In this paper we discuss the current status of publishing OGD in Brazil and some challenges to be met. Our experience demonstrated that information sharing is bounded by specific RDF vocabularies used during the process of creating the triple sets. The problem is aggravated when the official language is not English, as these cases create datasets that can only be linked to local data. We also discuss how we addressed the problem of correlating local data to that of other countries, a major motivation for using Linked Data in the first place.

1. Introduction
Towards the adoption of the Open Government Data (OGD) paradigm [Accar 2009], countries all over the world have started to publish information produced by their public bodies on the Web according with specific standards, such as Linked Data [Bizer 2009]. Brazil is no exception, and efforts by federal government bodies and academia have emerged since 2009, revealing incipient but promising results towards the setting of a general approach for creating Brazil’s Linked Open Government Data infrastructure. In what follows, we discuss the current status of OGD in Brazil and summarize the lessons learned from producing Linked Open Government Data in Brazil. Moreover, we discuss a strategy to create datasets that can be correlated to other relevant repositories of governmental data in the LOD cloud.

2. OGD in Brazil
In September 2011, Brazil became a member of the Open Government Partnership, a multinational initiative to promote worldwide adoption of OGD. As a member, Brazil is committed to public transparency and action in securing open publication of official data. The commitment comprises both political and technical landmarks and includes a presidential mandate for the launch of the Brazilian Open Government Data portal1. Notwithstanding recent legislation, efforts towards OGD publication can be traced back to 2009, when the Information Organizing Committee of the Presidency of Brazil (COI) started to

1 A beta version is available at http://www.dados.gov.br
gather large amounts of aggregated government data for digital publication. The goal of the committee was to create a central information catalog of public activity, with the intent of improving governance and monitoring government activity. This catalog was originally created to serve the then President of the Republic and his team of advisors, as a reliable source of official data. The project was so successful that, reflecting open data principles, the catalog was made available to the public in 2010. TheDadosGov information catalog is comprised of slightly over 1,300 historic data series representing 8 years of public records that reflect government actions during Luiz Inácio “Lula” da Silva’s presidency (2003–2010). As a standard, the COI management team proposed to classify the data on two dimensions: territorial (country, states, cities) and temporal (year or month). Data series were classified in several hierarchical thematic trees, which branch from general (e.g., infrastructure, citizenship, and social inclusion) to more specific subjects that define third- and fourth-level trees. The original data comes from spreadsheets, provided to the COI team by more than forty different government bodies. The total volume is approximately 2.5 million records. The data received is stored in a relational database and made available through a specific Web site\(^2\).DadosGov data is also published in XML and JSON. However, neither XML nor JSON offer the semantic expression of RDF or interoperability with the rest of the Linked Open Data (LOD) cloud.

The office of the W3C in Brazil has also played an important educational role in the process of OGD publication in our country by sponsoring training in OGD technologies to information and communications technology (ICT) professionals in the public sector. Their initiative has brought government officials closer to the research community, which is collaborating in the construction of the Infraestrutura Nacional de Dados Abertos, the Brazilian OGD infrastructure. As a consequence, a task group with representatives from all sectors in the federal government was created to discuss basic issues: from the definition of a common vocabulary to represent data, to the adoption of methodologies and tools.

### 3. Practice and Lessons Learned

In 2010, we conducted – in cooperation with the COI team – a first experiment in converting a subset of theDadosGov database into RDF datasets. The conversion of the data from relational format to RDF triples was done using Triplify [Auer 2009]. This tool takes as input a relational database (schema and data) and produces as output a set of RDF triples, according to a mapping file that defines how database schema concepts must be represented in terms of RDF classes and properties. We found, however, that neither the selected tool nor other similar ones offered the necessary support during the conceptual modeling stage. We tackled this problem by developing StdTrip, a process and tool for helping to identify vocabulary matches [Salas et al. 2011]. The approach emphasized the use of well known RDF vocabularies to facilitate integration with other datasets in the LOD cloud. As a result, the data was described reusing – when possible – well known RDF vocabularies, e.g., DBpedia\(^3\), GeoNames\(^4\) and Datacube vocabulary\(^5\), among others. Moreover, we could link our datasets to DBpedia, which contains detailed information about Brazilian cities and states (population, area, etc). It provided enough semantics to produce a comprehensible set of linked RDF triple sets that enabled the creation of relevant mashups, providing graphical visualizations of data series fromDadosGov integrated with the map of Brazil.

In spite of this effort, most of the resulting triples were represented using an RDF vocabulary specially crafted to describe Brazilian government issues. We noted that the use of such contextualized, though locally relevant vocabulary, was very restrictive. The integration of these data series was limited to other datasets created based on the same vocabulary. We realized that, to enjoy a true Linked Data experience, we needed to generate datasets represented using vocabularies that enable the creation of relevant local mashups, but also the correlation to other datasets

---

\(^2\) [http://i3gov.planejamento.gov.br/dadosgov](http://i3gov.planejamento.gov.br/dadosgov)

\(^3\) [http://wiki.dbpedia.org/Ontology](http://wiki.dbpedia.org/Ontology)

\(^4\) [http://www.geonames.org/](http://www.geonames.org/)

produced globally. According to [DiFranzo et al. 2011], the hardest and most time-consuming step in the creation of a data mashup consists of gaining a deeper working understanding of the datasets that will be the basis of the mashup. We argue that this understanding is fundamental to support the selection of the vocabularies to be used to represent the RDF tuples, because such vocabularies must be directly related to the datasets with which we want to correlate our data in a mashup.

In order to create mashups where data about Brazil could be compared with data about the United States, new RDF datasets were generated considering vocabularies defined in data.gov website to represent similar concepts contained in the database schemas of DataGov. As an outcome, we improved our ability to create relevant mashups and made available datasets that can be easily used by international institutions with similar purposes. In Figure 1, we illustrate the potential of this approach by showing the correlation between the population living in extreme poverty in Brazil (top right) and in the US (top left). In the maps of both countries, the coloring of each state represents the percentage of population in extreme poverty. The darker the green color of the state is, the greater is the associated percentage. The pop-up (over the Brazilian map) presents the exact data for a selected state. The bar graph on the bottom left provides a detailed view comparing the states in Brazil and in the US with the highest percentages of population living in extreme poverty. The bar graph on the bottom right compares the overall percentage of people in extreme poverty in both countries.

Figure 1. Mashup of data from DadosGov and Data.gov showing the percentage of the population living in extreme poverty in each state of Brazil and the US

4. Conclusion

To secure Open Government Data integration at a global level, it is necessary to promote the use of standard RDF vocabularies. Adequate tooling is thus necessary during the triplification process to help users map local concepts to existing RDF vocabularies, in used by other datasets in the LOD Cloud. Our research group has invested in the development of user friendly tools that help users map contextualized information to standard vocabularies [Salas et al 2010; Salas et al 2011].

Our greatest challenge today, however, is the diversity of information sources. Although a great part of the data is stored in relational databases (OLAP datacubes) to facilitate the manipulation of statistical data, the sources usually require some pre-processing before the conversion process can take place. Another issue that is under development is improving the overall quality of the provenance provided, so as to increases user’s trust and the overall reliability of the

---

*DadosGov datasets and other different visualizations are available at [http://www.lodzone.com](http://www.lodzone.com).*
published datasets. Finally, at multinational level, it would be a good strategy to create a centralized repository to capture, store, and annotate the RDF vocabularies used by different countries together with their mappings.

References


