An approach for managing and semantically enriching the publication of linked open governmental data

Article · January 2011

CITATIONS
5

READS
119

12 authors, including:

Fabricio Firmino de Faria
Federal University of Rio de Janeiro
18 PUBLICATIONS **28** CITATIONS
SEE PROFILE

Bianca Pereira
National University of Ireland, Galway
10 PUBLICATIONS **46** CITATIONS
SEE PROFILE

Rodrigo Calhau
Instituto Federal de Educação, Ciência e Tecnologia do Espírito Santo...
13 PUBLICATIONS **35** CITATIONS
SEE PROFILE

Veruska Zamborlini
University of Amsterdam
25 PUBLICATIONS **123** CITATIONS
SEE PROFILE

An approach for managing and semantically enriching the publication of Linked Open Governmental Data

Kelli de Faria Cordeiro[1], Fabricio Firmino de Faria[1], Bianca de Oliveira Pereira[1,2], André Freitas[2], Cristiano Expedito Ribeiro[1], João Vitor Villas Boas Freitas[1], Ana Christina Bringuente[3], Lucas de Oliveira Arantes[3], Rodrigo Calhau[3], Veruska Zamborlini[3], Maria Luiza Machado Campos[1], Giancarlo Guizzardi[3]

[1]Graduate Program in Informatics (PPGI)
Federal University of Rio de Janeiro (UFRJ) – Brazil

[2]Digital Enterprise Research Institute (DERI)
National University of Ireland – Galway – Ireland

[3]Ontology and Conceptual Modeling Research Group (NEMO) – Computer Science Department – Federal University of Espírito Santo (UFES) – Brazil

{kelli, firmino, mluiza}@ufrj.br, andre.freitas@deri.org, {bianca.oli.pereira, expedit, jvfreitas.freitas, lucasdeoliveira}@gmail.com, {acobringuente, rfcalhau, veruska, gguizzardi}@inf.ufes.br

Abstract. *With the growth of e-government programs, the available data to citizens is growing in volume every day. However, to make these data a useful source of information, to be referenced and integrated more easily by different applications, they should be published according to the best practices of Linked Open Data, using standards for description (RDF) and identification (URI) of data resources on the web. The main goal of this work is to propose a platform and approach to support the exposure, sharing and association of data resources in the form of Linked Open Data, offering a user-friendly environment to stimulate the publication of data and their association to other existing data. Central functionalities to be included are data cleaning, transformation, linking, annotation and referencing to terminology mechanisms.*

1. Introduction

There is a large amount of government data available on the Web, generated from the many e-government initiatives and open-government movements, which emphasize the broad dissemination of information to citizens and organizations promoting electronic citizenship (e-citizen) and "empowering citizens". Given the difficulty of consumption and reuse of those datasets, with interfaces designed only for ad-hoc consultation or extraction and high costs and problems involved in their analysis, the initiatives of Linked Open Data (LOD) propose the use of open standards, inspired by the success of the "Web of Documents". Supported by the W3C, Linked Data involves simple principles of the Semantic Web to interlink and annotate data reusing vocabularies or schemas.

The Linked Data Web, mainly composed of open governmental data (43%) (Bizer et. al 2010), is a powerful environment, which can be used, for example, to provide useful information to citizens about their community or to support decision makers. Governments are making considerable efforts to open and link their data (e.g., UK Government Data - http://data.gov.uk) following the standards and principles of Linked Open Data (LOD) (Berners-Lee, 2006), creating an eco-system of cooperation between government and population.

In the current scenario, Linked Data publication process does not happen in a standardized manner. Despite the current availability of a fragmented set of tools to support the Linked Data effort, the lack of integrated Linked Data transformation and publication frameworks still presents a barrier for government users, which demand solutions that could operate under the scale, heterogeneity and data quality standards present in government environments. To address some of these critical issues, a platform and an associated approach to support the publication process of linked data is the focus of the work described in this paper. Our platform constitutes an initial infrastructure, integrating a set of tools to support the exposure, sharing and association of data resources in the form of Linked Open Data, offering a user-friendly environment to stimulate the publication of governmental data. In this paper, we discuss the current version of the platform, which includes: an ETL (extraction-transformation-and-loading) tool to manage the publication process; plug-ins designed to facilitate the extraction, transformation and mapping of raw data into RDF (triplification) and linking; and a repository to manage and store data in RDF format.

The paper is organized as follows. In section 2, the basic concepts of Linked Open Government Data are presented with its perspective in the international and the Brazilian scenarios. In section 3, the proposed platform for Linked Open Data publication is described. Our approach of using a workflow management system (in the form of an ETL tool) to orchestrate and integrate a set of tools used in the publication process is outlined in section 4. Complementarily, the approach also includes a data semantics enrichment process, which is described in section 5. Finally, section 6 covers some study cases scenarios which are described to illustrate the proposed approach, followed by conclusion and future work.

2. Linked data: an approach for interrelating heterogeneous data

The vision behind *Linked Data*, proposed by Berners-Lee (2006), focuses on the reduction of the barriers for the publication and consumption of data on the Web. At the core of Linked Data resides the idea of interconnecting fine-grained information resources, which are not originally associated, across the Web, leveraging on the Web infrastructure by using HTTP, URIs (Uniform Resource Identifier) and RDF (Resource Description Framework) as a data representation framework. Four Linked Data principles are stated. The first principle describes the use of URIs to identify resources (referring to informational or concrete resources). For example, the URI *http://dbpedia.org/resource/Rio_de_Janeiro* represents the city of Rio de Janeiro. The second principle states that HTTP name lookup under the Domain Name System (DNS) authority should be used as a lookup protocol between the identifier (URI) and the representation of the resource. The third principle describes a standardized way to provide the resource description using the Resource Description Framework (RDF)

graph representation format and SPARQL (a query language for RDF graphs). The fourth and last principle covers the process of interlinking new information with existing resources, maximizing the reuse and interlinking among existing data.

RDF is a graph-based representation format, described as a W3C recommendation standard (Manola and Miller, 2004). A RDF graph consists of a set of triple statements in the <*subject, predicate, object*> form. The *subject* part of the statement identifies the resource that the statement is about (e.g. *http://dbpedia.org/resource/Rio_de_Janeiro*), where a resource can be a URI or a blank node. The *predicate* is a relationship or a property of this resource (e.g. *http://dbpedia.org/ontology/populationTotal*) and the *object* represents a value or a resource associated with the subject through a predicate (e.g. *"6186710 ^^xsd:integer"*). An additional typing structure can be associated with each resource (e.g. *http://dbpedia.org/resource/Rio_de_Janeiro* rdf:type *http://dbpedia.org/ontology/Place*). The open standard graph-based nature of RDF leveraged over the Web infrastructure of unique identifiers (URIs) and content access (HTTP) lowers the barriers for exposing data on the Web and allowing the interconnection between datasets.

2.1 Basic Linked Data Publication Process

The conversion task from raw data to Linked Data involves more than just transforming original data into RDF (Heath and Bizer, 2011). One fundamental part in the publication process includes determination and creation of vocabularies and ontologies which provide the data model behind linked datasets, aiming towards a more integrated view of data and a maximization of semantic interoperability between datasets and between data producers-consumers. Another important part in the definition of the data models is the maximization of the reuse and the extension of existing vocabularies and ontologies. This is a fundamental element in the reduction of the effort involved in the consumption and integration of the published linked datasets.

After the RDF representation of the original data is built, the entities in the dataset are linked to entities in external datasets passing through an entity reconciliation step. External datasets can include other datasets in the domain of an organization or linked open datasets on the Web such as DBPedia (Bizer et. al, 2009). Automated approaches and tools are already available to support the dataset interlinking process (Volz, 2009).

The final step consists of making the dataset available on the Web for consumption. Linked Data can be made available in three forms: (i) through dereferenceable URIs (where an RDF document is returned when an URI is dereferenced through an URI HTTP request); (ii) through a SPARQL endpoint (a SPARQL query interface for the RDF dataset which is exposed on the Web); and (iii) through RDF data dumps. The published datasets should be made discoverable on the Web by providing the appropriate high-level dataset and search engine descriptors.

One important aspect which is currently not present in most published linked datasets is the availability of provenance information associated with the generated data, where the historical trail behind the data could be traced back to the original data sources (Hartig, 2009). Provenance is a cornerstone element for allowing trustworthiness and data quality assessment in the scale of data reuse and aggregation

that the Linked Data enables. However, the effort for capturing, representing and managing provenance in the Linked Data publication process remains high.

Despite the availability of tools supporting parts of the Linked Data transformation and publication workflow there is still a gap in terms of approaches providing an integrated and structured view of this workflow. Another important gap is the lack of methodologies and good practices for the construction of ontologies and vocabularies suitable to the consistency, quality and interoperability levels required in government environments. A third important gap is the construction of a provenance-aware Linked Data transformation and publication workflow. These three gaps create barriers for the adoption of Linked Data within government organizations and are in the center of the contributions of this work.

2.2 Linked Open Governmental Data

The open government data effort consists of making public information available to the general public and across government sectors. In order to become usable the published data need to be made available in a format which could maximize its reuse in applications developed by government, companies or citizens. This principle is expressed by Eaves (2009), which created three laws for government data: (i) if the data cannot be found and indexed on the Web, it does not exist; (ii) if it is not already open and available in machine-understandable format, it cannot be reused, and (iii) if any legal provision does not permit their replication, it is not useful. The effort for governments to open and expose data to the public as Linked Data became known as Linked Open Government Data.

2.3 Linked Data Perspectives

Since the publication of Tim Berners Lee's memo at the W3C in 2007, listing some of the Linked Data basic principles, until today, the set of available data sources in the Linked Open Data (LOD) cloud grew considerably. According to a survey of September 2010 (Bizer, 2010), the LOD Cloud reached 203 sources, involving approximately 27 billion RDF triples, connected by about 400 million RDF outgoing links. Outgoing links refer to the links that are set from data sources within a domain to other data sources.

There have been a significantly greater number of international initiatives. In the Brazilian scenario the interest in the topic is increasing with some proposals and prototypes emerging in both the academic and government settings.

2.3.1 International Scenario

Efforts aimed to interlink government data are growing in some countries around the world. Interesting examples are given by the British Government, the United States and Spain. In 2009, the British Government began to adopt Linked Data as the official standard for publication of public domain data. In particular, the initiative of the Asociación Española de Linked Data (AELID) aims to stimulate research in Linked Data in Spain and Europe, and create a network of researchers in order to promote the exchange of knowledge and experience in the subject. Initiatives such as the Data-Gov Wiki (Rensselaer Polytechnic Institute) and GeoLinkedData (University of Madrid) are also efforts to extend the use of Linked Data, transforming public government data, such

as National Statistics Institute (INE) and National Geographic Institute of Spain (IGN-E), available in several formats, to Linked Data standards.

Due to the growth of the Linked Data initiative, in September 2010, the European Union started the LOD2 project[1], lasting approximately 4 years, and emerging as an integration project in large-scale, involving researchers, companies and information providers from 7 European countries, including precursors groups in the subject, such as DERI (Ireland) and the Freie Universität Berlin. They are handling some challenges of the LOD paradigm associated to intelligent information management: the exploitation of the web as a platform for data and information integration. Some of their main concerns are related to: coherence and quality of data published on the web, establishing trust on the Linked Data Web, methodologies for exposing, high-quality multi-domain ontologies, automatically interlinking and fusing data, standards and methods for reliably tracking provenance, ensuring privacy and data security as well as for assessing the quality of information. The expected results include tools, methods and collections in the form of Linked Datasets

The importance of the development and use of conceptual tools for semantic enrichment of models has also been recognized in the international scene. An example is the creation of the International Association for Ontology and Applications (IAOA), in 2009, an international organization that aims to promote research and education on ontologies in the world with a focus on aspects of conceptual modeling and semantic enrichment. An example of an institutional member of this organization is the National Coordination Office (NCO) for the Networking and Information Technology Research and Development (NITRD) of the U.S. Government (http://www.nitrd.gov/). This year, the NITRD NCO-IAOA, in collaboration with institutions such as NIST (National Institute of Standards and Technology), are organizing the Ontology Summit 2011 which aims to discuss international affairs with concrete and successful use of ontologies.

2.3.2. Linked Open Brazilian Governmental Data Perspectives

The international open data projects are aligned to the Brazilian Government transparency initiatives, such as, e-Gov (Brazilian Federal Portal) and e-Ping (Interoperability Standards for Electronic Government). For example, this year, the Consegi (International Open Software and Electronic Govern Conference), a Conference promoted by the Brazilian Government, had as its main theme “Open Data for Democracy on the Digital Age”.

In Brazil, in the academic context, different groups that were already active in the areas of semantic web and knowledge representation/ontologies, have engaged in this movement, seeking to adopt and develop projects that enable support for the publication and data access in the form of LOD, acting both in the prototyping of tools, and approaches to semantic interoperability and on investigation of many challenges still exist in this area.

[1] LOD2 – Linked Open Data 2 Project, http://lod2.eu

In government institutions, there seems to be a clear tendency to evolve from the current open data movement to Linked Open Data, stimulating the increasing use of standards, metadata and controlled terminologies, which allow for a more efficient exploration of the data already made available by organizations such as Prodasen, IBGE, etc. In the same way, the Ministry of Planning, the Information Technology Steering Committee and the W3C Brazil have played an important role, promoting meetings, courses and an intense debate on the subject.

3. LinkedDataBR Project

The publication process is the core of Linked Open Data life cycle: from one side there are lots of unlinked and distributed data and on the other side there are many applications that can make use of the available data. However there are large barriers for consuming and repurposing data which is heterogeneous, distributed, disconnected and non-standardized.

The publication process consists of three main steps: pre-processing, triplification and linking. The pre-processing step is responsible for extracting data from files in a variety of data formats, cleaning and normalizing them. The triplification step is the core of the publication process. In this step the properties and entities are created, expressed as triples and semantically annotated with an ontology or a vocabulary. Finally, the linking step finds links between items of a given dataset and items of third-party datasets in order to extend the information provided.

By exposing data in a fine grained format, annotating them with metadata, linking them and monitoring the entire publication process, it is possible to reduce the barriers to consume legacy data from different formats and ensure the quality and provenance of the data. The focus of the LinkedDataBR project is to reduce the barriers to publish data in Linked Open Data format. In order to allow this, a platform and approach have been proposed and developed to support the conversion of legacy data into interlinked triples.

3.1. The LinkedDataBR platform and approach

The LinkedDataBR platform aims to support all steps related to the transformation and publication process since the extraction from source files until the linking of annotated triples with other datasets. The entire process is tracked to provide provenance information in order to ensure the quality of data. The platform consists of a set of well-known tools on data processing and tools specially designed to improve the LOD publication process in an integrated environment.

3.2 Proposed Architecture

The proposed architecture aims to integrate tools available for publication of Linked Open Data, searching for the maximization of the reuse of existing tools. In an extensible platform, different features can be attached to the data treatment process to handle different raw data sources. To provide this extensibility, the tools are orchestrated and managed through a workflow infrastructure, which can be configured to allow the implementation of the architecture in different domains and computational environments.

Other concerns of the proposed architecture are the focus on data quality, provenance data capture and semantic treatment support, fundamental to governmental data. To address these issues, the proposed architecture is composed of three groups of elements disposed in two layers, as illustrated in Figure 1. The groups of elements are: *web elements*, *interface elements*, and the *LinkedDataBR infrastructure elements*. And the two layers are: *publication and access applications*, and *data repositories*.

The architecture layers interact with the web through a portal, index services and LOD consumption applications. The portal is an access point to the entire environment; the index services are used to make the published dataset known to the LOD Cloud (Ping Semantic Web) and also to search other datasets to create links to them; and LOD applications are those developed to consume the published dataset preferably together with other linked data.

The interface components of the architecture include a Data Input Interface on which the raw data to be processed, triplified and exposed is identified and an Access Interface to make the triples available. The Data Input Interface is composed of a File Transfer feature used to upload the raw data files, such as xml, csv, xls and relational database, and a Form feature for the files description. The Triple Access Interface is composed of: (i) a SPARQL Endpoint, which allows access to triples through SPARQL Queries; (ii) a feature to make RDF Triples available for download and a HTML interface to allow navigation through data; and (iii) a search interface over local repositories with index engine. Moreover, the Ping the Semantic Web feature uses external index services to make the triple store known to the LOD Cloud.

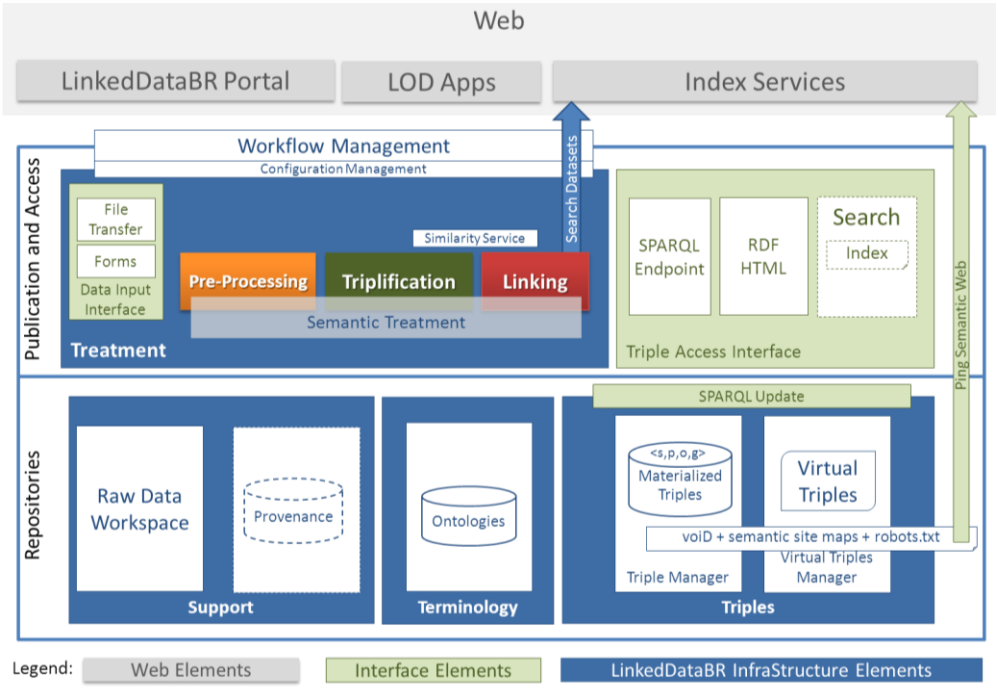

Figure 1. Linked Open Data Publication Architecture

After being mapped as an input data source, the data pass through a treatment process composed of steps which are supported by three main components: (i) Pre-processing, to extract and clean the relevant data of the different types and format files, including extracted data from relational databases; (ii) Triplication, to transform the

extracted data into triples and to map triples into RDF triples annotated with vocabulary or ontologies; and (iii) Linking, to identify the relations between entities across different datasets (entity reconciliation) based on similarity measures.

All data handled by the publication process is stored in repositories that support the semantic treatment process and the triples created. The repositories are used to store raw data (as a staging area), the workflow provenance data, the ontologies, the materialized RDF triples, and the mapping of relational data into RDF (Virtual Triples). Finally, the created dataset descriptors and crawler related information (semantic sitemaps and robots text file) feed external index engines. A standard vocabulary (voiD) (Alexander et. al, 2009) is used to describe the dataset and its interlinking.

3.3. Benefits of the platform and approach

The proposed platform allows monitoring the whole process, providing information about the generated triples. Questions like "where are these data from?", "which process generated this data?" or even "which agency is responsible for these data?" can be answered using provenance tracking. The implementation of the platform in a modular approach, supports easier integration with other tools currently in use by government agencies. The development based on open source software is in accordance with the current guidelines of government information infrastructure in Brazil.

The semantic data annotation provides knowledge about the published data and the possibility to make reasoning on top of them. The linking phase allows interconnecting data assets from different agencies, providing a partial entity-centric integration across different datasets. Agencies publishing governmental data with the proposed approach can more easily control and monitor the publication process and also increase reuse of data transformation and publication tasks. From the citizen's point of view, data consumer applications can make use of integrated and standardized datasets with richer semantics.

4. ETL Workflow Implementation for Linked Open Governmental Data

ETL processes were initially conceived with focus on Business Intelligence applications, which involve data integration from different data sources, in distinct formats, in order to generate a database in a standard format to enable the visualization or exploration of massive data under different perspectives (Kimball and Caserta, 2004). An ETL process actually comprises the main steps that are necessary to publish LOD data. The first step is the extraction of useful data from different existing sources. The second step is the transformation, in which data inconsistencies are eliminated and data is converted from one format to another. Due to the similarities between ETL and LOD publishing processes, in the LinkedDataBR project we have used an ETL tool to implement the publication workflow, more specifically, the open source tool Pentaho Data Integration – Kettle (Bouman and Dongen, 2009).

Kettle provides an intuitive graphical user interface, based on a drag and drop model, to build ETL workflows. Each Kettle ETL process is represented by a set of interconnected steps. These steps can be pre-defined actions as well as calls to services offered by other applications (Web Services). This feature allows a range of services to be reused. The ETL process is specified from two different types of objects, namely: Job

and Transformation[2]. Transformations can be characterized as data-oriented tasks, with the purpose to extract, transform and load data. Jobs can be viewed as collections of Transformations, and are task-oriented. This task may be related to the implementation of ETL, security level, implementation arrangements, among others.

Another important aspect is related to how Jobs and Transformations are stored: via files or via a repository. The storage via files (XML) represents the simplest form of persistence of Jobs and Transformations. In this model, Jobs and Transformations are stored locally with specific extensions, `.kjb` and `.ktr`, respectively. In the repository model, Jobs and Transformations are stored in a central database, where multiple users can collaborate during the development of an ETL workflow. Finally, it is noteworthy that Kettle allows the extension of functionality through plug-ins. Thus, if there is the need to include specific steps to the process of creation of Linked Data throughout the project, plug-ins can be developed through the API Kettle.

5. Semantic Treatment

With the expansion of the LOD cloud, the quality of the available information becomes a concern for organizations that want to publish data as LOD. It is especially true regarding governmental data, which demands a high level of reliability of data and their links. Nevertheless, the current LOD scenario is not completely favorable. Among some deficiencies we can mention: lack of conceptual description of datasets, absence of schema level links and lack of expressivity of data representations (Jain, 2010). These aspects can compromise the quality of data publication.

For this reason, there is a challenge on offering support to organizations to publish meaningful data. By meaningful data we refer to data which have their meaning made explicit and available. There are some vocabularies (light ontologies) used in LOD that have this role, like FOAF, Geonames and DBpedia. However their use is in general not enough to assign proper semantics, mainly because these vocabularies consist of a list of terms and their definitions, but do not comprise a proper conceptual model of the part of reality they represent. However, the enthusiasts of LOD argue that it is better to be agile on publishing data and have them ready for consumption, than to be very strict on quality and description, but have very little data available and frustrate citizen's expectations. For this reason, we propose the support of different semantic treatment levels according to the various needs of organizations that want to publish data in the LOD cloud. Following, we list four approaches that illustrate some of these levels.

1st Approach

The first and simplest approach to publish linked data is characterized by a **lack of concern with semantics** (Figure 2a). The publication process consists of: (i) extracting the relevant data from the raw data sources; (ii) triplifying them and thus (iii) linking them with other datasets in LOD (for example, using the *owl:sameAs* relation).

On the one hand it promotes ease and speed of the publication process, on the other hand we have a low reliability of data and links resulting from this process. This approach has no well-established criteria to consider the semantics of published data.

---

[2] Carte User Documentation, http://wiki.pentaho.com/display/EAI

2nd Approach

The second approach introduces a **low concern with semantics**, which is to give data some meaning annotating them with terminological instruments in the data cloud (Figure 2b). The publication process consists of: (i) extracting the data from the raw data sources; (ii) identifying the vocabularies with which data are to be annotated, and finally (iii) triplifying data and linking them with other datasets in LOD.

Similarly to the first approach, it promotes the ease and speed of the publication process. The difference is in the need to identify which are the appropriate terminological instruments to be used (ontologies or vocabularies used in LOD). However, the expressiveness of descriptors assigned to data in this process may still be low. Simply annotating data does not solve the aforementioned problem. In fact, the annotation process without a true concern about the intended meaning of the used terms can worsen the problem of reliability. The erroneous attribution of meaning is even worse than not having any meaning assigned.

3rd Approach

The 3rd approach is characterized by a **medium concern with the semantics** (Figure 2c). The aim is to solve the integration process in the conceptual level (technologically independent). The first step of the publication process is (i) *semantic pre-processing* and it consists of both (a) *extracting the conceptual model*, i.e., modeling of the raw data

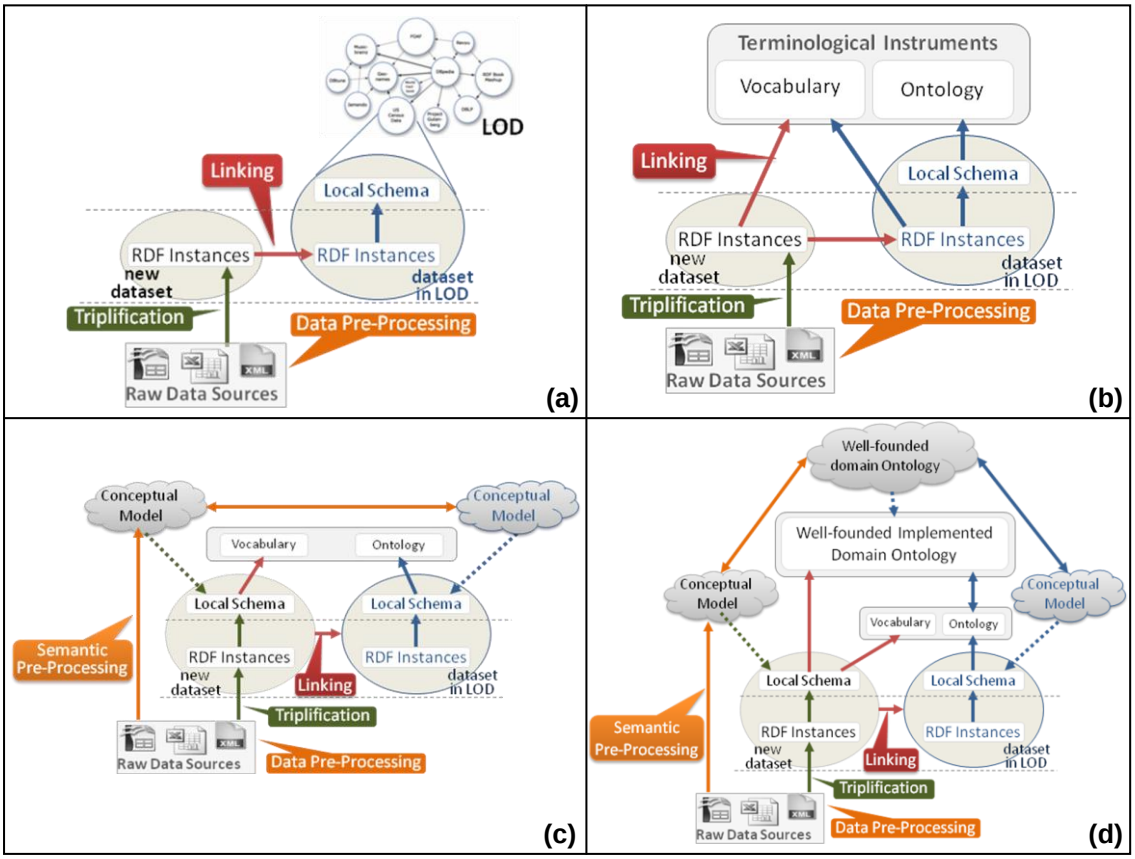

Figure 2: LOD approaches with (a) lack of (b) low (c) medium and (d) high semantics concern

sources and also of the existing databases with which connections are to be made (if the conceptual models are not available); (b) *integrating the conceptual models*, i.e., the concepts in each model are linked together. The next step is (ii) *data pre-processing* in which data are extracted from the raw data sources. Following, we have (iii) *triplification* in which a local rdf-schema is created from the conceptual model of data sources, and data are triplified and annotated with it. Finally, it comes step (iv) linking, in which the dataset is linked with other datasets in LOD, driven by the integrated conceptual models.

Although this approach potentially improves the reliability of the links between data and schema, there can still be semantic problems in the process of integrating conceptual models. It is due to the questionable quality of these models, or yet, to the ambiguity of linguistic terms, i.e., the same term with different meanings can be mistakenly integrated in different models (e.g. "bank" meaning "a financial institution" or "the edge of a river").

4th Approach

The 4th approach is characterized by a **high concern with the semantics** (Figure 2d). The aim is to make quite explicit the semantics of the data. To do so, this approach takes advantage of a well-founded domain ontology as the basis for the integration of conceptual models. A well-founded domain ontology is a domain-specific model articulated with a domain independent formal system of categories, called Foundational Ontologies (Guizzardi, 2005). The commitment to these ontological categories can bring many benefits. It helps to clarify the intended meaning of the adopted terms through a set of semantic distinctions, avoiding ambiguity.

This level of semantic treatment is composed of the same steps of the third approach. The difference is that the conceptual models of the datasets are mapped to a well-founded domain ontology in the semantic pre-processing step. This mapping is responsible for assigning a more precise semantics to the conceptual models of datasets. Besides, in the linking step, the local schemas of datasets are linked to the implementation of the domain ontology, driven by the aforementioned mappings; and, finally, the dataset is linked with other datasets in the LOD cloud, driven by the mappings between datasets and the implementation of the domain ontology.

The last approach seeks to improve the problem of data expressiveness in LOD. It is mainly based on the use of a conceptual level and well-founded domain ontologies. The use of a conceptual level is important because it abstracts technological aspects, provides a conceptual description of the datasets and improves human comprehension and semantics assignment. The use of well-founded domain ontologies, in turn, improves mainly the quality of data representation in LOD.

The objective is not to embrace all the possible approaches, but to illustrate scenarios of LOD publication that vary from a minimal level of semantic treatment to a very high one using formal ontologies. Moreover, we defend that the process of publishing data in LOD can be done in an incremental way, in the sense that data are published in a certain level of semantic treatment, but can still be improved to a higher level, promoting their semantic enrichment.

6. Study Case: Science and Technology Scenario

Our first study case scenario comprises data on research projects, researchers and their associated academic production. As an example, the Brazilian National Research Network - RNP, through its Working Groups (WG), enables the development of collaborative projects that show the feasibility of new protocols, services and network applications. As part of the selection of these projects, RNP performs a data survey about the researchers involved within the projects, in which data is obtained from the CNPq Lattes Curriculum (CV Lattes) database together with the analysis of the subject of each proposed project. Given these data, it is possible to extract information such as the project relevance, the actual relatedness of the researcher with the proposed project subject, among other information that can be considered relevant on the WG context.

The LinkedDataBR platform and approach have been applied to data associated to this scenario. Subsets of data from RNP projects are published and interlinked to datasets from CNPq CV Lattes researchers, CNPq research groups and from Brazilian Higher Education Institutions, using the 3rd treatment semantic approach (see section 5), which considers the conceptual models from data sources. Some steps of the triplification workflow for these data, implemented in Kettle, are illustrated in Figure 3. Likewise, an example of the resultant RDF dataset with some links is presented in a RDF Diagram (Figure 4) and in a RDF Code (Figure 5).

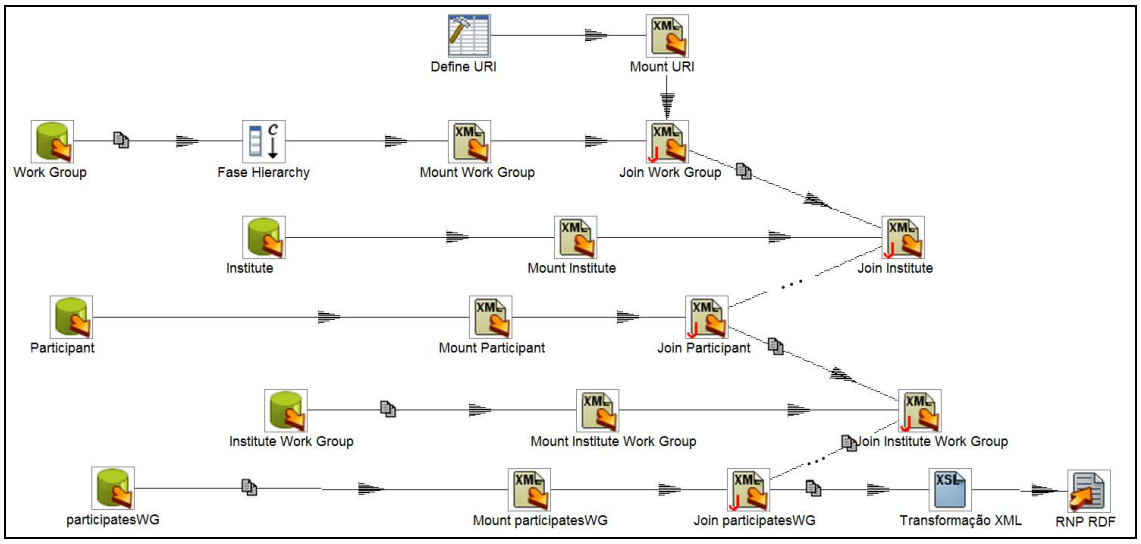

Figure 3. Some steps of RNP Triplification Process

The RDF entities in the diagram of Figure 4 are separeted in big boxes labeled with the source name (RNP projects, Brazilian Higher Educational Institutions, CNPq research groups and CNPq CV Lattes researchers). Moreover, those located out of this boxes represent terms of a vocabulary called SWRC (Semantic Web for Research Communities), available at LOV (Linked Open Vocabulary). Besides that, each diagram shape represents a different element: individuals - ellipses; data values - rectangular boxes; classes - rounded-corner boxes; properties - arrows. Particularly, the arrows connecting individuals and classes mean the instantiation property (rdf:type property) and, finally, the thicker arrows represent the links between individuals of different datasets.

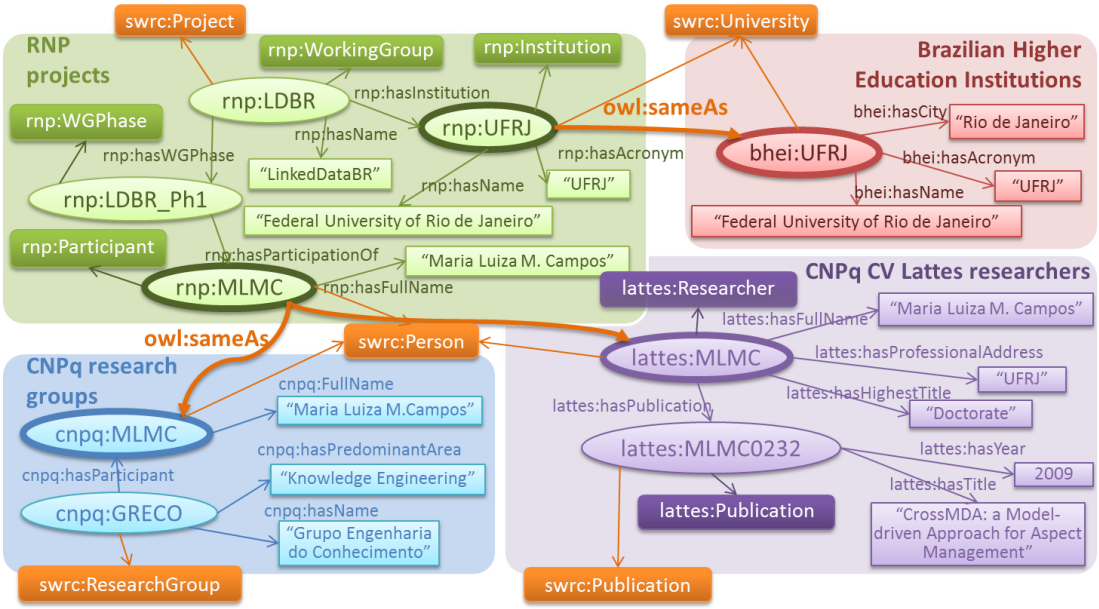

Figure 4. RDF Diagram linking example

```
...
<rdf:Description rdf:about="http://www.rnp.br/resource/LDBR_Ph1">
    <rdf:type rdf:resource="http://www.rnp.br/ontology/WGPhase"/>
    <rnp:hasParticipationOf rdf:resource="http://www.rnp.br/resource/MLMC"/>
</rdf:Description>
...
<rdf:Description rdf:about="http://www.rnp.br/resource/MLMC">
    <rdf:type rdf:resource="http://www.rnp.br/ontology/Participant"/>
    <rdf:type rdf:resource="http://swrc.ontoware.org/ontology/Person"/>
    <rnp:hasFullName> Maria Luiza  Machado Campos </rnp:hasFullName>
    <owl:sameAs rdf:resource="http://lattes.cnpq.br/resource/MLMC"/>
    <owl:sameAs rdf:resource="http://www.cnpq.br/resource/MLMC"/>
</rdf:Description>
...
```

Figure 5. RDF code linking example

Finally, a prototype application was built to visualize and explore the data. For the exploration of the published Linked Data, an SPARQL endpoint was defined, and an application was developed to allow navigation through the data. For example, from a list of RNP projects, one can list the participants of a selected project, and reach their CV Lattes or the research group to which they are associated and so on.

7. Conclusion

Governments around the world have been engaged in different initiatives towards decentralization and transparency of their actions, together with increasing citizens' participation. Open data play a fundamental role on these initiatives, but the real value of government assets is only truly revealed if data from various sources can be explored and used together.

LOD has emerged as a light-weight data interoperability and integration approach, exploiting already existing semantic web standards and technologies. Its potential was rapidly recognized, as an increasing number of new tools has been developed and there has been a substantial growth of the linked data cloud. But, when referring to governmental data, it is essential to consider some guaranties for the sustainability of the publication process and the quality assurance of the data interoperability strategy. In this work, we described a platform and an associated approach integrating different tools to facilitate LOD publication and to leverage

semantic interoperability. The platform was conceived to support governmental data publishers on managing the various phases of LOD life cycle, capturing provenance data along the process and allowing for various levels of conceptual enrichment.

The pay-as-you-go approach of linked data allows for an incremental data integration strategy and the flexibility to add new links stimulates associativity between resources. This can greatly contribute to increase the number of applications over governmental data, especially if we consider new opportunities for development teams and private sector. The potential of citizens' collaboration and the so called "wisdom of the crowds" will play an important role on mapping and linking. In this scenario, provenance management and mechanisms to assist in the data curation process become crucial and constitute future work.

Acknowledgements: This research is funded by RNP. Giancarlo Guizzardi and Maria Luiza M.Campos have research grants from CNPq.

References

Alexander, K., Hausenblas, M., Cyganiak, R., Zhao, J. (2009) "Describing Linked Datasets, On the Design and Usage of voiD, the Vocabulary of Interlinked Datasets", LDOW, Madrid, Spain.

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

Bizer, C., Jentzsch, A., Cyganiak, R. (2010) "State of the LOD Cloud", Freie Universität Berlin.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. and Hellmann, S. (2009) DBpedia: A crystallization point for the Web of Data, Web Semantics: Science, Services and Agents on the World Wide Web, vol. 7, no. 3.

Bouman, R., Dongen, J. (2009) Pentaho Solutions: Business Intelligence and Data Warehousing with Pentaho and MySQL, ISBN: 978-0-470-48432-6, 648 pages

Eaves, D. (2009) "The Three Laws of Open Government Data", available from: http://eaves.ca/2009/09/30/three-law-of-open-government-data.

Guizzardi, G. (2005) "Ontological foundations for structural conceptual models", PhD Thesis, CTIT, Centre for Telematics and Information Technology, Enschede.

Hartig, O. (2009) "Provenance Information in the Web of Data", LDOW, Madrid, Spain.

Heath, T., Bizer, C. (2011) "Linked Data - Evolving the Web into a Global Data Space" Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool

Jain, P.; Hitzler, P; Yeh, P; Verma, K; Shelt, A. (2010) "Linked Data is Merely More Data", Semantic Technology Conference 2010

Kimball, R., Caserta, J. (2004) The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data, Wiley, ISBN: 978-0-7645-6757-5

Manola, F., Miller, E. (2004) "RDF Primer", W3C Recommendation, available from: http://www.w3.org/TR/rdf-syntax/

Volz, J., Bizer C., Gaedke, M., Kobilarov, G. (2009) "Silk – A Link Discovery Framework for the Web of Data", LDOW, Madrid, Spain.