Tecnologias de informação e comunicação para disponibilização de dados abertos em formato semântico

Tecnologías de la información y la comunicación para proporcionar datos abiertos en formato semántico

Information and communication technologies to provide open data in semantic formats

Jose Eduardo SANTAREM SEGUNDO

Department of Education, Information and Comunication, Universidade de São Paulo – USP,
Ribeirão Preto - São Paulo – Brasil, santarem@usp.br

Resumen

La Ley de Acceso a la Información brasileña de 2011 ha despertado el interés de la comunidad en cuanto a la disponibilidad real y el acceso a la información de los datos proporcionados por el gobierno de Brasil. En este contexto, el gobierno de Brasil ha promocionado una iniciativa de datos abiertos, la Infraestructura de Datos Nacional Abierta (INDA), para crear estructuras que permitan el uso de esta información por parte de la sociedad. Junto con el proceso de publicación de los datos, también es necesario desarrollar tecnologías que permiten publicar y recuperación futura de este tipo de datos en formatos semánticos, que permiten cruzar datos que son semánticamente relacionados, pero que están en entornos distribuidos.

Palabras clave: Datos abiertos. Web Semántica. Ley de acceso a la información. RDF. SPARQL. Linked Data.

Abstract

The Brazilian Law on Access to Information of 2011 has attracted the interest of the community in relation to the actual availability and access to information provided by the Brazilian government. In this context, the Brazilian government has promoted an open data initiative, the Infraestructura de Datos Nacional Abierta (INDA), to create infrastructures that allow the use of this information by the society. Along with the process of publication of the data, it is also necessary to develop technologies to publish and recover such data in semantic formats that allow its interrelation in distributed environments.

Keywords: Open data. Semantic Web. Access to information act. RDF. SPARQL. Linked Data.

1. Introdução

A publicação no Brasil da Lei do Acesso à Informação, nº 12.527, de 18 de novembro de 2011, apesar de recente, já tem despertado um conjunto de ações e, principalmente, de pesquisas com a intenção de abrir caminhos para que o povo brasileiro tenha acesso às informações governamentais de forma mais ampla e clara.

Neste contexto, a Ciência da Informação deve aparecer como principal área de estudo, sobretudo no desenvolvimento e aplicação de conceitos e métodos visando a colaborar com a aplicação da Lei.

O primeiro artigo da Lei (Brasil, 2011) explana a respeito da disposição, aplicação e órgãos que estão diretamente subordinados a ela.

> Art. 1º Esta Lei dispõe sobre os procedimentos a serem observados pela União, Estados, Distrito Federal e Municípios, com o fim de garantir o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal.

Parágrafo único. Subordinam-se ao regime desta Lei (Brasil, 2011):

> I - os órgãos públicos integrantes da administração direta dos Poderes Executivo, Legislativo, incluindo as Cortes de Contas, e Judiciário e do Ministério Público;
>
> II - as autarquias, as fundações públicas, as empresas públicas, as sociedades de economia mista e demais entidades controladas direta ou indiretamente pela União, Estados, Distrito Federal e Municípios.

Portanto, trata-se efetivamente de dar conhecimento à população das informações e procedimentos adotados por órgãos na esfera Federal, Estadual e Municipal no Brasil.

O artigo 3º da lei aborda suas diretrizes (Brasil, 2011):

> I - observância da publicidade como preceito geral e do sigilo como exceção;

> II - divulgação de informações de interesse público, independentemente de solicitações;
>
> III - utilização de meios de comunicação viabilizados pela tecnologia da informação;
>
> IV - fomento ao desenvolvimento da cultura de transparência na administração pública;
>
> V - desenvolvimento do controle social da administração pública.

Ainda sobre a Lei do Acesso destaca-se o artigo 8º que diz o seguinte (Brasil, 2011):

> Art. 8º É dever dos órgãos e entidades públicas promover, independentemente de requerimentos, a divulgação em local de fácil acesso, no âmbito de suas competências, de informações de interesse coletivo ou geral por eles produzidas ou custodiadas.

O inciso III do parágrafo terceiro do artigo 8º diz (Brasil, 2011):

> III - possibilitar o acesso automatizado por sistemas externos em formatos abertos, estruturados e legíveis por máquina;

A amplitude e conjunto de artigos, parágrafos e incisos que a lei dispõe merecem e sugerem um elevado número de estudos e pesquisas, entretanto este trabalho está estruturado e baseado principalmente neste último inciso citado da lei, pois se associa a este a ideia de estruturar as informações de forma a serem legíveis por máquina, e também estar estruturadas dentro de um contexto semântico.

Assim como nos Estados Unidos, Inglaterra e França a Espanha, através da Lei 11/2007 (acesso eletrônico dos cidadãos aos serviços públicos) aplica planos nacionais de administração eletrônica, incluindo a obrigatoriedade da transformação da administração pública em administração eletrônica em benefício dos cidadãos. Ainda sobre a Espanha, foi publicado o Real Decreto 4/2010, o Esquema Nacional de Interoperabilidade, que compreende um conjunto de critérios, recomendações e decisões tecnológicas que deverão ser tomados pela administração pública para garantia da interoperabilidade (Hallo et al., 2012).

Portanto, o objetivo deste trabalho é apresentar o conjunto de tecnologias que proponham e facilitem o acesso automatizado por sistemas externos em formatos abertos, estruturados e legíveis por máquina, dentro de uma perspectiva de geração de dados governamentais em formato aberto, mas principalmente em uma estrutura semântica e interoperável.

2. Dados abertos

A Lei do Acesso à Informação nasceu a partir de um movimento chamado de Dados Abertos (Open Data) que tem se caracterizado desde 2009, quando países como Estados Unidos e Inglaterra iniciaram um 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.

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.

Certamente, o avanço tecnológico e o crescente aumento do acesso pela comunidade a novos dispositivos que permitem conexão à Internet têm 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 têm procurado publicar suas informações de modo que a sociedade possa consumir esses dados.

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.

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 publicas 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.

A Lei do Acesso à Informação é uma das iniciativas que corrobora a ideia de que o Brasil tem se preocupado e também tem se envolvido nas questões de disseminação de informações governamentais.

O Brasil lidera, juntamente com os EUA, um movimento chamado Parceria para Governo Aberto (OGP), iniciado em 2011, que envolve oito países e tem como objetivo promover governos mais transparentes e eficientes.

Atualmente a Secretaria de Logística e Tecnologia da Informação do Ministério do Planejamento, Orçamento de Gestão desenvolve a Infraestrutura Nacional de Dados Abertos (INDA). Segundo a Cartilha (2011):

> 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 – dados.gov.br, que tem o objetivo de ser o ponto central para a publicação, a busca e o acesso de dados públicos no Brasil.

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

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.

A estrutura em que o INDA se baseia está constituída a partir dos 8 princípios de dados abertos governamentais definidos pelo Open Government Working Group, grupo de trabalho que se reuniu em 2007 nos Estados Unidos. Os 8 princípio são:

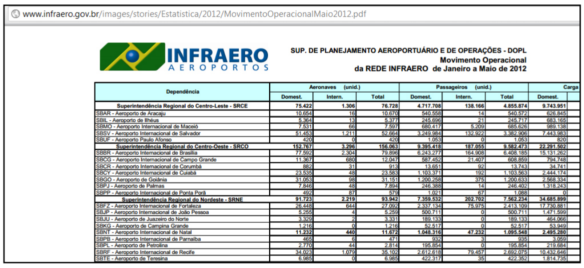
- *Completos*. Todos os dados públicos são disponibilizados. Dadas são informações eletronicamente gravadas, incluindo, mas não se limitando a documentos, bancos de dados, transcrições e gravações audiovisuais. Dados públicos são dados que não estão sujeitos a limitações válidas de privacidade, segurança ou controle de acesso, regulados por estatutos.
- *Primários*. Os dados são publicados na forma coletada na fonte, com a mais fina granularidade possível, e não de forma agregada ou transformada.
- *Atuais*. Os dados são disponibilizados o quão rapidamente seja necessário para preservar o seu valor.
- *Acessíveis*. Os dados são disponibilizados para o público mais amplo possível e para os propósitos mais variados possíveis.
- *Processáveis por máquina***.** Os dados são razoavelmente estruturados para possibilitar o seu processamento automatizado.
- *Acesso não discriminatório***.** Os dados estão disponíveis a todos, sem que seja necessária identificação ou registro.
- *Formatos não proprietários***.** Os dados estão disponíveis em um formato sobre o qual nenhum ente tenha controle exclusivo.
- *Livres de licenças*. Os dados não estão sujeitos a regulações de direitos autorais, marcas, patentes ou segredo industrial. Restrições razoáveis de privacidade, segurança e controle de acesso podem ser permitidas na forma regulada por estatutos.

Segundo Abella (2011), a Espanha pode ser considerada o segundo país mais desenvolvido nas questões de dados abertos na União Européia. Esta iniciativa poderia criar 45 mil postos de trabalho em 10 anos. Com estas iniciativas, a administração pública espanhola deve colocar à disposição do público os dados em formato aberto de forma a permitir a combinação destes dados objetivando a criação de novos serviços e melhoria dos já existentes.

Publicar os dados em formatos proprietários é um dos principais erros cometidos pela grande maioria dos órgãos do governo que tem procurado publicar dados em formato aberto. A utilização de formatos em que seja necessária a utilização de ferramentas do pacote Office da

Microsoft, por exemplo, pode inviabilizar o acesso aos dados pela comunidade.

Outro erro bastante cometido pelas instituições públicas é a publicação de seus dados em formato PDF. Apesar do leitor de arquivos no formato PDF ser encontrado sem custo na Internet, publicar dados neste formato inviabiliza a leitura por máquinas destes dados, ferindo o item 5 dos princípios dos dados abertos e também o inciso III do parágrafo 3 do artigo 8º. da Lei do Acesso à Informação, destacado nesse trabalho.



www.infraero.gov.br/images/stories/Estatistica/2012/MovimentoOperacionalMaio2012.pdf

INFRAERO AEROPORTOS

SUP. DE PLANEJAMENTO AEROPORTUÁRIO E DE OPERAÇÕES - DOPL
Movimento Operacional
da REDE INFRAERO de Janeiro a Maio de 2012

Dependência	Aeronaves (unid.)			Passageiros (unid.)			Carga
	Domest.	Intern.	Total	Domest.	Intern.	Total	Domest.
Superintendência Regional do Centro-Leste - SRCE	**75.422**	**1.306**	**76.728**	**4.717.708**	**138.166**	**4.855.874**	**9.743.951**
SBAR - Aeroporto de Aracaju	10.654	16	10.670	540.558	14	540.572	626.845
SBIL - Aeroporto de Ilhéus	5.364	13	5.377	245.696	21	245.717	683.165
SBMO - Aeroporto Internacional de Maceió	7.531	66	7.597	680.417	5.209	685.626	989.138
SBSV - Aeroporto Internacional de Salvador	51.453	1.211	52.664	3.249.984	132.922	3.382.906	7.443.983
SBUF - Aeroporto Paulo Afonso	420	0	420	1.053	0	1.053	820
Superintendência Regional do Centro-Oeste - SRCO	**152.767**	**3.296**	**156.063**	**9.395.418**	**187.055**	**9.582.473**	**22.291.502**
SBBR - Aeroporto Internacional de Brasília	77.592	2.304	79.896	6.243.277	164.908	6.408.185	15.131.262
SBCG - Aeroporto Internacional de Campo Grande	11.367	680	12.047	587.452	21.407	608.859	794.748
SBCR - Aeroporto Internacional de Corumbá	882	31	913	13.651	92	13.743	34.741
SBCY - Aeroporto Internacional de Cuiabá	23.535	48	23.583	1.103.371	192	1.103.563	2.444.174
SBGO - Aeroporto de Goiânia	31.053	98	31.151	1.200.258	375	1.200.633	2.568.334
SBPJ - Aeroporto de Palmas	7.846	48	7.894	246.388	14	246.402	1.318.243
SBPP - Aeroporto Internacional de Ponta Porã	492	87	579	1.021	67	1.088	0
Superintendência Regional do Nordeste - SRNE	**91.723**	**2.219**	**93.942**	**7.359.532**	**202.702**	**7.562.234**	**34.685.899**
SBFZ - Aeroporto Internacional de Fortaleza	26.448	644	27.092	2.337.134	75.975	2.413.109	17.730.881
SBJP - Aeroporto Internacional de João Pessoa	5.255	4	5.259	500.711	0	500.711	1.471.599
SBJU - Aeroporto de Juazeiro do Norte	3.329	2	3.331	189.133	0	189.133	464.066
SBKG - Aeroporto de Campina Grande	1.216	0	1.216	52.517	0	52.517	53.949
SBNT - Aeroporto Internacional de Natal	11.232	440	11.672	1.048.316	47.232	1.095.548	2.495.280
SBPB - Aeroporto Internacional de Parnaíba	465	6	471	932	3	935	3.059
SBPL - Aeroporto de Petrolina	2.770	44	2.814	195.854	0	195.854	219.684
SBRF - Aeroporto Internacional de Recife	34.023	1.079	35.102	2.612.618	79.457	2.692.075	10.432.646
SBTE - Aeroporto de Teresina	6.985	0	6.985	422.317	35	422.352	1.814.735

Figura 1. Dados da INFRAERO em formato PDF

A Figura 1 apresenta uma parte do conjunto de dados disponibilizados pela Infraero e que fazem parte do Portal Brasileiro de Dados Abertos. Apesar da disponibilização das informações, elas não podem ser processadas por máquinas porque são apresentadas em formato PDF. Esse tipo de iniciativa cumpre com parte dos requisitos para publicação de dados abertos tornando o cruzamento e redistribuição destes dados impossível.

A Cartilha Técnica para Publicação de Dados Abertos no Brasil (2011) sugere a utilização dos seguintes formatos de arquivos: JSON (JavaScript Object Notation), é um padrão aberto de estruturação de dados baseado em texto e legível por humano; XML (Extensible Markup Language), baseado em texto e tem como principais objetivos simplicidade, extensibilidade e usabilidade; CSV (Comma-Separated Values), ou valores separados por vírgula, e é um formato para armazenamento de dados tabulares em texto; ODS (Open Document Spreadsheet), é um formato não proprietário de arquivo baseado em XML, padronizado pela ABNT sob a norma NBR ISO/IEC 26300:2006 e o formato RDF (Resource Description Framework), é um modelo de dados estruturado em grafos e possui diversos formatos de serialização, tais como RDF/XML, Notation 3 e Turtle.

Dentre os formatos apresentados, o RDF com certeza é o único que permite estruturar de forma semântica as informações e isso torna sua utilização mais complexa, entretanto fortalecendo a qualidade da informação. Um dos eixos fundamentais da Web Semântica constitui-se na utilização do padrão RDF.

3. Web Semântica

Após o ano de 2001, quando Tim Berners-Lee, James Hendler e Ora Lassila publicaram na revista Scientific American o texto *The Semantic Web*, algumas áreas de pesquisa, em especial dentro da Ciência da Informação e da Ciência da Computação, têm direcionado esforços para colocar em prática os conceitos abordados no texto citado.

Nos últimos anos, vários elementos foram surgindo e ampliando o contexto da ideia original 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. 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 à Informação, abordado neste trabalho, que indicam a possibilidade dos dados serem processados por máquina, além de possibilitar o relacionamento 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.

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

Segundo o W3C, o RDF é uma linguagem de uso geral para representar informações na Web. O RDF tem como princípio fornecer interoperabilidade aos dados, de forma que possa contribuir com a recuperação de informações de recursos na Web.

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.

O modelo RDF é constituído de três objetos básicos: recursos, propriedades e declarações. Um recurso é uma informação (página web, livro, cd, pessoa, lugar, documento disponível em um repositório ou biblioteca digital) que pode ser identificada por uma URI (Universal Resource Identificator). Propriedades são as informações que representam as características do recurso, ou seja, são os atributos que permitem distinguir um recurso de outro ou que descrevem o relacionamento entre recursos. Nomes de propriedades devem ser associados a um esquema.

A declaração é a constituição da informação completa, que compreende um recurso com suas propriedades e valores para as propriedades. Uma URI pode ser um local ou página na WEB como uma URL (Unified Resource Locator) ou ainda outro tipo de identificador único.

Basicamente, a representação de uma sentença em RDF é feita utilizando-se um grafo. 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 e Vidotti, 2011).

Além de representar graficamente uma informação através de grafos, o modelo RDF pode ser representado através da sintaxe XML (Santarem Segundo e Vidotti, 2011).

Um dos esquemas mais utilizados para representação das propriedades do RDF em grande parte dos exemplos e modelos encontrados na literatura é constituído a partir de um vocabulário baseado no esquema de metadados Dublin Core (DC).

O esquema de metadados DC é apenas um dos exemplos. Outros bastante utilizados e conhecidos são VCard, para descrição de contatos entre as pessoas e FOAF (friend of friend) para informações que descrevam pessoas, suas atividades e relacionamentos sociais.

Vários são os modelos de esquemas e ontologias disponíveis para serem utilizados livremente, porém um grande conjunto desses esquemas pode ser encontrado e consumido. A ideia que agrega e congrega esses esquemas é chamada de Linked Data.

4. Linked Data e DBPedia

Pensando em um modelo associativo de publicação de dados estruturados na Web, foi constituído o Linked Data (dados ligados). Organizado por Tim Berners-Lee tem como característica principal o estabelecimento de links entre dados de fontes distribuídas.

Segundo 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 Linked Data, quando você tem um pouco de dados, você pode encontrar outros que estão relacionados.

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

- Usar URIs como nomes para os itens.
- Usar URIs HTTP para que as pessoas possam consultar esses nomes.
- Quando alguém consulta uma URI, prover informação RDF útil.
- Incluir sentenças RDF com links para outras URIs, a fim de permitir que itens relacionados possam ser descobertos.

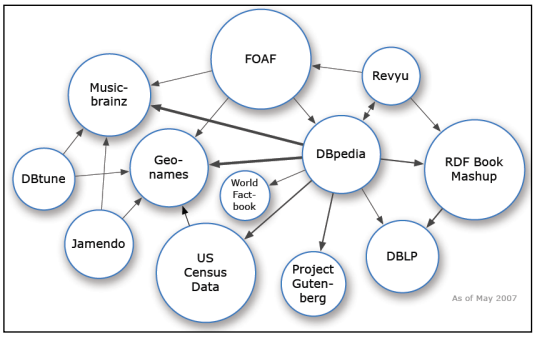


Figura 2. Linked Data (linkeddata.org, 2007)

O projeto tem crescido muito nos últimos anos, em 2007 o Linked Data era constituído de aproximadamente um bilhão de declarações RDF, interligados por 120.000 links RDF. Atualmente já são 52 bilhões de declarações RDF e o conjunto de ligações não para de crescer.

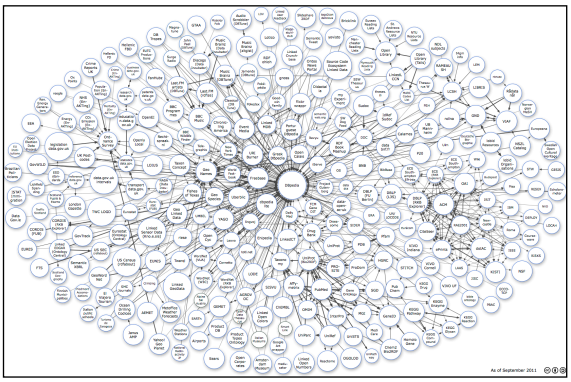


Figura 3. Linked Data ((linkeddata.org, Nov/2011)

A Figura 2 apresenta a situação de ligações no ano de 2007, quando o projeto foi iniciado. Apesar de ilegível, a Figura 3 tem a finalidade apenas de dar a dimensão do crescimento do conjunto de ligações que o Linked Data tem alcançado.

Com certeza, a base de maior tamanho e também maior destaque no Linked Data é a DBPedia.

DBPedia é uma interface orientada a dados, constituída a partir de um esforço da comunidade para extrair informação estruturada da Wikipédia e tornar esta informação disponível na web. DBPedia permite executar consultas sofisticadas no Wikipédia e vincular outros conjuntos de dados na Web para os dados da Wikipédia.

A base de conhecimento DBPedia atualmente descreve mais de 3,64 milhão de itens, das quais 1,83 milhões são classificados em uma ontologia consistente, incluindo 416.000 pessoas, 526.000 lugares, 106.000 álbuns de música, 60.000 filmes, 17.500 jogos de vídeo, 169.000 organizações, 183.000 espécies e 5.400 doenças. Os dados da DBPedia apresentam estes 3,64 milhões de itens em até 97 idiomas diferentes; 2,724,000 links para imagens e 6.300.000 links para páginas web externas; 6.200.000 ligações externas em outros conjuntos de dados RDF, 740.000 categorias Wikipédia, e 2.900.000 categorias Yago. A base de conhecimento DBpedia totalmente consiste de mais de 1,2 bilhões de declarações RDF, dos quais 335 milhões foram extraídos da edição em Inglês da Wikipédia e 865 milhões foram extraídos de edições em outros idiomas (DBPEDIA, 2010).

Outra base de destaque, principalmente porque praticamente 95% de seus dados têm precisão manual é a Yago. Essa base relaciona 2 milhões de entidades (pessoas, organizações, cidades, etc.). São aproximadamente 20 milhões de fatos sobre essas entidades.

Destaca-se ainda a integração de dados geográficos, tais como nomes de lugares em vários idiomas, altitude, população e outras fontes publicadas pelo banco de dados Geonames (1) que está disponível para download gratuitamente sob uma licença Creative Commons Atribuição. Ele contém mais de oito milhões de nomes geográficos e é composto por 6,3 milhões de recursos exclusivos. Todos os recursos são classificados em uma das nove classes de recurso e ainda subcategorizados em um dos 645 códigos de recurso. Os dados são acessíveis gratuitamente através de um número de webservices e uma exportação de banco de dados diariamente. O webservice Geonames.org já está servindo a mais de 3 milhões de solicitações de serviço da web por dia (Geonames, 2013).

Outra característica a ser abordada nos dados publicados no Linked Data é a disseminação de instrumentos de controle de vocabulário e controle de autoridade. O AGROVOC é um tesauro desenvolvido a partir de 1980 que atualmente é referência para temas ligados à agricultura, piscicultura, silvicultura entre outros assuntos relacionados ao meio ambiente. O AGROVOC é utilizado mundialmente por pesquisadores, bibliotecários, gestores de informação e outros profissionais para indexar, recuperar e organizar dados em sistemas de informações agrícolas. Utilizando-se da terminología SKOS, modelo apropriado para publicação de tesauros na web, o AGROVOC também está disponível no formato Linked Data e pode ser consultado assim como interligado com outros recursos para provimento de dados em formato semântico. (AGROVOC, 2013).

Há ainda, como pode ser observado na Figura 3, uma grande diversidade de bases de dados, que vão desde informações sobre Ciências da Vida (GeneID, PubMed, Geo Species, Gene Ontology, etc.), Dados Geográficos (Aeroportos, Earth, Linked GeoData, etc.), Dados de Uso Geral (Slideshare, Semantic Tweet, Delicious, Flickr, etc.), Mídia (BBC, Music Brainz, New York Times, Last.FM, etc.), Publicações (IEEE, ePrints, CiteSEER, theses.fr, etc.) e Dados Governamentais (Patentsdata.gov.uk, researchdata.gov.uk, transportdata.gov.uk, Políticos Brasileiros (2), etc.).

A esse último item citado no grupo de Dados Governamentais, encontra-se em um projeto brasileiro que apresenta informações a respeito de políticos brasileiros.

5. SPARQL

Essa grande estrutura informacional tem cada vez mais pertinência a partir do momento em que existem linguagens e ferramentas que permitem a recuperação dos dados.

SPARQL é um conjunto de especificações que fornece linguagens e protocolos para consultar e manipular o conteúdo publicado em RDF na Web. O padrão compreende as seguintes especificações: uma linguagem de consulta para RDF; uma especificação que define uma extensão do SPARQL Query Language para executar consultas distribuídas em diferentes terminais SPARQL; uma especificação que define a semântica de consultas SPARQL sob regimes de vinculação, como RDF Schema, OWL, ou RIF; um protocolo que define os meios para a transmissão de consultas SPARQL arbitrárias e solicitações de atualização para um serviço de SPARQL; uma especificação que define um método para descobrir e um vocabulário para descrever serviços SPARQL, e um conjunto de testes, para avaliação da especificação SPARQL 1.1 (SPARQL, 2013).

Através da Sparql é possível recuperar informações nessa grande base de dados estruturados e distribuídos.

Segundo DuCharm (2011, p. 19) "SPARQL é uma linguagem de consulta para dados que segue um modelo específico, estruturado no formato RDF". O nome SPARQL é um acrônimo de (SPARQL - Protocol and RDF Query Language).

A especificação SPARQL é um padrão recomendado pelo W3C desde janeiro de 2008 e permite ao usuário combinar dados de arquivos RDF advindos de diferentes fontes.

Um grande conjunto de ferramentas tem sido desenvolvido para que seja possível utilizar SPARQL. Esses ambientes que estão prontos e aptos a receber consultas SPARQL são chamados de Endpoints.

O princípio de uso da SPARQL como linguagem de consulta está baseado na linguagem SQL. Suas principais cláusulas são: SELECT [DISTINCT], FROM (opcional), WHERE (opcional), ORDER BY (opcional) e UNION (opcional – funcionamento diferente do SQL).

Basicamente a consulta SPARQL é construída sobre triple pattern, ou seja: subject, predicate e object, com base na mesma estrutura de construção de um arquivo RDF.

Ressalta-se que consultas deste tipo podem ser realizadas, desde que se conheça a estrutura semântica utilizada nas ontologias e vocabulários construídos que formam não apenas o Linked Data, mas também qualquer outra base de dados semântica baseada neste tipo de informação.

Muitas das informações distribuídas nas redes podem gerar um grande conjunto de resultados, porém em algumas situações, como na Wikipédia, por exemplo, agrupar algumas informações é uma tarefa um tanto árdua. O exemplo 1 apresenta um modelo de consulta, com resultado na Figura 4 que pode ser obtido no DBPedia e que sem a utilização de ferramentas de consulta levaria um tempo para ser realizada.

```
SELECT ?filme
WHERE {
?filme
<http://dbpedia.org/ontology/starring>
<http://dbpedia.org/resource/Robert_De_Niro;
<http://dbpedia.org/ontology/director>
<http://dbpedia.org/resource/Robert_De_Niro>
}
```

Exemplo 1. Consulta Sparql – Filmes Robert de Niro

No exemplo 1, a consulta busca os filmes que tiveram o ator Robert de Niro tanto como ator, quanto como diretor, realizando o cruzamento das informações.

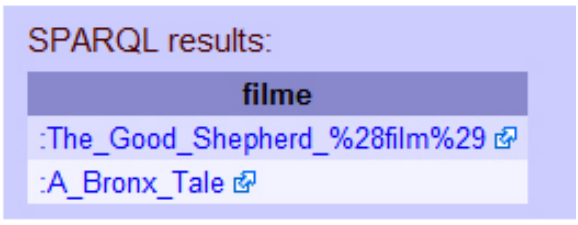


Figura 4. Resultado da Consulta Sparql – Exemplo 1

A especificação SPARQL é repleta de recursos e pode explorar de forma muito ampla o conjunto de dados ligados disponíveis pela Internet.

6. Considerações finais

Uma pesquisa no Portal Brasileiro de Dados Abertos, citado neste trabalho, nos leva a crer

que as iniciativas brasileiras para publicação de dados abertos governamentais ainda são tímidas. A Espanha já apresenta um grau mais avançado de desenvolvimento, iniciado com a publicação da Lei 11/2007. Quando se pensa em dados com estrutura semântica tem-se quase uma nulidade de resultados e também de iniciativas, principalmente dentro dos órgãos governamentais brasileiros.

A partir da apresentação desse conjunto de tecnologias é possível afirmar que as instituições públicas brasileiras podem, através de seus centros e polos de informática, tornar seus dados abertos, semânticos e livres para consulta, fortalecendo de forma clara a política de dados abertos do governo brasileiro.

Observa-se que, além das tecnologias apresentadas aqui, existe uma gama muito grande de ferramentas que favorecem a estruturação e publicação dos dados já existentes em bancos de dados alimentados pelos sistemas atuais para o formato aberto e semântico. Iniciativas como o Framework Jena, Plataforma D2RQ, OpenLink Virtuoso, Sesame2, Meronymy SPARQL Database Server, ARC2 e Mulgara RDF Database têm sido estudadas por este pesquisador e alimentado o desenvolvimento de trabalhos de pesquisa em torno publicação de dados abertos governamentais estruturados semanticamente.

Notas

(1) http://geonames.org

(2) http://ligadonospoliticos.com.br/?pag=home.

Referencias

Abella, Alberto (2011). Reutilización de información pública y privada en España. http://www.navarra.es/NR/rdonlyres/16750B44-0B82-4AB0-9DA7-17CDD0093216/189326/paper_reutilizacion_informacion_publica_privada_op.pdf (2013-06-10).

AGROVOC (2013). AGROVOC Linked Open Data. http://aims.fao.org/standards/agrovoc/linked-open-data (2013-06-10).

Berners-Lee, T. (2006). Linked Data Principles. http://www.w3.org/DesignIssues/LinkedData.html. (2013-01-11).

Berners-Lee, T.; Hendler J.; Lassila, O. (2011). The Semantic Web. // Scientific American. (May 2001) 29-37

Brasil. Lei nº. 12.527, de 18 de novembro de 2011. (2011) Regula o acesso à informação. http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm (2013-02-15).

Cartilha técnica para publicação de dados abertos no Brasil v1.0. ([2011?]). http://wiki.gtinda.ibge.gov.br/GetFile.aspx?Page=Tecnologia&File=Cartilha%20T%C3%A9cnica%20para%20Publica%C3%A7%C3%A3o%20de%20Dados%20Abertos%20no%20Brasil%20v1.pdf (2013-02-01).

Dbpedia Team (2010). DBPedia. http://thedatahub.org/dataset/dbpedia (2012-04-12)

Ducharme, B. (2011) Learning SPARQL. Sebastopol: O'Reilly, 2011.

Eaves, D. (2009). The three laws of Open Government Data. http://eaves.ca/2009/09/30/three-law-of-open-government-data/ (2013-01-08)

Geonames. (2013). Geonames Ontology. http://www.geonames.org/ontology/documentation.html (2013-06-08).

Hallo, María Asunción; Martínez González, M. Mercedes; Fuente Redondo, Pablo de la (2012). Las tecnologías de Linked Data y sus aplicaciones en el gobierno electrónico. // Scire. 18:1 (en.-jun. 2012) 49-61.

Lassila, O. (1999). Resource description framework (RDF) model and syntax specification 1.0. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/ (2013-01-08).

Manual dos dados abertos: desenvolvedores. (2011) 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).

Open Knowledge Foundation (2004). http://okfn.org/about/ (2013-02-23)

Santarem Segundo, J. E.; Vidotti, S. A. B. G. (2011). Rede de tags para recuperação da informação no contexto da representação iterativa. // InCID: Revista de Ciência da Informação e Documentação. 2:1 (jan./jun. 2011) 86-109. http://revistas.ffclrp.usp.br/incid/article/view/50/pdf (2013-03-10)

SPARQL (2013). Sparql 1.1 Overview. http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/ (2013-06-12).

Enviado: 2013-04-05. Segunda versión: 2013-07-15.
Aceptado: 2013-09-02.