Knowledge Organization and Cultural Diversity

Organizers: José Augusto Chaves Guimarães e Vera Dodebei

ISKO-BRASIL

Organizers

José Augusto Chaves Guimarães
Vera Dodebei

Knowledge Organization and Cultural Diversity

Marília, São Paulo, Brasil

Sociedade Brasileira de Organização do Conhecimento
(ISKO-Brasil)

2017

K73k Knowledge Organization and Cultural Diversity [recurso eletrônico] / José Augusto Chaves Guimarães, Vera Dodebei, organizadores. -- Pernambuco: ISKO-Brasil ; UFPE, 2017.
760 f. ; 30 cm.

ISBN: 978-85-415-0924-4

Livro digital

1. Organização do conhecimento. I. Título.

CDD 025.4

Editors

Isadora Victorino Evangelista

Gilberto Gomes Cândido

Rafael Aparecido Moron Semidão

Rafael Cacciolari Dalessandro

Suellen Oliveira Milani

Design

Maíra Fernandes Alencar

Translation

Natália Nakano

SUMMARY

THE EPISTEMOLOGIC DIMENSION OF KNOWLEDGE ORGANIZATION

THE SOCIAL, CULTURAL AND POLITICAL DIMENSION OF KNOWLEDGE ORGANIZATION

PROSPECTS IN KNOWLEDGE ORGANIZATION RESEARCH IN BRAZIL

RESEARCH IN DOCUMENTAL ANALYSIS ON BRAZIL: THE JEAN-CLAUDE GARDIN INFFLUENCE

A study on actions to make government datasets available in linked open data

Fernando de Assis Rodrigues
São Paulo State University
fernando@elleth.org

Ricardo César Gonçalves Sant'Ana
São Paulo State University
ricardosantana@marilia.unesp.br

Introduction

The principles of Linked Open Data (LOD) establish a new way of sharing datasets opened by the Internet, aiming to promote the wide distribution of structured data in languages, such as eXtensible Markup Language (XML) and in compliance with the recommendations of the Resource Description Framework (RDF) (BERNERS-LEE, 2009; BIZER; HEATH; BERNERS-LEE, 2009; HEATH, 2015; W3C, 2014, 2015).

In this scenario, government datasets play a prominent role: they represent 18.58% of the total number of existing LOD datasets and 41.54% of these government datasets have at least one relationship with ontologies or controlled vocabularies, according to the results of the mapping developed by Linking Open Data cloud diagram (SCHMACHTENBERG; BIZER; PAULHEIM, 2014a, 2014b).

However, according to Schmachtenberg, Bizer, and Paulheim (2014a), there are still characteristics in the LOD dataset structures at the moment of data retrieval that is not considered ideal nor adopted good practices, such as the absence of metadata and licenses information.

Actions to make public government data accessible are an integral part of discussions on trends in the modernization of public administration models, which seek to redistribute skills and resources among different intra-governmental and extra-governmental organizations, allowing greater institutional pluralism in public functions (MALIN, 2006; SANT'ANA; RODRIGUES, 2013).

The strengthening of transparency actions can be expanded by building information sharing environments that, among other characteristics, provide an

increase in information flows between public administration and society, thus ensuring greater visibility of the State activities (BOHMAN, 2000; MARCONDES, JARDIM, 2003). These environments become components of greater citizen participation, extending possibilities of participation beyond voting; and the State can improve the effectiveness and monitoring of the activities and results of its actions, in addition to complying with the obligation to publish government data (BRASIL, 2011; SANT'ANA; RODRIGUES, 2013).

Access to government datasets on the results of legislative votes is important in monitoring the activities of representatives, supporting the construction of analyzes, such as "[...] the identification of party clusters" and "[...] consistency of each of our representatives in the voting during their mandates" (SANT'ANA; RODRIGUES, 2013, page 58).

The objective of this paper is to explore the actions needed to provide government datasets in Linked Open Data, starting from an application of a model of recommendations for data publication "Linked Data Best Practices in Different Topical Domains", proposed by Schmachtenberg, Bizer And Paulheim (2014a), in databases available on legislative votes of the Brazilian Senate.

The research object was delimited to datasets available in Communication and Information Technology (ICT) tools of the Brazilian Senate, more precisely on the existing votes in the 'Portal e-Cidadania - Open Data', analyzed between January and March 2015.

Methodological procedures

The methodology adopted was the exploratory analysis of research object, with a qualitative approach, through the specification of the characteristics of the existing dataset (i.e. the location of the resource on the web site, information about the descriptive page and available files); and the data structures found at the time of data collection.

Those characteristics formed a set of information that served as a subsidy for the proposal of a strategy of actions necessary for restructuring this existing data, in compliance with the established recommendations and good practices of LOD datasets availability, proposed by Schmachtenberg, Bizer, and Paulheim (2014a).

Theoretical Background

Schmachtenberg, Bizer, and Paulheim (2014a, 2014b) propose a model with recommendations for data publication, with the objective of identifying the compliance to LOD concepts and good practices for data sharing by public datasets stored in various domains. These recommendations were elaborated from community practices and the results presented by the LOD dataset mapping developed by the Linking Open Data cloud diagram (JENTZSCH; CYGANIAK; BIZER, 2011).

The model is divided into nine recommendations:

Providing Provenance Information

In the process of data retrieval, it is necessary that datasets have unique identifiers to help data retrieval process by external agents, in compliance with the first principle of LOD (BIZER; HEATH; BERNERS-LEE, 2009). These unique identifiers must conform to rules established by the Uniform Resource Identifier (URI) and the RDF (JENTZSCH; CYGANIAK; BIZER, 2011; SCHMACHTENBERG; BIZER; PAULHEIM, 2014a, 2014b).

Defining links with other datasets

The dataset needs to have links to other datasets through the relationship rules established by the RDF. This procedure facilitates the automated data collection by external agents, including others datasets to which it was linked (JENTZSCH; CYGANIAK; BIZER, 2011; SCHMACHTENBERG; BIZER; PAULHEIM, 2014a, 2014b).

Use of controlled vocabularies and existing ontologies

As data is a basic element "[...] formed by a sign or finite set of signs that do not contain, intrinsically, a semantic component, but only syntactic elements" (SANTOS; SANT'ANA, 2002, np), and it is necessary to use controlled vocabularies and ontologies to extend the semantic load at the moment of data collection by external agents, such as: Dublin Core Metadata Set (DC), Friend of a Friend (FOAF), Simple Knowledge Organization System (SKOS), among others (JENTZSCH; CYGANIAK; BIZER, 2011, SCHMACHTENBERG; BIZER; PAULHEIM, 2014a, 2014b).

Definitions of terms, elements, and attributes in vocabularies and ontologies

The additional documents linked to the dataset, containing information about ontologies and controlled vocabularies, must have unique URIs for terms, elements, and attributes (SCHMACHTENBERG; BIZER; PAULHEIM, 2014a, 2014b). For example, in FOAF the definitions that describes terms like 'name' or 'birthday', it must

be accessible by unique URIs, either by using split symbols ('/'), or hash-tag (#) to differentiate access to each term (BRICKLEY; MILLER, 2014).

Linking terms between vocabularies

In case it is necessary to develop new and own vocabularies, is important that terms of this vocabulary are linked to existing vocabulary terms, such as DC, FOAF, SKOS, among others. The linking of new vocabularies with comprehensive vocabularies provides a greater repertoire of information on terms developed for the vocabulary of external agents (JENZYK; CYGANIAK; BIZER, 2011, SCHMACHTENBERG; BIZER; PAULHEIM, 2014a, 2014b).

Providing metadata

At the time of data retrieval, datasets must have metadata elements to ensure quality on data retrieval process, to identify the data source, and to ensure quality (JENZYK; CYGANIAK; BIZER, 2011). The metadata "[...] is a key factor to minimize search and retrieval problems in the various informational environments [...]" (SANTOS; ALVES, 2009) and it is recommended that: their elements be available in the root element; and the use of DC elements (JENTZSCH; CYGANIAK; BIZER, 2011; SCHMACHTENBERG; BIZER; PAULHEIM, 2014a, 2014b).

Use of license terms in metadata

The dataset metadata must contain licenses terms in its elements and attributes, such as Creative Commons, Open Data Commons Attribution License, Open Database License (ODbL), among others (JENTZSCH; CYGANIAK; BIZER, 2011; W3C, 2011).

Providing metadata about the dataset structure

In the dataset retrieval, there must be metadata containing information about its structure - made available with the data or in supplementary documents - delimiting elements, iterations, used terms and attributes (JENZYK; CYGANIAK; BIZER, 2011; SCHMACHTENBERG; BIZER; PAULHEIM, 2014a, 2014b).

Use of alternative methods for data retrieval

The most common form of providing structured datasets in the RDF is through a SPARQL Protocol and RDF Query Language Endpoint (SPARQL Endpoint) (JENTZSCH; CYGANIAK; BIZER, 2011), which enables external agents to perform structured searches in the SPARQL query language. However, it is recommended that dump files (RDF Dump), are also available explicit in the RDF/XML standard or equivalent (SEMANTICWEB.ORG, 2011).

Dataset Characteristics

Portal e-Cidadania aims to promote transparency of actions and activities of the Brazilian Senate, through access to government data (BRAZIL, 2015a). In January 2015, the website had forty-five datasets, divided into eight groups: ‘Projetos e Matérias’, ‘Plenário’, ‘Parlamentares’, ‘Composição’, ‘Comissões’, ‘LexML’, ‘Legislação’, and ‘Processo Legislativo’.

The group 'Plenário' contains eight subdivisions: ‘Diários do Senado e do Congresso’; ‘Legislaturase Sessões Legislativas’; ‘Matérias com prazos’; ‘Pronunciamentos de senador’; ‘Questões de Ordem’; ‘Sessões do Plenário’; ‘Tabelas de tipos relacionados a plenário’, and; ‘Votações nominais’ - this last containing data about votes recorded in plenary and information related to sessions, bills, votes, like subjects and votes of each member (BRAZIL, 2015b).

The subdivision 'Votações nominais' consists of 11 items: 9 dump files in XML, containing data on votes grouped annually; 1 hyperlink to a web service, and; 1 hyperlink to a voting search page - this last in HyperText Markup Language (HTML) format, is not the subject of this study.

Characteristics of data structure in retrieval process

Each dump file in XML format has a unique URL, formed by the composition: the domain/primary hierarchy ‘http://legis.senado.leg.br/dadosabertos/dados/’; the prefix ‘ListaVotacoes’ follow by year referring to the data, and; file extension '.xml'.

In dataset retrieval via web service, its present only queries grouped by daily results. For example, to perform data collection on votes in a given month, it is necessary to perform 'x' queries, where 'x' represents the count of days in month. This retrieval of datasets has a URL for each daily result, with an URL value formed by the composition: the domain/primary hierarchy ‘http://legis.senado.leg.br/dadosabertos/plenario/lista/votacao/’; the year, month and day.

In both cases - when collected the dump files and retrieved data via web service - the datasets are explicit in XML language and the collections of elements, attributes and terms available are identical (Table 1).

Element	Associated to element	Type of Data	Attributes
ListaVotacoes	None (Root Element)	Group Element	'*xmlns:xsi*'and'*xsi:noNamespaceSchemaLocation*'
Metadados	ListaVotacoes	Group Element	None
Votacoes	ListaVotacoes	Group Element	None
Versao	*Metadados*	Text	None
VersaoServico	*Metadados*	Integer	None
DescricaoDataSet	*Metadados*	Text	None
Votacao	*Votacoes*	Group Element	None
CodigoSessao	*Votacao*	Integer	None
SiglaCasa	*Votacao*	Text	None
CodigoSessaoLegislativa	*Votacao*	Integer	None
TipoSessao	*Votacao*	Text	None
NumeroSessao	*Votacao*	Integer	None
DataSessao	*Votacao*	Text	None
HoraInicio	*Votacao*	Text	None
CodigoTramitacao	*Votacao*	Integer	None
CodigoSessaoVotacao	*Votacao*	Integer	None
SequencialSessao	*Votacao*	Integer	None
Secreta	*Votacao*	Text	None
DescricaoVotacao	*Votacao*	Text	None
Resultado	*Votacao*	Text	None
TotalVotosSim	*Votacao*	Integer	None
TotalVotosNao	*Votacao*	Integer	None
TotalVotosAbstencao	*Votacao*	Integer	None
CodigoMateria	*Votacao*	Integer	None
SiglaMateria	*Votacao*	Text	None
NumeroMateria	*Votacao*	Integer	None
AnoMateria	*Votacao*	Integer	None

Element	Associated to element	Type of Data	Attributes
Votos	*Votacao*	Group Element	None
VotoParlamentar	*Votos*	Group Element	None
CodigoParlamentar	*VotoParlamentar*	Integer	None
NomeParlamentar	*VotoParlamentar*	Text	None
SexoParlamentar	*VotoParlamentar*	Text	None
Url	*VotoParlamentar*	Text	None
Foto	*VotoParlamentar*	Text	None
Tratamento	*VotoParlamentar*	Text	None
Voto	*VotoParlamentar*	Text	None

Source: Authors

The root element is named ‘ListaVotacoes’ and has two attributes and two elements. Its two attributes are responsible for binding the dataset with a supplementary document (XML Schema), containing the delimitation of available elements, content types, and attributes.

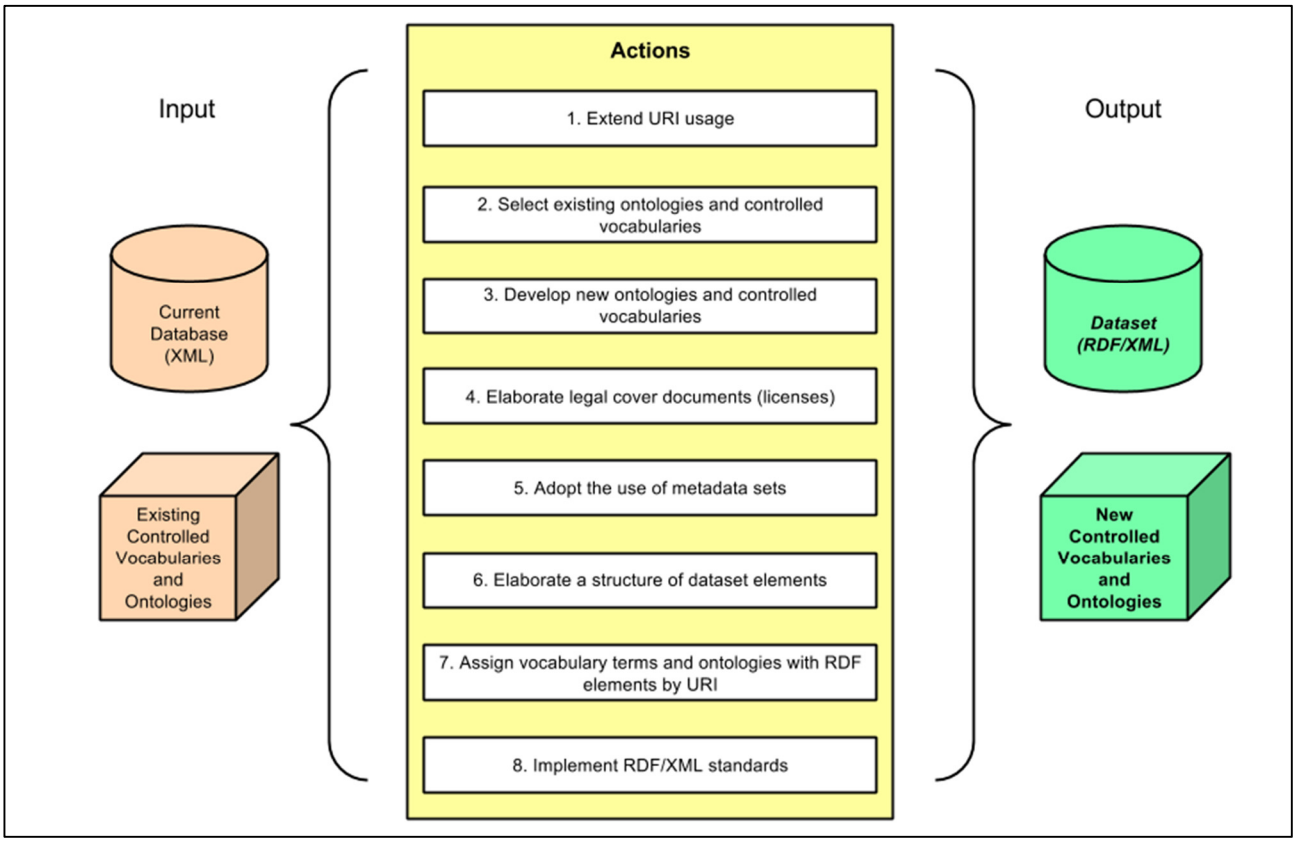
The elements 'Metadados' and 'Votacoes' are grouping elements with their value formed by a set of one or more elements; both with no attributes. The grouping element 'Metadados' has three elements; 'Votacoes' contains at least one or more elements 'Votacao'; and the element 'Votacao' has nineteen elements. None of the elements have attributes.

The grouping element 'Votos' (affiliated to the element 'Votacao') contains one or more elements 'VotoParlamentar', with no attributes. The element 'VotoParlamentar' has seven elements, also without attributes.

Results

From the analysis, eight actions were proposed to apply on existing data retrieval to publish government datasets in LOD (Figure 1).

Figure 1 - Synthesis of actions necessary for the development of government datasets



Sourde: the authors.

Previously to the implementation of these actions, it is important to be available information (input) about characteristics of existing databases and prior knowledge of available ontologies and vocabularies that may be part of the relationships and of the elements from new LOD dataset.

The actions identified in this study can be summarized in:

- Action 1: extend the use of URI identifiers for the identification of elements, attributes, and terms, explicit in additional documents, to facilitate dataset understanding and the rules of this system;
- Action 2: select ontologies and controlled vocabularies widely used by communities that can be useful to explicit relationships and elements of the new LOD dataset;
- Action 3: elaborate specific ontologies and vocabularies for relationships and elements that do not exist in ontologies and vocabularies adopted in Action 2;
- Action 4: develop additional documents containing legal coverage such as licenses of use and copyright, and link these documents with the dataset. It is important that these licenses be explicit in metadata (Action 5);

- Action 5: adopt the use of metadata elements sets of popular initiatives to inform more about datasets' content in the moment of data retrieval by external agents;
- Action 6: elaborate the logical structure of dataset, attributes, elements, values, and validation rules;
- Action 7: assign terms of selected ontologies and vocabularies to the dataset elements, linking them through URI, to extend the semantic load of these data;
- Action 8: implement RDF structures on XML markups, respecting the established forms for RDF/XML documents.

These actions should develop the LOD dataset (explicitly in RDF/XML format or equivalent) as output (result); and vocabularies and ontologies designed to meet the needs of the data context (Votações Nominais).

Conclusions

In the current form of dataset retrieval, it not considered important characteristics in the context of data, such as the use of controlled vocabularies and ontologies, directly interfering in independence between external agents and producers in data collection.

Although there are three metadata elements, there are no information about the content itself, such as author, license, source, date of creation, and date of publication.

The application of a recommendation model in this context served as a guideline for a development of actions proposed in this study, mainly by providing subsidies to elaborate the set of actions required for government datasets in LOD, allowing public managers to see important points that may be changed in available data that can be restructured into LOD datasets and, therefore, collectible by external agents.

It is expected that these actions applied on bills datasets will stimulate the application in other databases, in other spheres, and in other websites; but also stimulate an emergence of new Brazilian government datasets in this area.

References
The references was made following the ABNT rules.

BERNERS-LEE, T. The next Web. In: *TED2009*. Estados Unidos da América: fev. 2009.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked Data – The Story So Far: *International Journal on Semantic Web and Information Systems*, v. 5, n. 3, p. 1–22, 33 2009.

BOHMAN, J. *Public deliberation*: Pluralism, complexity, and democracy. [s.l.] MIT press, 2000.

BRASIL. *Lei No 12.527, de 18 de Novembro de 2011*. Regula 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; altera a Lei no 8.112, de 11 de dezembro de 1990; revoga a Lei no 11.111, de 5 de maio de 2005, e dispositivos da Lei no 8.159, de 8 de janeiro de 1991; e dá outras providências. Portal da Legislação, Brasília, 2011. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Acesso em: 3 abr. 2015.

BRASIL. *Portal e-Cidadania*, 2015a. Disponível em: <http://dadosabertos.senado.gov.br>. Acesso em: 3 abr. 2015.

BRASIL. *Votações Nominais*, 2015b. Disponível em: <http://dadosabertos.senado.gov.br/dataset/votaces-nominais>. Acesso em: 3 abr. 2015.

BRICKLEY, D.; MILLER, L. *FOAF Vocabulary Specification 0.99*, 14 jan. 2014. Disponível em: <http://xmlns.com/foaf/spec/20140114.html>. Acesso em: 27 mar. 2015.

HEATH, T. *Frequently Asked Questions*. In: Linked Data - Connect Distributed Data across the Web, 2015. Disponível em: <http://linkeddata.org/faq>. Acesso em: 22 mar. 2015.

JENTZSCH, A.; CYGANIAK, R.; BIZER, C. *State of the LOD Cloud*, 2011. Disponível em: <http://lod-cloud.net/state>. Acesso em: 24 mar. 2015.

MALIN, A. M. B. Gestão da Informação Governamental: em direção a uma metodologia de avaliação. *DataGramaZero - Revista de Ciência da Informação*, v. 7, n. 5, out. 2006.

MARCONDES, C. H.; JARDIM, J. M. Políticas de Informação Governamental: a construção de Governo Eletrônico na Administração Federal do Brasil. *DataGramaZero - Revista de Ciência da Informação*, v. 4, n. 2, abr. 2003.

SANT'ANA, R. C. G.; RODRIGUES, F. A. Visualização de afinidades entre parlamentares mediante dados de votações no Senado Brasileiro. *Informação & Sociedade*: estudos, v. 23, n. 1, p. 49–59, jan. 2013.

SANTOS, P. L. V. A. DA C.; ALVES, R. C. V. Metadados e Web Semântica para estruturação da Web 2.0 e Web 3.0. *DataGramaZero – Revista de Ciência da Informação*, v. 10, n. 6, dez. 2009.

SANTOS, P. L. V. A. DA C.; SANT'ANA, R. C. G. Transferência da Informação: análise para valoração de unidades de conhecimento. *DataGramaZero - Revista de Ciência da Informação*, v. 3, n. 2, abr. 2002.

SCHMACHTENBERG, M.; BIZER, C.; PAULHEIM, H. Adoption of the Linked Data Best Practices in Different Topical Domains. In: MIKA, P. et al. (Eds.). *The Semantic Web – ISWC 2014*. Cham: Springer International Publishing, 2014a. p. 245–260.

SCHMACHTENBERG, M.; BIZER, C.; PAULHEIM, H. *State of the LOD Cloud 2014*. University of Mannheim, 2014b. Disponível em: <http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state>. Acesso em: 25 mar. 2015.

SEMANTICWEB.ORG. SPARQL endpoint. In: *Semantic Web.org*, 2011. Disponível em: <http://semanticweb.org/index.php?title=SPARQL_endpoint&oldid=52541>. Acesso em: 04 jan. 2015.

W3C. *Data Licensing.* W3C, 2011. Disponível em: <http://www.w3.org/wiki/index.php?title=TaskForces/CommunityProjects/LinkingOpenData/DataLicensing&oldid=49411>. Acesso em: 20 mar. 2015.

W3C. *RDF 1.1 XML Syntax*. W3C, 2014. Disponível em: <http://www.w3.org/TR/2014/REC-rdf-syntax-grammar-20140225>. Acesso em: 22 mar. 2015.

W3C. *Linked Data Platform 1.0*. W3C, 2015. Disponível em: <http://www.w3.org/TR/2015/REC-ldp-20150226>. Acesso em: 24 mar. 2015.