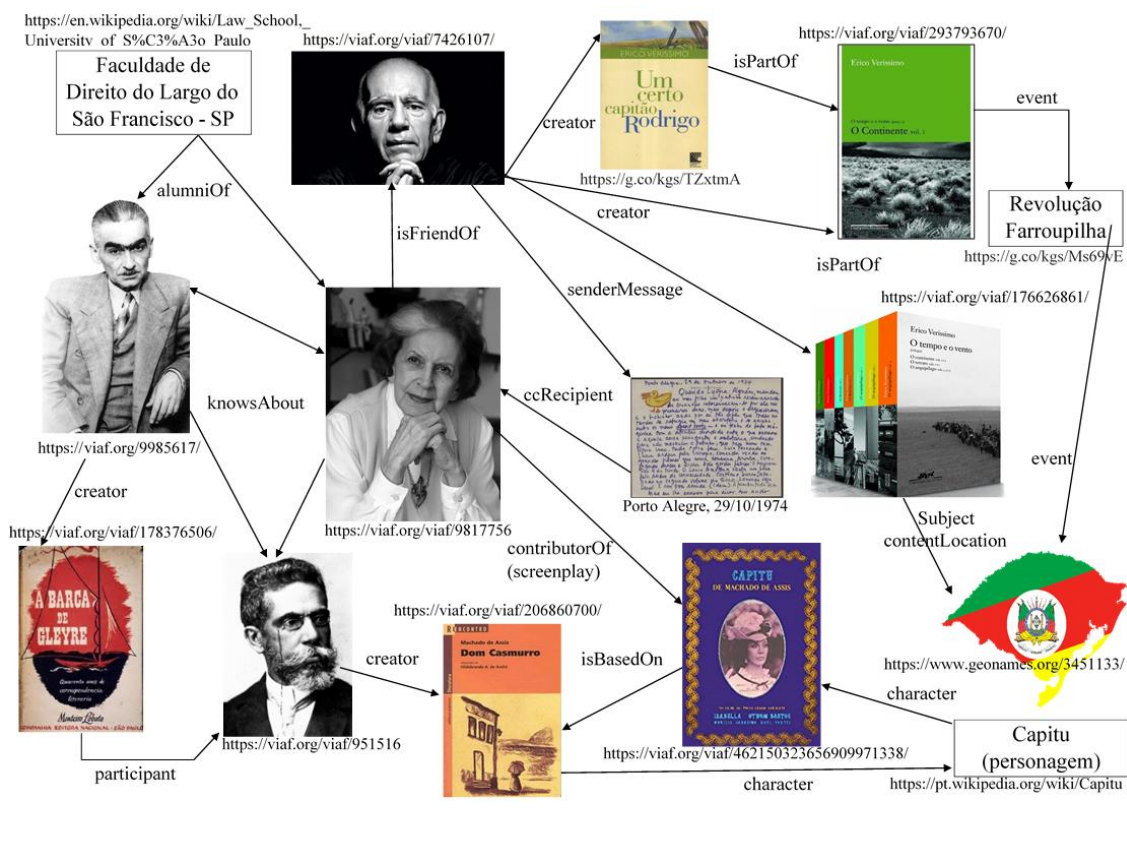
Fonte: elaborado pelos autores.

Nos exemplos das ilustrações acima, as relações existentes são apenas de criador e/ou contribuinte. Mesmo as obras de mesma autoria não possuem relações entre si. O usuário procura itens soltos no catálogo. É possível que ele encontre outras obras do mesmo autor ou com o mesmo assunto, mas os dados são estanques e opacos, sem vínculos semânticos, com as possibilidades de descoberta limitadas aos conhecimentos do usuário.

4 SILOS DO SILO

Ao aplicar a estrutura de dados da WS nos recursos que estão isolados no catálogo, podem ser criadas relações entre os dados, conforme a Figura 6..

Figura 6 - Grafo com relações de dados de registros presentes no catálogo



Fonte: elaborado pelos autores a partir da Figura 2, Figura 3, Figura 4 e Figura 5.

Observando a Figura 6 podem ser feitas inferências a partir de “Monteiro Lobato”, identificando que ele foi influenciado e escreveu o livro “A Barca de Gleyre” onde homenageia “Machado de Assis”, que por sua vez também influenciou a escritora “Lygia Fagundes Telles”. “Monteiro Lobato” e “Lygia Fagundes Telles” se conheceram e ambos estudaram na “Faculdade de Direito do Largo do São Francisco”, em épocas diferentes, em São Paulo. “Machado de Assis” escreveu “Dom Casmurro” onde apresenta a personagem “Capitu”. O filme “Capitu” teve roteiro adaptado por “Lygia Fagundes Telles”, que recebeu uma carta de seu amigo “Erico Veríssimo” em “29 de outubro de 1974”. “Erico Veríssimo” escreveu o livro “Um certo capitão Rodrigo”, que faz parte do livro “O Continente”, que tem como pano de fundo a “Revolução Farroupilha”. “O Continente” faz parte da trilogia “O Tempo e o Vento”, que é totalmente ambientada no “estado do Rio Grande do Sul”. Assim, é traçado um caminho que pode levar o usuário a saber que “Monteiro Lobato” conhecia “Lygia

Fagundes Telles”, que era amiga de “Erico Veríssimo”, que escreveu um livro que tem como pano de fundo a “Revolução Farroupilha” que faz parte da história do “estado do Rio Grande do Sul”, vinculando autoridades distintas e não relacionadas entre si, proporcionando outras descobertas.

Neste exemplo todos as inferências foram feitas nos dados existentes no próprio catálogo (os livros “A barca de Gleyre”, “Dom Casmurro”, “Um certo capitão Rodrigo”, “O Continente” e “O Tempo e o Vento”; o filme “Capitu”; e a carta enviada por Erico Verissimo para Lygia Fagundes Telles), englobando itens e autoridades que foram vinculados por relações estabelecidas por meio de vocabulários e identificadores retirados da *Wikipedia*, *VIAF* e *GeoNames*. Elementos externos foram incluídos no grafo, como por exemplo a “Faculdade de Direito do Largo São Francisco” e o “estado do Rio Grande do Sul”, proporcionando visão macro do universo analisado, além de poderem contribuir com a descrição temática das obras.

Na Figura 6 é possível observar que no catálogo tradicional não foi feita a relação entre as obras “Um certo capitão Rodrigo”, “O Continente” e “O Tempo e Vento”. Tampouco é dado destaque à personagem “Capitu”, de “Dom Casmurro”, cuja conduta até hoje faz os leitores questionarem se ela traiu ou não Bentinho e que já motivou a criação de outras obras, além de diversas críticas e estudos. Os registros estão descritos, porém são opacos, ou seja, não levam a outros registros e informações.

5 ENRIQUECIMENTO DE DADOS

O LD permite que dados do catálogo sejam enriquecidos por fontes externas. Este movimento proporciona mais informações aos usuários sobre as autoridades e registros bibliográficos, contribuindo para a descoberta de outros recursos e a realização de inferências. “A criação de vínculos pode proporcionar a capacidade de incorporação de informações biográficas, o que está (e de forma correta provavelmente) fora do escopo da catalogação” (SINGER, 2009, p. 120,

tradução nossa)⁶.

O enriquecimento de dados do catálogo pode ser feito por meio de *datasets*. Muitas fontes estão disponibilizadas na *Web* e não é necessário que estes dados estejam no modelo RDF, embora seja preferível adotar esta estrutura sempre que possível. Os vocabulários que não utilizam o modelo RDF não permitem expansão para outros *datasets*, porém proporcionam a obtenção de dados adicionais, mesmo que representados em literais e, portanto, opacos semanticamente. A Tabela 1 relaciona algumas fontes abertas que podem ser utilizadas para inclusão de identificadores (URIs) em dados de bibliotecas (PROGRAM..., 2018).

Tabela 1 - Algumas fontes de dados RDF e Não-RDF

Dados em RDF	Dados não-RDF
DBpedia	<i>AllMovie</i>
<i>Faceted Applications of Subject Terminology (FAST)</i>	<i>AllMusic</i>
<i>GeoNames</i>	<i>British Film Institute (BFI)</i>
AAT	<i>Canadian Geographical Names</i>
ISNI	<i>Discogs</i>
LCSH	<i>Geographic Names Information System (GNIS)</i>
MeSH	<i>Internet Broadway Database (IBDB)</i>
<i>Name Authority Cooperative Program of the PCC - NACO</i>	<i>Internet Movie Database (IMDb)</i>
<i>Open Researcher and Contributor (ORCID)</i>	<i>Legal Entity Identifier</i>
<i>RDA Vocabularies</i>	<i>Library of Congress Authorities</i>
VIAF	<i>ResearcherID</i>
<i>Wikidata</i>	<i>Scholar Universe</i>

Fonte: PROGRAM..., 2018, p. 1-2.

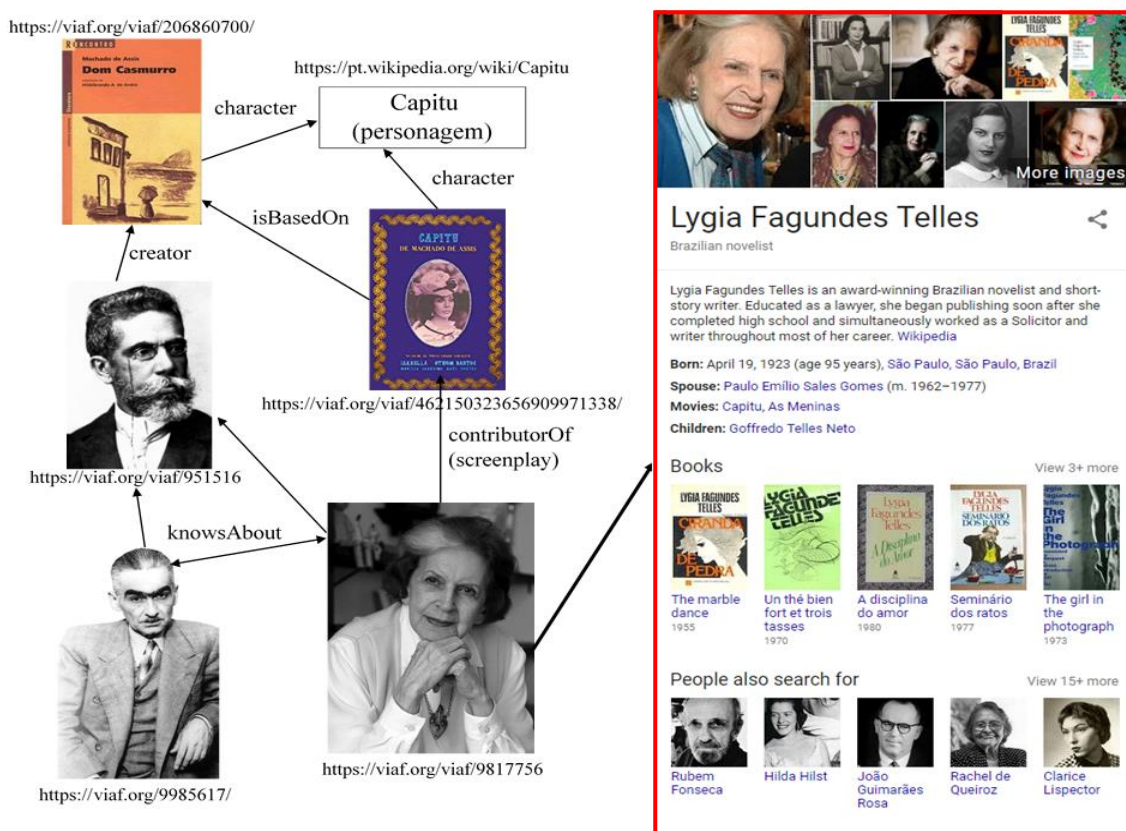
É recomendado que dados de biblioteca sejam abertos e passíveis de compartilhamento. Compreende-se que a geração e manutenção de metadados é onerosa às bibliotecas, afinal foi feito investimento na produção dos dados bibliográficos e de autoridades e é complexo disponibilizar este trabalho de forma gratuita na *Web*, podendo ser usufruído por instituições que não investem na capacitação de catalogadores. Outro aspecto que leva bibliotecas a inibir a exposição dos metadados é a qualidade deles, quando se opta por não os disponibilizar na *Web* por receio de comprometimento da imagem da instituição

⁶ *This linkade would provide the capability of incorporating biographical information, which is (and probably rightly so) outside the scope of cataloging (SINGER, 2009, p. 120).*

ou dos catalogadores envolvidos. Entretanto, a *Web* possui muitos dados que são confiáveis e a biblioteca pode selecionar os *datasets* com os quais trabalhará, além de ela mesma poder oferecer seu conjunto de dados.

No destaque da Figura 7 são identificadas informações adicionais de escritora “Lygia Fagundes Telles” que está registrada no catálogo de autoridades da biblioteca. Dados adicionais da escritora são informados como fotografias, o nome do marido e do filho, filmes e demais obras. Nem todas as informações são importantes para inclusão no catálogo. Também são apresentadas informações sobre outros autores, que podem ter obras no acervo ou não. Os dados externos podem ser consumidos e armazenados nas autoridades e/ou nos registros bibliográficos. Também podem ser consultados de forma dinâmica, enriquecendo a pesquisa somente no momento da busca, porém sem a inclusão de informações adicionais nos metadados de autoridades.

Figura 7 - Grafo com relações de dados de um autor complementado por fontes externas



Fonte: elaborado pelos autores a partir da Figura 4.

O enriquecimento de dados é uma tarefa complexa e detalhes sobre como isso pode ser feito depende dos *datasets* envolvidos. Por exemplo, dados selecionados de campos de registros podem ser copiados de um *dataset* para outro, convertendo e mesclando dados se necessário. [...] Registros de autoridades podem conter *links*, tanto internos quanto externos, que podem ser recursos valiosos para o enriquecimento de dados (BOJARS; ZOGLA; EGİTE, 2015, p. 187, tradução nossa)⁷.

Ao incluir identificadores nas autoridades, além da criação do vínculo entre *datasets*, também é possível facilitar a navegação do usuário entre as fontes utilizadas, conforme ilustrado na Figura 8.

Figura 8 - Autoridade vinculada a outros datasets



Fonte: elaborado pelos autores a partir da Figura 4.

Ao criar um link na autoridade “Lygia Fagundes Telles” no catálogo é possível orientar ao usuário que outras informações sobre a escritora podem ser obtidas nas fontes adicionais indicadas. Ao clicar nos *links*, a página da autoridade é aberta na fonte selecionada, como nova guia do navegador. No exemplo acima, o usuário pode descobrir no *Internet Movie Database* (IMDb) que outras obras de autoria de “Lygia Fagundes Telles” foram transformadas em filmes, novelas ou seriados de televisão e não somente que ela escreveu o roteiro do filme “Capitu”, que faz parte do acervo.

Porém, para que o OPAC tenha estrutura que permita a conexão com outros dados é necessário prepará-lo para isso, alterando sua estrutura e modelo

⁷ *Data enrichment is a complex task and details of how it can be performed depend on the datasets involved. For example, selected data record fields may be copied from one dataset to the other, converting and merging data as necessary. [...] Authority data records may contain links, both internal and external, that can be a valuable resource for data enrichment* (BOJARS; ZOGLA; EGİTE, 2015, p. 187).

de dados. Além do enriquecimento e possibilidades de inferências entre os registros do catálogo e a *Web*, a estrutura da WS também permite a geração de indicadores e estatísticas, uma vez que dados presentes em outros sistemas e que utilizam formatos diferentes podem ser reunidos e analisados em conjunto. Para Byrne e Goddard (2010) esta situação é compreendida como decisão baseada em evidências, quando, ao analisar os recursos, circulações e consultas realizadas em um período, é possível definir quais ações precisam ser tomadas para dar continuidade ou melhorar os serviços prestados pela biblioteca.

Os dados coletados internamente podem ser analisados com dados externos, governamentais ou não, abertos ou fechados, possibilitando a identificação de comportamentos e padrões em áreas diversas. Dados presentes em redes sociais como *Facebook* e *Twitter* também podem ser utilizados nesta tarefa, ampliando possibilidades de geração de recomendações, com base nos interesses do usuário e seus amigos.

Segundo Shieh e Reese (2015) as bibliotecas precisam enxergar que suas coleções agora são descritas com dados e não mais com registros e para realizar a conversão dos dados de MARC 21 para RDF é necessário quebrar os registros em pequenos pedaços que representam elementos ou entidades, e remontá-los em grafos usando predicados (propriedades) para expressar a relação entre o sujeito e o objeto. “O pilar da declaração em RDF (seu conjunto de triplas) é o identificador” (SHIEH; REESE, 2015, p. 210, tradução nossa)⁸.

É inviável às bibliotecas manter dois catálogos, sendo um o OPAC, para consumo dos usuários, e outro para consumo do LD. É necessária uma estrutura que permita a reunião de ambas as funcionalidades (COLE et al., 2013). Isso é alcançado com a conversão de metadados de bases de dados em MARC 21 legadas para o RDF.

⁸ *The cornerstone of an RDF statement (its triple store) is the Identifier* (SHIEH; REESE, 2015, p. 210).

6 CONSIDERAÇÕES FINAIS

A aplicação de *linked data* nos catálogos de bibliotecas proporciona o enriquecimento de dados permitindo que, além de conectar o catálogo com outras fontes disponíveis na *Web*, também seja possível a criação de relacionamentos entre os recursos presentes em uma mesma coleção.

O uso de *datasets* abertos dispostos na *Web* pode vir a facilitar o trabalho do catalogador ao prover mais elementos para identificação de autoridades e registros bibliográficos, poupando o tempo de pesquisa dispensado na elaboração dos metadados, abastecendo o catálogo de forma automática, ao invés de exigir a indicação manual de relacionamentos que podem existir entre os recursos. Também é possível enriquecer dados bibliográficos e de autoridades do catálogo sem que, necessariamente, estes dados sejam incluídos no banco de dados, com sua exibição ocorrendo somente no momento da busca, de forma dinâmica, dispensando a inclusão de metadados que podem ser úteis aos usuários, porém cujo armazenamento pode não ser necessário à construção de termos preferidos ou nas descrições dos recursos.

A interligação de dados proporcionada pelo *linked data* contribui com a transformação do catálogo de sintático para semântico, oferecendo contexto aos usuários e proporcionando alternativas de descoberta, com a realização de inferências ao interligar dados que até então eram opacos e estavam reunidos em um silo de dados que proporcionava opções de descoberta somente a partir de conhecimentos prévios dos usuários ou realização de pesquisas mais sofisticadas.

Um aspecto a ser considerado é que com o *linked data* os catálogos das bibliotecas podem alcançar outro posicionamento na *Web*, a partir do momento que podem ser considerados fontes de informações estruturadas e confiáveis e serem utilizados como *datasets* para prover dados a outras aplicações, favorecendo a *Web* dos dados como um todo. Assim, as bibliotecas podem ser consumidoras e provedoras de dados, porém não fornecendo dados somente a outras bibliotecas, mas para a *Web* e quaisquer usuários e aplicações que busquem um recurso de informação. Isso pode representar um novo paradigma

para bibliotecas e catalogadores, a partir do momento que os dados não são trabalhados somente para uso de uma comunidade específica, mas são disponibilizados a qualquer usuário, em qualquer ponto do planeta, para qualquer uso, não restritos ao universo do livro. Fazer parte da *Web* dos dados é a tendência esperada para as bibliotecas, alterando seu posicionamento e importância na rede e na sociedade.

AGRADECIMENTOS

Este estudo foi financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Brasil (CAPES) – Finança Código 001.

REFERÊNCIAS

- ALLEMANG, D. Semantic Web and the linked data enterprise. *In*: WOOD, D. (ed.). **Linking enterprise data**. New York: Springer, 2010. pp. 3-23. DOI 10.1007/978-1-4419-7665-9
- BERMÈS, E. Enabling your catalogue for the semantic web. *In*: CHAMBERS, S. (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.
- BOJARS, U.; ZOGLA, A.; EGLITE, E. Linking library data for quality improvement and data enrichment. *In*: INTERNATIONAL CONFERENCE ON DATA MANAGEMENT TECHNOLOGIES AND APPLICATIONS, 4., Colmar, France, 2015. **Proceedings** [...]. Colmar: INSTICC, 2015. p. 184-188, 2015. DOI: 10.5220/0005554901840188.
- BYRNE, G.; GODDARD, L. The strongest link: libraries and linked data. **D-Lib Magazine**, v. 16, n. 11/12, nov./dez. 2010. p. 1-9. Disponível em: <http://www.dlib.org/dlib/november10/byrne/11byrne.html>. Acesso em 19 nov. 2018.
- COLE, T. W. *et al.* Library MARC records into linked open data: challenges and opportunities. **Journal of Library Metadata**, v. 13, p. 163-196, 2013. DOI: 0.1080/19386389.2013.826074

COYLE, K. Library data in the Web world. **Library Technology Reports**, Chicago: ALA, feb./mar. 2010, p. 5-11.

DENAUX, R. *et al.* Knowledge architecture for organizations. *In*: PAN, J. Z. *et al.* (ed.). **Exploiting linked data and knowledge graphs in large organizations**. Cham: Springer, 2017. p. 57- 94. Disponível em: <http://dx.doi.org/10.1007/978-3-319-45654-6>. Acesso em: 05 nov. 2018.

DEWEESE, K. P.; SEGAL, D. **Libraries and the semantic web**. Williston, VT: Morgan & Claypool, 2014. (Synthesis lectures on emerging trends in librarianship). DOI: 10.2200/S00615ED1V01Y201411ETL003

FURNER, J. Conceptual analysis: a method for understanding information as evidence, and evidence as information. **Archival Science**, v. 4, p. 233-265, 2004. DOI: 10.1007/s10502-005-2594-8.

HYLAND, B. Preparing for linked data enterprise. *In*: WOOD, D. (ed.). **Linking enterprise data**. New York: Springer, 2010. pp. 51-64. DOI 10.1007/978-1-4419-7665-9.

MACEWAN, A. Project InterParty: from library authority files to e-commerce. **Cataloging & Classification Quarterly**, v. 39, n. 1-2, p. 429-442, 2004. DOI: 10.1300/J104v39n01_11

MALMSTEN, M. **Exposing library data as linked data**. IFLA, 2009. Disponível em: http://wtlab.um.ac.ir/images/e-library/linked_data/other/Exposing%20Library%20Data%20as%20Linked%20Data.pdf. Acesso em 15 nov. 2018.

PROGRAM FOR COOPERATIVE CATALOGING. **Formulating and obtaining URIs: a guide to commonly used vocabularies and reference sources**. 14 Feb. 2018. Disponível em: https://www.loc.gov/aba/pcc/bibframe/TaskGroups/formulate_obtain_URI_guide.pdf. Acesso em: 21 nov. 2018.

SHIEH, J.; REESE, T. The importance of identifiers in the new Web environment and using the Uniform Resource Identifier (URI) in subfield zero (\$0): a small step that is actually a big step. **Journal of Library Metadata**, v. 15, p. 208-226, 2015. DOI: 10.1080/19386389.2015.1099981.

SINGER, R. Linked library data now! **Journal of Electronic Resources Librarianship**, v. 21, p. 114-126, 2009. Disponível em: <https://www.tandfonline.com/doi/pdf/10.1080/19411260903035809?needAccess=true>. Acesso em: 14 dez. 2018

WORLD WIDE WEB CONSORTIUM. **RDF Primer**. Cambridge, 2014. Disponível em: <https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/>. Acesso em: 22 out. 2018.

WORLD WIDE WEB INCUBATOR GROUP REPORT. **Library Linked Data Incubator Group final report**. 25 out. 2011. Disponível em: <https://www.w3.org/2005/Incubator/llid/XGR-llid-20111025/>. Acesso em: 14 nov. 2018.

WU, H. *et al.* Understanding knowledge graphs. In: Pan, J. Z. *et al.* (ed.). **Exploiting linked data and knowledge graphs in large organizations**. Cham: Springer, 2017. pp. 147-180. Disponível em: <http://dx.doi.org/10.1007/978-3-319-45654-6>. Acesso em: 30 out. 2018.

FROM DATA SILOS TO WEB OF DATA

ABSTRACT

Objective: to show how catalogs are currently opaque, without semantic elements, and how they can be enriched with datasets present on the Web so that they become semantic and provide other possibilities of discovery. **Methodology:** the conceptual analysis technique was used, seeking to verify from studies how much the catalog is in fact a data silo and how this situation can be changed through linked data. **Results:** through examples of bibliographic records, it was possible to include relationships between the data present in the catalog, providing context to patrons and enriching the information retrieved from the catalog. **Conclusions:** library catalogs can be trusted sources of data and contribute to the Web of data, in addition to providing patrons with more possibilities of discoveries.

Descriptors: Catalog. Semantic Web. Linked Data. Data Silos.

DE LOS SILOS DE DATOS A LA WEB DE LOS DATOS

RESUMEN

Objetivo: mostrar cómo los catálogos son actualmente opacos, sin elementos semánticos, y cómo se pueden enriquecer con conjuntos de datos presentes en la Web para que se vuelvan semánticos y brinden otras posibilidades de descubrimiento para los usuarios. **Metodología:** Se utilizó la técnica de análisis conceptual, buscando verificar a partir de estudios hasta qué punto el catálogo es en realidad un silo de datos y cómo se puede cambiar esta situación a través de datos enlazados. **Resultados:** a través de ejemplos de registros bibliográficos, fue posible incluir relaciones entre los datos presentes en el catálogo, así como las páginas web, dando contexto a los usuarios y enriqueciendo la información recuperada en las búsquedas realizadas en el catálogo. **Conclusiones:** Los catálogos de bibliotecas pueden ser fuentes confiables de datos y contribuir a la Web de los datos, además de brindar a los usuarios más posibilidades de descubrimiento.

Descriptores: Catálogo. Web semántica. Datos Enlazados. Silos de Datos.

Recebido em: 14.09.2020

Aceito em: 22.02.2021