Supporting Competitive Intelligence with Linked Enterprise Data

Vitor Afonso Pinto[1], Guilherme Sousa Bastos[2], Fabricio Ziviani[1] and Fernando Silva Parreiras[1]
[1]*LAIS – Laboratory of Advanced Information Systems, FUMEC University, Belo Horizonte, MG, Brazil*
[2]*IESTI – Institute of Systems Engineering and Information Technology, UNIFEI, Itajuba, MG, Brazil*

Keywords: Semantic Web, Linked Data, Linked Enterprise Data, Competitive Intelligence.

Abstract: Competitive Intelligence is a process which involves retrieving, analyzing and packaging information to offer a final product that responds to the intelligence needs of a particular decision maker or community of decision makers. Internet-based information sources are becoming increasingly important in this process because most of the contents available on the Web are available free of charge. In this work the following research question was addressed: What are the concepts and technologies related to linked data which allow gathering, integration and sharing of information to support competitive intelligence? To answer this question, firstly, the literature was reviewed in order to outline the conceptual framework. Next, some competency questions were defined through a focus group in a study object. Finally, DB4Trading tool was built as a prototype able to validate the conceptual framework. Results point out that adoption of Semantic Web technologies enable to obtain the data needed for the analysis of external environments. Besides that, results indicate that companies use Semantic Web technologies to support its operations despite consider these technologies as complex. This work adds to the decision-making process, specially in the context of competitive intelligence. This work also contributes to reducing costs to obtain information beyond organization boundaries by using Semantic Web technologies.

1 INTRODUCTION

Organizations today seek to thrive in turbulent times. As environments become increasingly volatile, organizations are turning their gaze to the external horizon, watching and grappling with a confusion of signals, messages, and cues. Sensing and making sense of the environment have become the sine qua non for organizational growth and survival. (Bouthillier and Shearer, 2003). However, a great deal of data is available through direct and open channels. These channels include: 1) Published sources; 2) Governmental data; 3) Other public documents. (Gelb et al., 1991).

Semantic Web works to create an environment where people work together sharing knowledge and having tools for information management and analysis. (PAN, 2009). Linked Data initiative was proposed to remove the barriers to data access and sharing and also to enable data from different sources to be connected and queried. (Hu and Svensson, 2010). All these characteristics, if used in a corporate environment, could facilitate competitive intelligence.

1.1 Competitive Intelligence

In order to obtain the sufficient knowledge to understand both the internal and external environment as a whole, organizations perform coordinated actions to seek, treat, distribute and protect information. (Tarapanoff, 2006). The set of activities performed by organizations to gather information about competitors, products and markets, is called **Competitive Intelligence (CI)**. (Moresi, 2006). It is also seen as seeking any information which improves the organization positioning. (Tarapanoff, 2006). Intelligence gathering goes on everyday, without necessarily being called by its rightful name. More specifically, it implies legal research efforts by business studying their competitor's products, organizations and related matters. It is defined as the use of public sources to develop information on competition, competitors, and the market environment, including economic, regulatory, political, demographic influences, etc. (Cronin et al., 1994).

The importance of such external data will not be fully demonstrated if they are not combined with internal enterprise data and consumed in realtime busi-

Afonso Pinto V., Sousa Bastos G., Ziviani F. and Silva Parreiras F..
Supporting Competitive Intelligence with Linked Enterprise Data.
DOI: 10.5220/0005472604090415
In *Proceedings of the 17th International Conference on Enterprise Information Systems* (ICEIS-2015), pages 409-415
ISBN: 978-989-758-096-3

ness decision making. (Hu and Svensson, 2010). But, specially in large organizations, enterprise data is available through a highly complex enterprise-wide IT system, as several hundreds of interconnected systems may be employed. (Lindström et al., 2006). And the size of each single system may vary extensively from enterprise resource planning systems to custom-made niche products, making the system interconnections numerous and heterogeneous.

This way, in order to perform Competitive Intelligence, companies need to integrate data from internal systems (which are not always integrated) and also data generated outside its boundaries (which are not always structured or standardized). In this context, Semantic Web is presented as a solution for increase Competitive Intelligence through interlinking heterogeneous data sources.

1.2 Semantic Web

Semantic Web is a vision: the idea of having data on the web defined and linked in a way that it can be used by machines - not just for display purposes, but for using it in various applications. (Xu et al., 2004). It is a web that includes documents, or portions of documents, describing explicit relationships between things and containing semantic information intended for automated processing by machines. (Li, 2002). It also carries the promise to make web machine-understandable by enriching available information with logic-based semantics and provide us with a new paradigm for knowledge interchange and sharing. (Chen and Wu, 2003).

The fundamental concept of the Semantic Web is to make the information, available on the web, more meaningful by making it accessible to automated tools that can augment our experience. (Niles and Jeremijenko, 2001). It provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. (Chen et al., 2012). It also refers to a range of standards, modeling languages and tool development initiatives aimed at annotating Web pages with well-defined metadata, so that the semantics associated with web contexts can be effectively interpreted. (Abrahams et al., 2004).

1.2.1 Linked Data

Linked Data is a term describing a set of best practices to facilitate the publishing, accessing and interlinking of the data of the Semantic Web. Thus, Linked Data is an open framework for the loose integration of data in the Internet, where data sources can easily cross-link. (Feridun and Tanner, 2010). It is the data exposed, shared, and connected via URIs on the Web. It uses URLs to identify things as resources to facilitate people to dereference them. It also provides useful information about these resources, as well as links to other related resources which may improve information discovery. (Mi et al., 2009).

A pragmatic vision of the Semantic Web has emerged via the **Linking Open Data** project (LOD), focusing on translating various datasets available on the Web into RDF and interlinking them, following the Linked Data principles. Lots of different datasets have been provided via this LOD initiative, such as DBpedia (the RDF export of Wikipedia) or Geonames (a large geolocation database). All together, they form a complete Webscaled graph of interlinked knowledge, commonly known as the Linked Open Data Cloud.

1.2.2 Linked Enterprise Data

Another goal of Semantic Web is to apply its technologies for Enterprise Information Integration, creating a Corporate Semantic Web. (Eisenberg and Kanza, 2011). Today's business is based on huge amount of information and extracting right information at right time is a difficult and tedious task. By applying semantics within structured (ERP, Billing, Financial, HR systems) and unstructured data (email, fax, office documents) we can take business decision on the basis of overall organization knowledge base. (Khan and Hussain, 2009).

Linked Enterprise Data (also called as "Linked Data Enterprise") is an organization in which the act of information creation is intimately coupled with the act of information sharing. In a linked data enterprise, individuals and groups continue to produce and consume information in ways that are specific to their own business needs, but they produce it in a way that can be connected to other aspects of the enterprise. (Allemang, 2010).

This study intends to answer the following research question: What are the concepts and technologies related to linked data which allow gathering, integration and sharing of information to support competitive intelligence? The main objective of this study is to analyze the application of linked data technologies into gathering and integrating data generated both inside and outside an organization for supporting the competitive intelligence process, in the context of portfolio management.

To achieve this goal, initially a systematic literature review was performed aiming to propose and outline a conceptual framework to support this study. Next, a focus group was conducted intending to identify competency questions, that is, variables and data

sources that could be used to build and validate the proposed framework. Next, a prototype for integrated data visualization was built, based on proposed framework. Finally, the conceptual framework was tested and validated, using the competency questions.

2 CONTRIBUTION

In order to address its objectives, our research was divided in three major phases:

- Phase 1 - Outline the conceptual framework through a Systematic Literature Review
- Phase 2 - Identify competency questions through a Focus Group survey
- Phase 3 - Build, test and validate a prototype for integrated data visualization

In this section, methods used to produce the results of this study are described.

2.1 Phase 1 - Systematic Literature Review

In order to understand the applications of linked data for corporate environments, we performed a comprehensive literature review on the state-of-the-art in the research field of Linked Enterprise Data. A Systematic Literature Review (SLR) is a means of evaluating and interpreting available research relevant to a particular research question, topic area, or phenomenon of interest. (Kitchenham, 2004). This kind of study comprises three consecutive steps: planning, execution and results.

The main purpose of the **planning** step is to deliver a protocol which drives the literature review efforts. After analyzing the existing literature related to Semantic Web and its application on corporate environments, the research questions addressed by the systematic literature review were defined. Next, the strategy used to search for primary studies, including search terms and resources to be searched was defined. Finally, the exclusion criteria and the strategy to extract data from selected papers were defined.

In the **execution** step, activities to extract and synthesize data from papers were performed. First, the primary studies were selected according to the defined data extraction strategy and exclusion criteria. After selecting primary studies, the information needed to address the questions of this review were collected, that is, primary studies were examined. Selected papers were analysed considering the information required by each research question. Finally, the **results** of the included primary studies were collated and summarized.

2.2 Phase 2 - Competency Questions

In order to identify competency questions, that is, variables and data sources that could be used to build and validate the proposed framework, we applied a focus group methodology, which is understood as a way of collecting qualitative data, and involves engaging a small number of people in informal group discussions, 'focused' around a particular topic or set of issues. (Hyland et al., 2014).

In the **design** step, time and location were defined. Also, participants were selected based on their experience with competitive intelligence. Basically, three types of professionals were selected: strategy specialists, engineers and health, safety & sustainability professionals. We also generated the focus group script and other forms to support the overall process. Finally, the questions used in the focus group were carefully chosen aiming at achieving the general purpose of focus group.

In the **execution** step, participants were informed of the study and received a written information form, including information on the aim of the study, anonymity issues, and a field for signing informed consent. A topic guide was then followed, in order to evoke the research themes during the discussion. The focus group session was audiotaped and a member of the research team took detailed notes of the discussions.

Finally, the notes taken during the focus group discussions were used for the analysis of the collected data, generating the **results**. Whenever clarifications of these notes were needed, the audiotapes were consulted.

Participants mentioned the existence of specific and generic datasources. 'Specific Datasources' may vary depending on the country, business or competitor being analyzed. 'Generic Datasources' can be used in multiple analysis, independently of country, business or competitor. Participants also separated datasources applicable to multiple segments from those applicable uniquely to their segment. Table 1 presents a list of the identified datasets.

2.3 Phase 3 - Prototype for Integrated Data Visualization

In order to test if semantic web technologies could be used to support competitive intelligence a prototype for integrated data visualization was built. The prototype was implemented following the framework

Table 1: Generic Datasources.

DATASOURCE	WEBSITE ADDRESS
GOOGLE	https://www.google.com/
DOING BUSINESS PROJECT	http://portugues.doingbusiness.org/
UNITED NATIONS	http://www.un.org/
WORLD HEALTH ORGANIZATION	http://www.who.int/en/
WORLD TRADE ORGANISATION	http://www.wto.org/
CENTRAL INTELLIGENCE AGENCY	https://www.cia.gov/index.html
DELLOITE	http://www.deloitte.com/
LONDON METAL EXCHANGE	http://www.lme.com/
INTERNATIONAL MONETARY FUND	http://www.imf.org/external/index.htm
EDGARD	https://www.sec.gov/edgar/aboutedgar.htm
SEDAR	http://www.sedar.com/homepage_en.htm
REUTERS	http://br.reuters.com/
ABNT	http://www.abnt.org.br/
CRU	http://www.crugroup.com/
SNL	http://www.snl.com/
BROOKHUNT	http://www.brookhunt.com
SCIENCEDIRECT	http://www.sciencedirect.com/
SCOPUS	http://www.scopus.com/home.url
U.S. CENSUS BUREAU	http://www.census.gov/
U.S. GEOLOGICAL SURVEY	http://www.usgs.gov/
STEEL BUSINESS BRIEFING	https://www.steelbb.com/pt/?PageID=1
NASA	http://www.nasa.gov/
ESRI	http://www.esri.com/
GREENPEACE	http://www.greenpeace.org/brasil/pt/
FRASER INSTITUTE	https://www.fraserinstitute.org/
WIKIPEDIA	http://www.wikipedia.org/

identified in the literature review and had four layers: data, wrapper, integration and presentation.

The **data layer** was implemented in MS-SQL Server 2012. This implementation enabled creation of the dataset which was used later to validate competency questions. Dataset was populated with information gathered from various sources, as identified on focus group. From all information identified during focus group, only those available on web in open data format were selected to compound this dataset.

In this study, both the **wrapper** and the **integration** layers were implemented through D2RQ Platform, which is a system for accessing relational databases as virtual, read-only RDF graphs. D2RQ Platform has D2R Server, which is a tool for publishing the content of relational databases on the Semantic Web. (Cyganiak, 2012). After installing and configuring D2RQ Platform, the dataset presented in Data Layer section was made available through SPARQL endpoint.

Finally, the **presentation layer** was implemented. A web application called **DB4Trading** was built so that users could validate data from Semantic Repository using their own criteria. From the weight defined for each one of the variables in the categories area, the application was designed to identify countries whose information were more adherent. In order to identify countries, we used a specific Google API capable to highlight the country polygon. DB4Trading was designed to calculate the color of countries, marking in red those classified as more relevant and leaving in blank those classified as less relevant.

3 RESULTS

Initially, we identified that enterprises are using Semantic Web technologies for support their operations. However, according to studies, enterprises still believe that Semantic Web technologies are complex and require high specialized teams. Studies also pointed out that there is a pattern regarding the frameworks used for implementing Semantic Web in enterprises. This common framework enables interlinking of both internal and external data sources and was used as the conceptual framework.

Based on focus group, we also identified sixty-seven variables to support competitive intelligence,

from which twelve are generated internally and fifty-five collected from external databases. We created categories to organize variables based on Nettleton. (Nettleton, 2014). From all information identified during focus group, only those available on web in open data format were selected to compound the prototype's dataset. This categorization was performed following guidelines proposed by Berners-Lee. (Berners-Lee, 2009).

To validate the conceptual framework we decided to run specific queries against dataset and also create specific scenarios in visualization application. First part of the validation aims to show that the dataset was made available in RDF and was accessible via SPARQL queries. Second part of the validation aims to show that the proposed framework was able to answer business needs by integrating and presenting information from multiple data sources.

3.1 Semantic Repository

SPARQL queries were performed against dataset to validate the framework. The competency questions presented next were considered. Figure 1 presents the SPARQL queries results.

1. What are the countries with the highest human development index (HDI)?
2. What are the countries with the highest numbers of hospital beds?
3. What are the countries with the highest export trade volume?
4. What are the countries with the lowest tax burdens?
5. What are the countries with the lowest time for starting a business?

3.2 Visualization Application

In order to address requirements of competitive intelligence professionals who attended the focus group, we created three scenarios according to the focus group professionals specialties:

1. Scenario 1 - Engineering perspective, considering infrastructure variables as more relevant.
2. Scenario 2 - Strategic Management perspective, considering projection variables as more relevant.
3. Scenario 3 - Health and Environment perspective, considering human variables as more relevant.

Figure 2 presents results obtained for the first scenario.

4 DISCUSSION

It is hoped that this research may contribute to the clarification of concepts and technologies related to linked data, facilitating dissemination and usage in other fields, including in corporate environments. Also, the results of this research can contribute for reducing costs to obtain information beyond organization boundaries, since currently, a major effort is carried out in order to index and normalize information gathered from multiple data sources. Finally, governments and other institutions may also be encouraged to release their information using open data patterns in order to increase the percentual of information shared globally.

5 RELATED WORK

Chen, Chau e Zeng introduce the Competitive Intelligence Spider, a tool responsible for performing real-time collection of Web pages from sites specified by the user and applying indexing and categorization analysis on the documents collected. (Chen et al., 2002). These authors concluded that there exists strong evidence in support of the potentially significant value of applying the CI Spider approach in CI applications.

Yang et al evaluate 14 existent text mining and visualization tools. (Yang et al., 2008). In this study, each tool is discussed in some detail and their strengths and potential limitations are identified from a patent analysis perspective. Authors observed that tools able to deal with unstructured data are the most flexible due to their capabilities to process a broad range of data sources and use of advanced semantic technologies.

Ferrara et al provide a structured and comprehensive overview of the research in Web Data Extraction field and also provide an overview of most recent results in the literature. (Ferrara et al., 2012). In this study, authors recognize that Web Data Extraction is an important problem that had been studied by means of different scientific tools and in a broad range of applications. According to the study, the linkage of datasets coming from independent Web platforms fuels novel scientific applications.

The work of Hu and Svensson is directly related to this study because that work aims to integrate external and internal data sets to provide strategical information. (Hu and Svensson, 2010). Their work also approaches the existence of connection to multiple data sources, the possibility to perform queries considering data from multiple data sources.

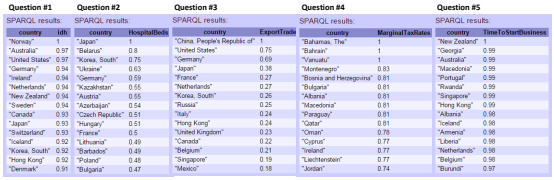

Figure 1: SPARQL Queries Results.

Figure 2: DB4Trading Scenario 1 - Results.

6 CONCLUSION

In this study we analyzed the application of linked data technologies into gathering and integrating data generated both inside and outside an organization for supporting the competitive intelligence process. Firstly, we outlined a conceptual framework. Next, we identified competency questions, that is, variables and data sources that were used to build and validate the framework. Next, we built a prototype for integrated data visualization, based on proposed framework. Finally, we tested and validated conceptual framework, using competency questions.

Based on the results of this research, we concluded that it is possible to use Semantic Web technologies to gather and distribute information for external environment analysis. Based on systematic literature review, we concluded that enterprises are already using Semantic Web technologies for support their operations, specially for internal data sources integration. However, enterprises still believe that Semantic Web technologies are complex and require high specialized teams.

6.1 Limitations

Regarding the **Systematic Literature Review**, we can point out the following limitations. First, only papers written in English had been considered. Additionally, during data extraction stage, it was necessary to interpret the subjective information provided by studies. Another potential threat to validity is the natural limitation of search engines. A **Focus Group** also has several limitations. Firstly, focus groups are susceptible to facilitator bias. Secondly, the discussions can be sidetracked or dominated by a few vocal individuals. Finally, information generated in focus group often has limited generalization to a whole population.

Regarding the **Semantic Repository**, only variables available in open data format were selected and related to entities. Also, no crawler was built to update the data after the information was inserted into the database. Regarding the **Visualization Application** we point out the following limitations. Although the site is dynamically constructed from the data files, no mechanism was generated to automate the export of data from the database to the files. Another limitation is related to countries polygon: for unknown reasons, the Google API was not able to create the polygon for the following countries: Tanzania, Senegal, Chile, French Guiana and North Korea.

6.2 Future Work

Although this study is limited to the examination of country data, the Semantic Repository was designed to allow the creation of classes to represent regions, cities, neighborhood, and also not geo-referenced entities. Further studies could consider the usage of the Semantic Repository generated in this study to create entities and variables applicable to other areas.

The Visualization Application (DB4Trading) generated in this research could incorporate some im-

provements: The tool should allow users to assign scores to each of the data sources because there is a consensus between participants that credibility may vary according to the data source. The tool also could save different scenarios's settings, allowing users to retrieve their preferences.

ACKNOWLEDGEMENTS

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

REFERENCES

Abrahams, B., McGrath, G. M., and Dai, W. (2004). A multi-agent approach for dynamic ontology loading to support semantic web applications. In Zhang, D., Grgoire, r., and DeGroot, D., editors, *IRI*, pages 570–575. IEEE Systems, Man, and Cybernetics Society.

Allemang, D. (2010). Semantic web and the linked data enterprise. In Wood, D., editor, *Linking Enterprise Data*, pages 3–23. Springer US.

Berners-Lee, T. (2009). Linked data. http://www.w3.org/DesignIssues/LinkedData.html. Accessed: 2014-03-10.

Bouthillier, F. and Shearer, K. (2003). *Assessing Competitive Intelligence Software: A Guide to Evaluating CI Technology*. Information Today.

Chen, H., Chau, M., and Zeng, D. (2002). {CI} spider: a tool for competitive intelligence on the web. *Decision Support Systems*, 34(1):1 – 17.

Chen, H. and Wu, Z. (2003). On case-based knowledge sharing in semantic web. In *ICTAI*, pages 200–207. IEEE Computer Society.

Chen, T., Zhang, Y., Zhang, S., Chen, C., and Chen, H. (2012). Building semantic information search platform with extended sesame framework. In *Proceedings of the 8th International Conference on Semantic Systems*, I-SEMANTICS '12, pages 193–196, New York, NY, USA. ACM.

Cronin, B., Overfelt, K., Fouchereaux, K., Manzvanzvike, T., Cha, M., and Sona, E. (1994). The internet and competitive intelligence: A survey of current practice. *International Journal of Information Management*, 14(3):204 – 222.

Cyganiak, R. (2012). Accessing relational databases as virtual rdf graphs. http://www.d2rq.org. Accessed: 2014-10-01.

Eisenberg, V. and Kanza, Y. (2011). Ruby on semantic web. In Abiteboul, S., Bhm, K., Koch, C., and Tan, K.-L., editors, *ICDE*, pages 1324–1327. IEEE Computer Society.

Feridun, M. and Tanner, A. (2010). Using linked data for systems management. In *NOMS*, pages 926–929. IEEE.

Ferrara, E., Meo, P. D., Fiumara, G., and Baumgartner, R. (2012). Web data extraction, applications and techniques: A survey. *CoRR*, abs/1207.0246.

Gelb, B. D., Saxton, M. J., Zinkhan, G. M., and Albers, N. D. (1991). Competitive intelligence: Insights from executives. *Business Horizons*, 34(1):43 – 47.

Hu, B. and Svensson, G. (2010). A case study of linked enterprise data. In Patel-Schneider, P. F., Pan, Y., Hitzler, P., Mika, P., 0007, L. Z., Pan, J. Z., Horrocks, I., and Glimm, B., editors, *International Semantic Web Conference (2)*, volume 6497 of *Lecture Notes in Computer Science*, pages 129–144. Springer.

Hyland, S., Haugen, A. S., and yvind Thomassen (2014). Perceptions of time spent on safety tasks in surgical operations: A focus group study. *Safety Science*, 70(0):70 – 79.

Khan, A. and Hussain, M. (2009). Application of semantic web in e-business and telecommunication. In *Advanced Computer Control, 2009. ICACC '09. International Conference on*, pages 513–517.

Kitchenham, B. A. (2004). Procedures for undertaking systematic reviews. http://www.inf.ufsc.br/ awangenh/kitchenham.pdf. Accessed: 2013-09-01.

Li, W. (2002). Intelligent information agent with ontology on the semantic web. In *Intelligent Control and Automation, 2002. Proceedings of the 4th World Congress on*, volume 2, pages 1501–1504 vol.2.

Lindström, A., Johnson, P., Johansson, E., Ekstedt, M., and Simonsson, M. (2006). A survey on cio concerns-do enterprise architecture frameworks support them? *Information Systems Frontiers*, 8(2):81–90.

Mi, J., Chen, H., Lu, B., Yu, T., and Pan, G. (2009). Deriving similarity graphs from open linked data on semantic web. In *IRI*, pages 157–162. IEEE Systems, Man, and Cybernetics Society.

Moresi, E. (2006). Memoria Organizacional e Gestao do Conhecimento. In Tarapanoff, K., editor, *Inteligencia, Informacao e Conhecimento*, pages 277–301. IBICT.

Nettleton, D. (2014). Chapter 3 - incorporating various sources of data and information. In Nettleton, D., editor, *Commercial Data Mining*, pages 17 – 47. Morgan Kaufmann, Boston.

Niles, C. and Jeremijenko, N. (2001). Collated path: A one-dimensional interface element to promote user orientation and sense-making activities in the semantic web. In *IV*, pages 555–562. IEEE Computer Society.

PAN, J. (2009). Resource description framework. In Staab, S. and Studer, R., editors, *Handbook on Ontologies*. Springer, second edition.

Passant, A., Laublet, P., Breslin, J. G., and Decker, S. (2009). Semslates: Improving enterprise 2.0 information systems using semantic web technologies. In *CollaborateCom*, pages 1–10. IEEE.

Tarapanoff, K. (2006). Informacao, Conhecimento e Inteligencia em Corporacoes: relacoes e complementaridade. In Tarapanoff, K., editor, *Inteligencia, Informacao e Conhecimento*, pages 19–36. IBICT.

Xu, J., Zhu, Q., Li, J.-Z., Zhang, P., and Wang, K. (2004). Modeling and implementation of unified semantic web platform. In *Web Intelligence*, pages 603–606. IEEE Computer Society.

Yang, Y., Akers, L., Klose, T., and Barcelonyang, C. (2008). Text mining and visualization tools Impressions of emerging capabilities. *World Patent Information*, 30(4):280–293.