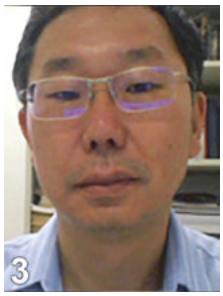
Transforming Open Data to Linked Open Data Using Ontologies for Information Organization in Big Data Environments of the Brazilian Government: the Brazilian Database Government Open Linked Data – DBgoldbr

Marcio Victorino*, Maristela Terto de Holanda**, Edison Ishikawa***, Edgard Costa Oliveira****, Sammohan Chhetri*****

*University of Brasilia, Faculty of Information Science, Central Library Building, Brasília- DF, CEP: 70.910-900, <mcvictorino@unb.br>
University of Brasilia, Department of Computer Science, Brasília- DF, CEP: 70.910-900, **<mholanda@unb.br>, ***<ishikawa@unb.br>
****University of Brasilia, Department of Production Engineering, Brasília- DF, CEP: 70.904-970, <ecosta@unb.br>
*****Brunel University London, Department of Computer Science, Kingston Lane, Uxbridge, Middlesex, UB8 3PH, <sammohan.chhetri@gmail.com>

Marcio Victorino graduated from the Military Institute of Engineering, Brazil, in computer engineering in 1994. He completed his MS in computer and systems engineering from the Military Institute of Engineering, Brazil in 2001. He received his PhD in information science from the University of Brasília (UnB), Brazil in 2011. Since 2017, he has been Professor in the Faculty of Information Science and Computer Science Department at the University of Brasília, Brazil. His current research interests include information organization, information retrieval, big data, NoSQL databases and data science.

Maristela Holanda received her BS in electronic engineering from the Federal University of Rio Grande do Norte (UFRN), Brazil in 1996. She completed her MS degree from the University of Brasília (UnB), Brazil in 1999. She received a PhD from the UFRN in 2007. Since 2009, she has been working for the UnB at the Department of Computer Science. Her current research interests include big data, NoSQL databases, transaction concurrency control systems, geographical database, biological databases and Internet of Thing databases.



Edison Ishikawa graduated from the Military Institute of Engineering, Brazil, in computer engineering in 1992. He received his PhD in computer and systems engineering from the Federal University of Rio de Janeiro, Brazil in 2003. Since 2014, he has been Professor in the Computer Science Department at the University of Brasília, Brazil. His current research interests include semantic computing, distributed systems and systems engineering.

Edgard Costa Oliveira is a professor in the Production Engineering Department at the University of Brasília (UnB). He graduated in language and literature from UniCEUB in 1995. He received his MSc in information science from the University of Brasília in 2001 and his PhD in information science from the University of Brasília in 2006 and a PhD in computer science from the Vrije Universiteit Amsterdam in 2004. He is currently teaching and researching themes such as semantic computing for text authoring, information systems for industrial contexts, risk management and IT governance.

Sammohan Chhetri graduated from City University of London, United Kingdom with BSc (Hons) in business studies in 2012. He received his MSc in business system integration (with SAP technology) from Brunel University of London, United Kingdom in 2016. Since 2017, he has been a PhD researcher in the Computer Science

Department at Brunel University of London, United Kingdom. His current research interests include free and open source software adoption challenges in developing countries.

Victorino, Marcio, Maristela Terto de Holanda, Edison Ishikawa, Edgard Costa Oliveira and Sammohan Chhetri. 2018. "Transforming Open Data to Linked Open Data Using Ontologies for Information Organization in Big Data Environments of the Brazilian Government: the Brazilian Database Government Open Linked Data – DBgoldbr." *Knowledge Organization* 45(6): 443-466. 36 references. DOI: 10.5771/0943-7444-2018-6-443.

Abstract: The Brazilian Government has made a massive volume of structured, semi-structured and non-structured public data available on the web to ensure that the administration is as transparent as possible. Subsequently, providing applications with enough capability to handle this "big data environment" so that vital and decisive information is readily accessible, has become a tremendous challenge. In this environment, data processing is done via new approaches in the area of information and computer science, involving technologies and processes for collecting, representing, storing and disseminating information. Along these lines, this paper presents a conceptual model, the technical architecture and the prototype implementation of a tool, denominated DBgoldbr, designed to classify government public information with the help of ontologies, by transforming open data into open linked data. To achieve this objective, we used "soft system methodology" to identify problems, to collect users needs and to design solutions according to the objectives of specific groups. The DBgoldbr tool was designed to facilitate the search for open data made available by many Brazilian government institutions, so that this data can be reused to support the evaluation and monitoring of social programs, in order to support the design and management of public policies.

Received: 18 September 2017; Revised: 21 February 2018; Accepted: 15 July 2018

Keywords: big data, open data, government data, DBgoldbr

1.0 Introduction

The Brazilian government has made a massive volume of structured, semi-structured or non-structured public data available on the web, in an effort to uphold the principle of transparency in the general administration. It is available online (Transparency Portal of the Brazilian Government—http://www.transparencia.gov.br) and provides access to data on public expenses of some of its representative public institutions, such as Anisio Texeitra National Institute of Educational Studies and Research, INEP (*Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira* - http://portal.inep.gov.br) or the Brazilian Institute of Geography and Statistics, IBGE (*Instituto Brasileiro de Geografia e Estatística*—http://www.ibge.gov.br). These are large public institutions for the management of massive data, such as national educational and demographic data, among others.

Within this context, the need arises to develop integrated applications that are able to quickly generate insights from a huge volume of data, in varied formats, known as "big data application environments." Processing this information resource demands new research from information and computer science, in all phases of the information cycle. Thus, due to its complexity, we propose the implementation of a "big data ecosystem" that can support the analysis of linked open data from the Brazilian government. A thorough review of the literature that addresses some of these issues revealed that the ecosystem model proposed by Demchenko et al. (2014) is also aligned to the goals for managing the type of government data in question. Along with this model, we also consider the use of ontologies to support semantic interoperability in the development of the ecosystem.

The "big data ecosystem" proposed in this study treats data storage from different sources, thereby providing valuable support for the evaluation of social programs and the management of public policies. The extension of this ecosystem consists of our proposed tool, built with a set of components that use an ontology-based semantic information classification, denominated DBgoldbr. Its main goal is to transform open data into linked open data and to facilitate the identification and localization of these data sources. This is done from the semantic description or from the metadata. Thus, this paper presents our proposal of a conceptual model, as well as the technical architecture of DBgoldbr and its interfaces.

2.0 Open data

According to the Open Knowledge Foundation (https://opendefinition.org/), "open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)." Eaves (2009), an open data activist and considered one of the main experts on the matter, and also a specialist in public policies, proposed three laws that were also adopted by the W3C:

– If it cannot be found or indexed, it does not exist
– If it is not available in open and machine readable format, it cannot engage
– If a legal framework does not allow it to be repurposed, it does not empower.

Many countries have promoted the active participation of society in public administration by making data open and accessible, as evidence of transperency in their activities. Thus, the general population can accompany and keep watch on their government leaders, aiding them by making suggestions and critiques, in an effort to promote excellence in public management.

2.1. Semantic web and open data

From the beginning of the web, an enormous volume of data has been published daily in a uncontrolled and disorganized way, which impedes the access, understanding and processing of these data. Given this situation, Berners-Lee (2001) presented what would become the semantic web. It deals with the evolution of the World Wide Web itself, in which data using standards established by the World Wide Web Consortium—W3C, among which are: eXtensible Markup Language (XML), Resource Description Framework (RDF), SPARQL Protocol and RDF Query Language (SPARQL), Uniform Resource Locator (URL), Web Ontology Language (OWL). In this way, the data published would be accessible and processible on machines and passable for organization processes that can facilitate the presentation of these data, the generation of new data, the link with other groups of data and increase the knowledge to support the decision.

Although there is no direct relationship between the concepts of open data and the semantic web, when semantic web resources are joined with the publication of open data, the goal is to reach a significant improvement in the organization and access of these data to make these open data available in a semantic structure is necessary to take into account the stratified structure proposed by Berners-Lee for the semantic web. In this structure, RDF is highlighted, as well as the use of metadata and mainly the use of ontologies.

One of the main objectives of the RDF is to construct a network of distributed information, where the nodes are semantically linked, forming a large global graph, with information originating from various different sources around the planet. On the other hand, ontology is the resource used to construct an organized relationship between terms within a domain, favoring the possibility of contexualizing the data, turning the process of interpretation of the data more efficient and facilitating the recovery of information through computational tools. According to Berners-Lee (2006), the global graph we created semantically linked and structured through the use of technologies previewed in the layers of the semantic web, among them, RDF and ontologies, are called linked data. Heath and Bizer (2011) affirm that linked data is a set of best practices for the publication and linking of structured data on the web, allowing the establishment of links between items from different sources of data to form one single space of global data. According to Berners-Lee (2006), linked open data (LOD) is linked data that is released under an open license.

2.2. Linked open data

Berners-Lee (2006) affirms that the construction of linked data and of linked open data is based on four principles—that it:

1) Use URIs as names for things;
2) Use HTTP URIs so that people can look up those names;
3) When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL);
4) Include links to other URIs so that they can discover more things.

There is a category of open data originating from government agencies, denominated "open government data." The W3C defines open government data as the publication and disclosure of public information on the web, shared in an open gross format, logically comprehensible, allowing the reuse of information in digital applications developed by society.

Based on the relevance of interoperability of open government data, Berners-Lee (2006) proposed the five-star system, which classifies the degree of data openness—the more open, the more number of stars, and the easier it is to enrich the data. The five stars of linked open data are:

– "One star": make it available on the web (whatever format) under an open license;
– "Two stars": make it available as structured data (e.g., Excel instead of image scan of a table);
– "Three stars": make it available in a non-proprietary open format (e.g., CSV instead of Excel);
– "Four stars": following all of the prior rules, but within the standards established by the W3C (Resource Description Framework (RDF) and SPARQL Protocol and RDF Query Language (SPARQL)), use of Uniform Resource Locator (URL) for the identification of things and properties, such that everyone can be directed to their publications;
– "Five stars": link your data to other data to provide context.

An analysis of Brazilian government data from public administration agencies—which is made available on the internet—showed that this massive data can be classified as a three-star level, because they are available on the Internet

in a structured way, in a non-proprietary CSV format, and any person or entity can download them.

2.3. Brazilian government open data

As previously mentioned, the Brazilian government is increasingly taking steps to provide as much transparency as possible of its administration by making massive amounts of information that is of interest to the public available on the web. In 2006, the Controller General of the Union (Controladoria-Geral da União—CGU), currently known as the Ministry of Transparency, Fiscalization and Control is the agency responsable for the internal control, of the federal government, together with the Ministry of Planning, Budget and Management (Ministério do Planejamento, Orçamento e Gestão—MPOG), established by the Interministry Ac CGU/MPOG n. 140, of March, 16 2006. This resolution determines that the agencies and offices of the Public Federal Administration are responsible for maintaining on their respective electronic sites the detailed information about specific aspects of operation, such as budget spending, bidding processes, contracting, partnerships with cities and states, among others. These must be maintained on specific pages managed by Controller General of the Union, called Public Transparency Pages (Portal da Transparência Ministério Da Transparência E Controladoria-Geral Da União http://www3.transparencia.gov.br/).

According to the CGU, the mission of the Public Transparency Pages is to promote visibility of the public spending and motivate social control to ensure that the Public Administration practices are founded and executed legally and ethically. In addition to the Public Transparency Pages, which present data referring to the spending of each agency and office of the Public Federal Administration, the federal government also provides information about the application of public resources, from the consolidation of millions of data originating from diverse federal agencies relative to the programs and actions of the government on a single site called the Transparency Portal of the Brazilian Government (2012).

Another important initiative of the Brazilian government in this is the National Infrastructure of Open Data (Infraestrutura Nacional de Dados Abertos—INDA; https://www.governoeletronico.gov.br/eixos-de-atuacao/cidadao/dados-abertos/inda-infraestrutura-nacional-de-dados-abertos), which consists of a set of standards, technologies, procedures and mechanisms of control necessary to attend disseminate and share public data and information modeled as open data. In other words, INDA is the policy of the Brazilian government for open data. Notably, INDA is aligned with the W3C standards. In this way, the Brazilian government provides access to massive volumes of structured, semi-structured and unstructured public data of interest to individuals and society in general. This scenario is characterized as a big data environment comprised of open data that are categorized by a three-star scale according to the criteria proposed by Berners-Lee (2006), in line with their open data policy known as INDA. However, in the model proposed in this research, we aim to raise the openness level of data from a level three to a level five, by making use of ontologies to transform open data into linked open data.

3.0 Ontologies

The term ontology originates from the Greek words ontos (being) and logos (treaty). The original term refers to the Aristotelian word "category," used to classify anything. Aristotle presented the categories that classify any entity and introduced the term "differentia" for properties that distinguish the different aspects of the same genre. The term "ontology" has been used in various fields, including language and cognition and computer and information sciences. In these areas, the term ontology is directly or indirectly related to the treatment and communication of information and/or knowledge. Corcho et al. (2003) affirm that the word "ontology" has its origins in philosophy and signifies the systematic explanation of being. This work highlights that "ontology" became a relevant term in the field of the engineering of knowledge, and that many definitions have been created, some having changed and evolved over time.

According to Sowa (1999), ontology is the study of the categories of things that exist or can exist in a particular domain. This study produces a catalogue of types of things that exist or can exist in a domain of interest D, from the perspective of a person who uses a language L, aiming to talk about D. Types of ontologies represent predicates—meanings of words or concepts and types of relations of the language L—when used to discuss issues in the domain D.

Guarino (1998) highlights the predominant use of the term ontologies in AI—artificial intelligence, defined as an engineering artifact of a specific vocabulary, used to describe a determined reality and a set of explicit suppositions, related to the intentional meaning of words in a vocabulary. Many definitions were created in the last decade but the most cited in this context comes from Gruber (1993):

– Definition 1: Gruber (1993) proposed that ontology is the specification of a conceptualization;
– Definition 2: Borst (1997) complemented Gruber by affirming that ontology is the specification of a shared conceptualization; and

– Definition 3: Studer et al. (1998) mixed both definitions by saying that ontology is a formal explicit specification of a shared conceptualization.

Uschold and Grüninger (1996) explain that ontology is the term used to refer to a shared understanding of a specific domain of interest that can be used as a unified structure to solve many types of problems, such as those related to knowledge sharing and interoperability. According to Hjørland (2017), ontologies are widely recognized types of knowledge organization systems (KOSs). According to the author, the objective of knowledge organization (KO) is to discover or interpret some type of configuration of order. Smiraglia (2017, 315) affirms that, "Knowledge Organization is the science of the order of knowledge, based on the central unit of the concept."

In a discussion on the notion of knowledge organization systems (KOS), Mazzocchi (2018) points out that ontologies aim to represent complex relationships between entities and add rules and axioms that support logical reasoning. According to this author, formal ontologies function as conceptual vocabularies and provide properties and instances. They are a valuable resource in the area of information retrieval, knowledge reuse and automatic inference of new knowledge. Given this, ontologies can be considered extremely useful tools for organizing knowledge. In this study, which deals with the organization of knowledge in a big data environment of open data from the Brazilian government, ontologies are used to contextualize open data, transforming them into linked open data, fomenting an increase of the level of openess of the Brazilian government data from three to five stars, as defined by Berners-Lee (2006).

4.0 Big data ecosystem

There are various definitions and understandings for the term "big data." One of the most widely accepted is the 3V definition presented by Laney (2001) and ratified by Beyer and Laney (2012, 2014): "Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." However, Demchenko et al. (2013) proposed the definition of big data as having the following 5V properties: volume, velocity, variety that constitute native/original big data properties, and value and veracity, acquired as a result of initial data classification and processing in the context of a specific process or model. Uddin and Gupta (2014) proposed a 7V model, which includes volatility and visualization. Volatility refers to data meaning, which is constantly changing and visualization relates to all means of expressing content that are easy to understand and read.

Boyd and Crawford (2012) propose a broader approach. These authors define big data as a cultural, technological and academic phenomenon, which is based on the interaction of:

– Technology: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.
– Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical and legal claims.
– Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity and accuracy.

Frické (2015) also observed researchers in the field recognizing that the big data phenomenon provided scientific research with results never before imagined. The author, after a thorough reflection concluded that the capacity to collect great quantities of data provides bigger samples and the tests of the theories tend to present more consistent results than previously obtained. However, Frické highlights that to do science we need problems, reflections, theories, proposed experiments. He concludes that be that as it may, ultimately, science is in need of more theories and less data.

Crawford (2013) warns of the risks of hidden bias in the data collecting and analysis stages in big data environments. The author exemplifies this type of problem citing the crossing of data of supermarket purchases and tweets collected from October 27th to November 1st of 2012, a little before, during and after Hurricane Sandy, which devastaded the east coast of the United States on October 28th, 2012. Due to the fact that the majority of tweets about Sandy were sent from Manhattan, which created the illusion that this city was the center of the hurricane. However, the cities most affected—Breezy Point, Coney Island and Rockaway—did not have any tweets, mainly because the number of people with smartphones, the popularity (or lack thereof) of Twitter, and the problems with power outtages impeded the use of smartphones and other electronic devices. Corroborating the controversial aspects of big data, Ekbia et al. (2015) present an analysis of the most varied conceptual and practical dilemas involving this environment. The authors provide a synthesis, based on the relevant bibliograpy from the sciences, humanities, politics and in the commercial literature.

There are some definitions of "big data ecosystems," such as Shin and Choi (2015) for instance, where big data is seen as an ecological ecosystem that involves the following aspects: technology, government, industry, market, users and society, taking into account the effects of big data

in all sectors involved. Big data is defined by Demchenko et al. (2014) as a complex system of technical facilities and components built around a source of specific data and its application. The complex of interrelated components is used for storage, processing, visualization and the delivery of the results obtained from the big data. This ecosystem comprises big data in general as well as the several categories of architectural components that follow.

4.1 Models and data structures

According to Demchenko et al. (2014), the many stages of big data transformation require different data structures, models and formats, including the possibility of processing both structured and non-structured data. It is possible that the data structure and corresponding models undergo changes during the different stages of data processing. However, it is important to keep the link between these structures.

Figure 1 shows examples of structures, models and links of data from the original figures from Demchenko et al. (2014). The top of the figure represents a data model containing information such as: structures and data types, as well as links between data—raw data into information, and from information to presentation. Raw data represents data in its original state, as it was brought from its original source. Below, the origins for data transformations during the processing cycle are represented with arrows.

4.2 Big data architecture

Big data architechture is formed by a set of technologies and componets for the big data processing and analysis. Demchenko et al. (2014) mention two groups of main technologies named big data analytics infrastructure (BDAI), which are: a general architecture formed by technologies and components for storage, computing, network, devices and operational support to big data; and the processing and analysis architecture, which is formed by tools for data presentation and visualizations, as well as analysis and processing.

4.3 Big data lifecycle management

Demchenko et al. (2014) emphasize the need to use scientific methods to obtain the benefits of new opportunities for data collection and mining in order to acquire the necessary information. The information lifecycle involves the storage and preservation of data in different stages, in order to allow the reuse of analytic research of processed data and published results. However, this is only possible if necessary solutions have been implemented to allow complete identification, crossed referencing and data linkage. Data integrity, control and auditing must be supported during all data lifecycle.

Figure 2 shows the top storage layer, where data are persisted and data models that represent them during the whole life cycle. In the lower portion of Figure 2, we present the phases of data transformation, starting with data collection (and data registration), data treatment (the filtering, enriching and classification of data) data processing and analysis (the analysis, modelling and prediction of data), and the delivery of results generated to the consumer data analytics application via the delivery and visualization of data process.

4.4 Big data security infrastructure

Big data security infrastructure gathers the necessary set of components and policies to provide data access control and a safe processing environment. The ecosystem to treat open data in the Brazilian government was developed by extending the big data ecosystem from Demchenko et al. (2014), with the addition of a structure for the semantic classification of sources in the category of components to manage the big data lifecycle, and thus providing data linkage via its semantic representation.

This semantic classification structure constitutes our proposal for the Brazilian Database Government Open Linked Data - DBgoldbr, implemented using the W3C standards, via the resource description framework (RDF), SPARQL Protocol and RDF Query Language (SPARQL), Uniform Resource Locator (URL) and Web Ontology Language (OWL). This allows information sources to achieve a four-star level for data openness. In addition, DBgoldbr can also create references to other open data sources of the government because it is structured to semantically connect open data as they are made available, with the help of an ontology of the state government that allows a semantic linkage between government sources, thereby achieving a five star level for the open data context.

Aware of the challenges in a big data environment, Crawford et al. (2014) highlight that the analysis of massive data must be ethical and respect civil liberties and privacy, among other characteristics. Respecting these principles, it is possible that access to the ecosystem of big data for open data of the Brazilian government contributes to improving public services and understanding of governmental activities. This access may also foment a more effective management of public resources, in addition to increasing civilian participation in public management.

5.0 Resource description framework

Resource Description Framework (RDF) is a directed, labeled graph data format for representing information on

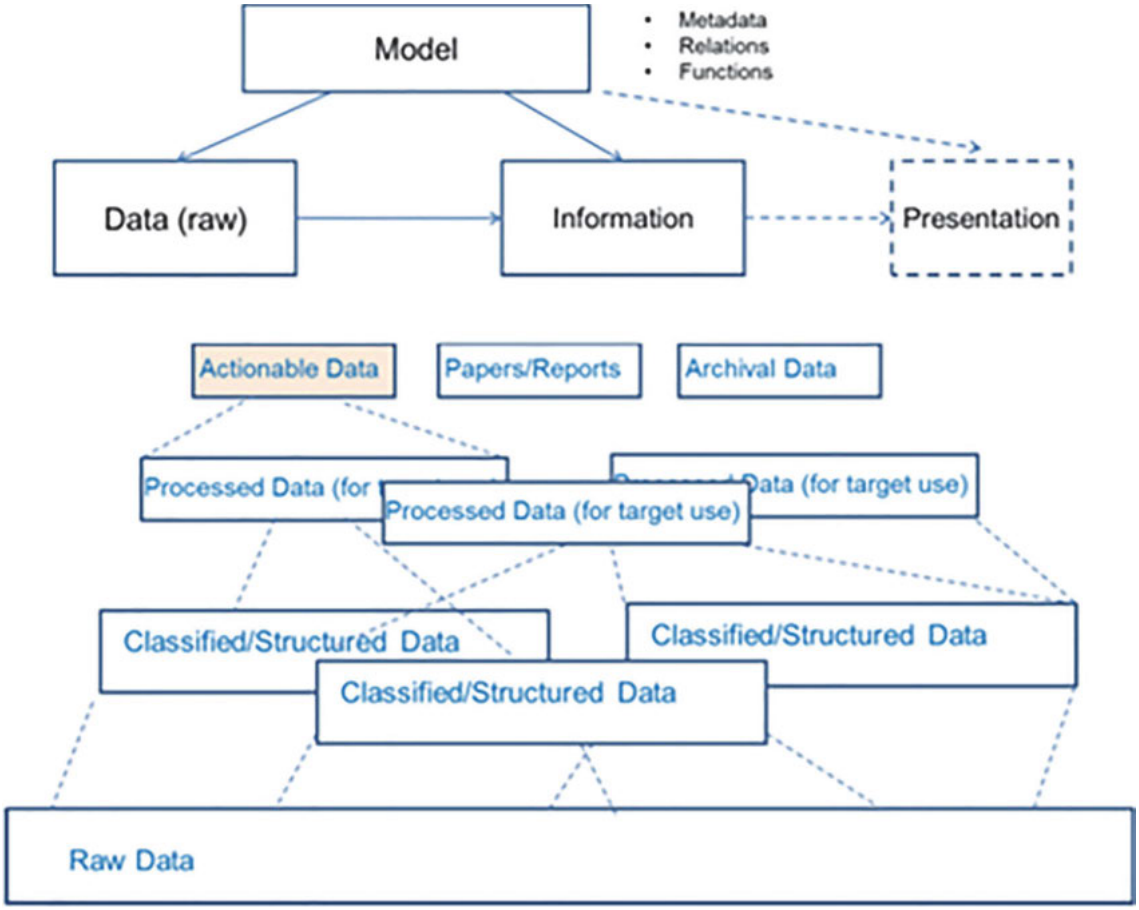


Figure 1. Big Data structures, models and their links at different processing stages.

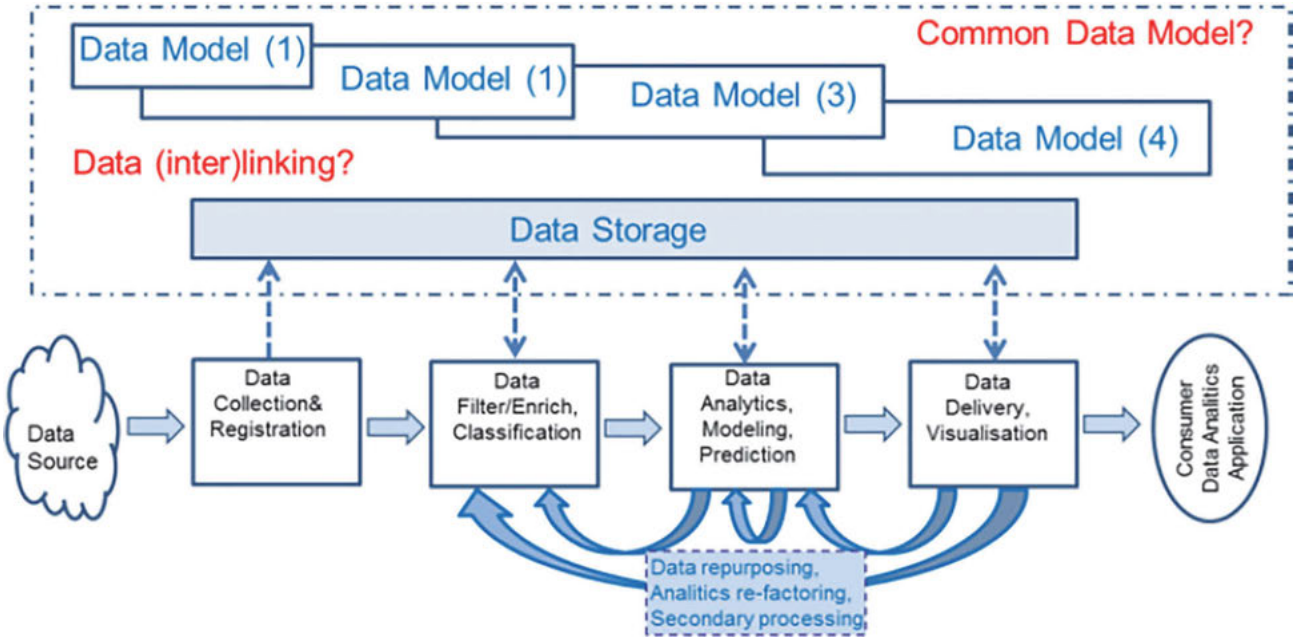


Figure 2. Big data lifecycle in big data ecosystem.

the web. RDF is often used to represent, among other things, personal information, social networks, metadata about digital artifacts, in addition to providing a means of integration over disparate sources of information, according to Prud'hommeaux and Seaborne (2008). On the other hand, ontology is a formal and explicit specification of a shared conceptualization and can be also considered as a common vocabulary inside a domain where everyone can understand the same semantics whenever interpreting the meaning of a term. Figure 3 shows the graphic representation of an RDF triple, according to the Jena Apache Project (2013).

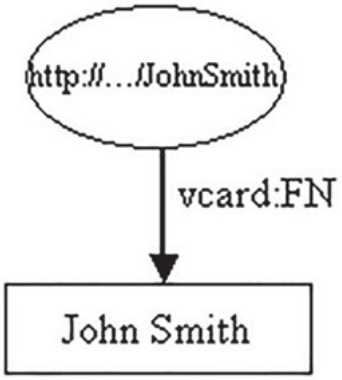


Figure 3. Graphic representation of an RDF triple.

As shown in Figure 3, RDF is best thought of as the form of node and arc diagrams. Each arc in an RDF Model is called a statement. Each statement asserts a fact about a resource. A statement has three parts:

– the subject is the resource from which the arc leaves;
– the predicate is the property that labels the arc;
– the object is the resource or literal pointed to by the arc.

In Figure 3, the resource, John Smith, is shown as an ellipse and is identified by a Uniform Resource Identifier (URI), in this case "http://.../JohnSmith." Resources have properties. This example shows only one property, John Smith's full name. According to Iannella and McKinney (2014), The term "full name," represented by "vcard:FN," belongs to the vCard ontology. vCard ontology is a specification developed by the Internet Engineering Task Force (IETF) for the description of people and organizations. Figure 4 shows the same triple of Figure 3 now in RDF XML, according to the Jena Apache Project (2013).

6.0 Methodology

According to Checkland (1981), Soft Systems Methodology (SSM) was one of the methods used to collect information of the needs of users and to design the DBgoldbr with the stakeholders. SSM is based on a systemic view, starting from the problem domain from a holistic perspective in order to recognize which parts of the system need to be interconnected. In this way, we were able to make changes in the system and to analyse the effect on other parts of the system. The SSM was followed, using Checkland's seven stages:

– Stage 1: Problem situation unstructured;
– Stage 2: Problem situation expressed;
– Stage 3: Root definition of relevant systems;
– Stage 4: Conceptual models;
– Stage 5: Comparing conceptual models (4) with reality (2);
– Stage 6: Defining changes desirable and feasible;
– Stage 7: Actions to improve the problem situation.

SSM stages one, two, five, six and seven refer to activities that involve people in the real world and stages three and four refer to the systemic thought.

6.1. Problem situation unstructured

At this stage in the application of the SSM, the user and publishers of open government data were consulted, and an analysis was done of the government sites that provide open data. Through this analysis, it was possible to confer

```
<rdf:RDF
   xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
   xmlns:vcard='http://www.w3.org/2001/vcard-rdf/3.0#'
>
   <rdf:Description rdf:about='http://somewhere/JohnSmith` >
   <vcard:FN>John Smith</vcard:FN>
   </rdf:Description>
</rdf:RDF>
```

Figure 4. RDF XML syntax of an RDF triple.

that various organs of the Brazilian government publish enormous volumes of open data; however, finding and processing it in an integrated way is very difficult.

6.2 Problem situation expressed

At this phase, a structured outline of the problem situation, as preconceived by the SSM, was sought. The outline was obtained through the analysis of initiatives carried out by the Brazilian government that promote the opening of data produced by government agencies. This analysis shows that, although the Brazilian government has been making efforts to disclose open government data since 2006, there is no infrastructure for semantic support during the process of publishing these data, nor tools for information retrieval that facilitate interpreting the semantics of the data available, to improve results.

6.3 Root definition of relevant systems

The objective of this stage was to elaborate the root definition (RD), which consists of a concise description of a specific vision of a system of human activity. This RD was elaborated from the identification and description of the relevant systems, identified in the structured problem-situation, with the goal of expressing the principles proposed in the system of human activity in question. We sought to capture the essence of the system, incorporating the activities considered significant to its performance. Thus, the RD was proposed for the system previously cited as being "a system that makes it possible for the open data currently published by the Brazilian government to be published as governament linked open data, through the use of technologies recommended by the semantic web, aiming to make these data passable for processes of organization that can facilitate the presentation of these data, the generation of new data, the linking with groups of data and the increase of knowledge to support the making of decisions."

6.4 Conceptual models

At this stage the conceptual model was developed, based on the process of transformation established in the root definition (RD), which represent the actions necessary for the intended transformation intended to be carried in an integral way. Section seven presents this model.

6.5 Comparing conceptual models (4) with reality (2)

At this stage in the methodology, a comparison of the aspects of current reality is made, expressed in stage two, with the conceptual model developed in stage four. The conceptual model was evaluated, and the conclusion drawn was that implementing this model is technically viable and benefits the publication and recovery of open government data.

6.6 Defining changes desirable and feasible

Considering the comparison between the conceptual model with reality, carried out in the previous stage, it is clear that various actions can promote desirable and viable changes, as follows:

- Using RDF for the creation of graphs of open government data;
- Using ontologies for the semantic description of open government data;
- Providing a tool to support the process of publication and recovery of open government data, turning this process more formal and standardized.

6.7 Actions to improve the problem situation

This last stage aims to detail how actions will be carried out and implemented. As the results of the actions are not entirely predictable, according to Chekland (1981), after their implementation, there is a verifiable need to restart the SSM process. During the carrying out of the present study, only one SSM cycle was completed, which culminated in the development of the tool denominated, Brazilian Database Government Open Linked Data—DBgoldbr, presented in Section 8.

7.0 Brazilian Database Government Open Linked Data—DBgoldbr conceptual model

The DBgoldbr became viable due to the Brazilian government's interest in sharing open data on the internet, on the Public Transparency Pages and the Transparency Portal, providing greater transparency of the administration. These web environments present public data of the federal government, from twenty-seven units of the federation and 5,570 Brazilian cities. Among the mandatory issues covered by law, there are data regarding budgetary spending, bidding, contracts and partnerships, as well as many other issues. Broadening these initiatives of the Brazilian federal government, various public federal, state and municipal agencies, provide access to information about diverse areas, such as education, social welfare, health, safety, demographics, economy and the environment. Following the policy directives of the Brazilian government with regard to open data, INDA, these data are accessible in a "CSV" format and normally have a file in "PDF" format associated with the description of its semantics. Each publishing agency describes its own data in its own particular

fashion, making it difficult for users to understand them. Another complicating aspect in the use of these data is the absence of a specific tool to help locate them. Aware of these difficulties, this work aims to provide access to a software for knowledge organization that helps users working with open data find the data they are interested in and understand their semantics, facilitating, thus, their use singularly or collectively in an integrated way.

A review of the literature presents initiatives with similar objectives. For example, Zeng et al. (2014) presents a study on computer-assisted semantic analysis, using OpenCalais, to verify its potential for generating access to the issue at the levels of "description" and "identification" for archival record groups and philosophy theses. Shiri (2014) presents a facet analysis of key topics and issues in big data, aiming to facilitate comprehension with respect to the particularities of a "big data environment"—its most important components and characteristics. Soergel (2015) highlights the omnipresence of knowledge organizaiton and discusses the use of knowledge organization resources in a variety of informational environments, among these, linked data, big data and data analytics. Souza et al. (2012) analyze various studies on the classification of knowledge organization systems, presenting their positive and negative points and proposing a taxonomy of KOS dimensions. The present work is based on the previously mentioned research, which served to inform the implementation of the first prototype of the DBgoldbr software. Corresponding to some of the concepts presented in the literature, this software made use of ontologies for the classification of open data from the Brazilian government.

The target audience for this prototype of the DBgoldbr are information professionals with basic knowledge in the field of classification of information, mainly ontologies, and who know how to use RDF triples to represent information available on the web. Later, it is intended to serve the needs of the end users through the implementation of interfaces that are more user-friendly and that encapsulate the RDF triples.

The Multimodal Digital Media (MDM) project, was one of the projects that inspired the implementation of DBgoldbr. This project is being developed through a partnership with the University of Brasilia (Brasília, Brazil) and the University of Brunel (London, United Kingdom), by a multidisciplinary team of researchers in the area of information sciences, computer sciences and journalism. The MDM project deals with computational semantic models in a context of convergence in journalistic publishing, according to humanistic and social projects. The model is centered on the citizen, through the interactions on the internet that facilitate the construction of intercultural and creative dialogues, according to the theory of collective intelligence and the semantic sphere by Lévy (2011).

During the development of the MDM project, it was observed that the journalists were using Google as a tool for finding sources of data from the Brazilian government to analyze and generate news. However, the results offered by Google did not present the desired quality. After verifying that the Brazilian government already had specific legislation and an open data policy—INDA—, which recommends a set of rules regarding the publication of open data, among them a standard "CVS" format, the concept of DBgoldbr began. The objective of DBgoldbr then was to provide a tool for consulting open data from the Brazilian government, based on semantics, with the capacity to provide better results than the quality of information gathered through Google.

In fact, this project intends to serve the needs of public agencies, news agencies and governmental and non-governmental organizations for finding sources of open data, through the semantic description. This will allow them to be used in the analysis of indicators of the most varied types, for example, educational, social, those related to health and economics or to announce successful and/or disasterous government policies of public interest. Notably, the objective of this study is not to develop ontologies, but rather to develop a flexible software, capable of generating and consulting RDF triples based on ontologies, taxonomies and metadata described in "OWL" format. Thus, the same source of data can be described by various distinct RDF triples, being able to be classified using distinct ontologies.

DBgoldbr aims to perform a semantic classification of previously published data sources. This semantic classification is based on the use of ontologies and RDF triples in order to transform open published data into open linked data. Figure 5 presents a conceptual view of how these resources are used.

As we can see in Figure 5, all data sources are published in the CSV format and are represented via URIs, as well as being part of the triples as subjects. The data sources, located on the web sites of the government institutions, will be part of the triples as objects. The predicates are based on controlled vocabularies and for that purpose we use ontologies. The generated triples can be stored in RDF XML, N-Triples, Turtle or JSON. Therefore, we expect a big volume of RDF triples to represent information. The environment uses an RDF database management system to store the triples, which is a persistency device available from the Jena Apache Project (2013).

By using DBgoldbr, we intend to enhance the quality of search results from queries that users carry out with the open data sources from the Brazilian government on the web. Currently, Google's search engine is one of the most popular search tools used. Figure 6 presents a simplified representation of the conventional search process by a

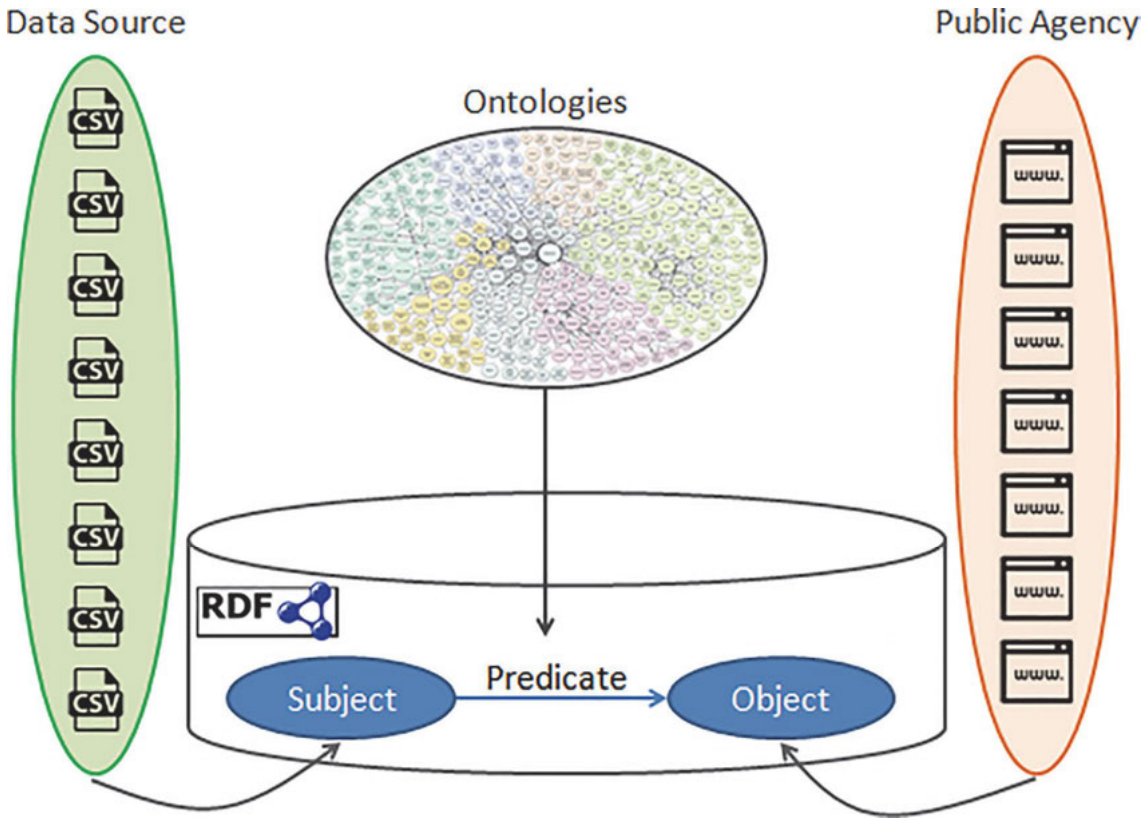


Figure 5. Conceptual View of Brazilian Database Government Open Linked Data—DBgoldbr.

user searching for open data sources published by the Brazilian government. Figure 6 shows how users type keywords and the search engine uses algorithms to find documents that fit this keyword. In this case Google uses the PageRank algorithm described by Page et al. (1999).

DBgoldbr aims to replace these conventional search tools of open data sources made available by the Brazilian government. For its construction, we built an RDF triple repository with a semantic description of the data sources. Initially, these triples were created in the instant the source is made available, followed by the use of one of the DBgoldbr interfaces. The definition of an automated classification technique, in order to perform data mining, is needed. Figure 7 presents the source search process using DBgoldbr.

In Figure 7, the DBgoldbr is shown to have a native repository of RDF triples that semantically describe the published sources (subject), the publishing entities (object) and the predicates which are the terms obtained from the ontologies. This feature offers users search options based on the available data sources, obtained from the institutions that have published data or vocabularies. In turn, this offers a common understanding within the domain and thus improves the quality of results. The DBgoldbr resources are presented in detail in the following sections.

8.0 Architecture and interfaces of the DBgoldbr

The logical architecture of the DBgoldbr is organized in two parts. The first part is composed of the registration of publishing entities, published sources and ontologies. The second part involves the necessary resources for the creation, storage and query of RDF triples, described as follows.

8.1. Part one of the DBgoldbr architecture

Figure 8 shows the first part of the logical architecture of the DBgoldbr, represented in layers: the first layer is the application developed in Java (jdk1.8.0_101), the second layer is the Application Program Interface (API) Java Database Connectivity (JDBC), the third layer is the connection driver of the Database Management System (DBMS) MySQL and the fourth layer is the DBMS MySQL (version 5.6.22.0). The second and third layers are used to allow the connection between the application and the database, based on Oracle and MySQL, respectively.

In this architecture, the Relational Database MySQL was the persistence device used to persist data over the publishing entities—government institutions or any other entity capable of generating open data for the government. The database also persists the published sources and the

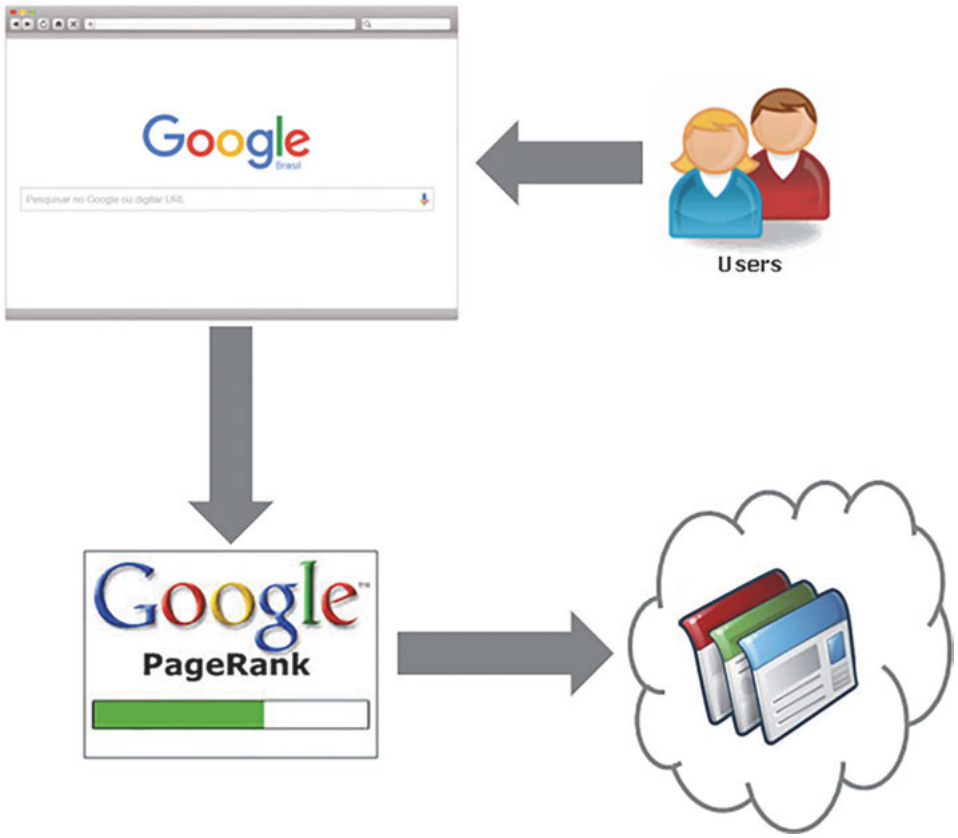


Figure 6. Conventional process of searching open data sources.

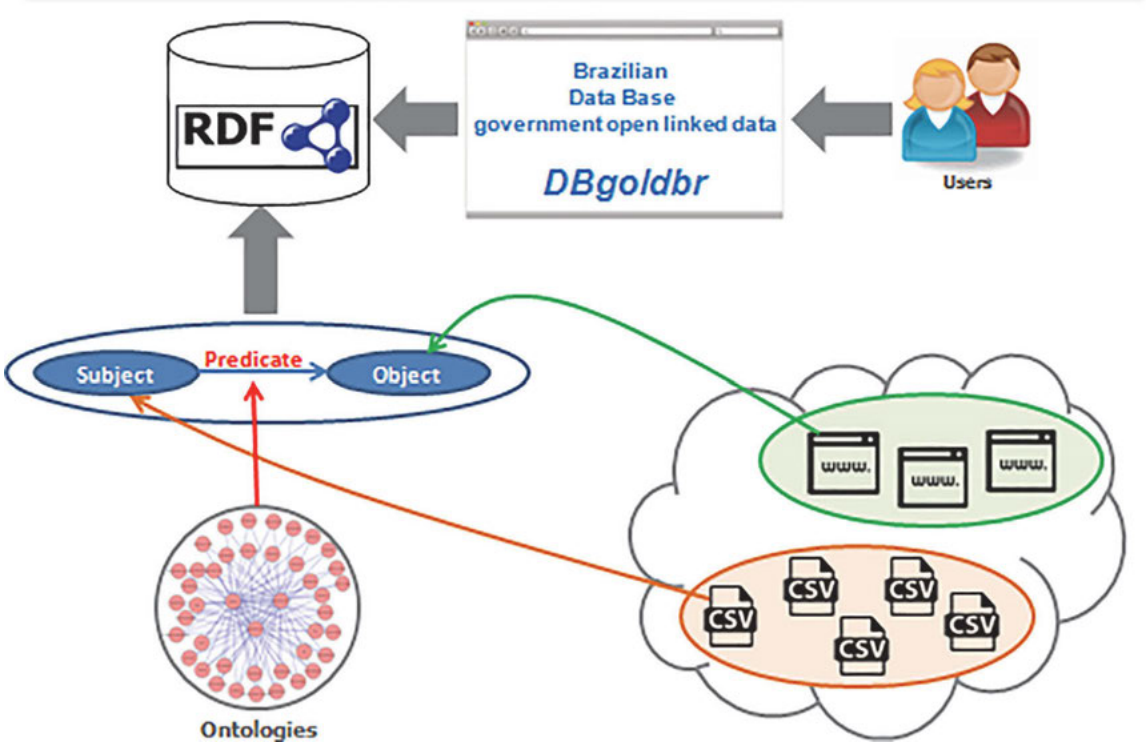


Figure 7. Using DBgoldbr to search for open data sources.

ontologies in relational tables. In this figure, the Java Application Client layer was developed via the implementation of a prototype of DBgoldbr in Java.

Figure 9 represents the homepage of the prototype, with the main menu "Registration" which represents the first part of the architecture proposed here. Via the menu option, it is possible to insert data about the publishing entities, published sources and ontologies in the DBMS MySQL.

Figure 10 presents each interface related to the menu options "Publishing Entity" (a), "Published Source" (b)

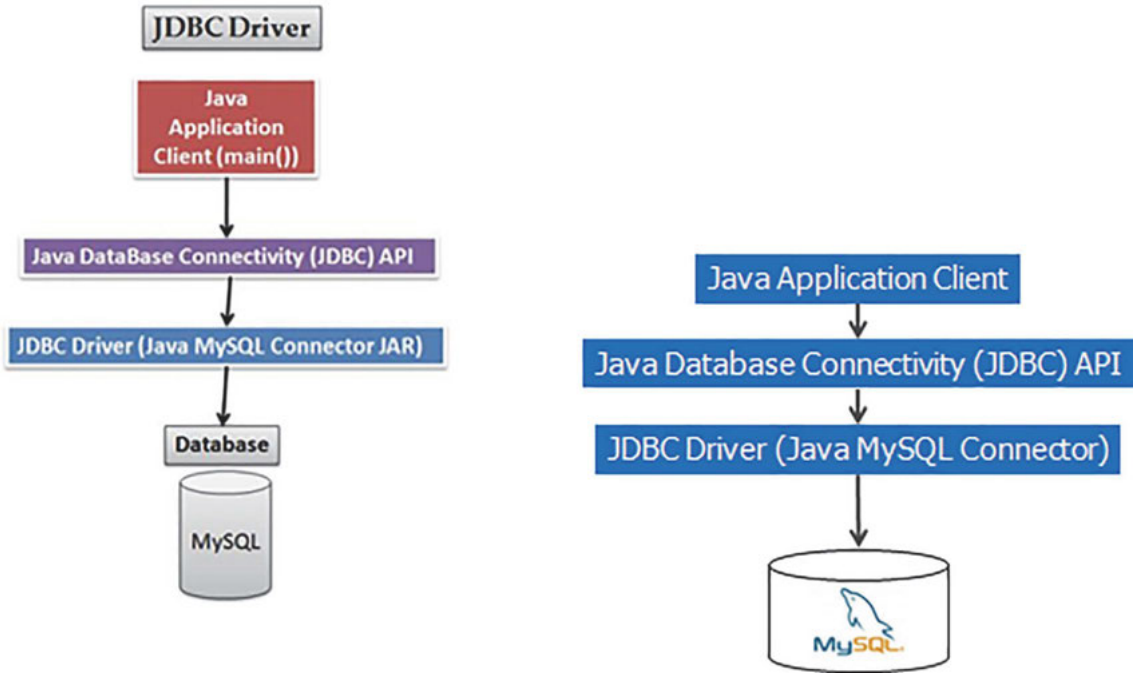


Figure 8. Part I of the lgical arhitecture of DBgoldbr.

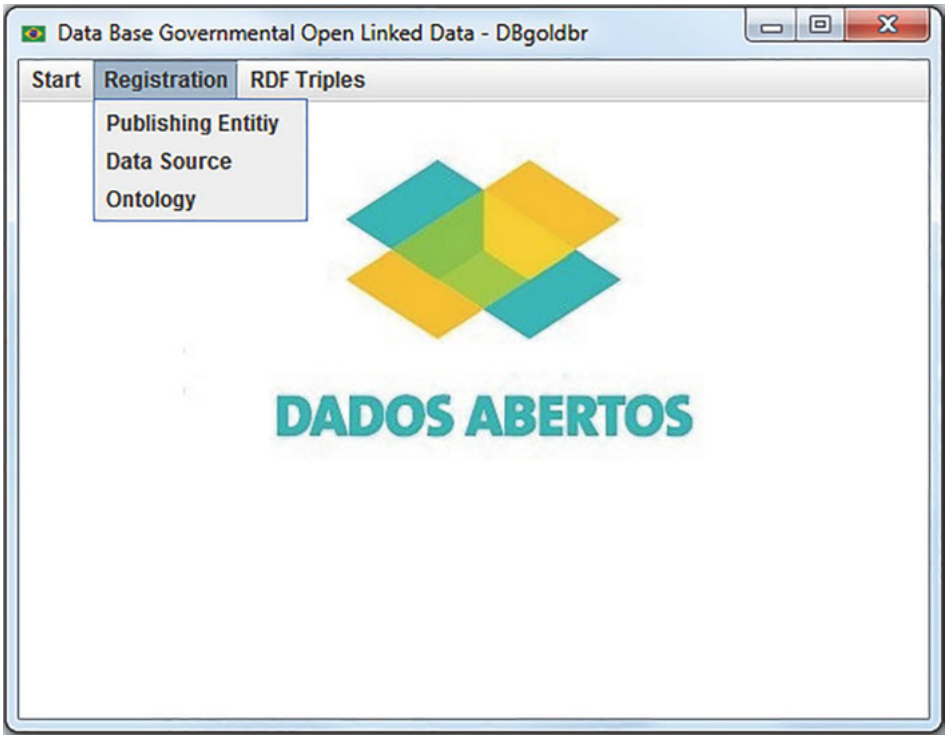


Figure 9. DBgoldbr prototype interface, rgistration menu

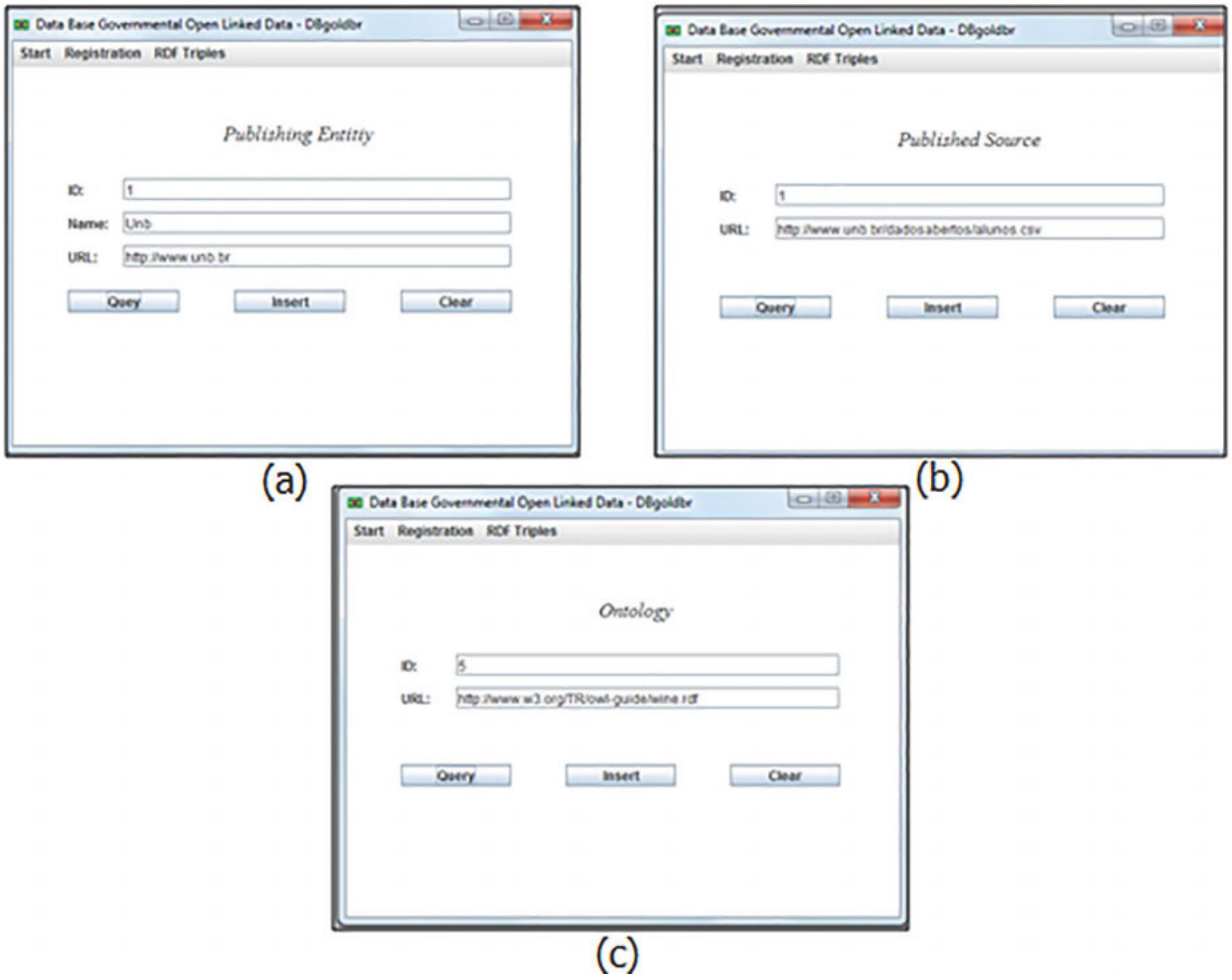


Figure 10. Interfaces with each menu option of the DBgoldbr.

and "Ontology" (c). All interfaces offer insertion of new data or data query to existing registers.

Figure 10 (a) presents the data stored about the publishing entities, their code, name and URL with a link to the official website of these entities. Along with other data, the URL and the link to the location of the source may also be included. Figure 10 (b) presents the data about the published sources, such as metadata in the Dublin Core Metadata Initiative (DCMI) (2012) standards. Figure 10 (c) shows the data regarding the ontologies, followed by the OWL files that contain the ontologies being used or the URL to its source.

Figure 11 presents the tables created by MySQL about the "Publishing Entity" (a), "Published Source" (b) and "Ontology," (c) which are processed by the interfaces shown in Figure 10.

8.2. Part two of the logic architecture of the DBgoldbr

Figure 12 presents the second part of the logic architecture of DBgoldbr, represented in layers, as follows: the first layer is a Java Application Client, representing the application, which is being developed in Java (jdk1.8.0_101), and the other layers are part of the Apache Jena framework. This framework has APIS to process RDF files, which represent the triples, the OWL files, which represent the ontologies and give support to SPARQL queries in the RDF triples. There is also an API to allow inferences and triple stores. DBgoldbr uses TDB to persist RDF triples.

The Java Application Client layer of Figure 13 shows the homepage of our DBgoldbr prototype developed in Java. In this interface, we have the third option Triples RDF that represent the second part of the proposed architecture. This option allows the persistence and queries of RDF via TDB.

ID	Name	URL
1	Unb	http://www.unb.br
2	Ufrj	http://www.ufrj.br
3	Ufp	http://www.ufp.br
4	Ufpr	http://www.ufpr.br
5	Ufmg	http://www.ufmg.br

(a)

ID	URL
1	http://www.unb.br/dadosabertos/alunos.csv
2	http://www.unb.br/dadosabertos/professores.csv
3	http://www.unb.br/dadosabertos/disciplinas.csv
4	http://www.ufg.br/dadosabertos/alunos.csv
5	http://www.ufg.br/dadosabertos/professores.csv
6	http://www.ufg.br/dadosabertos/disciplinas.csv

(b)

ID	URL
1	http://protege.stanford.edu/ontologies/camera.owl
2	http://protege.stanford.edu/ontologies/koala.owl
3	http://protege.stanford.edu/ontologies/travel.owl
4	http://protege.stanford.edu/ontologies/pizza/pizza.owl
5	http://www.w3.org/TR/owl-guide/wine.rdf
6	http://owl.man.ac.uk/2006/07/sssw/people.owl
7	http://protege.stanford.edu/plugins/owl/dc/dublincore.owl

(c)

Figure 11. Tables of "Publishing Entity" (a), "Published Source" (b) and "Ontology" (c)

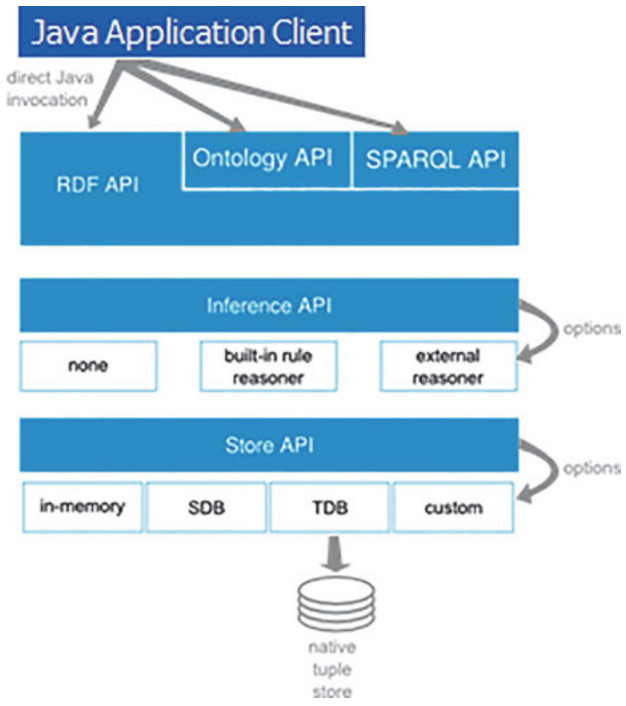


Figure 12. Second part of the logical architecture of DBgoldbr.

Figure 13. DBgoldbr prototype interface with RDF Triples option.

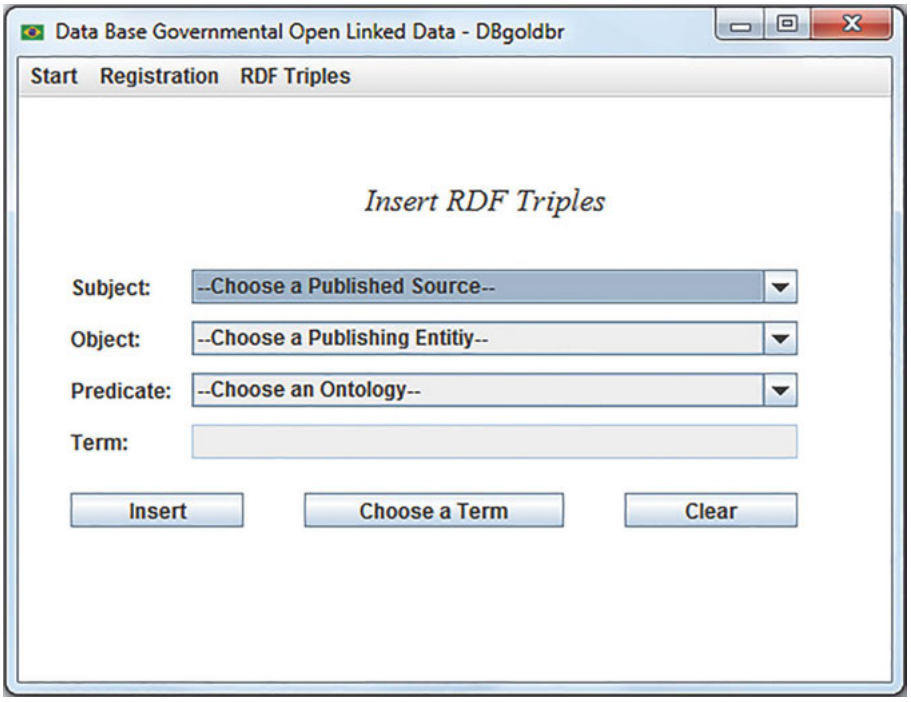


Figure 14. DBgoldbr screen to insert RDF triples.

Figure 14 shows the interface with the Insert option of the menu, used to create an RDF triple and to persist it. Users can choose a subject (the published source), an object (the publishing entity) and a predicate (the term in the ontology).

Figure 15 shows a user choosing a subject from the created triple. In this case, it is a published source, using a list of options. This list is stored in MySQL as represented in Figure 11 (b).

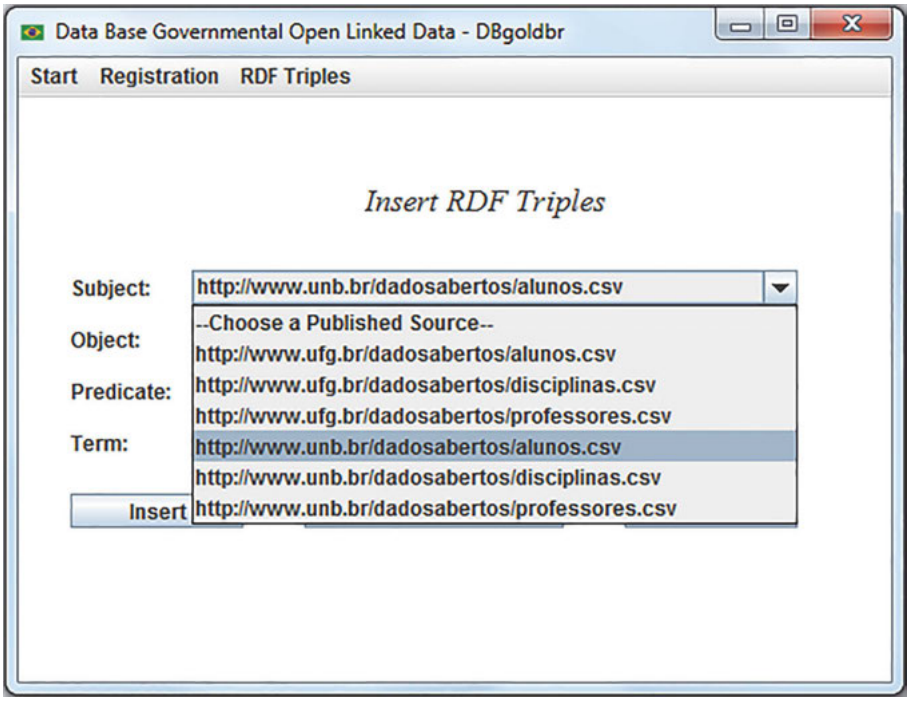


Figure 15. DBgoldbr interface to select subject of an RDF triple.

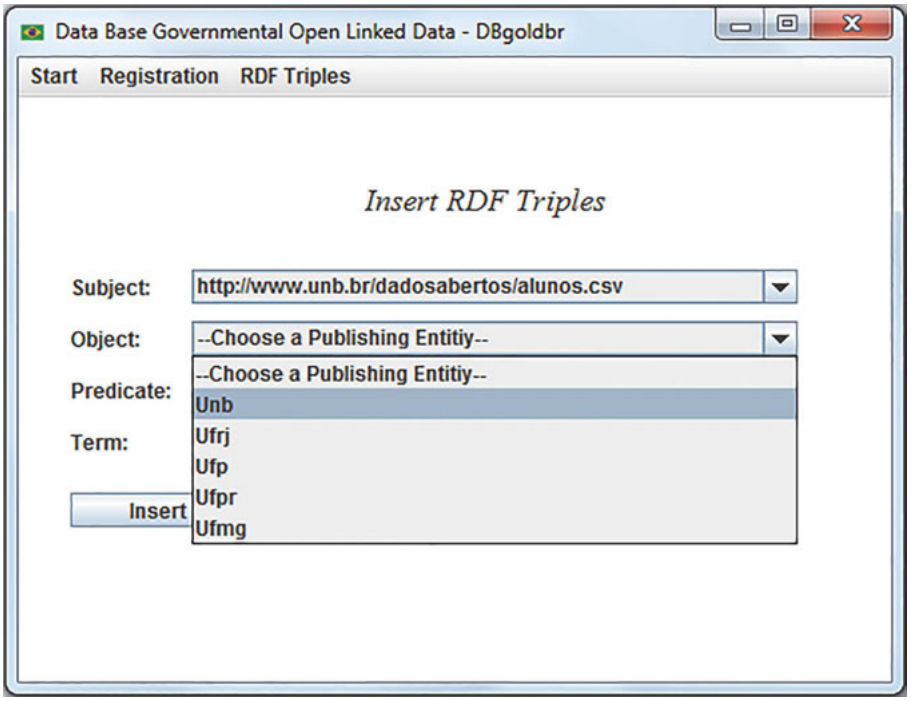


Figure 16. DBgoldbr interface to select object of an RDF triple.

Figure 16 shows the user selecting an object of a new triple, in this case the publishing entity, out of a list with options, generated from the MySQL table also represented in Figure 11 (a).

Figure 17 shows the option to select a "Term" that will represent the predicate of the RDF triple. This activity starts with the selection of the ontology where the term is registered, via a list of options from a MySQL table presented in Figure 11 (c).

After selecting the ontology, the user clicks on the button "Choose a term," as shown in Figure 18, so that the terms registered in the ontology are presented.

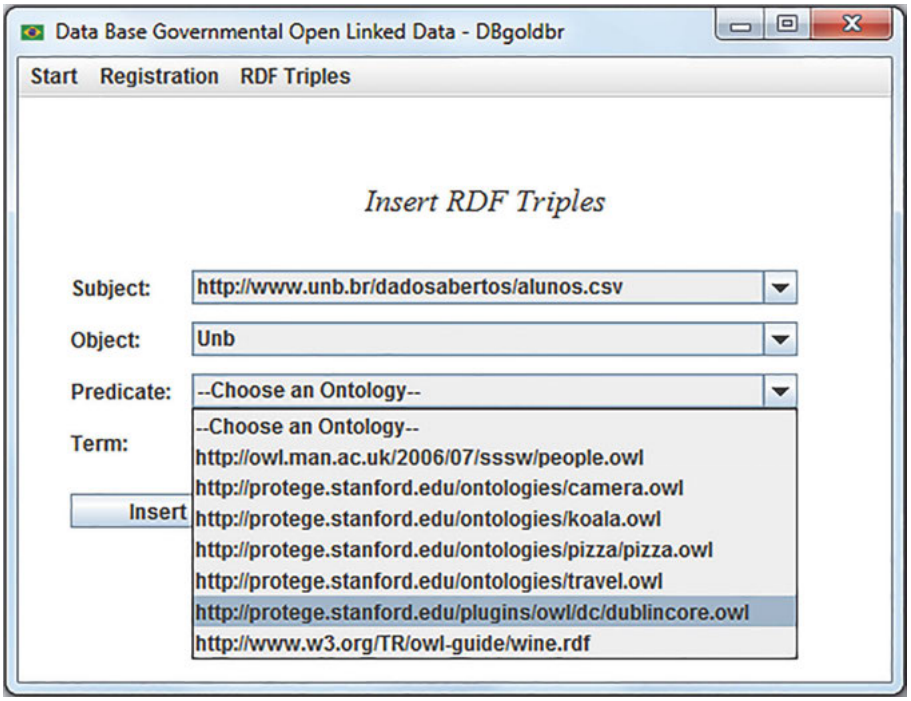


Figure 17. DBgoldbr interface to select predicate of a triple from an ontology.

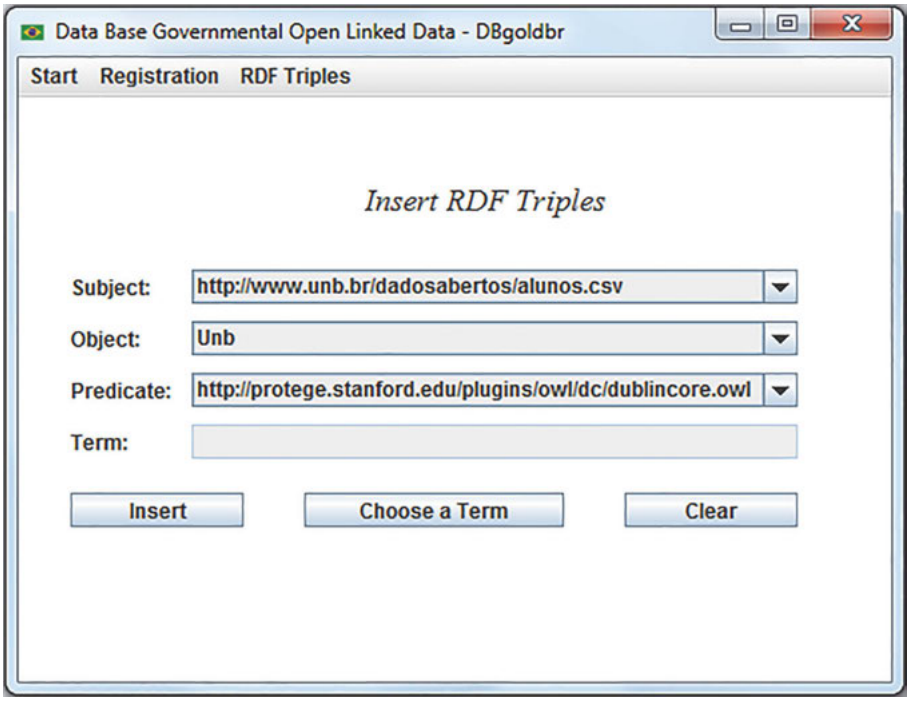


Figure 18. DBgoldbr screen to select a term from the ontology.

When a user clicks the "Choose a Term" button, the screen (Figure 19) is presented, where the terms of the ontology are shown. The ontology was previously selected, located at "http://protege.stanford.edu/plugins/owl/dc/dublincore.owl," as shown in Figure 18. In this case we chose the OWL file to use the standard term of the DCMI metadata, and then we presented the "Annotation Properties" of this file. To raise the number of ontologies available, it was necessary to insert a new link to the table of Figure 11 (c) via the interface of Figure 10 (c).

The term chosen from Figure 19 was then transferred to the filed "Term" of the Figure 20 interface. Figure 20

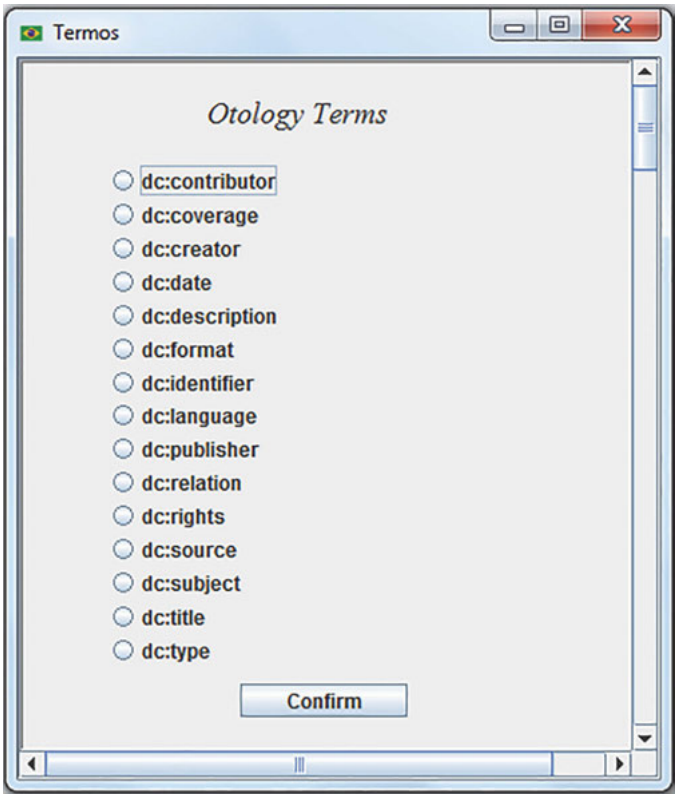


Figure 19. DBgoldbr interface to select term in the ontology.

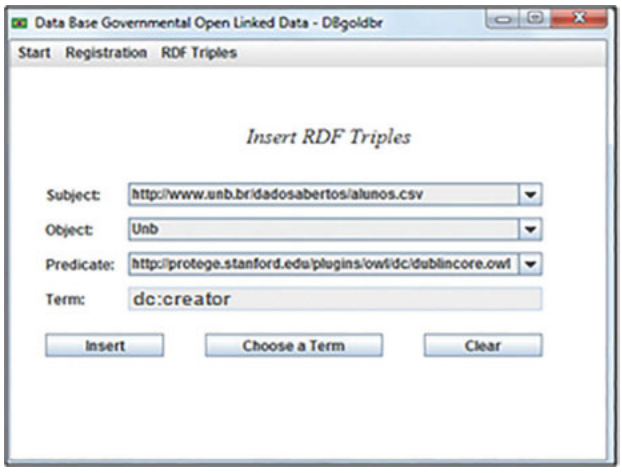


Figure 20. DBgoldbr interface with the RDF triple components.

shows the parts of the RDF triple whose XML RDF code is presented in Figure 21.

Figure 21 presents the RDF XML code from the Figure 20 triple. This triple will be persisted via the TDB resource Apache Jena, after clicking the "Insert" button shown in Figure 20.

Lastly, Figure 22 presents the interface with the menu option "Query" from Figure 13. This interface offers the user the option to query RDF triples persisted via the Insert option of Figure 13. Figure 22 shows that a user can choose a "Subject" (Published Source), an "Object" (a Publishing Entity) and a "Predicate" (a Term in the ontology used to create the triple).

Figure 23 presents the RDF XML code of a possible set of triples persisting via the use of the Figure 14 interface.

Figure 24, shows that a user may choose to select a subject, a predicate or an object to build a query of the persisted triples. Each of these three items are obtained via a list of options generated from the queries to the persisted

```
<?xmls version="1.0"?>

<RDF:RDF xmlns:RDF="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/">
       <RDF:Description about="http://www.unb.br/dadosabertos/alunos.csv">
            <dc:creator>http://www.unb.br</dc:creator>
       </RDF:Description>     <RDF:Description
</RDF:RDF>
```

Figure 21. RDF triple code created by DBgoldbr in the RDF XML format.

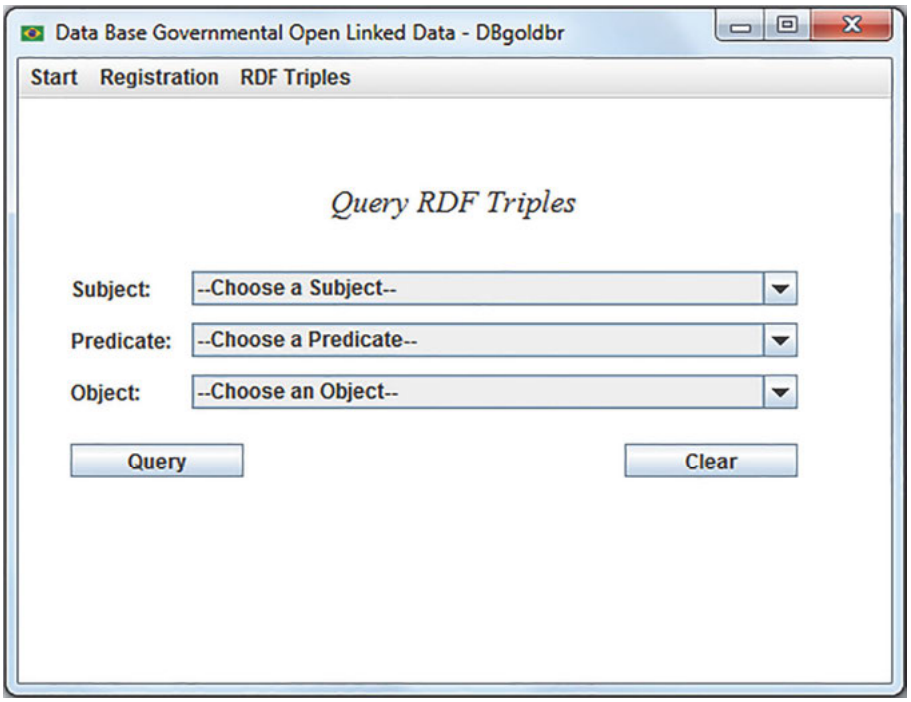


Figure 22. DBgoldbr interface to query RDF triples.

```
<?xmls version="1.0"?>
<RDF:RDF xmlns:RDF="http://www.w3.org/1999/02/22-rdf-syntax-s#"
    xmlns:dc="http://purl.org/dc/elements/1.1/">
      <RDF:Description about="http://www.unb.br/dadosabertos/alunos.csv">
        <dc:creator>http://www.unb.br</dc:creator>
      </RDF:Description>
      <RDF:Description about="http://www.unb.br/dadosabertos/professores.csv">
        <dc:creator>http://www.unb.br</dc:creator>
      </RDF:Description>
      <RDF:Description about="http://www.ufg.br/dadosabertos/disciplinas.csv">
        <dc:creator>http://www.ufg.br</dc:creator>
      </RDF:Description>
</RDF:RDF>
```

Figure 23. RDF XML code of a possible set of triples obtained with DBgoldbr.

triples. In this case, we considered the Figure 23 file to represent the set of RDF triples persisted.

Figure 24 (a) shows the user selecting a "Subject" out of the following values:

"http://www.unb.br/dadosabertos/alunos.csv,"
"http://www.ufg.br/dadosabertos/disciplinas.csv" and;
"http://www.unb.br/dadosabertos/professores.csv."

These values are obtained from an RDF XML file shown in Figure 23, where three RDF triples are persisted.

In Figure 24 (b), we can see the user's choice of "Predicate" "dc:creator," because it is the only predicate used by the triples.

Figure 24 (c) shows the user's selection of an "Object" from: "http://www.unb.br" and "http://www.ufg.br." These values are also obtained from the RDF XML (Fig-

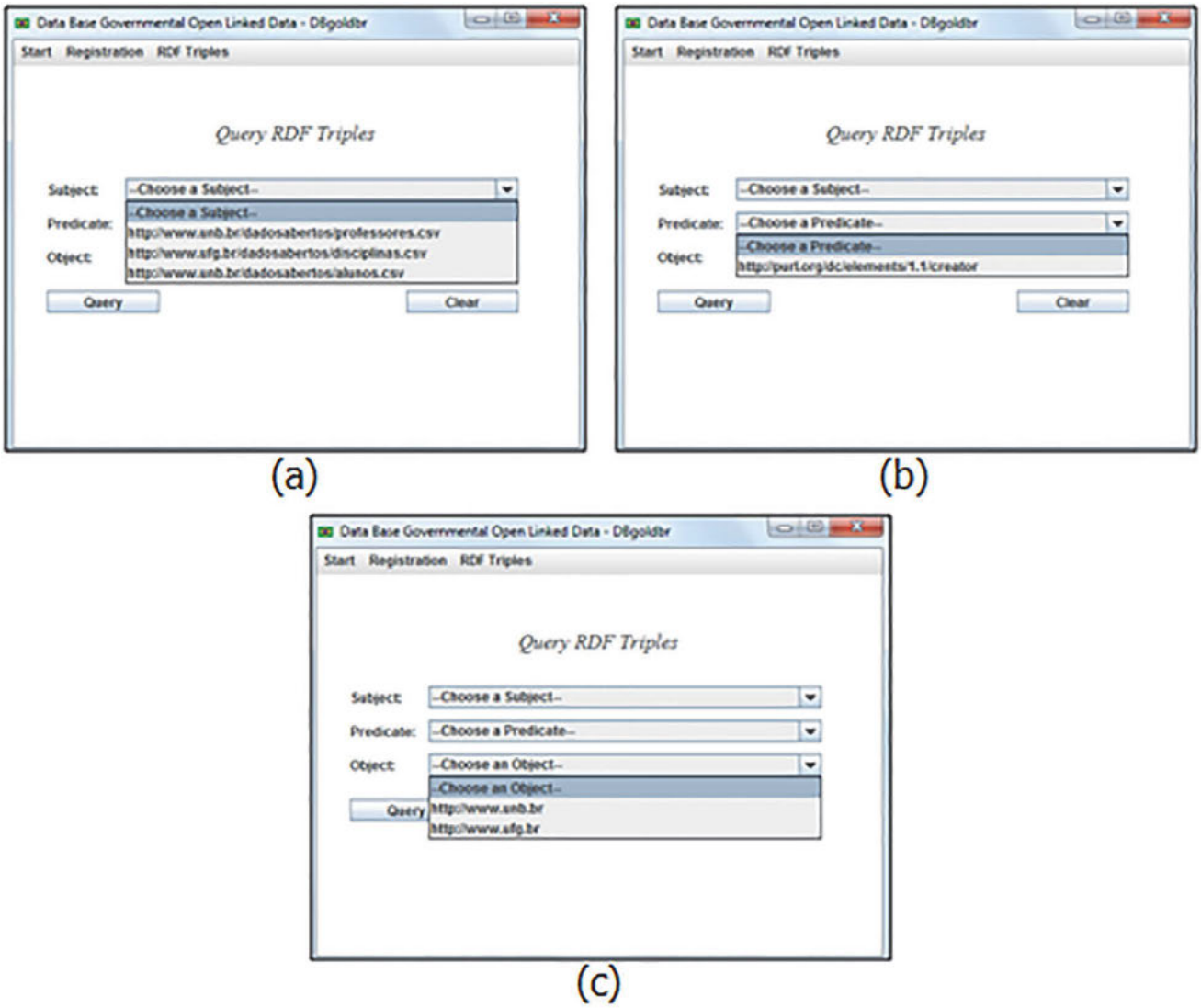


Figure 24. DBgoldbr interface to query RDF triples for Subject, Predicate and Object.

ure 23), where triples are persisted and the object "http://www.unb.br" appears twice.

Figure 25, shows the parameters selected by the user to answer the following question: "What are the resources created by the University of Brasília?" In this case, the predicate is the term "creator" of DCMI, the object is "http://www.unb.br," which represents the official website of the University of Brasília—Unb. When the user clicks the Query button, the application performs a SPARQL query with the chosen parameters and then executes via the Apache Jena API. The result of this query is in Figure 26.

The upper portion of Figure 26 shows the SPARQL query mounted by the application based on the user's choice from Figure 25. In the lower portion of the figures, we represent the resources, in this case the subjects, which correspond to the user's query: What are the resources created by the University of Brasília? Then the answer to this query is the Subject "http://www.unb.br/dadosabertos/alunos.csv" and "http://www.unb.br/dadosabertos/professores.csv." Figure 26 shows one of the many options that can be queries via the interface of Figure 25, because users can set up different combinations of the subject, predicate and object.

9.0 Conclusions

This paper presents a solution for the challenges in managing big data, specifically from the Brazilian government, so that it may be disclosed in an open and linked manner. It has been designed for the benefit of citizens and institutions, providing transparency of Brazilian government data and employing best practices. Furthermore, this solution proposes accessibility of the open data of the Brazilian government via the web.

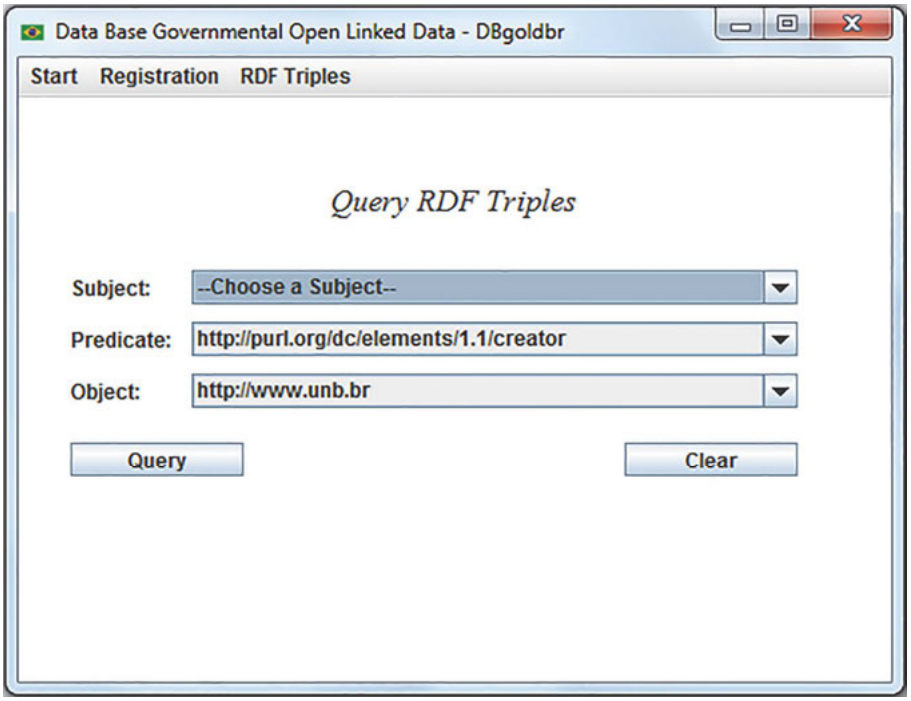


Figure 25. DBgoldbr interface with option to query to RDF triples.

Figure 26. DBgoldbr screen with the results of the RDF triples query.

Subsequently, with this study, we have gained a greater appreciation of the complexity in processing the massive volumes of data generated daily by Brazilian government institutions, in an environment of sharing knowledge that originates in different formats and a varied IT infrastructure. This processing is a key component in enhancing public efficiency and transparency, crucial for public managers and staff who are faced with making critical decisions daily, helping to ensure that they are making the safest decisions possible, based on the most amount of information possible.

Thus, this paper presented the conceptual view and technical architecture of the "big data ecosystem" DBgoldbr (Brazilian Database Government Open Linked Data) by illustrating the development of a prototype tool built with a set of resources to perform an ontology-based semantic classification of information. DBgoldbr aims to increase the openness level of data from the Brazilian government from three to five stars, allowing it to be linked via ontologies that can organize and represent huge volumes of massive data and their respective semantics.

This research constitutes a development plan, created and carried out by a team of researchers in the fields of computer and information science in which steps were taken to migrate software architecture to a web environment. The objective was to provide end users of DBgoldbr with the ability to query a huge volume of public data in many different areas of the government. The professional information may also identify relevant sources of information needed to prepare an appropriate decision-making environment, based on semantically represented, linked and open data mining. This solution is also designed for the public searches for data that can be integrated in order to help answer questions about the efficiency of public policies in social contexts, improving resource management and justifying public expenditure. The DBoldbr solution described here primarily benefits the Brazilian population in general, providing accessibility to public information, to all users, interoperably and intuitively (semantically), via user-friendly open data repository interfaces, which are also easily integrated.

References

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The Semantic Web." *Scientific American* 284, no. 5: 29-37.

Berners-Lee, Tim. 2006. "Linked Data-Design Issues." http://www.w3.org/DesignIssues/LinkedData.html

Beyer, Mark A. and Douglas Laney. 2012. "The Importance of 'Big Data': A Definition." Stamford, CT: Gartner. https://www.gartner.com/doc/2057415/importance-big-data-definition

Borst, Willem Nico. 1997. "Construction of Engineering Ontologies for Knowledge Sharing and Reuse." PhD diss., University of Twente.

Boyd, Danah and Kate Crawford. 2012. "Critical Questions for Big Data." *Information, Communication & Society* 15: 662-79. doi:10.1080/1369118X.2012.678878

Checkland, Peter. 1981. *Systems Thinking, Systems Practice*. Chichester: John Wiley & Sons.

Corcho, Oscar, Mariano Fernández-López and Asunción Gómez-Pérez. 2003. "Methodologies, Tools and Languages for Building Ontologies. Where is Their Meeting Point?" *Data & Knowledge Engineering* 46: 41-64. doi:10.1016/S0169-023X(02)00195-7

Crawford, Kate. 2013. "The Hidden Biases in Big Data." *Harvard Business Review* (blog). http://blogs.hbr.org/cs/2013/04/the_hidden_biases_in_big_data.html

Crawford, Kate, Kate Miltner, and Mary L. Gray. 2014. "Critiquing Big Data: Politics, Ethics, Epistemology." *International Journal of Communication* 8: 1663-72.

Demchenko, Yuri, Cees de Laat, and Peter Membrey. 2014. "Defining Architecture Components of the Big Data Ecosystem." In *2014 International Conference on Collaboration Technologies and Systems (CTS)* 104-12. doi:10.1109/CTS.2014.6867550

Demchenko, Yuri, Paola Grosso, Cees de Laat, and Peter Membrey. 2013. "Addressing Big Data Issues in Scientific Data Infrastructure." In *2013 International Conference on Collaboration Technologies and Systems (CTS) 20-24 May 2013* 48-55. doi:10.1109/CTS.2013.6567203

Dublin Core Metadata Initiative (DCMI). 2012. Dublin Core Metadata Element Set, Version 1.1: Reference Desciption. http://dublincore.org/documents/dces/

Eaves, D. 2009. "The Three Laws of Open Government Data." *eaves.ca* (blog), September 30. https://eaves.ca/2009/09/30/three-law-of-open-government-data/

Ekbia, Hamid, Michael Mattioli, Inna Kouper, G. Arave, Ali Ghazinejad, Timothy Bowman, Venkata Ratandeep Suri, Andrew Tsou, Scott Weingart and Cassidy R. Sugimoto. 2015. "Big Data, Bigger Dilemmas: A Critical Review." *Journal of the Association for Information Science and Technology* 66: 1523–45. doi:10.1002/asi.23294

Frické, Martin. 2015. "Big Data and Its Epistemology." *Journal of the Association for Information Science and Technology* 66: 651–61.

Gruber, Tom. 1993. "What is an Ontology." http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

Guarino, N. 1998. *Formal Ontology in Information Systems: Proceedings of the First International Conference (FOIS'98), June 6-8, Trento, Italy*, ed. N. Guarino. Frontiers in Artificial Intelligence and Applications. Amsterdam: IOS Press.

Heath, Tom and Christian Bizer. 2011. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web 1. [San Rafael, CA]: Morgan & Claypool.

Hjørland, Birger. 2017. "Classification." *Knowledge Organization* 44: 97-128.

Iannella, Renato and James McKinney. 2014. "vCard Ontology-for describing People and Organizations." W3C Interest Group Note 22 May 2014. https://www.w3.org/TR/vcard-rdf/

Apache Jena. 2013. https://jena.apache.org

Laney, Doug. 2001. "Application Delivery Strategies. 3D Data Management: Controlling Data Volume, Velocity, and Variety." *META Group, Stamford* (blog). http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lévy, Pierre. 2011. *The Semantic Sphere 1: Computation, Cognition and Information Economy*. Oxford: Wiley-Blackwell; London: ISTE.

Mazzocchi, Fulvio. 2018. "Knowledge Organization System (KOS): An Introductory Critical Account". *Knowledge Organization* 45: 54-78.

Page, Lawrence, Sergey Brin, Rajeev Motwani and Terry Winograd. 1999. "The PageRank Citation Ranking: Bringing Order to the Web." Stanford InfoLab. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.206.775&rep=rep1&type=pdf

Prud'hommeaux, Eric and Andy Seaborne. 2008. "SPARQL Query Language for RDF." https://www.w3.org/TR/rdf-sparql-query/

Shin, Dong-Hee and Min Jae Choi. 2015. "Ecological Views of Big Data: Perspectives and Issues." *Telematics and Informatics* 32: 311-20. doi: 10.1016/j.tele.2014.09.006

Shiri, Ali. 2014. "Making Sense of Big Data: A Facet Analysis Approach." *Knowledge Organization* 41: 357-68.

Smiraglia, Richard P. 2017. "Replication and Accumulation in Knowledge Organization: An Editorial." *Knowledge Organization* 44: 315-7.

Soergel, Dagobert. 2015. "Unleashing the Power of Data Through Organization: Structure and Connections for Meaning, Learning and Discovery." *Knowledge Organization* 42: 401-27.

Souza, Renato Rocha, Douglas Tudhope, and Maurício Barcellos Almeida. 2012. "Towards a Taxonomy of KOS: Dimensions for Classifying Knowledge Organization Systems." *Knowledge Organization* 39: 179-92.

Sowa, John F. 1999. "Building, Sharing and Merging Ontologies." http://www.jfsowa.com/ontology/ontoshar.htm

Studer, Rudi, V. Richard Benjamins, and Dieter Fensela. 1998. "Knowledge Engineering: Principles and Methods." *Data & Knowledge Engineering* 25: 161-97. doi: 10.1016/S0169-023X(97)00056-6

Uddin, Muhammad Fahim and Navarun Gupta. 2014. "Seven V's of Big Data Understanding Big Data to Extract Value." In *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education: Engineering Education: Industry Involvement and Interdisciplinary Trends; April 3 - 5, 2014 University of Bridgeport Bridgeport, Connecticut*, ed. Elif Kongar. [New York]: IEEE. doi:10.1109/ASEE-Zone1.2014.6820689

Uschold, Mike and Michael Gruninger. 1996. "Ontologies: Principles, Methods and Applications." *The Knowledge Engineering Review* 11: 93-136. doi:10.1017/S0269888900007797

Zeng, Marcia Lei, Karen F. Gracy and Maja Žumer. 2014. "Using a Semantic Analysis Tool to Generate Subject Access Points: A Study Using Panofsky's Theory and Two Research Samples." *Knowledge Organization* 41: 440-51.