A Fuzzy Ontology-Driven Approach to Semantic Interoperability in e-Government Big Data

Andreiwid Sh. Corrêa, Cleverton Borba, Daniel Lins da Silva, and Pedro Corrêa

Abstract—**With the increasing production of information from e-government initiatives, there is also the need to transform a large volume of unstructured data into useful information for society. All this information should be easily accessible and made available in a meaningful and effective way in order to achieve semantic interoperability in electronic government services, which is a challenge to be pursued by governments round the world. Our aim is to discuss the context of e-Government Big Data and to present a framework to promote semantic interoperability through automatic generation of ontologies from unstructured information found in the Internet. We propose the use of fuzzy mechanisms to deal with natural language terms and present some related works found in this area. The results achieved in this study are based on the architectural definition and major components and requirements in order to compose the proposed framework. With this, it is possible to take advantage of the large volume of information generated from e-Government initiatives and use it to benefit society**

Index Terms—**E-government, semantic interoperability, fuzzy, ontology, big data.**

I. Introduction

With the emergence of the Semantic Web [1], new perspectives have surfaced regarding the significance of the information provided. Having information available only to humans is not enough: it is necessary to make it available for processing by machines.

The concept of Big Data is essentially the existence of a very large, heterogeneous and dynamic volume of data from unstructured sources, and processing typically occurs only through human intervention with the use of very specialized tools, for which traditional relational databases systems are not well suited. The government sector, through its e-Gov initiatives, for example, is the major data generator because its function is to manage public administration, which is originally complex and heterogeneous due to its decentralized structure, based on roles and responsibilities (local, state and/or federal).

The large volume of data generated by e-Gov (e-Gov Big Data) generates interoperability problems, mainly related to the semantics of information. This matter increases when there is a scenario of multi roles and responsibilities of each government institution, in which civil society should be involved in all government levels to the provision of public services through e-government. Adopting an information pattern in the e-Gov services (semantic interoperability) is considered essential.

This paper presents a framework to promote semantic interoperability in e-Gov Big Data scenario, by automatically generating ontologies from unstructured information, supported by fuzzy logic techniques.

It is organized as follows. Section II provides a brief literature review as a background. Section III explains the reasons for this study. Section IV overviews related works found in the area. Section V presents the framework proposed. Section VI discusses our proposal and conclusion. Section VII presents the challenges and future works.

II. Background

A. E-Gov Big Data

The term Big Data is a new term, although the concept is not new [2]. Currently, there are several definitions of the term; we chose the one found in [3], which suggests that Big Data is characterized by "V" words, Volume, Variety and Velocity. It can be used to investigate situations and events whereby a large amount of data is involved.

Big Data has been applied in several areas of research. One of them relates to e-Gov. Some related experiences [2], [4] are examples of e-Gov Big Data use in U.S. experience. On the one hand, from the government point of view, Big Data may increase value of the existing unstructured data and bring new information to support decision-making processes. On the other hand, from the society point of view, better decision taken by their government, lead to social gains and to the rational use of public resources.

B. E-Gov Interoperability, Semantic and Ontology

The general term interoperability means the ability of different systems and organizations to work together. In the e-Gov context, interoperability means providing better services to promote efficiency, transparency and resource saving for society [5]. E-Gov interoperability is so important that the United Nations Development Program [6]-[8] encourages governments round the world to define their interoperability architectures, known as e-Government Interoperability Frameworks - e-GIFs

Currently, countries and institutions such as the United Kingdom, Brazil, Germany, Denmark e the European Union are some examples of governments that have their e-GIF defined [9]. An e-GIF aims to define an interoperability model that a government should follow at technical, semantic and organizational levels and includes technical and non-technical issues. Especially at the semantic level, an e-GIF should impose standards that lead to meaningful information, so that it can be understood by the parties

Manuscript received October 9, 2013; revised December 10, 2013.

Andreiwid Sheffer Correa is with the Computer Engineering Department, University of Sao Paulo, Brazil (e-mail: andreiwid@usp.br).

DOI: 10.7763/IJSSH.2015.V5.448

involved [7]. Meaningful information here is limited to simply adopting a standard for message exchange - such as XML - found in most existing e-GIFs [10].

The term semantic gained a new meaning with the advent of the Semantic Web, an idea proposed by Tim Berners-Lee, James Hendler and Ora Lassila [1]. The authors define the Semantic Web as "a web of data that can be processed directly and indirectly by machines". Thus, the semantic meaning goes beyond the language adopted to give way to ontologies.

Ontology is defined as an explicit specification of a conceptualization, which is a simplified view of a knowledge domain. There are several languages for formally representing ontologies, the main ones are OWL and RDF [11].

C. Fuzzy Reasoning

Fuzzy logic was proposed by Lotfi Zadeh in his Fuzzy Sets Theory [12]. Fuzzy logic contrasts with the bivalent logic to be multivalued, allowing proposition granularity. According to the fuzzy logic definition, something may be partially true or partially false. Bivalent logic is based on False and True propositions (0, 1); multivalued allows any value ranging from False to True (0...1). One of its core applications allows dealing with uncertainty. Uncertainty is an abstract concept that encompasses, among other things, vagueness and imprecision. Uncertainty is a key issue when working on decision-making and control processes.

According to [13], imprecision means the use of natural language data that are routinely used by humans. Vagueness means the existence of incomplete data, inconsistent or even non-existing.

In the context of e-Gov, for example, we could define the class Expenses as something that is related to government spending on its agencies staff members or acquisition of services from third-parties. If our working context addresses agencies staff expenditure, we could say that the meaning of "Expenses" has a 1.0 degree of confidence in the Expenses class. In the same way, if our context maps agencies spending on outsourcing, we could infer that the meaning of "Expenses" has a 0.75 degree.

Fuzzy reasoning has been widely used for a multitude of applications and there are several studies in this area, especially [14]-[16], in which Natural Language Processing (NLP) is involved.

III. Motivation

A survey [10] incorporated in the discussions of the OASIS Transformational Government Framework Committee pointed out that 53% of e-GIFs consider XML a standard for semantic interoperability. Although XML is the de facto standard in the industry, its use alone does not guarantee the semantics of the information exchanged. It is necessary to consolidate a formal ontology model into the desired context.

In this sense, it is necessary for public sectors to take advantage of ontology models with meaningful concepts to promote open government data and thereby enable the inclusion of the civil society in the public administration so that society feels like being a part of the government.

We argue that open government data can only be built through a new perspective whereby data, meaning and knowledge are present. Some other reasons to do so are:

- Need of transparency from the society point of view. Society must be aware of how and where their money is spent.
- Need of one-stop data portal where society can make inquiries and cross data according to their needs.
- A way to make good use of distributed legacy data and make them available through new technologies that allow interoperability.
- Follow good industry practices and mainly a worldwide effort to define government interoperability architectures (United Nations example).

IV. Related Works

The automatic generation of ontologies has been proposed in the literature for some years. Generation from free text and schemas are the most common [17]-[19].

Taking fuzzy reasoning approach as a mechanism to address the uncertainty in ontologies generation, we found papers [20]-[24] as examples of research. In [24], the authors propose FOGA (Fuzzy Ontology Generation Framework), which automatically generates a fuzzy ontology from uncertainty data based on Formal Concept Analysis. The authors' proposition is based on a new technique, such that linguist variables and linguistic terms are no longer necessary.

We take into account all related works described here to accomplish the proposition of this work.

V. A Fuzzy-ontology Framework

We propose a framework in which heterogeneous human-readable documents and information will be considered as data sources. The aim is to make them all meaningful and machine-readable data, based on the literature review exposed here.

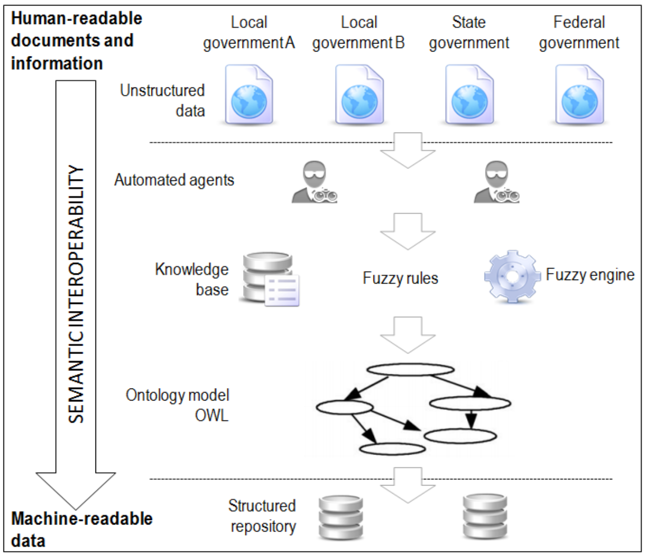


Fig. 1. An overview of the framework proposed.

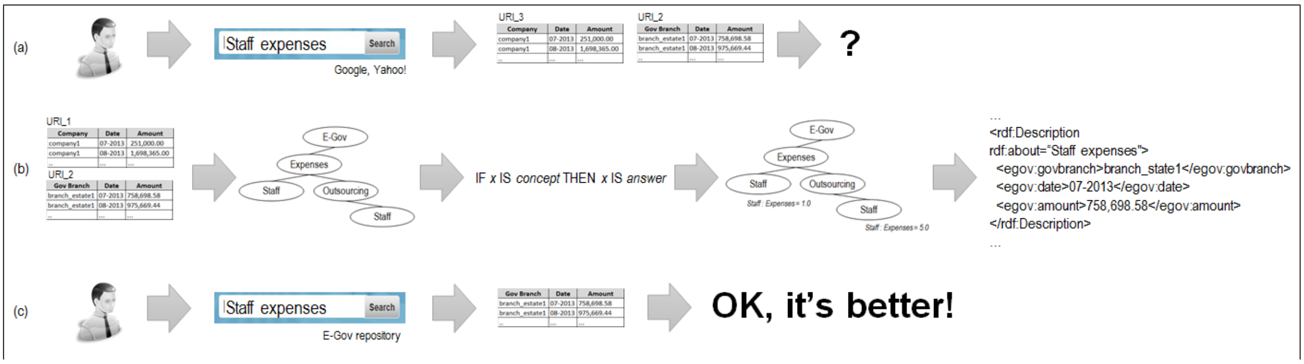
Fig. 1 depicts a conceptual overview proposed framework with five main components: Unstructured data source, Automated agents, Fuzzy engine and knowledge-base, Ontology model and Structured repository.

Some details of each component are provided as follows:

- **Unstructured data source** - all the data disclosed by governments on the Internet, such as HTML tags, PDF documents or flat text files. They are usually human-readable documents and information that depends on human beings to be interpreted and understood.
- **Automated agents** - comprises mechanisms that are responsible for discovering and for capturing the entire content in unstructured data sources. The idea is to have autonomous agents that look into the network sources in order to collect all the data.
- **Fuzzy engine and knowledge base** - involves all the mechanisms responsible for fuzzy processing, with its knowledge base. The knowledge base contains linguistic variables, terms and rules that will make the translation into NLP ontology conceptions.
- **Ontology model** - the ontology model generated from fuzzy rules processing to unstructured data content. At this stage annotations to data with semantic meaning with use of ontology languages (OWL/RDF).
- **Structured repository**-it is the database with machine-readable data. We intend to provide a repository to allow later access. The systems should control information update on a time basis.

To briefly exemplify the working process of our proposal, we introduce a hypothetical situation depicted in Fig. 2 where a user wants to have further information on staff expenses in his/her city (local government).

- In this situation, the user types key words in a search system (Google, Yahoo!) and possibly gets a huge amount of results. Of course, the user needs to click and read each returned page. Even after some selections, users may be in doubt about two results that match his/her search string, because "Staff expenses" may refer to an agency staff salaries or outsourcing expenses. Note that both situations match staff expenses. This situation is a typical search scenario where a human plays the essential role of deciding on what he/she wants to know about.
- With our proposed framework, automated agents would do prior search and discovery. The target content are documents and information freely available in the Internet. In this regard, agents should be smart enough to follow a network path as a starting point. From the pages returned, the ontology meta model will be applied, which was previously built according to ontology classes regarding e-Gov domain. Then, from knowledge base, the system will retry all the fuzzy rules defined in the context (e-Gov meta model with fuzzy annotations, such as FuzzyOWL). Fuzzy engine calculates membership degree (ranging from 0 to 1) for each ontology class and applies a label with its degree. At this step, we will use a known algorithm for matching semantic words from search string to fuzzy linguistic variables and terms. Finally, the system is ready to generate OWL and RDF descriptors and make them machine-readable data. We intend to cache descriptors in the repository component and make them available without direct access to the source of information.
- In this scenario, users would have access to what we called e-Gov repository, where cached OWL/RDF descriptors are stored. After typing users' search, the system will yield more effective results, because information has already processed according to its ontology and semantic meaning. In Fig. 2 (c) illustration, the user gets what he/she wants: "Staff expenses" corresponds (more precisely, in this case) to agency staff salaries.



Company	Date	Amount
company1	07-2013	251,000.00
company1	08-2013	1,698,365.00
...	...	...

Gov Branch	Date	Amount
branch_estate1	07-2013	758,698.58
branch_estate1	08-2013	975,669.44
...	...	...

```
...
<rdf:Description
rdf:about="Staff expenses">
  <egov:govbranch>branch_state1</egov:govbranch>
  <egov:date>07-2013</egov:date>
  <egov:amount>758,698.58</egov:amount>
</rdf:Description>
...
```

Fig. 2. (a) A typical search situation, (b) proposed framework working process and (c) search in e-Gov repository.

VI. Discussion and Conclusion

We present a conceptual framework to promote semantic interoperability in e-Gov Big Data scenario, through the automatic generation of ontologies from unstructured information.

To deal with uncertainty issues related to human-readable information (free texts, PDF documents, HTML, NLP and so forth), we rely on fuzzy logic techniques to process the meaning of ontology against search strings. As we have seen, there are some good experiences in the literature in favor of processing with fuzzy ontologies.

Our framework aims to take advantage of the large volume of information generated from e-Gov initiatives and use it to benefit society. Legacy data and system can be put to good use. Finally, with our approach, governments can offer better and effective services to society in the e-Gov context.

VII. Challenges and Future Work

This is an ongoing research. We point out some challenges involved in the continuation of this work:

- Defining a meta model ontology to characterize the domain knowledge related to e-Gov.
- Developing and implementing a prototype application that implements fuzzy reasoning to automatically generate ontologies.
- Developing and implementing a prototype of automated agents responsible for sweeping government data on the Internet, according to the knowledge domain defined.
- Comparing the results of automatic generation of ontologies from classical existing tools on the market that do not rely on fuzzy reasoning.

References

[1] T. B. Lee, J. Hendler, and O. Lassila, "The Semantic Web: Scientific American," *Scientific American*, no. 5, May 1, 2001.

[2] J. C. Bertot and H. Choi, "Big data and e-government: issues, policies, and recommendations," in *Proc. the 14th Annual International Conference on Digital Government Research*, New York, NY, USA, 2013, pp. 1-10.

[3] D. E. O'Leary, "Artificial intelligence and big data," *IEEE Intell. Syst*, vol. 28, no. 2, pp. 96-99, 2013.

[4] R. Joseph and N. Johnson, "The good, the bad, and the big: big data and t-government," *IT Prof.*, 2013.

[5] L. Guijarro, "Semantic interoperability in e-Government initiatives," *Comput. Stand. Interfaces*, vol. 31, no. 1, pp. 174-180, Jan. 2009.

[6] E. Lallana, *E-Government Interoperability: A Review of Government Interoperability Frameworks in Selected Countries*. Bangkok, Thailand: United Nations Development Programme, 2007.

[7] E. Lallana, *E-Government Interoperability: Guide*, Bangkok, Thailand: United Nations Development Programme, 2007.

[8] E. Lallana, *E-Government Interoperability: Overview*, Bangkok, Thailand: United Nations Development Programme, 2007.

[9] D. Ray, U. Gulla, S. S. Dash, and M. P. Gupta, "A critical survey of selected government interoperability frameworks," *Transform Gov. People Process Policy*, vol. 5, no. 2, pp. 114-142, 2011.

[10] C. Transform, "E-government interoperability-A comparative analysis of 30 countries," Feb. 2011.

[11] G. Antoniou and G. Antoniou, *A Semantic Web primer*, Cambridge, MA: The MIT Press, 2012.

[12] L. A. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, pp. 338-353, 1965.

[13] M. Negnevitsky, *Artificial intelligence: a guide to intelligent systems*. Harlow, England; New York: Addison-Wesley, 2005.

[14] L. A. Zadeh, "Fuzzy logic = computing with words," *IEEE Trans. Fuzzy Syst.*, vol. 4, no. 2, pp. 103-111, 1996.

[15] L. A. Zadeh, "From computing with numbers to computing with words from manipulation of measurements to manipulation of perceptions," *Appl Math Comput Sci*, vol. 12, no. 3, pp. 307-324, 2002.

[16] L. A. Zadeh, "Computing with words (CW) and its application to decision support and systems analysis," in *Proc. 2003 IEEE International Symposium on Intelligent Signal Processing*, 2003, pp. 1-2.

[17] N. Yahia, S. A. Mokhtar, and A. Ahmed, "Automatic Generation of OWL Ontology from XML Data Source," *Int. J. Comput. Sci. Issues IJCSI*, vol. 9, no. 2, pp. 77-83, Mar. 2012.

[18] I. Bedini and B. Nguyen, "Automatic Ontology Generation: State of the Art," University of Versailles, Technical report, Dec. 2007.

[19] N. Arch-int and S. Arch-int, "Semantic Ontology Mapping for Interoperability of Learning Resource Systems using a rule-based reasoning approach," *Expert Syst. Appl.*, vol. 40, no. 18, pp. 7428-7443, Dec. 2013.

[20] U. Straccia, "A fuzzy description logic for the semantic web," in *Fuzzy Logic and the Semantic Web, Capturing Intelligence*, ch. 4, 2005, pp. 167-181.

[21] G. Stoilos, N. Simou, G. Stamou, and S. Kollias, "Uncertainty and the semantic web," *IEEE Intell. Syst.*, vol. 21, no. 5, pp. 84-87, 2006.

[22] M. Nagy, E. Motta, and M. V. Vera, "Multi-agent ontology mapping with uncertainty on the semantic web," in *Proc. 2007 IEEE International Conference on Intelligent Computer Communication and Processing*, 2007, pp. 49-56.

[23] S. Calegari and D. Ciucci, "Fuzzy ontology, fuzzy description logics and fuzzy-owl," in *Applications of Fuzzy Sets Theory*, F. Masulli, S. Mitra, and G. Pasi, Eds. Springer Berlin Heidelberg, 2007, pp. 118-126.

[24] Q. T. Tho, S. C. Hui, A. C. M. Fong, and T. H. Cao, "Automatic fuzzy ontology generation for semantic Web," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 6, pp. 842-856, 2006.

Andreiwid Sh. Corrêa was born in Brazil, 1981. He has a degree in computer science (2004), MSc in electrical engineering (2011) and is currently a Ph.D. candidate in computer engineering at the University of Sao Paulo, Brazil. His research interests include e-government, open data, government transparency and fuzzy knowledge-based systems. He is a professor of computer science at the Federal Institute of Technology and works as a system development coordinator at the information technology division of the Campinas municipality, Brazil.

Cleverton Borba was born in Brazil. He has a degree in computer science (2005) and MSc in the same area (2010). He is currently a Ph.D. candidate in computer engineering at the University of Sao Paulo, Brazil. His research interests include distributed database systems, biodiversity and ecological informatics and artificial intelligence. He is a full professor at Centro Universitário Adventista de São Paulo, Brazil

Daniel Lins da Silva was born in Brazil. He has a degree in computer science (2007). He holds a MSc (2011) and is a Ph.D. candidate in computer engineering at the University of Sao Paulo, Brazil. His research interests include distributed systems, traceability systems and big data. He is a full time researcher at the University of Sao Paulo, Brazil. He coordinates projects related to biodiversity systems.

Pedro Corrêa was born in Brazil. He holds a BSc in computer science (1987), MSc in computer science (1992) and a Ph.D in electrical engineering (2002), all from the University of Sao Paulo. He is currently a professor at the Computer Engineering Department at the University of Sao Paulo, Brazil. He was a consultant for the United Nations Development Programme-UNDP. Project BRA/97/001-developed at the Treasury Department of the State of São Paulo for the modernization project of tax administration in the State–PROMOCAT. He has two books published, 12 periodic publications with editorial policy setting and over 40 works published in congresses.