A Europeana e a agregação de metadados na web: análise dos esquemas ESE/EDM e da aplicação de *standards* da web semântica a dados de bibliotecas

Helena Simões Patrício
Biblioteca Nacional de Portugal
Campo Grande, 83
1749-081 Lisboa
Tel.: 217982000
E-mail: hpatricio@bnportugal.pt

RESUMO

Este artigo tem por objetivo apresentar os esquemas de metadados definidos pela Europeana para a agregação de dados, no contexto da evolução do esquema *Europeana Semantic Elements* (ESE) para o *Europeana Data Model* (EDM), que visa o enriquecimento dos dados agregados, a normalização de valores e a respetiva formalização como *Linked Open Data* (LOD). A evolução da Europeana de um esquema plano de metadados, para um modelo em que se pretende a integração dos dados agregados em redes semânticas de recursos surge no contexto da web semântica e dos objetivos de um serviço de agregação de dados nesse ambiente. Assim, começa-se por apresentar o conceito e os princípios da web semântica. Analisa-se depois as vantagens da respetiva aplicação aos recursos criados por bibliotecas e descreve-se os desafios decorrentes da representação de dados bibliográficos num formato que permita a sua utilização interativa na web, partindo da análise de iniciativas já existentes de bibliotecas que disponibilizam os seus dados como LOD. No contexto da agregação de dados como serviços para a descoberta e a reutilização de recursos num sistema de informação global; descreve-se o esquema ESE definido para o mapeamento de dados provenientes de diversos setores do património cultural num conjunto comum de elementos, que são depois recolhidos e indexados pela Europeana numa base de dados central, para pesquisa e apresentação de resultados ao utilizador final. Apresenta-se, como caso prático, o mapeamento de campos UNIMARC para o formato ESE, efetuado no contexto da implementação do serviço RNOD (Registo Nacional de Objetos Digitais), para agregação de informação bibliográfica de entidades portuguesas. Este artigo introduz os requisitos e princípios fundamentais do EDM, enquanto novo modelo de dados da Europeana que, não se substituindo ao ESE e sendo compatível com esse esquema, pretende resolver problemas suscitados pelo efeito redutor do esquema ESE sobre a riqueza semântica dos metadados dos vários fornecedores. Explica-se de que forma o EDM possibilita que cada fornecedor da Europeana estruture os seus dados sem que a respetiva riqueza original se perca e, por outro lado, de que modo o EDM permite a agregação desses metadados na Europeana e o seu enriquecimento com ligações a outros recursos de contextualização na web. Neste contexto, o artigo apresenta as classes e propriedades próprias do EDM, analisando a aplicação do EDM que está a ser realizada num projeto piloto, em que a Biblioteca Nacional de Portugal (BNP) participa, e que consiste na constituição de um repositório de dados LOD.

PALAVRAS-CHAVE: Esquema de metadados / Europeana / Linked Open Data / Web semântica / Informação bibliográfica

INTRODUÇÃO

Nas últimas duas décadas o ambiente em que o conhecimento e os recursos de informação são criados, comunicados e utilizados mudou radicalmente. Com o surgimento da Internet, os factos deixam de ser unidades isoladas de conhecimento, passando a fazer parte de uma rede de ligações que os tornam úteis e compreensíveis. Estas ligações para a origem e o destino de determinada informação estão na base do aparecimento da tecnologia dados *Linked Data*, que visa facilitar a apresentação de dados de forma a que possam ser úteis para outros utilizadores ou serviços, de forma não antecipada (Weinberger, 2011). Os desafios que se colocam hoje às bibliotecas não se limitam à transformação de dados bibliográficos com a tecnologia *Linked Data*, mas sim a criação de um sistema novo para acesso e utilização dos dados bibliográficos, que seja compatível e totalmente integrado com a Web. Com efeito, os dados ligados não propiciam em si mesmos nada de novo, permitem sim a criação de novos serviços ao utilizador final, com base em formatos mais flexíveis de representação de dados (Hawtin, 2011; Coyle, Feb. 2010, 2012).

Nesta fase inicial de transformação de dados bibliográficos, não estão ainda disponíveis ao utilizador final serviços que reutilizem dados ligados, sendo também impossível prever que inovações podem derivar da reutilização de dados bibliográficos abertos e ligados. Por esse motivo este artigo é uma análise introdutória desta temática, começando-se por apresentar os conceitos base da Web Semântica, que contextualizam o surgimento da tecnologia *Linked Data*. São depois apresentadas as vantagens que esta tecnologia pode trazer aos serviços prestados pelas bibliotecas e os contributos que esta comunidade pode dar para o desenvolvimento da Web Semântica. Neste contexto, são referenciados alguns casos práticos de transformação de dados bibliográficos em dados ligados, por bibliotecas nacionais e analisa-se o modelo de dados do agregador

Europeana no contexto do projeto piloto Europeana Linked Open Data, na dupla perspetiva da publicação de informação bibliográfica segundo os princípios *Linked Data*, colocando este tipo de dados à disposição de qualquer utilizador ou serviço na Web, e da ligação dos dados bibliográficos a outros recursos da Web, para melhoria dos serviços prestados pelas próprias bibliotecas.

WEB SEMÂNTICA

A maior parte do conteúdo que atualmente integra a Web destina-se a ser compreendido por seres humanos e não a ser manipulado, quanto ao seu sentido, por aplicações informáticas. Os computadores podem, quando muito, processar esses conteúdos para os exibir em páginas Web ou para processamentos simples de rotina, identificando por exemplo o título ou as ligações para outras páginas, sem contudo conseguirem identificar o autor da página ou os conteúdos de destino das ligações feitas na página (Berners-Lee, 2001).

Tim Berners-Lee introduziu o conceito de Web Semântica como uma extensão da Web original, que, dando estrutura ao sentido dos conteúdos das páginas web, cria um ambiente em que os agentes de software conseguem percorrer as páginas e desempenhar tarefas sofisticadas para os seus utilizadores, sem necessidade de recurso à inteligência artificial. Com efeito, a Web Semântica não é uma realidade separada da Web original, basta que a informação das páginas seja codificada para que as aplicações informáticas possam compreender o sentido dos conteúdos em vez de se limitarem a apresentá-lo. A Web desenvolveu-se originalmente como um meio de transmissão de documentos para pessoas, mas tende cada vez mais a integrar dados e informação, codificados de modo a que o seu sentido seja formalizado e os mesmos possam ser processados de forma automática.

É pois necessário evoluir de uma Web de documentos para uma Web de dados (Herman, 2012). O que tornou a Web atual importante foi o facto de os documentos criados pelas pessoas terem um endereço (URI) e de estarem acessíveis para outros na Web. Quando os outros descobrem um site e fazem ligações para o mesmo, a "rede tem um efeito nos sites", quantas mais ligações houver para o site mais importante ele se torna. O mesmo precisa de acontecer os dados puros (efeito de "rede nos dados"): se os dados estiverem expostos na Web, as pessoas podem fazer ligações para esses dados e as aplicações que podem ser feitas para os utilizar tornam-se imprevisíveis.

Na Web tradicional, os seres humanos são tidos em conta implicitamente. Cada ligação tem um contexto (sentido), que o ser humano implicitamente utiliza. Ou seja, o ser humano compreende o sentido dessas ligações. Contudo, as máquinas não conseguem perceber o sentido de uma simples ligação. É pois necessário adicionar informação às ligações, essa informação tem de ser legível por máquina, deve referir-se tanto à ligação como ao seu objeto e deve permitir raciocínios básicos. Ou seja, para evoluir de uma Web de documentos para uma Web de dados, os dados puros devem estar disponíveis em formatos *standard* na Web; os conjuntos de dados devem estar ligados; as ligações, dados e *sites* devem estar descritos formalmente. A Web Semântica é o conjunto de tecnologias que permite construir a Web de dados (Herman, 2012).

A Web Semântica é uma rede interligada de informação codificada em documentos disponíveis na Web. A palavra "semântica" não deve ser entendida linguisticamente como o "sentido ou significação das palavras", uma vez que as máquinas nunca compreenderão o sentido do texto, mas sim como "linguagem formal" ou sintaxe, i.e. como conjunto de regras utilizadas pelos computadores para desempenhar operações (Coyle, Jan. 2010). No próximo ponto descreveremos brevemente essas linguagens formais.

Tecnologias e linguagens

O desafio da Web Semântica é propiciar uma linguagem que expresse dados e regras lógicas para raciocinar sobre esses dados, de forma interoperável na web. Adicionar lógica à Web significa utilizar regras para fazer inferências, escolher ações e responder a questões. As tecnologias da Web Semântica são o XML e o RDF. O XML permite que qualquer pessoa adicione etiquetas para anotar as suas páginas web ou secções de texto nessas páginas. Contudo, para os programadores poderem escrever scripts que usem essas etiquetas de forma sofisticada, têm de conhecer previamente o que as etiquetas significam, i.e. o sentido dessas estruturas. Este sentido é expresso em RDF, que o codifica em conjuntos de triplos, frases elementares com sujeito, verbo e predicado, sendo cada um desses elementos identificados com URIs (Universal Resource Identifier). Os URIs vão permitir a utilização desses sujeitos e predicados como ligações em qualquer página web e que qualquer pessoa defina um novo verbo ou conceito na Web (Berners-Lee, 2001).

Numa base de dados precisamos de saber, por exemplo, que campos correspondem ao código postal e ao nome da pessoa, se quisermos saber que pessoas vivem em determinada área. O RDF especifica que determinado campo da base de dados é do tipo "código postal", usando um simples URI, em vez de frases para cada termo. Ou seja, recorrendo ao xml ou bases de dados, para utilizar essa informação de forma automática o programador (pessoa) tem de conhecer o sentido dos campos ou etiquetas, i.e. o sentido é conhecido pelo ser humano que depois programa a aplicação que vai usar os dados. Com o RDF, os agentes de software (máquina) conseguem usar a informação sabendo automaticamente o sentido de cada campo, desde que esses campos/etiquetas estejam codificados em RDF, i.e. fazendo referência a URI em que o sentido de cada um está expresso. Com efeito, os URIs são utilizados para recuperar informação sobre a coisa identificada, usando apenas o protocolo http, logo não sendo necessária programação adicional. O URI permite que um recurso esteja na Web e seja acionável na Web. Estes identificadores permitem que um agente de software atue nos elementos de uma declaração da Web Semântica sem compreender o seu significado humano (Coyle, Jan. 2010).

Ontologias

Uma ontologia é um ficheiro ou documento que define formalmente relações entre termos. No contexto da Web, as ontologias são compostas por uma taxonomia, que define as classes de objetos e as relações entre eles, e por um conjunto de regras de inferência que permitem que os programas manipulem os termos de forma mais eficiente do que os seres humanos, dando-lhes a utilidade e sentido que forem necessários. Os programas não compreendem verdadeiramente essa informação, apenas aplicam as regras que permitem manipulá-la. As ontologias são um componente básico da Web

Semântica, pois permitem comparar ou combinar informação de bases de dados/esquemas distintos, declarando que campos ou etiquetas diferentes têm o mesmo significado, i.e. criando relações de equivalência. As ontologias podem melhorar o funcionamento da Web de muitas formas, nomeadamente no que respeita ao rigor dos resultados de pesquisa, na medida em que podem facilitar o desenvolvimento de programas capazes de responder a questões complexas cuja solução não se encontra numa única página web (Berners-Lee, 2001).

Agentes

O verdadeiro potencial da Web Semântica será realizado quando forem criados agentes de software capazes de coligir conteúdo Web de diferentes fontes, processar essa informação e trocar os resultados com outros programas. Este potencial será tanto mais forte quantos mais conteúdo legível por máquina for disponibilizado e quanto mais serviços automatizados (agentes) surgirem. A Web Semântica permite que mesmo agentes não desenhados expressamente para trabalharem em conjunto possam trocar dados, se esses dados tiverem semântica. Outro aspeto importante é a descoberta não só de recursos, mas também de serviços automatizados por outros agentes, através da utilização de linguagens comuns para a descrição de serviços ou de funções desempenhadas. Atualmente a descoberta de serviços depende da descrição de um conjunto pré-determinado de funcionalidades, não havendo antecipação de necessidades futuras. A Web Semântica torna a descoberta de agentes mais flexível, pois diferentes agentes conseguem, através das ontologias, ter um vocabulário comum para discutirem, podendo até alcançar novas capacidades de raciocínio (Berners-Lee, 2001). Na medida em que os serviços passam a poder autodescrever-se e autodescobrir-se, as utilizações são ilimitadas e imprevisíveis.

Linked Open Data

Os dados estruturados de acordo com as regras da Web Semântica designam-se por dados ligados (*linked data*). Os dados ligados são publicados de acordo com princípios que facilitam ligações entre conjuntos de dados, conjuntos de elementos e valores de vocabulários. Os dados ligados recorrem a URIs como identificadores únicos de recursos, de forma análoga aos termos de controlo de autoridade da comunidade de bibliotecas. Os dados ligados são expressos em RDF, linguagem que especifica relações entre recursos. Essas relações podem ser utilizadas para navegar ou integrar informação de múltiplas fontes (Baker, 2011).

A tecnologia *Linked Data* não exige que os dados sejam abertos, mas o potencial pleno apenas é atingido se estes dados forem tecnicamente interoperáveis, podendo ser livremente usados, reutilizados e redistribuídos (Baker, 2011).

Os *Linked Open Data* (*LOD)* são, portanto, dados ligados disponibilizados com uma licença aberta, que não impeça a sua reutilização gratuita. Tim Berners-Lee (2006) definiu 4 princípios/regras da tecnologia *Linked Data*:

1) Usar URIs como nomes das coisas: nomear objetos e recursos de forma inequívoca
2) Usar URIs http para que as pessoas possam procurar por esses nomes: usar a estrutura da Web
3) Quando alguém procurar por um URI, fornecer informação útil utilizando *standards* (RDF, SPARQL), quer ligando os dados às ontologias que possam dar informação sobre as respetivas propriedades e classes e sobre as relações entre os diversos termos da ontologia, quer ligando-os a outros dados ligados.
4) Incluir ligações para outros URIs, de modo a que se possa encontrar mais coisas.

As primeiras duas regras são essenciais para assegurar maior visibilidade dos dados reutilizáveis. Permitem uma identificação não ambígua dos recursos e a sua localização e referenciação na Web, o que é fundamental para qualquer método de partilha de dados. A terceira regra existe porque a correta interpretação dos dados ligados exige o acesso a outros dados que os descrevem (ontologias), especificando a sua forma e semântica. As ontologias podem ser expressas em RDFS (simples definições e relações hierárquicas) ou em OWL (restrições às relações entre entidades). É isto que torna os dados ligados autodescritivos, i.e. a partir de um URI pode chegar-se à definição do que esse URI significa. Esta é uma característica muito poderosa dos dados ligados (Hawtin, 2011).

LINKED OPEN DATA E DADOS BIBLIOGRÁFICOS

Os dados bibliográficos são texto codificado, mas não seguem os *standards* da Web Semântica, para poderem ser identificados e utilizados nesse contexto. A transformação dos dados bibliográficos num formato compatível com a Web Semântica está facilitada pelo facto de os dados já serem codificados (MARC) e por já existir um elevado grau de controlo de vocabulário. Na segunda metade do século XX as bases de dados relacionais foram aplicadas aos dados bibliográficos, uma vez que eram usadas pelos sistemas que utilizavam os dados em bibliotecas. Mas essas bases de dados são desadequadas pois destinam-se a dados tipicamente comerciais, que são menos textuais e mais compactos do que os dados bibliográficos; têm menos elementos; menor diversidade de conteúdo por elemento; maior repetição de valores. O XML foi, depois, utilizado essencialmente para a apresentação dos dados bibliográficos textuais, não configurando nenhuma alteração radical para uma utilização mais interativa dos dados bibliográficos na Web. A necessidade de mudança levou, entretanto, a algumas alterações tanto ao nível da conceção de um novo modelo conceptual para os dados bliográficos (FRBR), como da elaboração de novos modelos de dados (RDA). Neste momento há, pois, um novo modelo para os dados bibliográficos, mas falta transformar a forma como os mesmos são expressos/codificados, de modo a assegurar que os dados são processáveis por máquina e facilmente integráveis em serviços Web e aplicações informáticas. Esta transformação implica uma alteração da própria natureza dos dados bibliográficos (Coyle, Jan. 2010):

a) Transformar descrições textuais em conjuntos de elementos de dados processáveis por máquina
b) Assegurar que estes dados são compatíveis com a tecnologia Web

O RDF difere do atual sistema de registos bibliográficos, porque permite que as descrições bibliográficas interajam, ao nível da declaração RDF, com outros dados de fontes bibliográficas e não bibliográficas. O formato de registo cria um contentor que mantém um conjunto de elementos de dados juntos, para fins aplicacionais. Diferentemente, o formato de triplos (triplos RDF) permite que declarações individuais de dados possam interagir com outras declarações de dados (Coyle, Feb. 2010).

Por outro lado, os valores do objecto, sujeito e predicado de uma declaração RDF são URIs, o que significa que podem ser combinados com outros dados, sem perder o seu significado exato. Isto facilita o processamento por máquina, tornando fácil criar novos serviços, como por exemplo encontrar outras edições do mesmo livro, sem o utilizador ter de fazer uma nova pesquisa. A principal razão para organizarmos os nossos dados como elementos de dados separados e inequívocos é permitir que esses dados possam ser usados fora do contexto do catálogo e do registo bibliográfico, podendo ser combinados com outros dados (Coyle, Feb. 2010).

Os catálogos são sistemas isolados que usam a tecnologias de bases de dados. O acesso Web ao catálogo é feito por um "túnel" que liga a rede à base de dados do sistema da biblioteca. Os dados bibliográficos estão na Web. i.e. são utilizados pelo Google, Wikipedia, Open Library, etc, quer importando os dados da biblioteca ou criando os seus próprios dados. Em qualquer dos casos, esses sistemas *online* não usam os registos MARC. Pode também haver ligações de outros sistemas para os dados do catálogo, mas é para a página que mostra o registo, não para elementos do registo. Ou seja, correntemente a integração de dados bibliográficos com outros sistemas na Web é apenas parcial. Pois esses dados "estão na Web, mas não são da Web" (Coyle, Feb. 2010).

Vantagens para as bibliotecas e seus utilizadores

Como decorre do ponto anterior, transformando-se em dados ligados, os dados bibliográficos passam a ser partilhados na Web, o que expande o seu contexto, pois passam de dados isolados a conjuntos de dados ligados, capazes de interagir com os outros recursos de informação disponíveis na web (Coyle, Jan. 2010).

Os dados ligados são uma extensão natural dos modelos colaborativos subjacentes à criação e disponibilização de dados pelas bibliotecas, pois são partilháveis, extensíveis e facilmente reutilizáveis. Os recursos podem ser descritos em colaboração com outras bibliotecas e ligados a dados de outras comunidades ou indivíduos. Tal como as ligações que atualmente são feitas entre documentos, as ligações entre dados permite que o conhecimento de qualquer pessoa possa ser reutilizado e recombinado com os conhecimentos de outras pessoas. Por outro lado, as bibliotecas podem enriquecer o valor dos seus dados, ligando-os a dados de outras fontes que sejam de confiança. Este valor é maior do que a mera soma dessas fontes tomadas individualmente (Baker, 2011).

Outra vantagem da disponibilização de dados bibliográficos como dados ligados advém da atribuição de identificadores únicos a obras, sítios, pessoas, acontecimentos e assuntos. Ao disponibilizarem desta forma os seus dados, as bibliotecas permitem que esses recursos sejam referenciados por um conjunto muito amplo de fontes externas e, portanto, enriquecem as formas de acesso aos seus metadados (Baker, 2011).

Por outro lado, a utilização de identificadores únicos e a possibilidade de cada entidade fornecer declarações individuais (em vez de registos inteiros) sobre determinado recurso, permitem que os fornecedores de dados contribuam com porções dos seus dados como declarações. Por exemplo, uma biblioteca pode contribuir com uma identificação única de bibliografia nacional de um recurso, enquanto outra biblioteca poderá fornecer esse título traduzido. Os serviços de biblioteca podem aceitar essas declarações tal como atualmente ingerem capas de livros associando-as aos seus recursos (Baker, 2011).

Os dados de controlo de autoridade das bibliotecas podem ajudar a reduzir a redundância das descrições bibliográficas na Web, identificando claramente as entidades chave que são partilhadas na nuvem de dados ligados (Baker, 2011).

Na perspetiva do utilizador de bibliotecas, os dados bibliográficos ligados podem possibilitar uma navegação mais sofisticada entre recursos da biblioteca ou de outras entidades, podendo até eliminar os principais problemas dos motores de pesquisa federada: falta de granularidade, pesquisas pouco sofisticadas,lentidão e ausência de *ranking* por relevância (Byrne, 2010; Baker, 2011).

Com efeito, a criação de serviços a partir de dados bibliográficos que sigam os princípios *Linked* Data, pode evitar processos longos e entediantes de pesquisa em que o utilizador tem de consultar uma grande quantidade de sítios Web, todos diferentes em estilo, finalidade e língua, e integrar mentalmente toda essa informação. Apesar de haver serviços web especializados e APIs que agregam e combinam informação de outras fontes, esses serviços controlam a forma como se vê essa informação e que informação é mostrada, ora por vezes o utilizador quer personalizar: aceder aos dados originais (muitos dos quais não estão à superfície ou visíveis pelo ser humano nos sites originais) e combiná-los da sua própria forma (Herman, 2012).

Para além destas vantagens, a disponibilização de dados bibliográficos abertos e ligados, permite que a biblioteca dê resposta às novas expectativas dos utilizadores (personalização, reutilização por programadores), conheça melhor a atividade do utilizador, aumente a exposição desses dados aos motores de pesquisa e beneficie do desenvolvimento de serviços por terceiros, que serão sempre uma fonte de tráfego para os dados da biblioteca (JISC, 2010).

Contributo das bibliotecas para a Web de Dados

Os benefícios da disponibilização dos dados de bibliotecas na nuvem de dados ligados são em grande número e de grande relevância para as outras comunidades da Web. Efetivamente, passam a estar representadas na Web grandes coleções de material publicado e não publicado que as bibliotecas possuem, com a vantagem de a maioria desses materiais consistirem em termos controlados de nomes de pessoas ou organizações, descrições físicas, assuntos e classificações temáticas. Obedecendo aos princípios *Linked Data*, esses dados podem interagir com quase toda a informação da Web, uma vez que as coleções de

bibliotecas cobrem praticamente todos os aspetos da ação humana (Coyle, Jan. 2010).
Por outro lado, no contexto da Web Semântica, em que neste momento, se disponibilizam resultados de experiências e projetos pontuais, que não são alvo de verificação regular de rigor e de atualizações de manutenção, é muito relevante a experiência das bibliotecas na curadoria e preservação a longo prazo de conjuntos de dados ligados, contribuindo para o controlo de qualidade desses dados e assegurando a sua manutenção a longo prazo (Baker, 2011).
Por último, os bibliotecários podem facilmente transformar o seu conhecimento e experiência em metadados em conhecimento e experiência de trabalho com ontologias e modelação de dados. Enquanto os metadados tradicionais das bibliotecas visava ajudar os seres humanos a encontrar e utilizar informação, as ontologias da web semântica visam ajudar as máquinas a encontrar e utilizar informação, fazendo com que as máquinas compreendam o seu sentido e consigam agir sobre eles (Hellman apud Coyle, 2010, Feb. 2010).

Metodologias de conversão para reutilização

Relativamente à estratégia de disponibilização de dados de bibliotecas como dados ligados, apresenta-se em seguida as recomendações do grupo *Library Linked Data Incubator Group* (Consórcio W3C) (Baker, 2011) e a metodologias de trabalho propostas por Karen Coyle (Feb. 2010) e Gordon Dunsire (2012).

Recomendações estratégicas. O *Library Linked Data Incubator Group* (Baker, 2011) recomenda às bibliotecas que optem por uma estratégia gradual, identificando os dados mais prioritários e que requerem menor esforço de transformação, evitando num primeiro momento questões mais complexas. Sugere ainda que se comece pelos registos de autoridade e pelas listas controladas de termos. Os dados devem ser disponibilizados com licenciamento aberto, devendo atribuir-se um URI a cada recurso descrito e definir-se as políticas de *namespaces* usados. Como princípio geral, o Grupo sugere o alinhamento com os vocabulários de outras comunidades, utilizando os *standards* da web semântica, e a utilização de dados ligados provenientes de fontes externas, de modo a que se possa atingir mais utilizadores.

Metodologias de trabalho. Segundo Karen Coyle (Feb. 2010), no processo de aplicação dos princípios *Linked Data* aos dados bibliográficos, deve seguir-se as seguintes etapas:

1) Definição de um modelo conceptual, que especifique as entidades e as relações a que os metadados se referem.

No contexto dos dados bibliográficos, o modelo conceptual mais indicado é o modelo FRBR e restantes modelos funcionais. Os modelos funcionais (FRBR, FRAD, etc) não foram desenvolvidos para os *standards Linked Data*, mas usam conceitos como entidades e relações, que são conceptualmente similares aos conceitos básicos da Web Semântica. É portanto um bom modelo para a transformação dos dados bibliográficos em dados ligados.

2) Definição de um modelo de dados ou ontologia, que especifique os elementos dos dados: classes e propriedades.

Karen Coyle aconselha a utilização do esquema *Resource Discovery and Access* (RDA), uma vez que este esquema é uma implementação do modelo FRBR e que os elementos de dados identificados no RDA foram definidos, em articulação com a Dublin Core Metadata Initiative, usando as técnicas atuais da Web Semântica. Esses elementos estão no registry, acessíveis de forma aberta na Web, podendo ser reutilizados por quem quiser descrever dados bibliográficos.

3) Definição dos vocabulários de valores para os dados.

Consiste na definição do sentido que cada elemento pode ter, através da utilização de listas controladas de termos. Cada termo do vocabulário tem um identificador único, portanto cada sentido tem também um identificador diferente. Karen Coyle refere que, também no âmbito dos vocabulários de valores, se pode extrair do RDA a informação necessária para a criação de vocabulários para os metadados. Se esses valores não forem simplesmente texto, mas forem entradas controladas de um vocabulário, esse vocabulário deve ser formalizado em RDFS/SKOS e cada entrada ter um URI. O RDA define 70 vocabulários controlados de valores. A Biblioteca do Congresso está a preparar a disponibilização dos seus vocabulários bibliográficos em formato compatível com a Web Semântica, pelo que os mesmos também podem ser utilizados na disponibilização de conjunto de dados ligados.

4) Desenvolvimento de regras para a aplicação dos dados.

Definição de um perfil de aplicação que especifique restrições, como a obrigatoriedade da utilização de um elemento, a sua repetibilidade ou os valores que os elementos podem ter. Especifica também a seleção dos elementos de dados utilizados. Um perfil de aplicação inclui orientações, boas práticas para a criação de metadados por determinada comunidade. Deve utilizar-se o Dublin Core Description Set Profile (DCSP), o formato *standard* para a criação de perfis de aplicação.
Relativamente à definição de modelos de dados, Coyle demonstra preferência pelo RDA, mas essa não é a única ontologia que pode ser usada para a representação de dados bibliográficos ligados. Com efeito, tem-se verificado três tendências na utilização de ontologias pelas bibliotecas que querem disponibilizar os seus dados usando a tecnologia *Linked Data* (Dunsire, 2012):

- Usar vocabulários que a comunidade das bibliotecas está a tentar *standardizar* como ontologias específicas para dados bibliográficos (MARC21 em RDF, RDA ou ISBD em RDF). Estas iniciativas têm tido, no entanto, um desenvolvimento muito lento. Os vocabulários do RDA, por exemplo, começaram a ser desenvolvidos em 2008 e ainda não estabilizaram. Esta abordagem é mais complexa e formal do que a que se explica a seguir.
- Selecionar a informação mais importante dos dados e representá-la em RDF, reutilizando ontologias comuns da web, como Dublin Core, Bibliographic

Ontology (BIBO), Friend of a Friend (FOAF). Esta é a estratégia da British Library, LIBRIS, Universidade de Cambridge, etc. Algumas destas bibliotecas modelaram ainda classes FRBR para representar dados de responsabilidade (autoria e publicação) ou dados sobre os assuntos (pessoas, locais, assuntos, tempo).

- Uma mistura das duas tipologias de vocabulários - por exemplo Biblioteca Nacional da Alemanha reutiliza o RDA, FOAF, DC e SKOS.

Todas estas abordagens são válidas, desde que se assegure o mapeamento RDF ou relacionamento semântico entre os elementos dos diferentes vocabulários, de modo a ser possível o processamento por máquina de dados expressos em diferentes elementos de dados, uma vez que pela inferência será possível associar diferentes propriedades e classes. O mesmo se aplica aos vocabulários de valores. O que é importante é usar o RDFS para formalizar essas relações entre diferentes propriedades e valores. Este alinhamento ou interoperabilidade entre vocabulários, difere da utilização de crosswalks entre estruturas de metadados baseadas em registos (XML), pois consiste na identificação das equivalências e de outros tipos de relacionamentos entre elementos individuais, para tornar possível a aplicação dessas propriedades fora do contexto dos vocabulários de origem (Dunsire, 2012).

No capítulo seguinte, poderemos observar exemplos concretos de definição de ontologias para dados bibliográficos disponibilizados por algumas bibliotecas nacionais de acordo com os princípios *Linked Open Data*.

EXEMPLOS DE DADOS BIBLIOGRÁFICOS LIGADOS

Há poucos conjuntos de dados bibliográficos ligados. A maioria dos dados de bibliotecas ligados são conjuntos de valores de vocabulários (cabeçalhos de assunto da Biblioteca do Congresso, Classificação Decimal Dewey, etc) e de elementos (DCMI metadata terms, FRBR). Com efeito, segundo dados recolhidos em Novembro de 2011, no contexto de inquérito a organismos do sector cultural, realizado pelo Projeto *Linked Data*, 89,7% desses organismos declararam que não publicam dados ligados (McKenna, 2011).

O mesmo estudo permite concluir que a maioria dos dados ligados disponíveis na *Linked Data Cloud* não pertencem a bibliotecas. Assim, os pacotes de dados abertos com mais triplos são, por ordem decrescente: Linked GeoData; UK Legislation; Linked Sensor Data; Data Gov.uk; DBPedia e Open Library. Por outro lado, os pacotes de dados que são alvo de mais ligações por outros pacotes são, também por ordem descrescente: DBPedia; GeoNames Semantic Names (McKenna, 2011).

Não se apresentará neste artigo exemplos de conjuntos de valores para vocabulários (VIAF, LCSH, etc), nem em vocabulários de elementos de dados (RDA, DC, ISBD, etc); concentrando-nos antes nas experiências existentes de disponibilização de dados bibliográficos como dados ligados, pois é esse também o âmbito do modelo EDM e do projeto piloto LOD Europeana, que serão descritos em capítulo próprio. Nesta secção apresentaremos, portanto, alguns exemplos de bibliotecas nacionais, que poderão ajudar na análise dos dados ligados disponibilizados pela Europeana.

Os exemplos serão apresentados de forma breve e com a indicação das respetivas ontologias e de outros elementos que poderão guiar a sua análise mais detalhada.

Optámos por apresentar quatro casos distintos, selecionados pelas seguintes razões: LIBRIS, pois procede à transformação de dados não bibliográficos; British Library, pelo volume e homogeneidade de dados; Biblioteca Nacional de França, porque procedeu à seleção de elementos concretos dos registos de dados, transformou dados com diferentes esquemas de origem e disponibiliza um serviço concreto ao utilizador final; Biblioteca Nacional de Espanha, porque tem o FRBR como modelo conceptual, sem precisar de recorrer ao RDA como modelo de dados e pela excelente ferramenta de visualização de grafos.

LIBRIS (Biblioteca Nacional da Suécia)

Foi o primeiro catálogo coletivo ou de uma biblioteca nacional a ser completamente (i.e. com todos os seus registos, relações e ligações) exposto na web como conjunto de dados ligados (Rightscom, 2009).

Os dados do catálogo já estavam expostos de forma legível por máquina através de OAI-PMH, SRU/W e por z39.50, mas a sua disponibilização como dados ligados veio permitir que:

- Não ficassem expostos apenas os dados MARC, mas também os recursos ligados e outros dados como as anotações de utilizadores
- Não fosse necessário que cada utilizador (sobretudo os que não pertencem à comunidade das bibliotecas) tivesse de adquirir conhecimentos específicos das APIs

Têm mais de 12.000 links para os cabeçalhos de assuntos da Biblioteca do Congresso, o que permite fazer inferências com base em relações que não existem nos cabeçalhos de assuntos Libris (Rightscom, 2009).

British National Bibliography (British Library)

Conjunto de dados ligados: 3 milhões de registos da Bibliografia Nacional desde 1950

Exemplos de vocab. de elementos usados: RDA; ISBD; Skos; Foaf; DC

Exemplos de vocab. de valores usados: LCSH; MARC Country and Language; Dewey.info

Links para outros conjuntos de dados: VIAF, DBPedia

Licença: CC0 Public Domain. Pedem que se dê os créditos à BL

Esquema de dados: BLTerms[1]

Registado em: http://thedatahub.org/dataset/bluk-bnb

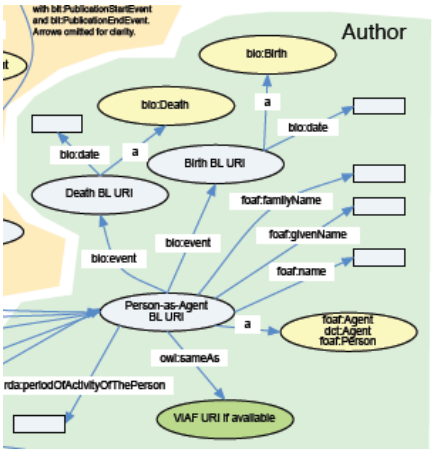

Figura 1 - Modelo de dados BLT

Metodologia de transformação BL LOD:

- Pré-processamento automático de registos MARC21 BNB: criação de uris e acrescento de URIs externos
- Transformação em RDFXML, usando XSLT
- A partir do ficheiro XML/RDF é gerado um Dump de triplos RDF
- Os dados ligados são carregados no data hub e em outras plataformas de publicação de dados ligados

Ao combinar dados ligados genéricos (GeoNames, FOAF) com dados ligados do domínio das bibliotecas (Marc Country records, VIAF), os dados ligados da British Library são colocados num contexto mais alargado.

Ainda não há serviços ao utilizador final conhecidos que reutilizem estes dados, mas têm 2 milhões de transações desses dados por mês. Também ainda não existem serviços ao utilizador final prestados pela própria BL.

Biblioteca Nacional de França

Conjunto de dados ligados: Dados da BNF criados em diferentes formatos: InterMarc (catálogo), EAD (inventários de arquivo) e DC (biblioteca digital)
Exemplos de vocab. de elementos usados: Rdf; DC; SKOS; FOAF; RDA
Exemplos de vocab. de valores usados: DDC
Links para outros conjuntos de dados: VIAF, DBPedia
Licença: LO (Licence ouverte - data.gouv.fr), Corresponde a CC BY; Necessária atribuição; Reutilização gratuita não comercial
Esquema de dados: Ontologia BNF[2]
Registado em: Dump completo está acessível em: http://echanges.bnf.fr/PIVOT/databnf_all_rdf_xml.tar.gz?user=databnf&password=databnf

Não foram convertidos em RDF todos os elementos dos registos, apenas a informação necessária para a disponibilização agregada de informação sobre o autor, obra e assuntos. É usado o FRBR como modelo conceptual.

Há um serviço web para o utilizador final, em que os dados são agrupados de forma automática numa página web, por obra, autor e assunto. Não mostra ligações para fontes externas, mas agrupa semanticamente a informação da BNF expressa na fonte em formatos diferentes, expressando os elementos de outros esquemas em RDF. Estes dados ligados ficam disponíveis para serem referenciados inequivocamente por fontes externas. No fundo, foi feita uma transformação em RDF de alguns elementos provenientes de diferentes fontes de dados. Essa transformação permitiu formalizar de forma legível por máquina as relações entre os diferentes dados, permitindo mostrar por exemplo os autores nascidos no mesmo ano de Vítor Hugo (v. Figura 2). Nestas páginas Web os dados foram todos expressos usando um esquema/ontologia comum, esse mesmo efeito poderia ser obtido mesmo que esse esquema não fosse RDF. Mas é o facto de serem dados ligados que vai permitir que os mesmos possam ser reutilizados para fins similares por outros serviços Web ou ser facilmente agregados com outros dadas Web, sem que se tenha previamente de conhecer o esquema de dados ou fazer mapeamentos para esquemas diferentes.

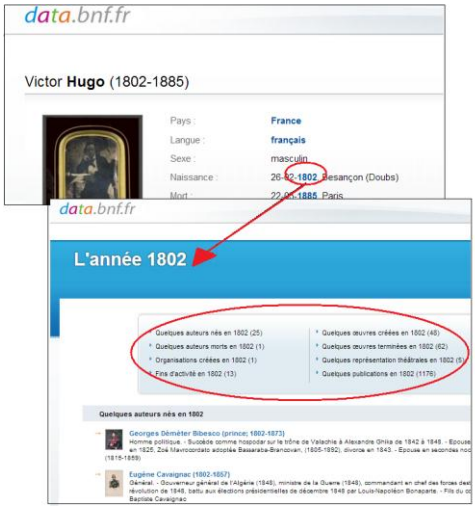

Figura 2: Exemplo de ligação a partir da página de autor no serviço http://data.bnf.fr

Biblioteca Nacional de Espanha

Conjunto de dados ligados: 8 milhões de registos bibliográficos e de autoridade
Exemplos de vocab. de elementos usados: ISBD; FRBRs
Links para outros conjuntos de dados: VIAF, DBPedia
Licença: CC0
Esquema: Ontologia BNE
Notas: Visualizador de dados: http://bne.linkeddata.es/ e exemplo de visualização de recursos http://datos.bne.es/page/resource/XX2348711
Registado em: http://thedatahub.org/pt_BR/dataset/datos-bne-es

Metodologia de transformação:

- Registos de autoridade transformados em instâncias RDF dos esquemas FRBR, FRAD e FRSAD, representando pessoas, obras ou expressões
- Registos bibliográficos transformados em instâncias RDF do esquema ISBD, representando manifestações
- Cada instância RDF tem um URI

Coyle (2012) salienta a importância de, neste caso concreto, se utilizar as autoridades (obras e expressões) como entidades *linking*, separadas das descrições bibliográficas (manifestações) que permanecem como texto. Ou seja, as ligações entre recursos de bibliotecas e recursos de outras comunidades podem acontecer sem ser necessário modificar grandemente a catalogação descritiva. Conclui esta autora que a atenção deve estar nos dados de autoridade, que são os que mais propiciam oportunidades de ligação e que a catalogação descritiva vai ser menos útil do que as entidades que são representadas pelos nossos dados de autoridade (ex: VIAF) (Coyle, 2012).

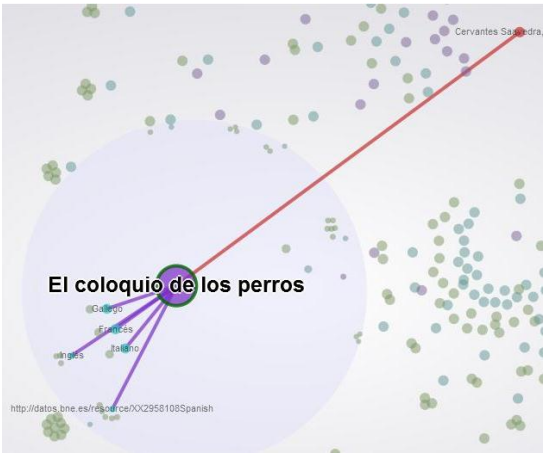

Figura 3: Exemplo de visualização em http://bne.linkeddata.es/

LINKED OPEN DATA E SERVIÇOS AGREGADORES

A inclusão de coleções em agregadores é um meio de propiciar um melhor acesso e visibilidade das coleções de uma biblioteca. Com efeito, a representação de coleções em agregadores expõe e promove os respetivos metadados em canais comumente usados, como são os motores de pesquisa (Groat, 2009).

Para além destas razões, há outros benefícios propiciados pela agregação de dados (Walk, 2011):

- Apoiar a descoberta de recursos, resolvendo problemas de latência de rede ou de sistemas, através da utilização de caches
- Aumentar a visibilidade dos recursos da Web, aproveitando o efeito dos rankings do Google que colocarão os resultados do agregador à frente dos do fornecedor de dados
- Montra de recursos
- Infraestrutura para suportar serviços de terceiros, i.e. mais do que agregar dados, oferecer serviços de dados.

Estas vantagens da agregação de dados são naturalmente incrementadas se, à agregação, associarmos a sua disponibilização como conjuntos de dados ligados, de acordo com os princípios da Web Semântica. Nos pontos que se seguem abordaremos, em primeiro lugar, o esquema utilizado pela Europeana para a representação de informação descritiva proveniente dos fornecedores de dados e, em segundo lugar, o processo de transformação desses recursos em dados ligados, no âmbito do projeto piloto Europeana Linked Open Data.

Europeana: do esquema ESE ao modelo EDM

Para garantir um nível mínimo de interoperabilidade entre dados de diferentes fornecedores que têm formatos de metadados distintos e que, sendo provenientes de comunidades diversas, se fundamentam em modelos conceptuais com perspetivas diferentes, optou-se por definir para a Europeana um esquema de metadados cujos elementos fossem comuns a todas as comunidades de fornecedores de dados e que, portanto, permitisse o mapeamento dos diversos conjuntos de dados para uma representação uniforme dos dados (Isaac, 2012).

Este esquema inicial, Europeana Semantic Elements (ESE)[3], consiste num perfil de aplicação Dublin Core, que utiliza um subconjunto de elementos do esquema Dublin Core e elementos específicos da Europeana, definidos para permitir a implementação de funcionalidades específicas do Portal Europeana. O ESE é o formato de metadados utilizado no sistema Europeana atualmente em produção (Isaac, 2012).

O esquema ESE apresenta vários problemas que dificultam a simples transformação em dados ligados. Por um lado, o ESE é um esquema "plano", i.e. agrega num único registo de metadados campos que se podem aplicar a diferentes entidades. Por outro lado, os valores dos dados representados em ESE são, na sua maioria, valores textuais, o que impede a sua ligação a outros objetos ou a entidades contextuais. O desenvolvimento do esquema RDF EDM (Europeana Data Model) foi uma resposta a estas fragilidades do esquema ESE (Isaac, 2012).

A Biblioteca Nacional de Portugal (BNP) participa na Europeana desde o lançamento do primeiro protótipo do Portal, em Novembro de 2008, fornecendo os dados bibliográficos relativos a objetos digitais de acesso público na Biblioteca Nacional Digital (BND). Estes dados são agregados pela The Euroepan Library (TEL), que os converte no formato ESE e os expõe para recolha pela Europeana por protocolo OAI-PMH. Não tendo este mapeamento para ESE sido efetuado diretamente pela BNP, abordaremos as questões colocadas por esta conversão a partir da experiência de conversão para ESE obtida pela BNP enquanto entidade responsável pelo agregador RNOD.

Em Maio de 2011, a BNP lançou publicamente o Registo Nacional de Objetos Digitais (RNOD)[4], como agregador de conteúdos digitais e digitalizados disponibilizados em rede por entidades portuguesas que visa a coordenação e difusão desses recursos, a nível nacional e internacional, designadamente através do Portal Europeana. O RNOD conta atualmente com 14 membros, disponibilizando um ponto único de acesso aos dados de 11 entidades participantes: Assembleia da República, Biblioteca Municipal do Porto, Biblioteca Nacional de Portugal, Biblioteca Municipal de Figueiró dos Vinhos, Biblioteca Municipal de Montalegre, Direção Regional da Cultura dos Açores, Hemeroteca Municipal de Lisboa, Instituto Camões, Instituto dos Museus e da Conservação, Universidade de Coimbra e Universidade de Lisboa. Não sendo um agregador exclusivo para bibliotecas, o RNOD está aberto à participação de quaisquer entidades que queiram disponibilizar material bibliográfico, pelo que os fornecedores de dados disponibilizam os seus dados nos formatos de origem, sendo os mesmos convertidos pelo RNOD para o formato UNIMARC e, posteriormente, para o formato ESE, de modo a poderem ser recolhidos para o Portal Europeana. Não sendo possível abordar neste artigo os detalhes do mapeamento

efetuado de UNIMARC para ESSE, remetemos para a consulta da respetiva documentação disponível no sítio Web do RNOD http://rnod.bnportugal.pt/mapeamento. Focaremos, portanto, a nossa análise apenas nos problemas suscitados pela conversão de dados UNIMARC para um formato mais simples como o ESE e em algumas questões de agregação de dados para a Europeana, no contexto da adoção dos princípios Linked Data pela Europeana.

O mapeamento para o formato ESE implica a perda de informação contida nos dados na origem, pois dada a diversidade de esquemas utilizados pelos fornecedores de dados, foi necessário criar um conjunto de elementos comuns, ignorando muitos elementos e valores específicos das diferentes comunidades que disponibilizam os seus dados na Europeana. Um dos motivos para a transição na Europeana para esquemas de dados RDF, consiste justamente na extensibilidade desses esquemas, que permitirá a especificação de elementos de dados derivados do esquema principal e a adoção de valores controlados por fontes externas. Relativamente ao UNIMARC e no caso específico do mapeamento para ESE efetuado pelo RNOD, esta fragilidade do ESE manifestou-se na representação de informação codificada como o ISBN, ISSN e ISMN (campos 010 $a, 011 $a e 013 $a) que, não tendo no ESE elementos específicos, foram mapeados indistintamente para europeana:unstored. Por outro lado, os campos relativos ao local de publicação (102 $a e 210 $a) também não têm mapeamento no esquema ESE. O campo 675 (CDU) também não tem mapeamento direto para ESE , sendo a notação incluída no elemento dc:subject, de forma indiferenciada e misturada com todos os outros assuntos mapeados do bloco 6. Por último, não existe no formato ESE forma de representar registos de autoridade, nem tão pouco diferenciar entradas de autoridade do bloco 7 ou do bloco 5 do UNIMARC. No caso do mapeamento RNOD todas as entradas de autor do bloco 7 foram mapeadas para os elementos dc:creator e dc:contributor, deixando portanto de ser identificáveis como entradas de autoridade.

Como resulta dos exemplos acima referenciados, as dificuldades no mapeamento de elementos com valores codificados ou normalizados internacionalmente (ISBN, país de publicação, CDU, etc) e a ausência de ligação a ficheiros de autoridade da BNP que estão representados no VIAF (The Virtual International Authority File), empobrece os dados agregados pela Europeana, nomeadamente no que respeita à sua disponibilização como dados ligados.

Europeana Data Model (EDM)

Não existe uma substituição do esquema ESE pelo Europeana Data Model (EDM). Com efeito, os dados continuam a ser agregados em formato ESE, podendo depois ser estruturados de acordo com o modelo EDM que visa reverter o efeito redutor do esquema ESE relativamente à riqueza dos metadados na sua origem e, bem assim, enriquecer esses mesmos metadados ligando-os a outros dados de fontes terceiras, sejam elas internas (outros fornecedores de dados) ou externas à Europeana (outros dados na Web). O EDM visa, por outro lado, suportar a representação de objetos mais complexos ou de estrutura hierárquica (partes de um livro, arquivos, etc). Ou seja, o formato ESE continua a ser o esquema base da Europeana, podendo aplicar-se o EDM como um modelo de dados mais flexível e extensível do que esquema ESE, que permite melhorar qualitativamente os processos como a Europeana recebe, gere e publica os dados agregados (Isaac, 2011).

O esquema EDM aplica os princípios da Web Semântica e, por isso, é a base do projeto de disponibilização dos dados da Europeana como dados ligados, que abordaremos no ponto seguinte deste artigo.

Requisitos do EDM. Para que o EDM permitisse representar diferentes perspetivas sobre um objeto cultural, objetos complexos e informação de contexto, foram seguidos os seguintes princípios no desenho do modelo de dados (Isaac, 2011, 2012):

- Distinguir o objeto real que é descrito das suas representações digitais
- Distinguir o objeto ou item dos registos de metadados que o descrevem
- Permitir a ingestão de múltiplos registos para um mesmo item, mesmo que contenham declarações contraditórias sobre o mesmo
- Suportar objetos compostos por outros objetos
- Compatibilidade com níveis diferentes de abstração das descrições
- Propiciar um esquema de metadados extensível, i.e. que possa ser especializado
- Suportar informação de contexto, incluindo conceitos de vocabulários controlados

Modelo conceptual. Visando o EDM suportar a integração dos diferentes modelos usados para dados no âmbito do património cultual, a definição das entidades e das relações a que os metadados se referem, efetuou-se com base no modelo OAI-ORE (Open Archives Initiative Object Reuse and Exchange) [5], relativo à descrição e partilha de agregações de recursos Web. Com efeito, para este modelo, as agregações são consideradas como objetos digitais compostos e podem referir-se a recursos de diversa natureza: texto, imagem, dados, vídeos, etc. O objetivo é expor os conteúdos dessas agregações a aplicações que possam apoiar o seu depósito, partilha, visualização, reutilização e preservação.

As entidades definidas no modelo EDM são, pois, o próprio objeto de património cultural (um livro, filme, pintura, etc), as representações digitais desse objeto e as agregações das várias descrições do objeto (Isaac, 2011). Este modelo conceptual permite que os vários elementos EDM não tenham de ser usados agrupados num mesmo registo, ao contrário do que acontecia no esquema ESE, podendo distinguir-se claramente a que entidade (objeto real, objeto digital ou descrição) se aplica determinada propriedade Isaac, 2011).

Vocabulário de elementos de dados. Os elementos do EDM têm classes e propriedades próprios e, também, reutilizados de outros esquemas: Open Archives Object Reuse and Exchange Model (OAI-ORE), Dublin Core e SKOS. A especificação de todos os elementos consta do documento "Definition of the Europeana Data Model elements"[6] e na ontologia Owl EDM[7].

Classes de recursos EDM (Isaac, 2012):

- Provided Cultural Heritage Objet (edm:ProvidedCHO)
  Representa o objeto descrito pelos dados agregados na Europeana. Corresponde ao objeto real digitalizado ou ao objeto nascido digital.
  É a partir do item que serão feitas as ligações para outros dados sobre o mesmo objeto, usando a propriedade owl:sameAs.
  Cada CHO tem um URI, atribuído pela Europeana.
  Não há metadados descritivos diretamente relacionados com o item. Essas descrições aplicam-se ao proxy que representa uma determinada visão do objeto.
- Proxies (ore:Proxy)
  Este recurso representa uma descrição do objeto. Há um proxy distinto para cada descrição do item. Um proxy está ligado a um recurso pela propriedade ore:proxyFor e a uma agregação pela propriedade ore:ProxyIn
  - Proxy do fornecedor de dados
    Classe usada para separar o "item" das declarações descritivas ESE feitas pelo fornecedor de dados. Esta separação permite que possam ser distinguidas diferentes descrições de um mesmo item.
  - Proxy da europeana
    Classe usada para separar o "item" das declarações descritivas ESE feitas pela Europeana, nomeadamente ligações para locais, pessoas, conceitos ou períodos de tempo contidos em conjuntos de dados externos. Cada proxy europeana tem um URI atribuído pela Europeana.
- Agregações (ore:aggregation)
  Cada agregação representa o conjunto dos elementos descritivos do "item" criados pelo fornecedor de dados ou pela Europeana. Cada agregação só pode ter um proxy por objeto (edm:ProvidedCHO)
  - Agregações do fornecedor de dados
    Representam o conjunto de informação descritiva do "item" e das suas representações digitais, agregada por fornecedor de dados. Inclui informações de direitos e dados de proveniência.
  - Agregações da Europeana (edm:EuropeanaAggregation)
    Representa cada conjunto de todas as informações descritivas do "item", criadas por todos os fornecedores de dados e pela própria Europeana. Cada agregação europeana tem um URI atribuído pela Europeana.

Os metadados descritivos agregados pelo EDM podem ter a seguinte tipologia:

- Descrições centradas no objeto: propriedades descritivas do objeto. Podem ser utilizadas propriedades Dublin Core (dc ou dcterms), como subpropriedades de propriedades EDM. Como as propriedades DublinCore/ESE podem ter na sua origem apenas valores textuais, para enriquecer essas descrições com ligações a outras entidades, a Europeana usará instâncias das classes EDM de recursos contextuais (local, pessoa, tempo) abaixo referidas. No entanto, o EDM recomenda que, sempre que possível, as propriedades ESE/DC sejam usadas com valores que sejam recursos (por exemplo, o elemento dcterms:creator ter com valor um recurso VIAF).
- Descrições contextuais: não descrevem o objeto em si mesmo, mas sim outros recursos da descrição. As classes EDM que representam estas descrições são edm:Agent (representação de pessoas ou organizações), edm:Event (acontecimentos), edm:Place (locais); edm:timeSpan (data ou períodos de tempo); skos:Concept (thesauri, classificações, etc)
- Descrições centradas em eventos: propriedades edm:waspresentAt (ligação do recurso a um evento em que o mesmo esteve envolvido); edm:happenedAt (ligação entre o recurso e um local); edm:occurredAt (ligação entre o recurso e um período de tempo)

Projeto LOD Europeana. O projeto piloto Linked Open Data consiste na primeira aplicação prática do modelo EDM, através da conversão de conjuntos de registos ESE no formato EDM. Em Junho de 2011, foram disponibilizados os primeiros dados ligados da Europeana, relativos a 2,4 milhões de objetos de fornecedores de dados que aderiram ao projeto piloto e em que se incluem os objetos da Biblioteca Nacional Digital (BND) agregados pelo TEL. Os dados ligados abertos da BND estão disponíveis em formato RDF/XML em http://data.europeana.eu/download/1.2/datasets/rdf/92039_Ag_EU_TEL_a0493_Portugal.rdf.gz
A adoção imediata do LOD no sistema Europeana que está em produção não foi possível porque não existiam metadados expressos em EDM, não havia ligações para outros recursos e porque era necessário alterar os acordos de fornecedores de dados para garantir o respetivo licenciamento aberto. Por este motivo, foi feita uma experiência piloto de disponibilização de dados agregados pela Europeana como LOD separadamente do sistema em produção, estando os mesmos disponíveis em http://data.europeana.eu. (Isaac, 2012).
O processo de transformação de dados consistiu na extração de um subconjunto de dados ESE XML, no seu mapeamento para EDM RDF, com base na criação das entidades EDM (CHO, agregações e proxies) e na atribuição de URIs a essas entidades.

Análise geral do protótipo EDM-LOD

Na análise desta primeira implementação do modelo EDM feita pela própria equipa técnica da Europeana (Isaac, 2012), foram identificados problemas de conectividade dos dados, originados pelo facto de os dados de origem mapeados para ESSE terem valores literais, i.e. não consistirem em recursos contextuais (autoridades, thesauri, etc):

- Perda de conectividade interna
  A conectividade interna é muito baixa, não havendo ligações entre os "items" ou os proxies que os

representam, por não estarem representados os recursos contextuais que podem propiciar ligações internas entre "items" (por exemplo, locais e assuntos). As formas de ultrapassar este problema estão a ser analisadas pela Europeana.

- Perda de conectividade externa

  A Europeana está a proceder ao enriquecimento dos dados agregados, através da geração automática de ligações para fontes externas a partir de elementos específicos ESE, como por exemplo ligações para o VIAF. Estas ligações externas integrarão a versão 1.2 dos dados ligados Europeana, estando um primeiro conjunto de exemplos demonstrativos disponível em: http://data.europeana.eu/download/1.2/. Nestes primeiros exemplos são feitas ligações para recursos contextuais externos de lugar (Geonames), tempo (Semium), conceitos (Gemet) e pessoas (Dbpedia).

O modelo de dados EDM, ao separar as diferentes fontes de dados da descrição do recurso, associando-as a agregações e proxies, tem como resultado a criação de redes complexas de agregações, proxies e outros recursos, levantando obstáculos ao acesso e reutilização desses dados e aumentando a complexidade dos grafos. A comunidade de utilizadores de dados ligados não está habituada a uma mediação de proxies, estando neste momento a ser equacionada uma solução para este problema que, por um lado, evite a duplicação de declarações dos associadas aos vários proxies e que, por outro, permita ligar directamente essas declarações ao recurso descrito (Isaac, 2012). A "Figura 4" apresenta um exemplo de mediação por proxy, no caso da representação do "local" associado ao registo descritivo de objeto da BND agregado na Europeana. Neste exemplo, a Europeana efetuou uma ligação externa para o valor "Portugal" no Geonames. Contudo, esta ligação efetua-se não directamente a partir do elemento dct:spatial que integra a informação descritiva do CHO (identificada com (1) na Figura 4), mas sim a partir do proxy que representa essas informações descritivas (identificado com (2) na Figura 4).

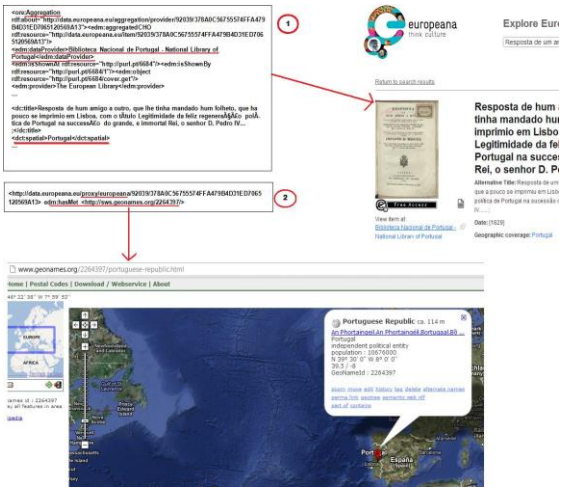

Figura 4 - Exemplo de registo BND ligado a elemento data

Por outro lado, há problemas de persistência dos URIs utilizados no protótipo EDM LOD, pois foram utilizados URIs derivados dos URIS do sistema Europeana que está em produção (para evitar que se perdesse a única ponte entre os dois sistemas) e estes têm manifestados problemas de persistência sempre que uma colecção é actualizada (Isaac, 2012).
Em conclusão, é necessário simplificar o modelo conceptual, aumentar a conectividade interna e externa, partindo da agregação de dados transformados em EDM logo na origem, i.e. sem serem convertidos em ESE (Isaac, 2012).

Análise do protótipo EDM-LOD pela comunidade de bibliotecas

O TEL tem, através do projeto Europeana Libraries, vindo a acompanhar a elaboração e a analisar o modelo EDM, na perspetiva do alinhamento dos dados produzidos pela comunidade de bibliotecas com o EDM, criando um perfil de aplicação específico para este tipo de dados. Seguidamente serão apresentadas as principais questões suscitadas pela implementação do EDM, no contexto da criação de um perfil EDM para as bibliotecas e que constam do relatório D5.1 do projeto Europeana Libraries[8].
No contexto das bibliotecas, os autores deste relatório afirmam ser necessário clarificar a definição da classe edm:ProvidedCHO no contexto do modelo conceptual FRBR. Com efeito, no caso dos recursos bibliográficos esta entidade não se refere ao "Item", mas sim à "Edição" ou "Manifestação". Para a representação de livros raros ou únicos, sugere-se a utilização da classe edm:PhysicalThing. O ideal seria que a Europeana definisse de forma standard a representação de entidades FRBR no modelo EDM (Angjeli, 2011).
É também necessária a especificação de extensões às propriedades EDM, que permitam, por exemplo, a utilização de propriedades de esquemas reconhecidos no domínio das bibliotecas relativamente ao edm:ProvidedCHO. O EDM deve incluir elementos que permitam representar o local de publicação. Relativamente à descrição de recursos Web (nascidos digitais ou digitalizados), é necessário acrescentar propriedades como data e formato, bem como a possibilidade de representar sequências de ficheiros digitais. Ainda neste contexto, a declaração de direitos de acesso também tem de ser aplicável aos recursos web e não apenas às agregações, como atualmente está especificado no EDM (Angjeli, 2011).
Por último é abordado o problema da utilização de valores literais em propriedades Dublin Core contrariamente ao especificado no DCAM. Esses valores literais terão de ser mantidos, porque foi assim que foram mapeados da origem para o esquema ESSE. Para "remediar" esta situação, serão atribuídos URIs relativamente a esses valores, num processo de enriquecimento de dados que está em desenvolvimento no âmbito do projeto Europeana Libraries e que, conforme já foi acima referido a propósito do primeiro teste LOD, será disponibilizado na versão 1.2 dos dados ligados Europeana (Angjeli, 2011).

CONCLUSÃO

A adoção de uma arquitetura LOD para a Europeana visa facilitar a agregação de dados de diferentes setores culturais, o enriquecimento dos dados agregados com recursos de contextualização disponíveis na web e a reutilização de dados da Europeana como recursos de contextualização para outros recursos da web. No caso específico dos dados bibliográficos, a sua disponibilização na web semântica implica decisões ao

nível da formalização de metadados que só agora começam a ser equacionadas e que, a par de alterações já em curso, tornarão os dados criados pelas bibliotecas e os recursos por eles descritos mais visíveis e reutilizáveis fora do seu contexto original. Apesar de a tecnologia *Linked Data* não estar ainda completamente madura e de as bibliotecas e a Europeana estarem ainda numa fase inicial e experimental de disponibilização dos seus dados de acordo com aqueles princípios, esta é uma oportunidade para a informação bibliográfica deixar de apenas "estar na web e passar a ser da web". Porque é na web que os nossos utilizadores tradicionais procuram informação, cada vez mais em primeiro lugar e numa multiplicidade de sistemas; e é na web que estão os sistemas que podem entre si reutilizar e interligar informação, que enriquecida melhor serve o utilizador. Este artigo pretendeu, a partir do estudo de caso da Europeana, apresentar os modelos e formatos de metadados da web semântica, com base nos quais se poderão construir serviços que tornem os conteúdos digitais do sector cultural mais fáceis de encontrar e utilizar, mas visou também, mais do que encontrar respostas e soluções, abordar as questões e os desafios colocados por esta transformação.

REFERÊNCIAS

ANGJELLI, Anilla [et al.] - **D5.1 Report on the alignement of library metadata with the Europeana Data Model (EDM)**. S.l.: Europeana Libraries Project, 2011. Disponível em: http://www.europeana-libraries.eu/documents/868553/1eade085-34ac-487f-82af-d5cd2545e619

BAKER, Thomas [et al.] - **Library Linked Data Incubator Group final report**. S.l.: W3C, 2011. Disponível em: http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora - The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. "Scientific American". May 2001. Disponível em: http://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf

BERNERS-LEE, Tim - **Linked Data**. September 27, 2006. Disponível em: http://www.w3.org/DesignIssues/LinkedData.html

BYRNE, Gillian; GODDARD, Lisa - The strongest link: libraries and linked data. "D-Lib Magazine". Vol. 16, n.º 11/12 (Nov/Dec 2010). Disponível em: http://www.dlib.org/dlib/november10/byrne/11byrne.html

COYLE, Karen - **FRBR, FRAD, ISBD in LD by BNE.** May 2012. Disponível em: http://kcoyle.blogspot.pt/2012/05/frbr-frad-isbd-in-ld-by-bne.html

COYLE, Karen - Understanding the Semantic Web: bibliographic data and metadata. "Library Technology Reports". Chicago: American Library Association. ISSN 0024-2586. Vol. 46, n.º 1 (Jan. 2010)

COYLE, Karen - RDA vocabularies for a Twenty-First-Century Data Environment. "Library Technology Reports". Chicago: American Library Association. ISSN 0024-2586. Vol. 46, n.º 2 (Feb./Mar. 2010)

DUNSIRE, Gordon [et al.] - Linked data vocabulary management: infrastructure support, data integration, and interoperability. "SQ Information Standards Quarterly". Vol. 42, n.º 2/3 (Spring/Summer 2012). Disponível em: http://www.niso.org/publications/isq/2012/v24no2-3/dunsire/

GROAT, Greta de - **Future directions in metadata remediation for metadata aggregators.** S.l.: Digital Library Foundation, 2009. ISBN 978-1-933645-07-5. Disponível em: http://old.diglib.org/aquifer/dlf110.pdf

HAWTIN, Rob [et al.] - **Review of the evidence for the value of the "linked data" approach: final report to JISC.** September 20, 2011. Disponível em: http://repository.jisc.ac.uk/559/

HERMAN, Ivan; BEEMAN, Hatley - **Open Data in practice: tutorial at the W3C Track at WWW2012.** April 17, 2012. Disponível em: http://www.w3.org/2012/Talks/0417-LD-Tutorial/Intro.pdf

ISAAC, Antoine; CLAYPHAN, Robina; HASLHOFER, Bernhard - Europeana: moving to Linked Open Data. "SQ Information Standards Quarterly". Vol. 42, n.º 2/3 (Spring/Summer 2012). Disponível em: http://www.niso.org/apps/group_public/download.php/9407/IP_Isaac-etal_Europeana_isqv24no2-3.pdf

ISAAC, Antoine; CLAYPHAN, Robina - **Europeana Data Model Primer v.1.0.** S.l.: Europeana, 2011. Disponível em: http://pro.europeana.eu/documents/900548/770bdb58-c60e-4beb-a687-874639312ba5

JOINT INFORMATION SYSTEMS COMITTEE - **Open bibliographic data guide.** S.l., JISC: 2010. Disponível em: http://obd.jisc.ac.uk/aggregation

MCKENNA, Gordon; STEIN, Regine - **Best practice report on cultural heritage linked data and metadata standards.** S.l.: Linked Heritage Project, 2011. Disponível em: http://www.linkedheritage.eu/getFile.php?id=229

RIGHTSCOM - **Information gathering exercise for the Resource Discovery Task Force: final report**. London: Rightscom, 2009. Disponível em: http://rdtf.jiscinvolve.org/wp/files/2009/09/jisc-resource-discovery-report-final-20090908.pdf

WALK, Paul - **Building metadata aggregation services for resource discovery.** S.l.: UKOLN, 2011. Disponível em: http://tinyurl.com/6zl8363

WEINBERGER, David - **Too big to know: rethinking knowledge now that facts aren't facts, experts are everywhere and the smartest person in the room is the room.** New York: Basic Books, 2011

[1] http://www.bl.uk/schemas/bibliographic/blterms-v1-2.rdf

[2] http://data.bnf.fr/ontology/bnf-onto/

[3] http://pro.europeana.eu/documents/900548/dc80802e-6efb-4127-a98e-c27c95396d57

[4] http://rnod.bnportugal.pt

[5] http://www.openarchives.org/ore/1.0/datamodel.html

[6] http://pro.europeana.eu/documents/900548/bb6b51df-ad11-4a78-8d8a-44cc41810f22

[7] http://europeanalabs.eu/browser/europeana/trunk/ROOT/src/main/webapp/schemas/edm/rdf/

[8] http://www.europeana-libraries.eu/documents/868553/1eade085-34ac-487f-82af-d5cd2545e619