A Linked Open Data Approach for Visualizing Flood Information
A Case Study of the Rio Doce Basin in Brazil

Patricia Carolina Neves Azevedo[1,2], Guilherme Sousa Bastos[3] and Fernando Silva Parreiras[2]
[1]*CPRM – Companhia de Pesquisa de Recursos Minerais, Av. Brasil 1731, 30140-002, Belo Horizonte, MG, Brazil*
[2]*LAIS – Laboratory of Advanced Information Systems, FUMEC University, Av. Afonso Pena 3880, 30130-009, Belo Horizonte, MG, Brazil*
[3]*Institute of System Engineering and Information Technology – IESTI, Federal University of Itajubá, Av. BPS 1303, 37500-903, Itajubá, MG, Brazil*
patricia.neves@cprm.gov.br, sousa@unifei.br, fernando.parreiras@fumec.br

Keywords: Linked Open Data, Geographical Information System, Flood, Semantic Web.

Abstract: The availability of open government data offers an easy way to mix and match these data to create new knowledge. Geographic Information Systems powered by Semantic Web technologies and linked data result in an integration of data from multiple sources, facilitating its use and enhancing the discovery and dissemination of new knowledge. In this work, we present a prototype application that integrates heterogeneous data located in various public organizations, related to flooding in Rio Doce Basin – Brazil. For this purpose, data were converted to RDF format, linked and displayed on a Geographic Information System, through SPARQL queries. We validate our approach using a proof of concept. The results show that our proposal of liking open data about flood information is able to answer the identified competency questions.

1 INTRODUCTION

The Brazilian federal government, through responsible agencies, adopts actions to minimize the damage caused by floods in river basins, such as collecting and analyzing data. However, despite the amount of information available, these are spread out over several data sources in multiple institutions (eg, government agencies, private companies and academic institutions), databases, schemas and heterogeneous formats. Some data are available only in PDF or scanned image files in non-compliance to the Brazilian Information Access Law (Law No. 12,527) and are causing rework in agencies and entities that use these files. The diversity of formats and data models hampers the interpretation, integration and reuse. Moreover, there is not possible to display them for a interested user in following up the history of water levels in the rivers of the Rio Doce basin.

In this context, the following question unfolds: What are the concepts and technologies that allow the integration and make available the data related to floods in the Rio Doce Basin?

When dealing with floods, one realizes that visualization, interaction and dissemination of these data can assist in disaster management. In this context, the principles of linked data (Bizer et al., 2009) are a means to make the information shared on the web available in an standardized way, publishing and linked datasets.

This paper presents a framework able to (1) receive, from different sources, data about floods in Rio Doce Basin, (2) integrate them using semantic web (Berners-Lee et al., 2001) technologies and standards and (3) make them available visually to interested users.

Thus, by viewing the integrated data from the Rio Doce basin, it will be possible to identify vulnerable communities and develop emergency and preventive actions, contributing to disaster management on the basin of the Rio Doce.

The Brazilian government encourages the publication of data to the public through the Internet, aiming to inform the population and support the transparency of government data. However, the publication of unstructured data is insufficient to achieve the goals of efficiency, transparency and accountability. Semantic Web technology can contribute to achieving these goals by providing data integration of heterogeneous sources.

The paper is organized as follows: Section 2 contextualizes the research problem. Section 3 describes

Carolina Neves Azevedo P., Sousa Bastos G. and Silva Parreiras F..
A Linked Open Data Approach for Visualizing Flood Information - A Case Study of the Rio Doce Basin in Brazil.
DOI: 10.5220/0005465502270232
In *Proceedings of the 1st International Conference on Geographical Information Systems Theory, Applications and Management* (GISTAM-2015), pages 227-232
ISBN: 978-989-758-099-4

the background. Section 4 details the proposed solution for visualizing linked data about floods in the Rio Doce basin, presenting the conceptual framework. Section 5 describes the implementation. Section 6 discusses the related work and Section 7 concludes the paper by highlighting its contribution and future lines of action.

2 SCENARIO

When analyzing the current situation of data from the Rio Doce basin, we observed that they are not in a format available for reuse. Nowadays, only reports with measurement data from sensors installed along the Rio Doce basin are available on the Internet, using technical language, not appropriate for lay users. In this scenario, in which citizens do not have access to information about the historical and monitoring of water levels from the Rio Doce basin, answering the following questions can expand the vision of managers and interested citizens:

Q1: What was the level of the river the monitoring points which recorded flood in the day X?

Q2: What is the region with the largest population affected by floods in the day X?

Q3: What are the municipalities most affected by rain with HDI below of X?

Q4: The works against floods from Brazilian Acceleration Plan (PAC) are developed in the most affected areas?

Q5: On which areas occurs more floods and diseases related to floods?

Q6: In which areas the occurrence of flooding happened in areas of low altitude?

3 BACKGROUND AND OBJECTIVES

When analyzing data about natural disasters in Brazil during the period 1980-2010, provided by the main database used by the UN, the International Disaster Database (EM-DAT), one might notice that the main recurring natural hazards are floods. Brazil is the seventh country in the global ranking on the number of flood victims. The study obtained data from 97 countries between 1980 and 2000, and reported that more than 29 million Brazilians live at risk of being affected by flooding (Collins, 2004).

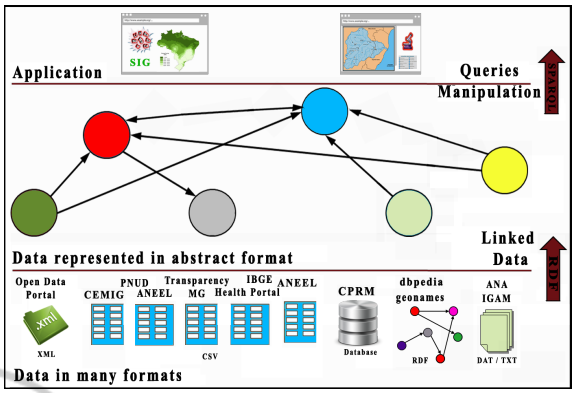

Figure 1: Overview of the proposed architecture, based on (Herman, 2012).

This work is supported by the Brazilian government initiative as regards about opening and dissemination of public data, according to the Brazilian Information Access Law (Brasil, 2011). Considering the interest by the government and the demand for solutions in the Rio Doce Basin, often afflicted by floods that causes economic, human and material losses, the focus will be the use of Geographical Information Systems (GIS) (Burrough et al., 1998) and semantic web tools as framework to generate information about the dynamics of the phenomenon in the Rio Doce Basin.

Government data published on the Web, by itself, already has great value for the population, as they contribute to increased transparency. But making such information available in open and accessible formats allows them to be machine-readable, facilitating the discovery, consumption and adding value, allowing linkage of data to other datasets.

In the scope of this work, we developed a prototype application which receives, from different sources, data about floods in the Rio Doce basin, integrates it and makes it available to interested users.

4 CONCEPTUAL FRAMEWORK

The Figure 1 depicts the decomposition of components that are part of the proposed solution, and the relations between them. With this architecture, it is possible, through linked data technologies and principles (Bizer et al., 2009), to receive data from different organizations, to integrate them and to make them available visually.

The Figure 1 is divided according to the following layers:

(a) Data. The data were obtained from various public agencies in different formats (txt, dat, csv, xml,

rdf), and open data from the Linked Data Community available on the Internet. These data were stored in a database and converted to standard RDF (Manola and Miller, 2004).

(b) Dataset. The dataset generated from the conversion is already one of the results of this research. It concerns any information of the levels of the rivers that comprise the Rio Doce Basin, as well as levels of attention and alert and information from municipalities connected. To answer the research questions, the SPARQL queries (PrudHommeaux et al., 2008) were engineered and the result forwarded to GIS.

(c) Visualization in a GIS. The application layer is on top of the architecture, where the information is displayed through the GIS, in a friendly interface and able to answer the questions suggested initially.

The Figure 1 shows three layers of the proposed solution architecture, where the first layer are the datasets. These data, relating to floods in the Rio Doce basin, are in different formats and will be converted to standard RDF with the aim of being interconnected and thereby generate the RDF graph, which is illustrated in the second layer of the architecture. In the last layer, we will use SPARQL [1] to query on this data. The result is the combination of all data, and a geographic visualization in a GIS. Geographic information is distinguished from other information by referring to objects or phenomena in a specific location in space and, therefore, has an spatial address (Kraak and Ormeling, 2003).

As one of our goals is to create a new dataset with data from flooding from the Rio Doce basin, it has become necessary to collect data from different sources, including government databases. In this case, there were collected data from ANA (National Water Agency), ANEEL (National Energy Agency), Cemig (Energy Company), IGAM (State Institute for Water Management) and CPRM (Mineral Resource Research Company) through FTP sites or directly through the organization's Web site.

After structuring the data, RDF is used [2] to represent the information, as proposed by the W3C to publish linked data on the web.

[1] As systems databases make use of SQL to query records in databases, SPARQL is a query language for retrieving information in RDF graphs (PrudHommeaux et al., 2008).

[2] Resource Description Framework (RDF) is a language for representing information on the Web and designed for situations where information needs to be processed by applications, rather than simply being shown to people (Manola and Miller, 2004).

5 IMPLEMENTATION

5.1 Data

Data used in this study came from many sources, including governmental agencies. These data were in unstructured formats and have undergone harmonizing, rescaling, and cleaning before its use in the prototype. Given the effort to promote the semantic web (Berners-Lee et al., 2001), we tried to follow open standards as recommended by W3C, representing datasets as linked data (Bizer et al., 2009).

The dataset creation involved two lines of action: the extraction of collected data through FTP sites or HTML pages of organizations and data conversion from relational databases to RDF model.

Several measurement stations operate in different parts of the rivers and municipalities. Data extracted from these were in TXT format and were converted to CSV using MS Excel software. Other data also related to river levels were collected directly from a CPRM's server, with an employee help. These were in DAT format and were also converted into CSV using the same software. Data were collected from the government website in XML format [3], regarding the Growth Acceleration Program (PAC) in Minas Gerais. Data about HDI of the municipalities was obtained through the PNUD website [4], and were in CSV format. In Brazilian Health Portal website, we collected data about the occurrence of the following diseases related to floods: tetanus, dengue, leptospirosis, malaria, hepatitis A and C, typhoid and cholera. These data were also found in CSV format. Population and altitude data of each municipality was collected directly from the Brazilian Statistics Bureau (IBGE) [5] website in CSV format.

Other sources of data were used aiming to aggregate information, as shown in Table 1.

5.2 Dataset

To convert spreadsheet, CSV files, XML files, relational databases and other documents to RDF format we used D2RQ platform (Bizer and Seaborne, 2004).

The D2RQ was chosen for use in this study because of some factors, among which stands out the flexibility of mapping language, the simplicity of the commands and the generation of RDF dumps, making possible the reuse of the dataset created.

The next step was to generate RDF dump file from the mapping file and through the dump-rdf

[3] http://dados.gov.br/

[4] http://www.pnud.org.br/

[5] http://cidades.ibge.gov.br/

Table 1: Source, description and format of data used in the study.

Source	Description	Format
ANA	Precipitation and Rivers Levels	DAT
ANEEL	Precipitation and Rivers Levels	CSV
CEMIG	Precipitation	CSV
IGAM	Precipitation	TXT
CPRM	Rivers Levels	Database
Transparency Portal of MG	Onlending of Investments	CSV
IBGE	Population and Altitude	CSV
Health Portal	Diseases	CSV
PNUD	HDI	CSV
Open Data Portal	PAC Works	XML
Geonames	Geographic Names, latitude and longitude	RDF
DBPEDIA	General data of the cities	RDF

D2RQ platform tool. The command provides the following types of output format: Turtle, RDF/XML, RDF/XML-Abbrev, N3 or N-Triple. In this work, the RDF/XML has been used.

5.3 Visualization in a GIS

The result of SPARQL queries was displayed into the GIS, a web application implemented using JavaScript language, where the user selects data about the Rio Doce basin, to be viewed on the map. Different combinations can be made with the objective of linking data from multiple sources simultaneously, for example, you can see if the places with high occurrence of floods are the same with occurrences of diseases related to floods or low HDI.

6 VALIDATION

In order to validate the proposed approach, a proof of concept through competency questions was conducted, as presented in Section 2. Following, the demonstration of queries use in the application and its results.

6.1 Data

With the RDF dataset created about the floods in the Rio Doce Basin and aggregate data, such information becomes part of the Web of Data, where machines and humans can search and use this data set as one of its data sources.

Is believed that the availability of open and standardized data enables discovery of new knowledge through reuse of this data in new applications. Publishing data about floods in Rio Doce Basin follows the principles of linked data and enables discovery, integration and searches for other sources of data.

The figure 2 shows the RDF graph, generated from the RDF file, where the classes Town and River inherit from the upper class Thing.

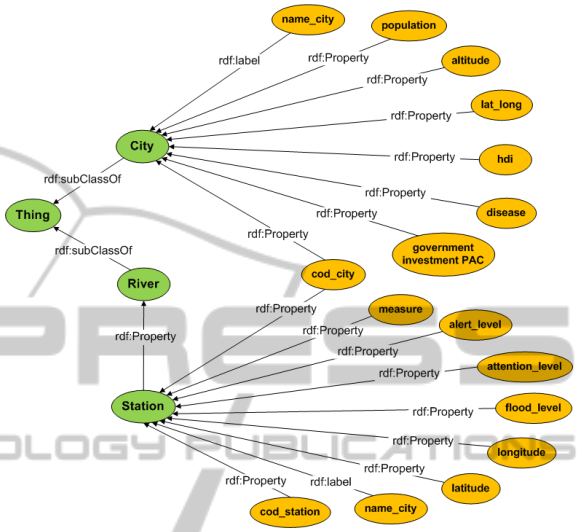

Figure 2: RDF graph representing the dataset created.

After generating data in RDF model, the file is validated according to Linked Data principles. This verification was taken through the online validation tool W3C RDF Validation Service which was executed successfully. RDF files are available in RDF/XML and N/Triple formats at the following links:

RDF/XML: https://db.tt/pJ0r78qw - N/Triple: https://db.tt/DKx7dkK4. Thus, data are ready to be consumed as linked data through browsers, search engines or applications for specific domains.

6.2 Dataset

The table 2 shows queries and results, limited to 10 lines and no sorting, of the competency questions Q1 and Q2 respectively, as a form of validation of the concepts mentioned before.

6.3 Visualization in a GIS

Visualization and interaction of linked data is a question that has been recognized since the beginning of the Semantic Web (Geroimenko and Chen, 2003). When applying techniques of information visualization, semantic beb assists users in exploration and interaction data. The processing and visual presentation of these data are the main goals of information visualization, so that users can get a better understanding of

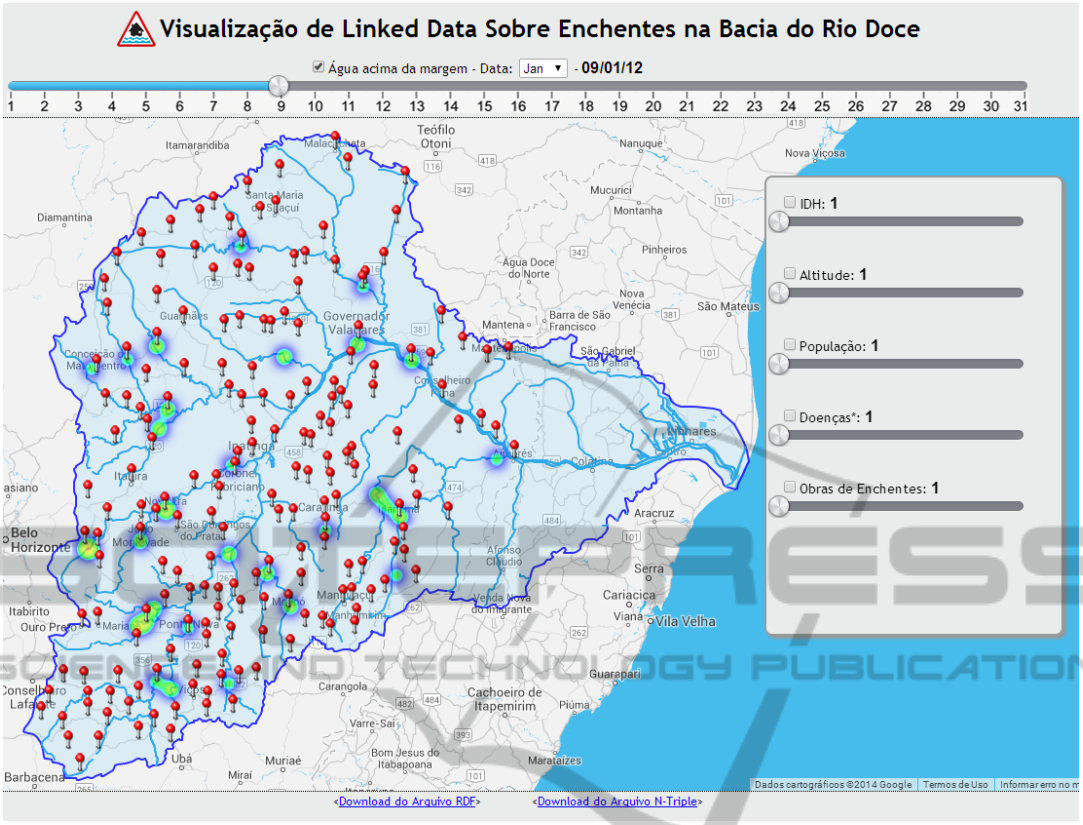

Figure 3: Visualization of query Q1.

Table 2: Query Q1.

Which stations recorded flood on 01/09/2012?
SELECT ?resource ?cod_station ?station ?measure ?alert_level ?date WHERE { ?resource geonames:featureCode ?cod_station . ?resource paoli:Open_stream_water_level_recorders ?station . ?resource dbpprop:date ?date . ?resource loa:WATER_LEVEL_2 ?measure . ?resource ontosem:flood ?alert_level . FILTER (?date = ”2012-09-01”^^xsd:date)}

the data (Card et al., 1999). Visualizations are useful for obtaining an overview of the datasets, their main types and the relationships between them.

This application of data visualization provides two main contributions: the visualization of information into a map and the proof that it is possible to make consistents applications from the dataset created in this research. The figure 3 illustrates the resulting prototype, which presents the return of SPARQL queries on a map.

In accordance with the Figure 3, the most populous municipalities which have suffered from flooding on 20/01/12 were Governador Valadares, Caratinga, Timteo and Coronel Fabriciano.

7 RELATED WORK

An early example of using geographic information system was performed by John Snow showing relation between water supply and cholera outbreaks in London in 1854, achieved by linking public data about contaminated water and disease (Johnson, 2006).

In the research of Nurefşan Gr, Laura Diaz e Tomi Kauppinen, was used linked open data to publish health-related data, such as diseases, disorders, genes, and drugs into a technology of visualization referred to geo web. For this, the use case studied was RCPH - Research Center of Public Health, based on three conceptual domains: health, spatial and statistical and following the linked data principles. Finally was used an infrastructure integrating geospatial and semantic web technologies to show mortality rates for specific diseases in a spatio-temporal format. (Gür et al., 2012).

Finally, (Vilches-Blázquez et al., 2010) presented

a sequence of procedures used to develop an application that used multiple heterogeneous public datasets, about Spain, which are specifically related to administrative units, hydrography and statistical units. The application aims to analyze existing relations between the Spanish coastal area and different statistical variables such as population, unemployment, housing, industry, commerce and construction. Besides providing methodological guidelines for the generation, publishing and exploitation of Linked Data from these datasets, it was used resources to handle the geometric information of data.

Can be observed that all related work generate an RDF file and visualization in a GIS, however none combined data from a specific topic with statistical data from the location involved, as seen in this study. It is important to note the use of government data in all related work.

8 CONCLUSION

With the RDF dataset created about floods in Rio Doce Basin and aggregate data, such information becomes part of the Web of Data, where machines and humans can search and use this data set as one of its data sources.

Thereby, the contribution of this experiment encompasses the use of methods and tools for publishing data as the principles and standards linked data. It is believed that the availability of open and standardized data enables discovery of new knowledge, of this data through reuse in new applications. For the citizen, the developed application allows a user-friendly visualization of data involved in the research and knowledge discovery from them.

In future, it is suggested adding data from year 2013 to compare to 2012 data, identify advances in government measures against floods, disease control and river levels in the same seasons. In addition, other lines of future action are highlighted: a) Expansion of the Dataset: The inclusion of pertinent data improves the relevance, especially when link with existing data; b) Improvements in data visualization application: Extending the dataset enables new ways of representing data more user friendly way. Therefore, the visualization of information can be enhanced with a larger amount of data, making it more dynamic and interactive for the end user application.

ACKNOWLEDGEMENTS

This work is partially supported by the Brazilian Funding Agencies FAPEMIG, CNPq and CAPES.

REFERENCES

Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001). The semantic web. *Scientific american*, 284(5):28–37.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far. *International journal on semantic web and information systems*, 5(3):1–22.

Bizer, C. and Seaborne, A. (2004). D2rq-treating non-rdf databases as virtual rdf graphs. In *Proceedings of ISWC2004*, volume 2004.

Brasil (2011). *Law on Access to Public Information*. Law Number 12.527/2011.

Burrough, P. A., McDonnell, R., Burrough, P. A., and McDonnell, R. (1998). *Principles of geographical information systems*, volume 333. Oxford university press.

Card, S. K., MacKinlay, J. D., and Schneiderman, B. (1999). *Readings in information visualization: using vision to think*. Interactive Technologies Series. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Collins, T. (2004). Disaster risk for floods.

Geroimenko, V. and Chen, C. (2003). *Visualizing the Semantic Web: Xml-Based Internet and Information Visualization*. Springer-Verlag GmbH.

Gür, N., Díaz, L., and Kauppinen, T. (2012). Gi systems for public health with an ontology based approach. In *AGILE2012*, Avignon, France.

Herman, I. (2012). Tutorial on semantic web technologies. Presentation. Available on http://www.w3.org/People/Ivan/CorePresentations/SWTutorial/.

Johnson, S. (2006). *The Ghost Map: The Story of London's Most Terrifying Epidemic–and how it Changed Science, Cities, and the Modern World*. Riverhead Books.

Kraak, J. and Ormeling, F. J. (2003). *Cartography: visualization of geospatial data*. Prentice Hall.

Manola, F. and Miller, E., editors (2004). *RDF Primer*. W3C Recommendation. W3C.

PrudHommeaux, E., Seaborne, A., et al. Sparql query language for rdf. W3c recommendation, W3C.

Vilches-Blázquez, L. M., Villazón-Terrazas, B., Saquicela, V., de León, A., Corcho, O., and Gómez-Pérez, A. (2010). Geolinked data and inspire through an application case. In *SIGSPATIAL 2010*, GIS '10, pages 446–449, New York, NY, USA. ACM.