ConvertView
A Tool for Conversion and Visualization of Open Heterogenic Governmental Data to the RDF Standard

Clóvis H. Nascimento[1], Ricardo A. Afonso[2], Vinicius C. Garcia[1] and Carlo M. Revoredo da Silva[1]
[1]*Informatics Center – Cin, Federal University of Pernambuco (UFPE), Caixa Postal 7851, 50.740-560, Recife, Pernambuco, PE, Brazil*
[2]*Federal University of Alagoas (UFAL), Campus Arapiraca, ARACOMP, Arapiraca, Alagoas, AL, Brazil*

Keywords: Open Data, Structured Pattern, Data Extraction, Cloud Computing.

Abstract: The diversity and relevance of the information made available by many governments, using the principles defined by the open data manifest, creates an opportunity for the populations aid in many fields of governmental administration, such as security, transportation, health care and education. However, most of this information was generated in unstructured digital files, making them hard to use. This paper presents an approach for the structuring of data, improving the processing and extraction of information. Thus, we created a tool based on a cloud computing environment, which has high processing and storage ability, to create files in a homogeneous and structured format, allowing for ominous and specific queries; contributing to the access of the information.

1 INTRODUCTION

When the manifesto named Open Data (Accar 2009) appeared through the governments of many countries like United States and Brazil, to provide their legacy and current data to the world through the internet, created positive expectations in the population. The manifesto aimed, total and permanent transparency of their data produced in several areas, such as security, transportation, health, education, among others. In addition the use of these data, enables a better political and economic contribution from the population, for example, identifying the areas that require a greater investment of resources by the government.

In Brazil's case, these expectations were not fully met, not because of the data itself, but because of the difficulty in using them. According to (Breitman, K. et al. 2011) our biggest challenge today comes from the diversity of data sources and the provenance of these sources. Furthermore, most of the data provided by the Brazilian government are not in the Linked Data standard (Bizer, C. 2009), which allows the interconnection of different data sources, being the format recommended by the W3C consortium. The available Open data is unstructured (documents, spreadsheets etc.), despite several government initiatives supported by the W3C, the results aren't sufficient to define a consistent approach, which would facilitate the creation of an infrastructure of Governmental Data in the Linked Data standard.

Despite the listed problems, the exponential growth in available data on the web and other open data initiatives creates new challenges, such as the processing of large volumes of information and minimizing the high costs of hardware and software. To solve these problems, new technological approaches emerged, such as big Data (Wu et al. 2013), designed to process large volumes of distributed, heterogeneous data and cloud computing (Qi Zhang et al. 2010), which was used in this work. Cloud Computing provides hardware and software as a service for a low cost combined with various benefits such as high processing power, storage, scalability and so on. According to (Simmhan et al. 2010), scalability and ease of integrating data from multiple heterogeneous networks, justifies the adoption of cloud computing.

In this context, this paper aims to contribute in the research of heterogeneous processing for the RDF (Resource Description Framework) data standard, thus, in an attempt to fill this gap, we propose a tool that operates in the cloud computing environment called ConvertView containing options for the user to

H. Nascimento C., A. Afonso R., C. Garcia V. and Revoredo da Silva C..
ConvertView - A Tool for Conversion and Visualization of Open Heterogenic Governmental Data to the RDF Standard.
DOI: 10.5220/0005456902530258
In *Proceedings of the 11th International Conference on Web Information Systems and Technologies* (WEBIST-2015), pages 253-258
ISBN: 978-989-758-106-9

convert XML file to RDF, SPARQL queries and display (Simple Protocol and RDF Query Language), in the destinations files. It is expected to evolve to the state of art in the literature.

The paper was divided into six sections, starting with the introduction to the subject, describing the problem and the motivation. Section 2 presents the related work, section 3 presents the proposed solution and its implementation, section 4 presents the results, section 5 the conclusion and finally in section 6 the references.

2 TOOLS AND RELATED WORK

This section describes briefly some recent references on the state of the art, since many works have the same goal: to solve the problems of data transformation for semantic format, but follow different approaches.

A tool called StdTrip, developed by (Salas, P. et al. 2011), generates RDF files from relational databases of the Brazilian government. This approach requires the triplication of the data set, in other words, the conversion to RDF of database schemas and their instances. A key issue in this process is deciding how to represent concepts of database schema in terms of RDF classes and properties by mapping database concepts and an RDF vocabulary.

The Rich Internet Application (RIA) called WebSmatch, created by (Coleta, Remi. et al. 2012), supports the entire process of importing, refining and integrating the data sources and uses third-party tools for high quality display. This application uses a common situation for the integration of public data unresolved by current tools: the input files are poorly structured (excel xls in this case) and occurs only a superficial view of these data.

(Edgard Marx et al. 2012) Created a tool that generates RDF files, database schemas and mappings, providing a plug-in for use with the Eclipse program and connects to the DBMS via JDBC (Java Database Connectivity). The tool offers a range of mapping languages, including R2RML (RDB to RDF Mapping Language), however it does not support SPARQL queries to perform.

The tool implemented by (Giacomo, G. et al. 2012) provides access to any data management system through the use of ontologies for semantically representing the data source. The connections with the DBMS are made via JDBC and semantic mappings are created using OWL (Web Ontology Language). The tool called Mastro provides a tool called OBDA environment (Ontology Based Data Access), with resources for creating and querying, as well as enabling consistency and reasoning checking on OWL ontology.

Despite the quality and relevance of the work described in this section, they do not use cloud computing. It´s important to recall that according to (Mutavdzic 2010) cloud computing can reduce costs, simplify management, improve service, provide transparency and allow interaction with the flexible citizen, helping to fulfil their governments wish to transform the way they serve their citizens. Government agencies and departments have difficulty to identify and implement opportunities for shared services that have a good chance of successfully implementing new services to citizens.

One of the advantages of the proposed tool is the ability to query the RDF files without the necessity of previous knowledge of the SPARQL query language. The intuitive interface allows anyone to use filters that will generate the query, and the tool shows the query command created by the filter selection. Table 1 shows a comparison of the tools described in this paper:

Table 1: Related works comparative table.

Project	Objective	Entry
StdTrip	Generate RDF, database schemes and mappings. Suggests vocabulary for the concepts.	Database schemes
WebSmatch	Generate XML and CVS files from XLS Excel sources.	XLS Excel
RDB2RDF	Create RDF, database schemes and mappings. Provides plugins to use with Eclipse and connects the bases through JDBC	Connection with the database through JDBC
Mastro	Generate a well-structured file from a database	Relational database
ConvertView	Generate RDF, mappings, visualization and consultation from sources	

Table 1: Related works comparative table (cont.).

Automation Level	Mapping Language	Supports SPARQL	Using Cloud Computing
Manual/ Automatic	SQL, RDF, R2RML	No	No
Manual	XML	No	No
Manual/ Automatic	R2RML	No	No
Automatic	OWL	No	No
Manual	XML	Yes	Yes

3 PROPOSED SOLUTION

In this section, a proposal for solving the problems described above is presented. For this reason, a tool that enables users to generate RDF databases from datasets in the XML (eXtensible Markup Language) format was created. The process of creating the RDF file is shown in Figure 1, starting with the user selecting a dataset file that served as basis for RDF. Then the terms in the datasets are presented on the screen and the user chooses the words and the corresponding vocabularies, where the available vocabularies are FOAF (Friend of a Friend) and DC (Dublin Core), and finally the RDF file and mapping are generated.

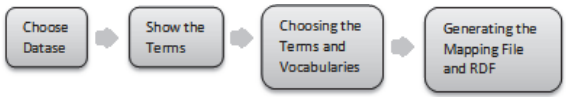

Figure 1: Steps mapping and generation of RDF file.

In state of art research, the majority of studies found use as a base XML, XLS, databases and other formats for the generation of RDF files. We observed that all of them operate interactively in the selection of the used vocabularies, making it up to the user to decide the vocabulary used in the process of generating RDF files. Another observation made was related to the languages of mappings: there is no standard, ie, each job uses a different representation formats that are worth highlighting: XLS, SQL, RDF and OWL. The consortium W3C (World Wide Web Consortium) recommends the use of the mapping language R2RML (RDB to RDF Mapping Language) since late 2012.

3.1 Implementation

The development of a tool for generation of semantic data, depending on the complexity of data sources and the degree of formality to be generated, can be an arduous and lengthy process. However, the previous clear definition of the steps should to be followed until their conclusion, in order to facilitate their implementation.

After defining the purpose and scope of the tool, the technologies needed for implementation are chosen. In this section, the technologies and development of the environment, which were used in the creation of the ConvertView tool, was presented.

3.2 Used Technology

The programming language and computing platform Java, version 1.7, was chosen because of its scalability and multi-platform aspects. In this work Jena Semantic Web Framework 2.7.4 was used to generate RDF files and perform queries SPARQL (Protocol and RDF Query Language) endpoint.

3.3 Development Environment

The chosen Open Source environment was the IDE (Integrated Development Environment) provided by Eclipse, used in this work to write, debug, and run the source code of the ConvertView tool. The Google App Engine platform offered as a service (PaaS) of cloud computing for the development of web applications, enabled the creation, implementation and management of the developed application developed and was chosen to provide plugins for the Eclipse IDE.

3.4 System Architecture

Observing the architecture of the tool in Figure 2, it is possible to observe that it has only two layers, one for the client and one for the cloud server. In the diagram we can see the main components of each layer and its interfaces.

4 RESULTS

In this section, the results obtained from the use of the Convertview tool will be presented. For this, some datasets that were used and various combinations of terms and vocabularies were selected, in order to demonstrate the use and efficacy of the tool. As shown in Figure 3, a dataset called Professores.xml was selected and the tool showed the terms presented

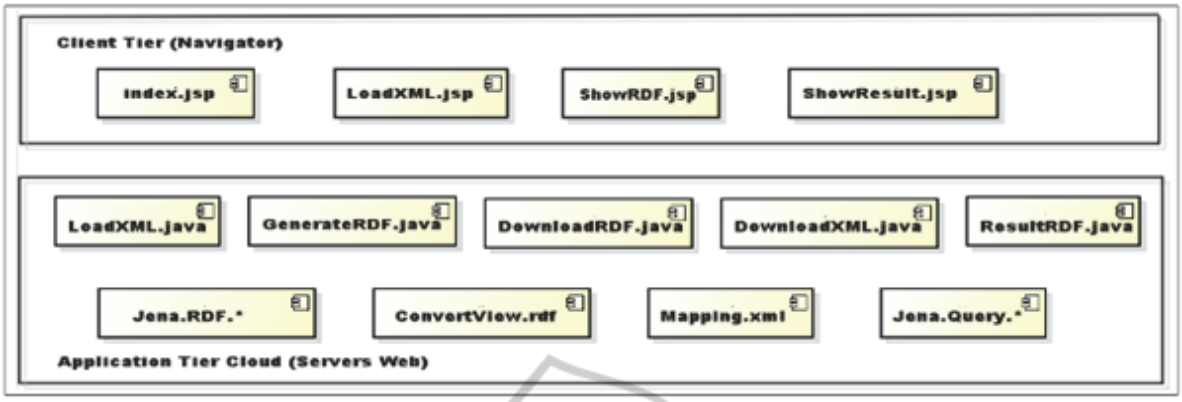

Figure 2: Architecture Diagram of ConvertView Tool.

in the dataset (in this case, Email, Matricula, Nome and Sala). Then, the user picked the three first terms, informing the desired vocabularies (accountname, identifier and name).

Figure 3: Main Screen of the tool ConvertView.

After selecting the terms and vocabularies, clicked in the button generate, to make possible complete the process of creating the RDF. Figures 4 and 5 show the mapping files and RDF file, respectively.

```
<?XML version="1.0" encoding="UTF-8"
standalone="no"?>
<mappings> <mapping> <term>Email</term>
<vocabulary>accountname</vocabulary>
<term>Matricula</term>
<vocabulary>identifier</vocabulary>
<term>Nome</term><vocabulary>name</vocabulary>
</mapping></mappings>
```

Figure 4: Mapping File generated by the ConvertView tool.

```
<RDF:RDF
XMLns:RDF="http://www.w3.org/1999/02/22-RDF-
syntax-ns#"
   XMLns:j.0="http://XMLns.com/foaf/0.1/"
XMLns:dc="http://purl.org/dc/elements/1.1/">
<RDF:DescriptionRDF:about="http://www.cin.ufpe.br
/thing/Professor2">
<j.0:name>Vinicius Cardoso Garcia</j.0:name>
<dc:identifier>2</dc:identifier>
<j.0:accountName>vcg@cin.ufpe.br</j.0:accountNam
e>
</RDF:Description>
<RDF:DescriptionRDF:about="http://www.cin.ufpe.br
/thing/Professor1">
<j.0:name>Silvio Romero de LemosMeira</j.0:name>
<dc:identifier>1</dc:identifier>
<j.0:accountName>srlm@cin.ufpe.br</j.0:accountNa
me>
</RDF:Description>
</RDF:RDF>
```

Figure 5: RDF file generated by the ConvertView tool.

The mapping file was created in the XML format, since semantic expressiveness with logical constructors is not needed as it is in the OWL language, with simple correspondence of terms and vocabularies taking its place.

After creating the RDF file, it is also possible to check your data inside the tool, using the SPARQL language, as shown in Figure 6.

The tool itself generates the prefixes of the vocabularies and the user can enter any“select” commandto search in the generated RDF file. Figure 7 shows the use of the select command, which showes on the screen all the data contained in the RDF file sorted by name.

5 CONCLUSIONS

This work aimed to advance the state of the art concerning the limitations of use of open data, due to the absence of structured modelling. For this, the Convertview tool was developed, showing several advantages due to its approach in a cloud computing environment, with high processing power, availability, storage and scalability. A summary of the issues involved in its development and use, the technologies and platforms used, as well as the reasons these choices was exposed.

The tests made allowed to confirm the success in the generation of RDF files and the option to make SPARQL queries complements the usefulness of the tool, which eliminates the need for another program to search the generated data. However, the tool is still in its early stage of development and the next steps of development and maturity predict the following developments in the following software:

1. Formats dataset sources: Increase the amount of file types it is able to read (cvs, json, xls, etc), not just XML;

2. Increase the amount of vocabularies: Use also other famous and stable vocabularies, such as EVENT, Geonames, AKT, BIBO, SWC and SWRC with all its terms, to include a wider variety of areas;

3. Suggestions of terms and vocabularies: Suggest terms and vocabularies, based on the content of the dataset, assisting in the choice of the user;

4. Mapping Languages: Add the option to choose the mapping language, to allow better interoperability with other tools.

5. Allow the creation of RDFS: If the user wishes, RDFS files could be generated as well, with classes and subclasses defined by the user.

6. Allow the input of simultaneous sources: Enable the reading of various heterogeneous sources simultaneously, allowing to correlate equivalent terms in order to enrich future consultations.

Finally, it is evident that the demand for such tools has increased due the amount of unstructured data that has been provided, especially on internet files,thus revealing that there is much work to be done. In this context, this work was developed to help to fill this gap.

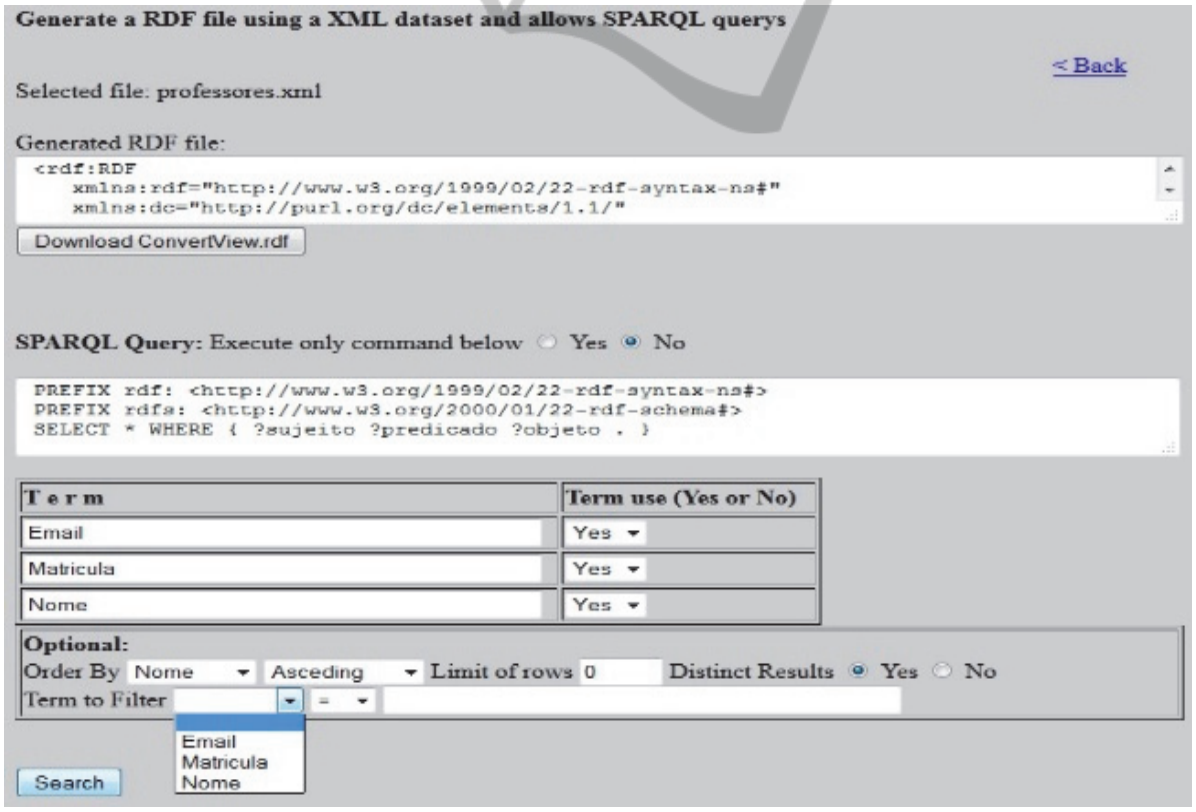

Figure 6: Screen filter search automatic or manual.

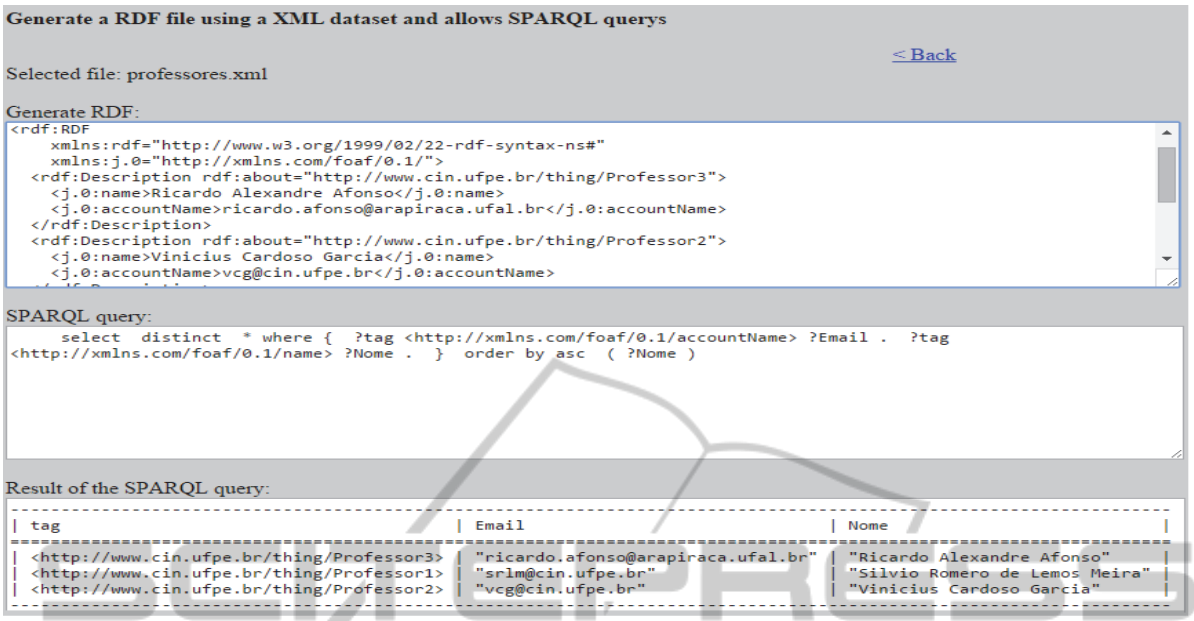

Figure 7: Screen queries in SPARQL.

REFERENCES

Accar, S., Alonso, J., Novak, K. (eds). Improving Access to *Government through Better Use of the Web (2009).* W3C Interest Group.

Bizer, C. "The Emerging Web of Linked Data"; *IEEE Intelligent Systems* (Sep/Oct. 2009).

Breitman, Karin K., Salas P., Saraiva D., Gama V., Casanova Marco A., Open Government Data in Brazil, Department of Informatics, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2011.

Coleta, Remi.,Castanier, E., Valduriez P., Frisch C., Ngo, D., Belllahsene Z., Public Data IntegrationwithWebSmatch, *INRIA and LIRMM*, Data Publica, Frannce, 2012.

Edgard Marx, Percy Salas, Karin Breitman, José Viterbo and Marco Antonio Casanova, RDB2RDF: *A relationalto RDF plug-in for Eclipse*, Informatics Department, Pontifícia Universidade Católica do Rio de Janeiro, 2012.

Giacomo, G.,Lembo, D., Lenzerini, M., Poggi, A. Riccardo Rosati, Ruzzi, M., Savo, D. F, Mastro: *A Reasoner for EfectiveOntology-Based Data Access*, University Sapienza of Roma, Italia, 2012.

MUTAVDŽIC, R. "Cloud Computing Architectures for National, Regional and Local Government".*Microsoft Croatia. 2010.*

Qi Zhang Lu Cheng Raouf Boutaba The Brazilian Computer Society. *Cloud computing: state-of-the-art and research challenges*, 2010.

Salas, P., Viterbo, J., Breitman, K. and Casanova, M.A. StdTrip:Promoting the Reuse of Standard Vocabularies in Open Government Data. In: D. Wood (ed.) *Linking Government Data*, Springer Verlag (2011).

Simmhan, Y., Giakkoupis, M., Cao, B., Prasanna, V. K. (2010) "On Using Cloud Platforms in a Software Architecture for Smart Energy Grids," *IEEE International Conference on Cloud Computing.*

Wu, X., Zhu, X., Wu, G.-Q., and Ding, W. (2013). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering, 99*(PrePrints):1.