XVI ENANCIB 2015
Informação, Memória e Patrimônio: do documento às redes
26 a 30 de outubro - João Pessoa - PB

XVI Encontro Nacional de Pesquisa em Ciência da Informação **(XVI ENANCIB)**
**ISSN 2177-3688**

**GT 8 – Informação e Tecnologia**
Comunicação Oral

# WEB SEMANTICA, DADOS LIGADOS E DADOS ABERTOS: UMA VISÃO DOS DESAFIOS DO BRASIL FRENTE ÀS INICIATIVAS INTERNACIONAIS[1]

## *SEMANTIC WEB, LINKED DATA AND OPEN DATA: AN OVERVIEW OF BRAZIL'S CHALLENGES AHEAD INTERNATIONAL INITIATIVES*

**Jose Eduardo Santarem Segundo, USP**
santarem@usp.br

**Resumo:** Publicar dados, sejam eles de origem pública ou privada, em formato aberto e baseado em estrutura semântica tornou-se um dos desafios não apenas das pesquisas em Ciência da Informação, mas também da comunidade que tem a necessidade de consumi-los. Junta-se a isso uma onda de publicação de dados abertos governamentais e necessidade de acesso a objetos digitais de cultura e patrimônio, tendência europeia, que se alastra pelo mundo. O objetivo principal deste trabalho é fazer uma análise da situação da publicação de dados no Brasil e apresentar iniciativas internacionais que tem despontado como vanguarda neste processo. A metodologia utilizada é parte descritiva, baseada em levantamento bibliográfico, e parte exploratória, baseada em análise de projetos e propostas de publicação de dados abertos e em formato semântico ou direcionados ao Linked Open Data (LOD). Com esta pesquisa foi possível entender que o Brasil ainda tem muito a melhorar na questão de abertura de seus dados, principalmente quando refere-se a modelos semânticos. Foi possível identificar também que iniciativas como Open Data Monitor e a estrutura organizacional do modelo EDM da Europeana podem ser utilizados como referência para o desenvolvimento dos projetos nacionais que pretendem fazer com os que dados cheguem aos usuários finais.

**Palavras-chave:** Linked Open Data. Web Semântica. Dados Abertos. EDM. Open Data Monitor.

**Abstract:** Post data, whether from public or private sources, in an open format and based on semantic structure has become one of the challenges not only of research in information science, but also of the community who have the need to consume them. Joins that a wave of publishing open government data and the need for access to digital objects of culture and heritage, European trend that is spreading around the world. The aim of this study is to analyze the data publication of the situation in Brazil and present international initiatives that has emerged as the vanguard in this process. The methodology used is descriptive, based on literature review and exploratory, based on analysis of open projects and data publishing proposals and semantic format or directed to the Linked Open Data (LOD). This research was possible to understand that Brazil still has much to improve on the question of opening

[1] O conteúdo textual deste artigo, os nomes e e-mails foram extraídos dos metadados informados e são de total responsabilidade dos autores do trabalho.

up your data, especially when it relates to semantic models. It was determined that initiatives such as Open Data Monitor and the organizational structure of EDM Europeana model can be used as reference for the development of national projects they want to do with that data arrives to end users

**Keywords:** Linked Open Data. Semantic Web. Open Data. EDM. Open Data Monitor.

## 1 INTRODUÇÃO

"Informação, memória e patrimônio: do documento as redes", tema do Enancib 2015, nos leva a refletir a questão das tecnologias como referência transformadora no processo de construção do conhecimento e também na contribuição para o acesso e preservação da memória e patrimônio de todas as comunidades ao redor do mundo.

O GT-8, constituído a partir de 2008, apresenta-se mais uma vez como elemento de suma importância na constituição das pesquisas que alicerçam o tema deste Enancib.

Esta pesquisa nasceu a partir do interesse deste pesquisador em conhecer mais profundamente a realidade de publicação de dados abertos e semânticos no contexto brasileiro e na busca por encontrar modelos internacionais que pudessem servir como referencia para o desenvolvimento brasileiro.

Entende-se que cultura de publicação de dados abertos no Brasil é muito tímida e quando acontece é carente de melhor infraestrutura, tanto do ponto de vista do armazenamento quanto da real possibilidade de acesso pelos usuários finais, a sociedade civil.

Este artigo tem como objetivo principal apresentar a realidade que envolve a publicação de dados abertos e o uso de tecnologias e conceito da Web Semântica no Brasil, assim como indicar iniciativas internacionais que tem se destacado e aparecido como vanguarda neste processo. Objetiva também destacar pontos importantes das iniciativas internacionais que podem ser utilizadas como referencia para o desenvolvimento de novos projetos de publicação de dados no Brasil.

A metodologia utilizada é parte descritiva, baseada em levantamento bibliográfico, com consultas às principais referências de dados abertos e Web Semântica, e parte exploratória, baseada em análise de projetos, navegação e estudo das propostas de publicação de dados abertos e em formato semântico ou direcionados ao Linked Open Data (LOD).

## 2 OS DADOS ABERTOS

O acesso a informação tem sido pautado como grande propulsor do desenvolvimento no século XXI. As instituições, sejam elas públicas ou privadas, tem investido na organização e no acesso a informação como o grande diferencial na tomada de decisão em várias de suas instâncias.

Há também no mundo uma tendência de publicação de dados governamentais, com o objetivo de criar a cultura de participação do cidadão na gestão do Estado, construindo um modelo conhecido como transparência.

A Lei de Acesso a Informação, nº 12.527, de 18 de novembro de 2011, no Brasil, constituiu-se a partir do movimento chamado de Dados Abertos (Open Data). Desde 2009 que países como Inglaterra e EUA avançam em modelo de gestão que tem como premissa ampliar a visibilidade de informações governamentais a fim de produzir efeitos que conduzam a população a contribuir com a eficiência e a transparência de seus governos, e principalmente de fortalecer a participação da sociedade em sua gestão. (SANTAREM SEGUNDO, 2013)

Atuando desde 2004 a Open Knowledge Foundation tem se dedicado a trabalhar com projetos que envolvem o conceito de conhecimento aberto. Segundo eles "Conhecimento Aberto é qualquer informação, conteúdo ou dados que as pessoas são livres para utilização, reutilização e redistribuição - sem qualquer restrição legal, tecnológica ou social".

Devido a fatores tais como o barateamento de computadores, dispositivos de armazenamento e o próprio desenvolvimento contínuo das TIC's, o volume de dados disponível por meio da infraestrutura da internet aumentou de forma muito expressiva. (RODRIGUES; SANT'ANA, 2012)

Santarem Segundo (2013, p. 34) diz que: "Certamente o avanço tecnológico e o crescente aumento do acesso a novos dispositivos que permitem conexão a Internet, tem qualificado a sociedade civil a acompanhar a publicação de dados pelo governo através de seus ambientes digitais, e em contra partida, ainda de forma tímida, alguns países tem procurado publicar suas informações de modo que esses dados possam ser consumidos pela sociedade".

Quando citamos a sociedade estamos falando não apenas de pessoas isoladamente, mas também de grupos organizados dentro da iniciativa privada, das organizações não governamentais, da esfera jornalística, da academia e de qualquer outra instância que tenha interesse nesse conjunto de informações, sejam elas para quaisquer fins, incluindo cruzamento e republicação destes dados.

Rodrigues, Sant'ana e Ferneda (2015, p. 34), diz que "A transparência das atividades e ações do Estado tem como uma de suas premissas fortalecer a participação dos cidadãos nesse novo modelo de administração pública. O fortalecimento pode ser garantido com a construção de ambientes democráticos que, dentre outras características, criem possibilidades de novos fluxos informacionais entre a administração do Estado e sociedade, garantindo assim uma maior visibilidade."

Segundo o MANUAL (2011), se houver boa utilização da tecnologia existente e

iniciativa da sociedade, os dados governamentais poderão ser cada vez mais benéficos para todos. Sua reutilização poderá garantir maior:

- Transparência: provendo melhor acesso aos dados.
- Participação: facilitando a educação pública, a democratização do conhecimento e a inovação.
- Colaboração: proporcionando contínua realimentação da sociedade e disseminação colaborativa do conhecimento.

O movimento de abertura de dados governamentais está embasado em 3 leis propostas pelo especialista em políticas públicas David Eaves:

- Se o dado não pode ser encontrado e indexado na web, ele não existe.
- Se não estiver aberto e em formato compreensível por máquina, ele não pode ser reaproveitado.
- Se algum dispositivo legal não permitir sua reaplicação, ele não é útil.

Apesar da clara necessidade de uso, dados abertos, especialmente os governamentais, constituem-se como um ótimo recurso, ainda timidamente explorado. Muitos indivíduos e organizações coletam uma ampla gama de diferentes tipos de dados para executar suas tarefas. O governo é particularmente importante nesse contexto, tanto por causa da quantidade e da centralidade dos dados que coleta quanto pelo fato de que tais dados são públicos, um direito garantido no artigo 5º da Constituição Federal brasileira (MANUAL, 2011).

A contextualização da necessidade de disponibilizar dados pode ser melhor entendida neste exemplo: O leitor sabe onde, na sua região, pode encontrar as melhores oportunidades de emprego e os locais mais arborizados? Sabe quando e como influenciar leis ou decisões públicas sobre temas com os quais se preocupa? Novas tecnologias tornam possível a construção de serviços para responder automaticamente a essas perguntas. Muitas pessoas, e não apenas os governos, seriam capazes de construir serviços assim. Mas, infelizmente, os dados necessários para a criação de projetos que atuem nesse sentido não estão disponíveis ou não são liberados em formato que torne possível o seu uso pela sociedade (MANUAL, 2011).

Quando se pensa na organização e posteriormente na recuperação desses dados, e há o conhecimento e entendimento das tecnologias e conceitos da Web Semântica, imediatamente se constrói um mundo de possibilidades acerca da publicação de dados.

Mesmo sabendo que os conceitos de dados abertos e de Web Semântica não têm relacionamento direto, quando se agrega as possibilidades que tangem esses dois conceitos,

vislumbra-se perspectivas sublimes de organização e acesso a dados.

Desde a proposta inicial de 2001, a Web Semântica vem ganhando força e agregando novas tecnologias, funcionalidades e evoluindo para tornar real o processo de construção de ambientes semânticos. Tecnologias como RDF (Resource Description Framework), XML (eXtensible Markup Language), OWL (Web Ontology Language) e todos os conceitos que as envolvem, ganham novas versões, descritos com clareza no W3C (World Wide Web Consortium), e tornam possível a materialização do conceito da Web Semântica. As tecnologias citadas estão diretamente relacionadas ao processo de construção da informação e armazenamento das mesmas, constituindo assim ambientes que possam ter conjunto de dados ligados semanticamente. (SANTAREM SEGUNDO, 2014).

## 3 LINKED OPEN DATA (LOD) E WEB SEMÂNTICA

Nos últimos anos vários elementos foram surgindo e ampliando o contexto da ideia original de Web Semântica de Berners-Lee. O W3C iniciou um processo de publicar, efetivar e disseminar um conjunto de tecnologias que foram se agregando em busca da Web Semântica ideal. Vários projetos ao redor do mundo também foram evoluindo de forma a constituir ambientes semânticos, tanto do ponto de vista de estrutura informacional quanto da possibilidade de recuperação semântica da informação.

Estruturar dados abertos de forma semântica não é apenas uma das formas de estabelecer a ligação entre o conceito de Dados Abertos e de Web Semântica, mas sim de estabelecer um modelo de estrutura de dados que favoreça o atendimento ao quinto princípio de dados abertos e também ao inciso da Lei de Acesso a Informação, que indicam a possibilidade dos dados serem processados por máquina, além da ligação entre informações de bases diferentes através de relacionamentos semânticos.

Essa característica torna os dados não apenas acessíveis e processáveis por máquinas, mas passíveis de processos de organização que podem facilitar a geração de novos dados, apresentação de resultados, relação com outros grupos de dados, aumento do conhecimento para tomadas de decisão, novos modelos de dados gerados a partir do relacionamento e cruzamento de dados de várias esferas governamentais, além da geração de novos modelos mentais de apresentação da informação de forma a facilitar o acesso dos dados pela sociedade civil.

Para disponibilizar dados numa estrutura semântica é necessário pensar em partes do modelo descrito por Berners-Lee em 2001, no chamado bolo de noiva, estrutura de camadas que apresenta a Web Semântica. Destaca-se neste quesito a linguagem RDF, também indicada para

representação de dados abertos, o uso de metadados e principalmente a construção e aplicação de ontologias de domínio.

Um dos principais objetivos da linguagem RDF é justamente criar uma rede de informações a partir de dados distribuídos.

Construir uma rede de informações onde os nós estejam semanticamente ligados, formando um grande grafo global, com informações advindas de várias fontes diferentes ao redor do planeta, é o conceito central da chamada Web de Dados. Um grafo é um modelo matemático muito poderoso que pode ser aplicado na resolução de um conjunto de problemas. É composto por um conjunto de vértices e arestas/arcos (SANTAREM SEGUNDO, 2010).

A Web de Dados pode compreender um conjunto imenso de possibilidades de oferta de dados, entretanto nessa pesquisa buscamos justamente entender o uso dos conceitos da Web Semântica aplicados a dados abertos.

Apesar das similaridades entre a Web de Dados e os conceitos e a estrutura do LOD, é importante ressaltar que são coisas diferentes.

O LOD, que atualmente apresenta-se como a melhor forma de materialização dos conceitos e tecnologias da Web Semântica, é um projeto, com um conjunto de normas a serem seguidas, que usa os mesmos princípios de ligação semântica da Web de Dados, entretanto tem particularidades específicas, indicando um grau de exigência maior na constituição de sua rede de interligações.

Segundo Heath e Bizer (2011), o LOD é um conjunto de melhores práticas para publicação e conexão de dados estruturados na Web, permitindo estabelecer links entre itens de diferentes fontes de dados para formar um único espaço de dados global.

Para Berners-Lee (2006),

> a Web Semântica não trata apenas de depósito de dados na web. Trata-se de fazer ligações, de modo que uma pessoa ou máquina possa explorar esse conjunto de dados. Com *LOD*, quando você tem um pouco de dados, você pode encontrar outros que estão relacionados.

A construção do *LOD* está baseada em quatro princípios publicados por Berners-Lee (2006):

(a) Usar URIs como nomes para os itens.

(b) Usar URIs HTTP para que as pessoas possam consultar esses nomes.

(c) Quando alguém consulta uma URI, prover informação RDF útil.

(d) Incluir sentenças RDF com links para outras URIs, a fim de permitir que itens relacionados possam ser descobertos.

O modelo de ligações do LOD e da Web de Dados, baseados principalmente no uso dos conceitos e aplicação da linguagem RDF, é muito similar ao conjunto de associações que os cérebros humanos executam para constituir e organizar a memória.

O RDF tem se posicionado como uma das principais tecnologias articuladoras no processo de construção da informação pela sua capacidade de associar sujeito, predicado e objeto, ou ainda como alguns nomeiam a tríade: recurso, propriedade e valor.

Segundo Lassila (1999),

> "RDF é uma aplicação da linguagem XML que se propõe ser uma base para o processamento de metadados na Web. Sua padronização estabelece um modelo de dados e sintaxe para codificar, representar e transmitir metadados, com o objetivo de torná-los processáveis por máquina, promovendo a integração dos sistemas de informação disponíveis na Web".

Apesar da constituição estrutural do modelo de organização dos dados ser realizado por triplas RDF, para que se possa publicar dados e principalmente torna-los disponíveis para serem recuperados com mais eficiência e eficácia, é necessário que o esquema lógico esteja sob uma ontologia, e de preferência utilizando-se de vocabulários de representação de conhecimento padronizados e reconhecidos internacionalmente.

Utilizar ontologias é uma das maneiras de se construir uma relação organizada entre termos dentro de um domínio, favorecendo a possibilidade de contextualizar os dados, tornando mais eficiente e facilitando o processo de interpretação dos dados pelas ferramentas de recuperação da informação.

Construir ontologias nem sempre é o melhor ou mais rápido caminho para disponibilizar dados em formato semântico. Em grande parte das vezes, utilizar-se de ontologias prontas e universalmente conhecidas, e reconhecidas por uma determinada comunidade, acelera o processo de publicação de dados, mas principalmente favorece o processo de recuperação da informação em ambientes semânticos.

Não utilizar esquemas lógicos como as ontologias e os principais vocabulários para descrever dados a serem publicados certamente é o pior caminho para publicação de dados em formato semântico, principalmente do ponto de vista da recuperação da informação.

A disponibilização de dados abertos em formato semântico tem na seleção e/ou construção de ontologias, seu principal desafio. Definir o melhor modelo lógico para estruturar e apresentar informações é uma tarefa que ainda desafia as pesquisas no mundo todo. Há uma infinidade de ontologias que já foram construídas e estão disponíveis para serem utilizadas nos mais variados domínios do conhecimento.

Da forma como a publicação de dados semânticos vem sendo realizada, não há uma regra geral para utilização de ontologias padronizadas, talvez esse nem seja o ponto principal para uma organização e interligação global de dados, entretanto o processo de recuperação dar-se-á com maior facilidade para os agentes computacionais de busca semântica, se houver uma mínima sintonia lógica entre conjuntos de dados do mesmo tipo.

No contexto de criação e uso de ontologias, é importante destacar o uso de vocabulários reconhecidos internacionalmente, esse com certeza é o ponto nevrálgico para que haja a constituição de modelos de dados semânticos, onde os significados dos recursos seja compreendido de forma global.

Quando nos referimos ao uso de "vocabulários", estamos utilizando o termo que o próprio W3C e as equipes do LOD e de Tim Berners-Lee utilizam para nomear os elementos que compõe os esquemas de metadados.

Contextualizando, apesar da criação de etiquetas para representação das propriedades do RDF ser livre, há um grupo de vocabulários que são utilizados em larga escala nas principais ontologias conhecidas e também em grande parte dos exemplos de publicação de *datasets* em formato semântico disponíveis na Web de Dados, por força de um reconhecimento imediato e global do significado que se pretende dar a ligação construída. Entende-se por *dataset* o conjunto de dados (base de dados) publicado por uma organização.

Apesar do enfoque dado neste texto seja em relação ao uso e construção de modelos semânticos de dados, com ênfase em dados abertos, há ainda um grande volume de dados sendo publicados de forma aberta que não são organizados utilizando-se dos recursos de ontologias e da linguagem RDF, entretanto estruturam-se sob modelos adaptados de metadados, conhecidos como Application Profiles.

Estes modelos adaptados de esquemas de metadados, para atender as necessidades de um conjunto específico de dados, é muito comum e torna difícil não apenas a interoperabilidade mas também a recuperação dos dados. Há pesquisas dedicadas exclusivamente a construir mecanismos automáticos que possam entender as regras e elementos usados nesses modelos.

Honma et al. (2014), construíram uma ferramenta exclusivamente para lidar com os Applications Profiles, com a proposta de "um método para extrair automaticamente as restrições e estruturas de esquemas adaptados, a partir de instâncias de metadados, com objetivo de reduzir o custo da extração e entendimento de esquema de metadados adaptados". Ou seja, o uso de Application Profiles é muito comum e tem se tornado um entrave para a recuperação e interligação de dados.

Retornando a ideia de uso das ontologias e dos vocabulários internacionalmente conhecidos, é importante destacar o uso destes vocabulários nos *datasets* publicados pelo LOD em sua ultima atualização, realizada em agosto de 2014, principalmente em comparação com sua versão anterior de 2011. Nesta versão, o LOD apresenta em seu diagrama 570 datasets, em quase dobrando o número de 295 datasets na versão de 2011.

Os vocabulários RDF[2], FOAF[3], RDFS[4], DC[5], e OWL[6] são os vocabulários mais utilizados pelos datasets. O vocabulário RDF aparece em 98,22% dos datasets. Comparando com o relatório de 2011, podemos afirmar que há uma tendência para a adoção de vocabulários reconhecidos internacionalmente. Por exemplo, enquanto o vocabulário FOAF foi usado por 27,46% de todos os datasets em 2011, é utilizado por 69,1% dos datasets em 2014. O mesmo acontece com o Dublin Core que é usado em 2014 por 56,01% dos conjuntos de dados e foi utilizado por apenas 31,19% em 2011. (SCHMACHTENBERG; BIZER; PAULHEIM, 2014).

Este conjunto de estatísticas nos leva a deduzir que há um esforço internacional para que a interligação semântica entre dados de fontes diferentes seja fortalecida e utilizada.

## 4 ESTADO DA ARTE DA PUBLICAÇÃO DE DADOS ABERTOS E SEMÂNTICOS NO BRASIL

No Brasil há uma tendência positiva para a publicação de dados abertos, apesar de ainda serem tímidos e isolados os esforços se comparado com outros países, principalmente do continente europeu.

Atualmente a Secretaria de Logística e Tecnologia da Informação do Ministério do Planejamento, Orçamento de Gestão desenvolve a Infraestrutura Nacional de Dados Abertos (INDA).

A INDA é um conjunto de padrões, tecnologias, procedimentos e mecanismos de controle necessários para atender às condições de disseminação e compartilhamento de dados e informações públicas no modelo de Dados Abertos. O principal projeto da INDA é o Portal Brasileiro de Dados Abertos[7], que tem o objetivo de ser o ponto central para a publicação, a busca e o acesso de dados públicos no Brasil (CARTILHA, 2011).

A arquitetura da Infraestrutura Nacional de Dados Abertos compreende todos os órgãos do governo, em todas as esferas e poderes, disponibilizando dados públicos à toda a sociedade,

[2] http://www.w3.org/1999/02/22-rdf-syntax-ns#
[3] http://xmlns.com/foaf/0.1/
[4] http://www.w3.org/2000/01/rdf-schema#
[5] http://purl.org/dc/terms/
[6] http://www.w3.org/2002/07/owl#
[7] http://dados.gov.br/

incluindo instituições privadas, organizações não governamentais e o próprio governo. A INDA servirá de referência para que os mais diferentes órgãos do governo sejam capazes de publicar de forma sistemática e padronizada, dentro de um conceito de boas práticas para disseminação da informação, o conjunto de dados que pretende disponibilizar. (Santarem segundo, ibersid)

Ao acessar o site do Portal Brasileiro de Dados Abertos é facilmente notado um trabalho inicial deste grupo, onde já é possível verificar dados de algumas instituições governamentais, porém em formato pouco amigável para consulta da população em geral. Atualmente existem 1047 datasets publicados, incluindo publicações constantes e recentes de novos datasets, o que indica que o trabalho de publicação de dados vem sendo efetivamente realizado.

Importante registrar que grande parte destes dados é apresentada em formato pouco amigável para recuperação e consulta pela população e para interligação com outros datasets. Os principais formatos encontrados são CSV e HTML, ou seja, todo o contexto de ligações semânticas, o uso de tecnologias e conceitos da Web Semântica e principalmente de esquemas de metadados discutidos até agora é simplesmente ignorado pela grande maioria das bases de dados.

O Brasil conta também com a INDE, Infraestrutura Nacional de Dados Espaciais, que tem o propósito de catalogar, integrar e harmonizar dados geoespaciais existentes nas instituições do governo brasileiro, produtoras e mantenedoras desse tipo de dado, de maneira que possam ser facilmente localizados, explorados e acessados para os mais diversos usos, por qualquer cliente que tenha acesso à Internet.

Parte dos dados publicados no portal de dados abertos do Brasil, que se refere a informações geoespaciais, é apresentada em um ambiente gráfico, o visualizador do INDE. O visualizador permite e facilita as consultas relacionadas a dados geográficos.

A figura 1 apresenta a área do visualizador onde o usuário pode selecionar os dados que deseja visualizar. Neste exemplo selecionou-se o item “Farmácia Popular no Brasil”. Como pode ser visto na Figura 1, a área destinada à seleção de dados pelo usuário permite ainda selecionar dados por tema, por instituição, e apresenta ainda abas para verificar quais dados foram selecionados e também as legendas disponíveis.

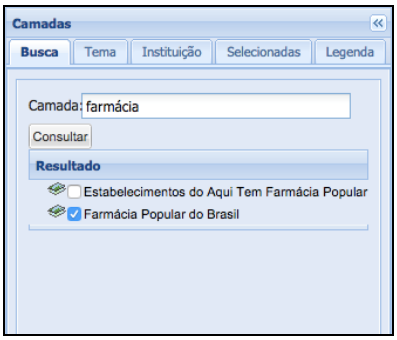


Figura 1 – Área para seleção de dados – Visualizador INDE.

Fonte: organizado pelo autor.

A Figura 2 apresenta a imagem que com as informações plotadas sobre o mapa do Brasil. Neste exemplo selecionou-se o item "Farmácia Popular do Brasil".

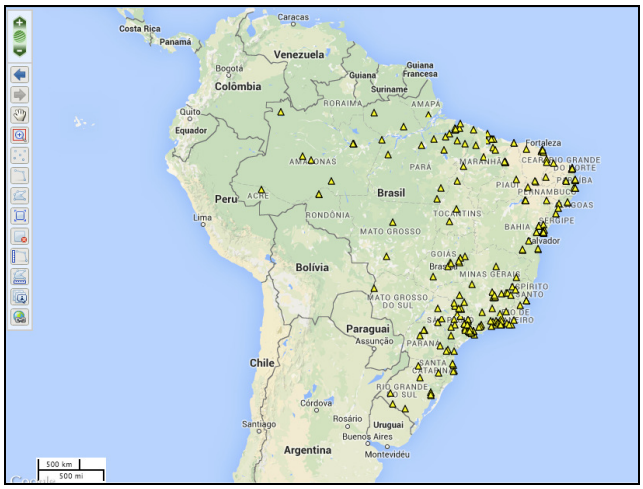


Figura 2 – Dados da Farmácia Popular do Brasil – Visualizador INDE.

Fonte: organizado pelo autor.

Além dos dados publicados no portal nacional, há outras iniciativas espalhadas pelo Brasil com o intuito de publicar dados abertos. O estado de Alagoas, mantem seu próprio portal de dados abertos[8]. O portal conta 130 conjuntos de dados e aparece referenciado no portal[9] de instâncias ativas do CKAN, software mais usado para publicação de catálogo de dados abertos ao redor do mundo. Outros estados e cidades brasileiras também mantém seus portais de dados abertos, alguns com dados bem recentes, outros com dados desatualizados, entretanto percebe-se que há inúmeras iniciativas para publicação de dados abertos no Brasil, entretanto ainda não se constituiu a cultura para que este fenômeno faça parte do dia a dia das organizações.

[8] http://www.dados.al.gov.br/
[9] http://ckan.org/instances/#

Ao analisar parte desses portais de dados abertos no Brasil, facilmente percebe-se uma preocupação exclusiva com a disponibilização dos dados, entretanto sem um formato pré-definido, uso de esquemas de metadados ou conceitos e tecnologias da Web Semântica, isso com certeza leva os dados brasileiros a terem representatividade quase nula no LOD.

## 5 INICIATIVAS INTERNACIONAIS PARA PUBLICAÇÃO DE DADOS ABERTOS E SEMANTICOS

A última parte desta pesquisa, desenvolvida de forma exploratória, apresenta iniciativas internacionais de publicação de dados. Apresentaremos aqui duas iniciativas relacionadas à publicação de dados: o projeto Open Data Monitor e também a estrutura de publicação de dados da Europeana.

### 5.1 OPEN DATA MONITOR

O continente europeu destaca-se frente a outros continentes em relação a publicação de dados abertos, o projeto Open Data Monitor[10] é uma iniciativa que visa qualificar a produção de dados abertos dos países da Europa propondo uma solução técnica abrangente, incluindo monitoramento, análise, recursos de relatórios e visualização dos dados publicados.

Além do lado técnico de contribuição com a melhoria na publicação de dados abertos, o Open Data Monitor visa oferecer aos visitantes uma visão geral dos recursos disponíveis de dados abertos, permitindo-lhes analisar e visualizar os catálogos de dados existentes usando tecnologias inovadoras. (OPEN DATA MONITOR, 2015).

O projeto Open Data Monitor é gerido por um consórcio formado pelas seguintes organizações: SYNYO, que é uma companhia de inovação com sede em Viena; o Open Data Institute (ODI), fundado por Tim Berners-Lee e o professor Nigel Shadbolt e responsável pelo portal de dados abertos do Reino Unido; o Athena Research and Innovation Center (GR) da Grécia; a University of Southampton (UK), reconhecida como uma das universidades de mais prestígio do Reino Unido; Potsdam eGovernment Competence Center (DE), instituto de pesquisa independente da Alemanha; a Cidade de Munique e a Red.es que é uma entidade publica atrelada ao Ministério da Industria da Espanha.

O funcionamento do Open Data Monitor é baseado em um framework que coleta dados de diversas fontes de dados abertos. Os dados e metadados recolhidos são então processados, harmonizados e organizados permitindo aos usuários finais entender mais sobre os repositórios de dados abertos regionais, nacionais e do continente europeu. Esse processo permite a

[10] http://opendatamonitor.eu

agregação de catálogos de uma mesma região ou país, permitindo comparações com outras regiões. (OPEN DATA MONITOR, 2015).

A análise dos dados e principalmente da estrutura de metadados recolhida, permitirá com que sejam identificadas lacunas que podem ser melhoradas através de uma proposta de harmonização da estrutura de metadados, com objetivo de tornar os metadados dos repositórios fonte mais interoperáveis e multilíngues possível para que possam ser mais consistentes e uteis.

Entende-se por processo de harmonização de metadados o trabalho com a heterogeneidade dos metadados recolhidos. Ele transforma e mapeia diferentes conjuntos de metadados a um esquema comum interno e, portanto, faz com que seja possível interpretar, agregar e utilizar os metadados de uma forma consistente e significativa para análise. A heterogeneidade pode envolver os metadados com diferentes esquemas, diferentes descrições de atributos ou diferentes representações de valores, por exemplo, abreviaturas usadas.

Todo o trabalho tem sido desenvolvido com o uso de padrões abertos, e a ferramenta CKAN foi a escolhida para constituir o catálogo de dados. Portanto todas as contribuições de desenvolvimento de software será integrada posteriormente como plug-in do CKAN. Isso garantirá com que tudo que for desenvolvido neste projeto esteja disponível para uma comunidade muito maior.

O Open Data Monitor permitirá uma variedade de funções de análise, tais como (OPEN DATA MONITOR, 2015):

- comparar órgãos públicos (nacional / local), indicando alterações e atualizações de catálogos;
- mostrar a qualidade dos metadados;
- realizar triagem de catálogos disponíveis para os domínios temáticos específicos;
- apresentar informações de licenciamento;
- tornar os dados disponíveis em formatos abertos específicos;
- mostrar a percentagem da população que tem acessado os dados;
- apresentar uma pontuação automática (score) para analisar o nível de abertura dos dados;

O Open Data Monitor apresenta os dados em dois níveis diferentes de agregação, um deles em nível nacional e outro em nível geral, ou seja, colocando os países em nível de comparação. A tela principal que apresenta o nível europeu de agregação oferece ao usuário um conjunto de dados baseados nas principais métricas utilizadas: licenças abertas, dados legíveis por máquina, acessibilidade e adequação dos metadados a um núcleo comum. A Figura 3

apresenta o nível europeu de informações sobre os dados, onde pode ser visto o mapa ao centro, com informações sobre o uso de licenças abertas em cada país. Na borda inferior da Figura 3 apresentam-se os indicadores de qualidade, ao clicar em cada um deles o mapa apresenta os números relativos ao indicador selecionado. Ainda na figura 3 é possível verificar no lado direito as informações globais sobre os dados do Open Data Monitor.

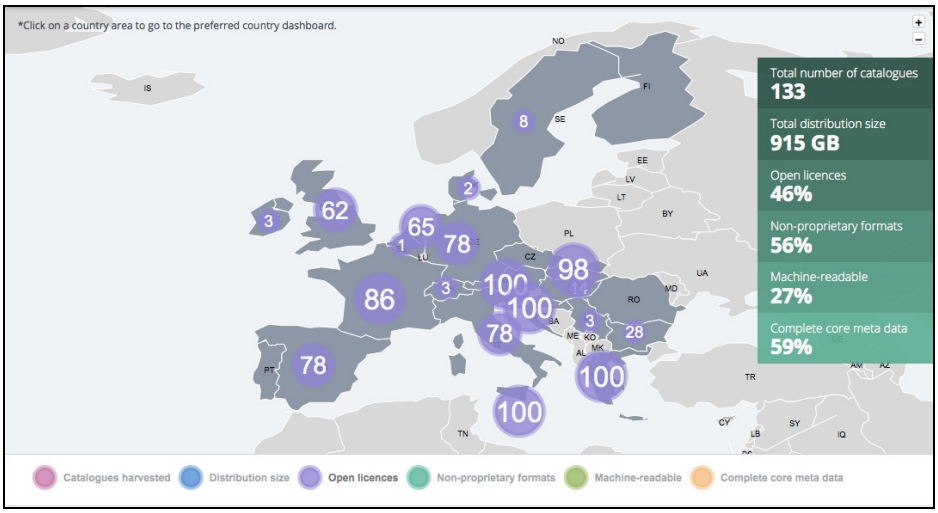


Figura 3 – Open Data Monitor: agregação de dados europeus de licenças abertas, por país.

Figura 3: organizado pelo autor.

A Figura 4 apresenta uma parte da tela disponível na plataforma principal do Open Data Monitor, nela é apresentado um "ranking" de abertura e disponibilização de dados baseado nas métricas definidas, observa-se que a "classificação final" é orientada pelo último item, nomeado de "Overall quality score".

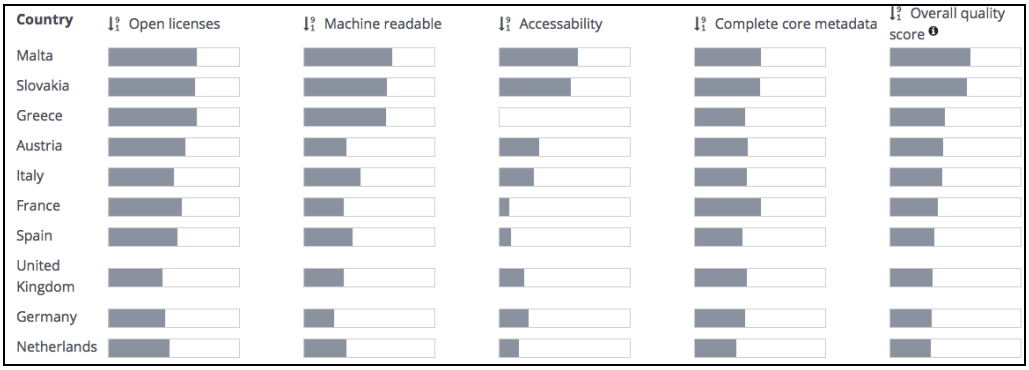


Figura 4 – Ranking com uma pontuação geral, baseado nas métricas do Open Data Monitor.

Fonte: organizado pelo autor.

Ao selecionar alguns dos países do continente europeu, é possível visualizar os dados no nível de agregação nacional. A figura 5 apresenta os dados específicos da Alemanha. Na Figura

5 é possível identificar que o país atende a 50 de 100 das métricas de qualidade, e também é possível verificar o quanto os repositórios alemães atendem individualmente as métricas definidas pelo Open Data Monitor.

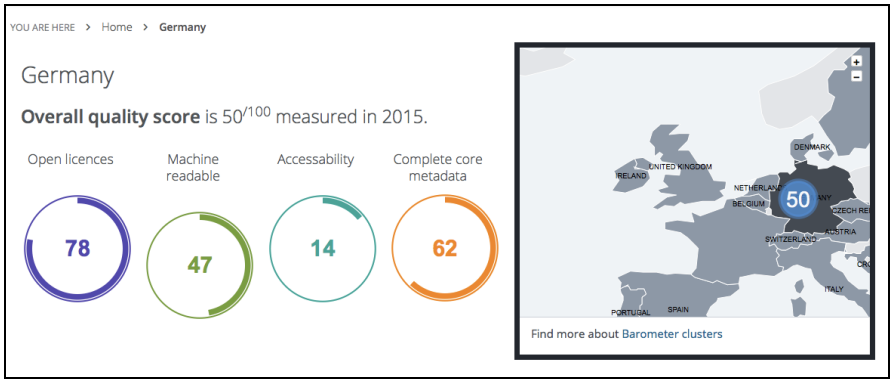


Figura 5 – Dados da Alemanha no Open Data Monitor.

Fonte: organizado pelo autor.

O projeto ainda encontra-se em fase inicial, entretanto verifica-se que a iniciativa tem apresentado resultado, estimulando os repositórios de dados abertos a se adequarem às métricas propostas e principalmente dando feedback às instituições governamentais no sentido de evoluir com a publicação de dados em todo continente europeu.

## 5.2 MODELO SEMÂNTICO DE DADOS DA EUROPEANA – EDM (EUROPEANA DATA MODEL).

Humanidades Digitais é o termo utilizado para descrever uma área em expansão, de pesquisa e ensino, que atua sobre a intersecção e aplicação de tecnologias às ciências humanas. Uma de suas premissas é associar técnicas computacionais como mineração de dados, recuperação da informação e visualização digital com os produtos gerados pelas ciências humanas em áreas como: história, arte, filosofia, música, literatura entre outras.

A Europeana[11] é uma biblioteca digital europeia que dá acesso a mais de 30 milhões de bibliotecas, arquivos, museus e objetos audiovisuais de 36 países diferentes. Diferente de todas as iniciativas apresentadas neste texto até agora, este projeto tem uma particularidade que é a disponibilização de dados relacionados ao conceito das Humanidades Digitais.

Os objetos disponíveis na Europeana são digitalizados e descritos por provedores de conteúdo em diferentes formatos de metadados. Agregadores nacionais ou de domínio entregam o objeto de metadados para a Europeana no formato EDM (Europeana Data Model). (EDM PRIMER, 2013).

[11] http://www.europeana.eu/

O EDM é resultado da melhoria do Europeana Semantic Elements (ESE), o modelo de dados com que a Europeana começou sua vida. Cada um dos diferentes setores representados na Europeana utiliza modelos de dados diferentes, e o ESE reduzia a representação destes a um denominador comum. O EDM reverte este abordagem redutora e é uma tentativa de transcender as perspectivas de informação dos setores que estão representados na Europeana - os museus, arquivos, coleções audiovisuais e bibliotecas. (EDM PRIMER, 2013)

É muito importante ressaltar que o EDM adota uma estrutura baseada em modelos da Web Semântica, usando vocabulários reconhecidos internacionalmente, que possibilitam ao modelo uma riqueza capaz de acomodar normas comunitárias específicas como LIDO para museus, EAD para arquivos e METS para bibliotecas digitais.

A estrutura de funcionamento do EDM está principalmente na criação de um núcleo de descrição de cada objeto, baseado em três classes, nomeadas de Core Classes. Para aproveitar a riqueza de descrição de alguns objetos, que vem de fontes que trabalham exaustivamente na descrição destes objetos, o EDM tem em sua constituição o que chama de Contextual Classes. Esse modelo tem sido a chave de sucesso para a descrição e interligação semântica entre recursos na Europeana.

A Europeana, com base no uso do EDM, apresenta-se no LOD como um dos datasets com maior quantidade de registros, disponibilizando inclusive uma interface do tipo SPARQL EndPoint para consulta a seus dados através de queries SPARQL.

O SPARQL é um conjunto de especificações que fornecem linguagens e protocolos para consultar e manipular o conteúdo publicado em RDF na Web. (SPARQL, 2013)

O uso de um modelo de dados construído com base na Web Semântica permite uma imensidade de ligações entre os dados vindos de bases diferentes, disponíveis dentro da própria base da Europeana, e também permite com que os dados sejam ligados a bases de dados que estão fora da Europeana, como é possível notar no LOD. Dessa forma é possível vislumbrar ligações entre obras de um mesmo artista que estejam em locais diferentes, ou então expressões diferentes de uma mesma obra interligadas através das propriedades do EDM.

Há um conjunto de documentos disponíveis no portal da Europeana (Europeana Data Model Documentaion[12]) que apresenta toda a estrutura de funcionamento e mapeamento do EDM. Estes documentos estão livres para serem utilizados como padrão de referencia para outros ambientes similares ao redor do mundo.

[12] http://pro.europeana.eu/page/edm-documentation

A Figura 6 apresenta um pequeno exemplo de estrutura da Europeana baseada em informações da obra Mona Lisa, de Leonardo da Vinci. Observa-se que estão agregados na mesma informação disponível pela Europeana os dados pertencentes ao Jaconde Database e a base de dados do Museu do Louvre. Na apresentação dos dados pela Europeana, as informações vindas das duas bases se completam usando o modelo EDM.

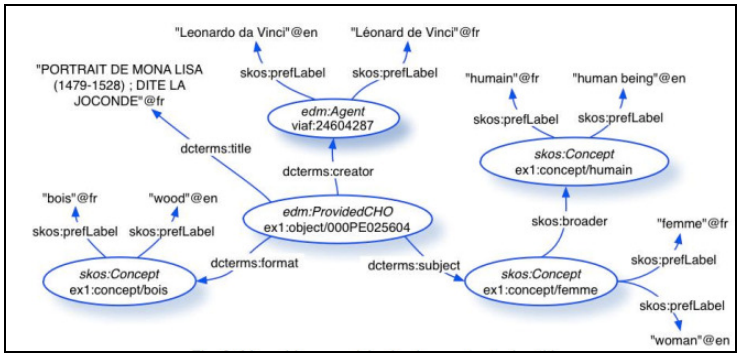


Figura 6 – Dados da obra Mona Lisa de Leonardo da Vince apresentados no padrão EDM.
Fonte: Europeana Data Model Primer, 2013.

O uso de um modelo construído sob os conceitos da Web Semântica favorece não apenas a disponibilização de dados à comunidade, mas principalmente a interligação entre esses dados, permitindo com que a recuperação da informação seja realizada de forma mais rica.

O EDM torna a estrutura e política de dados da Europeana um exemplo a ser seguido, colocando-a como vanguarda na publicação de dados de Humanidades Digitais no LOD.

## 6 CONSIDERAÇÕES FINAIS

A pesquisa realizada mostra efetivamente que o Brasil tem trabalhado na questão dos dados abertos, com projetos governamentais e também vindas da iniciativa privada, entretanto fica claro que há uma forte tendência de disponibilização dos dados sem a preocupação com o acesso e consumo destes dados.

Verifica-se que grande parte dos conjuntos de dados brasileiros disponibilizados publicamente não foi estruturada seguindo esquemas e padrões de metadados reconhecidos internacionalmente. Também não houve nenhuma preocupação em estabelecer ontologias ou aplicar tecnologias da Web Semântica, para que pudessem ser disponibilizados e integrados ao LOD.

As iniciativas internacionais apresentadas são referências importantes para serem seguidas e utilizadas como bons exemplos para o desenvolvimento dos dados abertos e semânticos no Brasil.

O projeto Open Data Monitor apresenta-se como um marco para a publicação dos dados abertos europeus, pois integra várias bases de dados, permitindo a agregação de bases oriundas de setores e países diferentes.

Outro ponto de destaque do Open Data Monitor, é o retorno e a integração dentre o projeto e os repositórios que são fontes de dados, visto que há um *feedback* no sentido de que os repositórios evoluam na publicação dos dados. A questão da integração dos dados também cria o sentimento da necessidade de evolução, pois apresenta os pontos frágeis e quanto se pode evoluir.

A Europeana apresenta um novo conceito em integração de dados, e destaca-se por ter criado um padrão totalmente alinhado aos conceitos de Web Semântica e prontos para o LOD. O EDM é sem dúvida uma evolução para toda a área de Humanidades Digitais, conceito que vem se alastrando por todo o mundo. O padrão EDM ainda precisa de pequenos ajustes, entretanto é referencia para o mundo quanto a organização de objetos digitais na área cultural.

Tanto o Open Data Monitor como a Europeana tem suas bases que trabalham pela melhoria dos dados organizados por consórcios que são compostos por governos, iniciativa privada e grupos do terceiro setor. Esse é outro ponto a ser pensado para o modelo brasileiro de infraestrutura de dados. Não é possível que apenas o governo tenha a iniciativa, ou que nesta iniciativa as universidades, institutos de pesquisa e iniciativa privada sejam mal representados. É necessário que se construam consórcios que realmente tenham interesses reais em conceitos tecnicamente viáveis para a evolução de publicação de dados no Brasil. É preciso pensar semanticamente nos dados abertos.

## REFERÊNCIAS


BERNERS-LEE T.; LASSILA, O.; HENDLER, J. The semantic web. ***Scientific American***, New York, v. 5, May 2001.

BERNERS-LEE, T. **Linked data principles**. 2006. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html> Acesso em: 10 abr. 2015.

CARTILHA técnica para publicação de dados abertos no Brasil v.1.0. 2011. Disponível em: < http://dados.gov.br/cartilha-publicacao-dados-abertos/>. Acesso em: 08 jun. 2015.

EUROPEANA data model primer. 2013. Disponível em: <http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf >. Acesso em: 19 jun. 2015.

HEATH, T.; BIZER, C. **Linked data**: evolving the web into a global data space. Morgan & Clarepool, 2011.

HONMA, T. et al. Extracting description set profiles from RDF datasets using metadata instances and SPARQL queries. DCMI GLOBAL MEETINGS & CONFERENCES, DC-2014, AUSTIN TEXAS, 2014. Disponível em: <http://dcevents.dublincore.org/IntConf/dc-2014/paper/view/249/235>. Acesso em: 18 jun. 2015.

LASSILA, O. **Resource description framework (RDF) model and syntax specification 1.0**. 1999. Disponível em: <http://www.w3c.org/TR/ REC-rdf-syntax>. Acesso em: 2 maio 2015.

MANUAL dos dados abertos: desenvolvedores. São Paulo: Comitê Gestor da Internet no Brasil, 2011. Cooperação técnica científica entre Laboratório Brasileiro de Cultura Digital e o Núcleo de Informação e Coordenação do Ponto BR (NIC.br), 2011. Disponível em: <http://www.w3c.br/pub/Materiais/PublicacoesW3C/Manual_Dados_Abertos_WEB.pdf>. Acesso em: 10 abr. 2015.

OPEN DATA MONITOR. 2015. Disponível em: <, http://www.opendatamonitor.eu/frontend/web/index.php?r=site%2Fabout>. Acesso em: 11 jun. 2015.

OPEN Knowledge Foundation, 2004. Disponível em: <http://okfn.org/about/>. Acesso em: 25 abr. 2015.

RODRIGUES, F. A.; SANT'ANA, R. C. G. Restrições tecnológicas e de acesso a dados disponíveis sobre destinos de repasses financeiros federais para a saúde pública em ambientes informacionais digitais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 13., 2012. **Anais**... Rio de Janeiro: FIOCRUZ, 2012. Disponível em: <http://www.eventosecongressos.com.br/metodo/enancib2012/arearestrita/pdfs/19435.pdf>. Acesso em: 10 jun. 2015.

RODRIGUES, F. de A.; SANT'ANA, R. C. G.; FERNEDA, E. Análise do processo de recuperação de conjuntos de dados em repositórios governamentais. **InCID:** revista de Ciência da Informação e Documentação**,** v. 6, n. 1, p. 38-56, abr. 2015. Disponível em: <http://www.revistas.usp.br/incid/article/view/73496>. Acesso em: 02 jun. 2015.

SANTAREM SEGUNDO, J. E. **Representação iterativa:** um modelo para repositórios digitais. 2010. 224 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília. 2010.

SANTAREM SEGUNDO, J. E. Tecnologías de la información y la comunicación para proporcionar datos abiertos en formato semántico. ***Ibersid,*** v. 7, p.33-40, 2013. Disponível em: <http://ibersid.eu/ojs/index.php/ibersid/article/view/4075/3744>. Acesso em: 03 maio 2015.

SANTAREM SEGUNDO, J. E. Web Semântica: introdução a recuperação de dados usando SPARQL. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO: além das nuvens, expandindo as fronteiras da Ciência da Informação, 15., 2014. Belo Horizonte. **Anais...** Belo Horizonte: UFMG/ ECI, 2014. p. 3863-3882.

SCHMACHTENBERG, M.; BIZER, C.; PAULHEIM, H.; Adoption of the Linked Data best practices in different topical domains, 2014. Disponível em: <http://dws.informatik.uni-

mannheim.de/fileadmin/lehrstuehle/ki/pub/SchmachtenbergBizerPaulheim-AdoptionOfLinkedDataBestPractices.pdf>. Acesso em: 09 jun. 2015.

SPARQL 1.1 Overview. W3C, 2013. Disponível em <http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/>. Acesso em: 19 jun. 2015.