# Using the Web of Data in Competitive Intelligence Process

Leandro Dal Pizzol[1], José Leomar Todesco[1], Bernardo P. R. Todesco[1]

[1] Universidade Federal de Santa Catarina, Laboratório de Engenharia do Conhecimento, Florianópolis, Santa Catarina, Brasil
{Leandro,tite,bernardo}@egc.ufsc.br

**Abstract.** This work proposes align the Competitive Intelligence (CI) process to the Web of Data (WoD). For this end, we propose a model for identify, select and classify information based on economic sectors to facilitate the retrieval and use of data in the collection step of the CI cycle. The proposed approach proved to be interesting since it reduced the time of the data collection phase and improved selection of the data source.

## 1 Introduction

Whit the rapid evolution of Web-based technologies, agility to gather information has become the key to success. Web brings great variety of data sources that can be freely accessed. Over the past twenty years, various solutions have appeared, Business Intelligence, Knowledge Management, Competitive Intelligence. They all attempt to tackle the problems raised by data proliferation. Undoubtedly, these tools provide operating benefits, but in most cases, none of them offers a genuine solution to the challenges of information growth.

Today, we need to think about corporation access to information within a unified space that receives data, not only from internal but also from the web. However, web data are usually unstructured, fragmented and its contents present ambiguity and heterogeneity problems, restricting the information retrieval and making the knowledge capture particularly difficult [1]. One way around these problems is the Web of Data (WoD) formed by the principles of Linked Data. Proposed by [2], the basic premise behind the WoD says that the utility and value of data increases through your access and recombination [3].

One of the abovementioned processes that can make use of this new source of information is the Competitive Intelligence (CI). Linked Data offers a designed fulfill form to Web information access. It creates a unified informational space that draws on all documents and data, whether structured or unstructured. Linked Data gives the unique opportunity to create new CI applications efficiently targeted to specific business needs reusing free information available on the web combined with traditional CI Systems.

Thus, the proposal developed in this paper aims to use the WoD in the IC process, especially in the collection stage. As the WoD enables explicit connections between

the datasets using representation formats and standard access mechanisms, generic tools such as browsers and search engines can be used to access and process data [4]. As a result, organizations will be able to explore new sources of knowledge, reduce capture and consequently analysis efforts due to the structuring of information. In the following sections of this article we presents, related works, theoretical concepts of CI and WoD, the proposed model and its verification, and at the end conclusions.

## 2 Related Works

According to [5], Web experiences suggest that there is a way for enterprises to build a sustainable architecture for enterprise information, transforming it into a "Enterprise Linked Data" where the act of creating information is closely connected with the act of sharing information. This approach creates a comprehensive and unified information space from which new business information is created to address the needs of traditional processes, such as CI.

In [6] Linked Data brings to the enterprise technologies that represent a response to the challenge of integrating a traditional information process in an open and flexible way, where internal data sources are connected and eventually consolidated with external data. This work is directly related to [6] by integrating sets of external and internal data in a strategic process to get information. The relationship is characterized because both works use the paradigms of Linked Data to create an open space for metadata. However, this paper advances to propose a classification by economic activity before storing the information of WoD. According to [6] business activity is based on large amounts of data and extract the correct information just in time it is a difficult and tedious task. In this case, the classification by economic activities drastically reduces the amount of information analyzed in a CI process.

## 3 Competitive Intelligence

From [7], the basis for most disciplines is found in its origin and history. Intelligence activities for military purposes dated three thousand years ago. Explicitly applied to business, the use of intelligence began in 1960. The 1980's has the introduction of formal functions of collection and analyze information about competitors [8]. Other prominent events were the end of the Cold War and the rise of capitalism where industries and services directly affect the future of citizens through the products and jobs they offer. Finally in the 1990´s the Internet increases and transform de CI into a dynamic and complex entity in [9].

For [13], there are many CI definitions and probably none of them is accurate and universally accepted. According to [10] the concept is rather vague despite several attempts to give meaning to the CI. Thereby, generally, formal definitions on the theme or about specific aspects highlight that CI should support decision-making. The Table 1 shows some of these concepts.

**Table 1** – Concepts and definitions about CI.

| Concept | Author |
|---|---|
| Systematic program of collection and analysis of information about competitors activity and business trends to assist in the company's goals. | [11] |
| Systematic process that transforms data and parts of competitive information into strategic knowledge for decision-making. | [14] |
| Systematic and ethical process of collection, analysis, dissemination and management of information about the external environment that may affect the plans, decisions and operation of the organization. Effective CI is an ongoing process involving legal and ethical collection of information. | [15] |

### 3.1 Competitive Intelligence Process

According to [11] and [12], the result of CI process allows management changes in strategic planning and offers greater organizational effectiveness. Moreover, [16] highlight the importance of systematizing these activities, making the CI process a continuum within the company. Also known as, Intelligence Cycle due to its cyclical and incremental aspect, the CI process usually follows four to five steps but the steps of collection, analysis and dissemination are common to all authors in the CI process. For this work, only the collection step will be addressed in detail, since it is directly linked to the proposal.

#### 3.1.1 Collection

The collection step is characterized by the search for data and information necessary to create knowledge about the competitive environment. The collection task is essentially practical and consisting of identification of sources, collection, treatment, and storage of information. In [11] the information is classified by the origin, domain and type, as described in Table 3.

**Table 3** - Classification of Information.

| *Classification* | *Type* | *Description* | *Examples* |
|---|---|---|---|
| Origin | Primary | Originates in own company | Annual reports, speeches, interviews |
| | Secondary | Originates in ohter sources that observing competitors | Newspapers, analyst reports |
| Domain | Public | Information made public by competitors | Organizational balances, web publications |
| | Not-public | Information that is not public by competitors | Research in fair and sales forces |
| Type | Hard | Information based in quantitative data | Statistical and financial reports |
| | Soft | Information based in qualitatice data | Interviews, gossip and rumours |

## 4 Web of Data and Linked Data

Data is open when can be freely used, modified, and shared by anyone for any purpose. That setting results in a global space we call the Web of Data [17]. The WoD forms a giant global graph, which, in consonance with [05] consists of a billions Resource Description Framework (RDF) triples and covers different domains.

Proposed by Sir Tim Bernes-Lee in 2006, the term Linked Data refers to a style of publishing and linking structured data on the Web. Linked Data does not represent a new technology, but rather a set of best practices for publishing and interlinking structured data on the Web [3] called "*Linked Data Principles*" [18]: 1. Use URIs as names for things; 2. Use HTTP URIs so that people can look up those names; 3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL); 4. Include links to other URIs, so that they can discover more things.

Behind these four principles, the goal of Linked Data is to use the Web architecture, to share structured data on a global scale [3]. As the web formed by hypertext links, Linked Data is constructed with Web documents, however, connections are made using *HyperData* links where information contained in documents can be connected [18].

## 5 Proposed Model

The proposed model brings together the steps and elements necessary to make the collection of Web of Data information to support the Competitive Intelligence cycle. The model consists in the arrangement of technologies and concepts in a Knowledge Engineering tool to feeding the collection phase of the CI process with information from WoD.

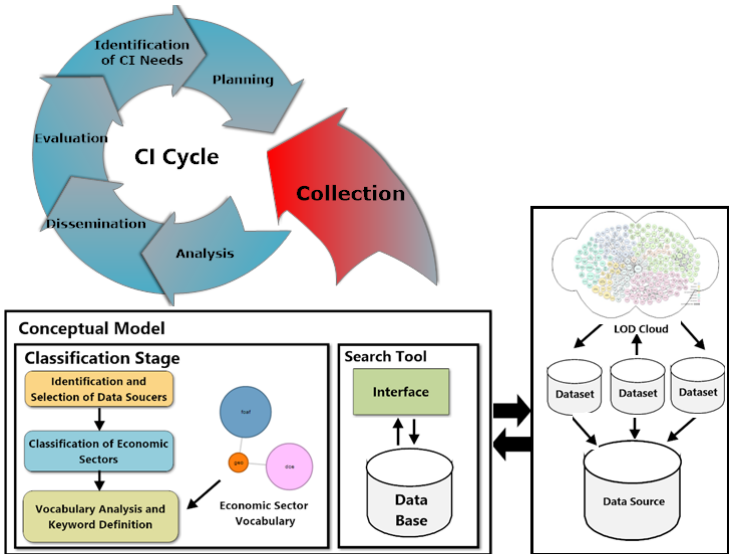


**Figure 1** –Conceptual Model.

The conceptual model provides core issues for CI focused on data collection in WoD related to a specific economic sector. More detailed information about a specific aspect or even proprietary information, such as profitability, consumers, intellectual property, trade secrets and proprietary methods, strategic plans, internal management procedures should be inferred and cataloged in the analysis phase.

### 5.1 Classification Stage

The classification stage comprises the initial steps for the use of WoD in the CI process. At this stage were developed tasks such as the identification and selection of data sources, the classification of economic sectors as well as the terms used to represent and return the information. In the following topics each of these tasks will be discussed.

#### 5.1.1 Data Source Identification and Selection task

The starting point for the classification task was the 295 data sources which form the 2011 *LOD Cloud* diagram[1]. Currently there is a new version of the WOD updated in 2014 with approximately one thousand datasets. This version was not used because the classification occurred before the release of the final version. From this initial set were selected at the end of the process 135 data sources, which according to the following specified criteria, are relevant to the CI process:

1. Active links presence: were eliminated data sources that even represented in the diagram does not have active links. This analysis resulted in the exclusion of 26 data sources;
2. Availability: 27 data sources have been excluded because have no recoverable data available;
3. Duplicity: 36 duplicate sources were excluded. These are represented more than once or were contemplated within datasets from other data sources;
4. Relevance: the content of data sources was analyzed for its relevance to the process of CI and those who do not have relevant information was excluded such as religion and repositories about cartoons. At this stage 71 sources were excluded.

It is noteworthy that the resulting data sources are composed by one or more datasets, a collection of published information, maintained or aggregated by a single provider [19]. Due to the connection between them, a dataset may appear in more than one data source simultaneously. However, based on information presented in the data sources identified approximately 150,000 datasets composed of around 50 billion resources[2].

[1] http://lod-cloud.net/state/

[2] Information present in a Dataset. e.g: triple RDF, an XML file, CSV, spreadsheet, among others.

### 5.1.2 Selection of Economic Sectors task

The data sources were manually classified according to the Brazilian CNAE (National Classification of Economic Activities), issued by the National Commission of Classification and provided by the IBGE (Brazilian Institute of Geography and Statistics). Into the CNAE classification, economic activities are organized into 21 sections and 99 divisions. "Public Administration, defense and social security" is the one with the highest quantity of information, with more than 100,000 datasets in 14 data sources. Initiatives of federal administration like the data portal of the USA Government[3] and the United Kingdom[4] respectively include around 87,000 and 20,000 datasets. The existence of approximately 20,000 datasets of "Extractive Industries" and "Manufacturing industry", and 18,000 of "Agriculture, Livestock, Forestry Production, Fisheries and Aquaculture" is also noteworthy.

## 5.2 Search tool

The developed search tool aims to support the recovery of data sources used in the collection step of the CI cycle. Thereby, when selecting an economic sector, only the corresponding data sources are presented. In turn, the data sets are arranged so that they can be directly retrieved using keywords, by selecting one of the economic sectors or simply by indexed text search. So only the datasets belonging to the data source are made available. Figure 3 shows one dataset for "Electricity and Gas" sector.

In Figure 2(a), the column that contains the name of the data source is a link to the interface where the registered datasets that compose it are located. By clicking the URI, the user is directed to the home page of the dataset. Figure 2(b) shows the page of South America power plants[5] in *Enipedia*, a data source of the energy sector structured in a similar way to *Wikipedia*.

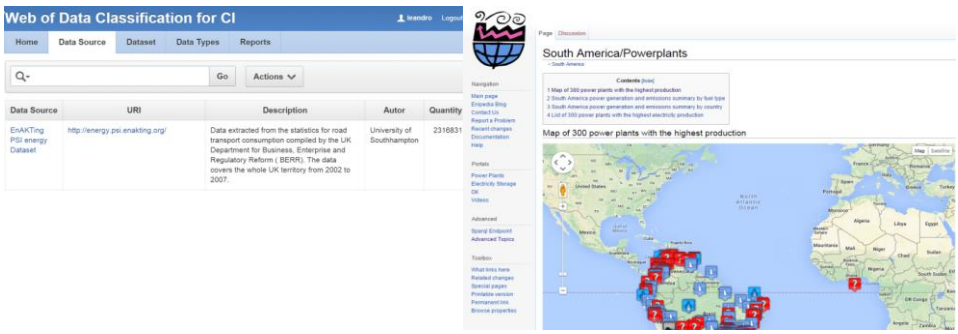


**Fig. 2** – (a) Data Sources Interface. (b) South America Power Plants

[3] http://www.data.gov/

[4] data.gov.uk

[5] http://enipedia.tudelft.nl/wiki/South_America/Powerplants

The same icons that point the location of the power plants also represent the type of fuel used to generate electricity. When you select one of these points, information about the specific power plant is displayed.

# 6 Model Verification

The Electricity and Gas sector has been chosen for the verification of proposed model. Especially in Brazil, this sector is characterized by high government regulation. Competition between companies and investments require detailed analysis of technical, economic, financial and environmental aspects before its effective implementation forcing developing corporate governance policies, fulfill their business plans and investments, as well as plan future energy demand.

The choose of this sector is due to the quantity and quality of data sets present in the WoD, the social and economic importance and the fact that it demands information from other areas present in the data cloud, such as: sustainability, spatial data, demographics, and business activity. Six data sources that directly address the Electricity and Gas sector are present in the WoD. These data sources comprise about two thousand datasets and approximately 23.2 million records.

## 6.1 Identification of CI goals

The analysis of works such as [20] and [21] can be seen that electric companies have their information needs framed into three functional categories: actions and strategic decisions, early alerts and description of the main competitors. Based on these concepts we list three information needs for the progress of an intelligence project in the electricity sector for actions and strategic decisions:

1. Identify needs of production expanding and energy demand;
2. Identify alternative energy sources;
3. Identify investments its destinations and the amounts involved, fusions and shareholdings.

## 6.2 Data Collection

This section presents the information datasets identified for each needs. The identification of datasets was done using metadata and keywords presents in the datasets descriptions. For this study are relevant to how information is made available, the data format, source, reliability and relevance. Table 5 presents an overview of the topics and the number of datasets that address each of the identified needs using de application.

**Table 5** – CI goals and identified datasets.

| CI goals | Topic | Qty. |
|---|---|---|
| Identify needs of production expanding and energy demand | Generation | 109 |
| | Consumption | 81 |
| | Demand | 10 |
| Identify alternative energy sources | Geothermal | 185 |
| | Solar | 85 |
| | Renewable | 42 |
| | Eolic | 32 |
| | Sustainable | 26 |
| | Hydrogen | 8 |
| | Biomass | 4 |
| Identify investments its destinations and the amounts involved, fusions and shareholdings | Indicators | 45 |
| | Bioenergy | 23 |
| | Efficiency | 14 |

Font: Authors (2015).

Altogether, 664 datasets with data about the three proposed CI needs were identified. To answer the need of "*Identify needs of production expanding and energy demand*", we found 200 datasets. These datasets address topics such as: the annual electricity[6]demand and consumption about each country between the years 1980 and 2009;

Data about demand for renewable energy in the period 2005-2009 and the energy consumption of a particular country; Datasets about the hydropower generation in Brazil, fossil fuel energy generation such as oil and coal, and renewable like *The Wind Power*[7] where data of wind farms capacity in megawatts of 102 countries is listed.

To answer the goal "*Identify alternative energy sources*", 405 datasets were catalogued. These datasets generally deal with the use of renewable energy and the involved technologies, as well as the potential of the world's renewable resources such as bioenergy, biomass and wind. For example, The *Center for Energy Research*[8] provides information about the average wind potential at 50 meters above the ground in Brazilian territory.

The last goal to reach, "*Identify investments its destinations and the amounts involved, fusions and shareholdings*", includes datasets like the *worldwide summary on energy efficiency*[9]. This summary presents indicators, statistics and trends in the power sector, as well as initiatives such as smart grids which seek to make a smarter chain of production and distribution of energy.

Lastly, it is noteworthy that all datasets that met the identified goals can be easily recovered by keywords or by text search within the application.

[6] http://en.openei.org/datasets/node/877

[7] http://www.thewindpower.net/country_list_fr.php

[8] http://en.openei.org/datasets/node/608

[9] http://en.openei.org/datasets/node/468

## 7 Conclusions

The problem presented in this paper deals with the use of the Web of Data in the collection step of the Competitive Intelligence process. To answer it, a conceptual model was developed, able to identification, selection and classification of information and a search tool to assist the collection of information applied in the electricity and gas sector.

The main aim of this work resides in the use of WoD as an external source of information, structured and easy to retrieval in the CI process. The Web of Data adds an additional semantic layer strongly interconnected with the traditional Web documents. With the use of WoD solution, a CI process can be expanded without altering the existing model, databases or mechanisms already in place. Unlike traditional data integration, WoD provides a comprehensive view of the data as a whole and can be used to create new information and goes beyond a document-based framework by creating informational objects contextualized to business objectives.

Among the contributions for CI professionals, the most evident is the model proposed for information collection and the application that resulted from this work. Once the information is grouped by economic activities, following the linking business objects are themselves connected. This makes navigation easier, allowing resource discovery and improving understanding. The proposed model gives a macro view of relationships between these objects to create new information. For this the proposed model must use data and documents from the WoD.

The main difficulties encountered during the development of this study were the lack of references that address the entire scope as well as the cataloging of data sources in the application. For this proof of concept, the cataloging task of datasets was done manually, which would require significant time for everyone to be cataloging. For this reason, only the datasets about Electricity and Gas sector were registered in its entirety. However, these difficulties did not harm the outcome of the current study and this process can be automated in the future.

In conclusion, the information present on the Web of Data can be used with significant gains for the CI collection step. Since this information is made available in structured formats and comes from reliable and specialized sources on the subject, its use may represent a great saving of time in the collection and analysis steps. Other identified benefits include prior data treatment, its connection with other sources, the preparation of reports and creation of applications that can help not only in the collection, but also in the analysis step of the CI cycle. In addition, the use of an application such as proposed in this paper facilitates the information retrieval job.

## References

[1] Passant, A. et al. Enhancing Enterprise 2.0 Ecosystems Using Semantic Web and Linked Data Technologies:The SemSLATES. In: WOOD, D. (Ed.). Linking Enterprise Data: Springer US, 2010. cap. 5, p.79-102. ISBN 978-1-4419-7664-2.

[2] Berners-lee, T. Giant Global Graph 2007.

[3] Heath, T.; Bizer, C. Linked Data: Evolving the Web Into a Global Data Space. Morgan & Claypool, 2011. ISBN 9781608454303.
[4] Jentzsch, A. et al. Enabling Tailored Therapeutics with Linked Data. Proceedings of the WWW2009 workshop on Linked Data on the Web, 2009, Madrid, Spain.
[5] Allemang, D. (2011) "Semantic Web and the Linked Data Enterprise", In: Wood, David. (Ed.), Linked Enterprise Data, Springer.
[6] Hu, B., Svensson, G. (2010) "A Case Study of Linked Enterprise Data", in ISCW2010, Shanghai, China.
[7] Juhari, A. S.; Stephens, D. Tracing the Origins of Competitive Intelligence Throughout History. Journal of Competitive Intelligence and Management, v. 3, n. 4, p. 61-82, // 2006. ISSN 1703-5147.
[8] Prescott, J. E. The Evolution of Competitive Intelligence: Designing a Process for Action. PROPOSAL Management, 1999.
[9] Colby, W. E. Competitive Intelligence in the New World of the 1990s. Global Perspectives on Competitive Intelligence, p. 5, 1993.
[10] Gilad, B. Strategy Without Intelligence, Intelligence Without Strategy. Business Strategy Series, v. 12, n. 1, p. 4-11, 2011.
[11] Kahaner, L. The basics of competitive intelligence. In: Competitive Intelligence: How to Gather Analyze and Use Information to Move Your Business to the Top. 1. New York: Simon & Shuste, 1996. 300 ISBN 0684844044.
[12] Sawka, K. Whither Analysis? Competitive Intelligence Mag., v. 9, n. 2, 2006.
[13] Abreu, A. F.; Cobral, E.; Ogliari, A. Gestão integrada da inovação: estratégia, organização e desenvolvimento de produtos. In: ATLAS (Ed.). Atlas. São Paulo: Atlas, v.1, 2008. p.269. ISBN 978-85-224-4976-7.
[14] Tyson, K. W. M. The Complete Guide to Competitive Intelligence. Division of Kirk Tyson Associates: Chicago, 1998.
[15] SCIP. Strategic and Competitive Intelligence Professionals. 2013. Disponível em: < www.scip.org >. Acesso em: 30/03/2013.
[16] Fan, W. et al. Tapping the power of text mining. Commun. ACM, v. 49, n. 9, p. 76-82, 2006. ISSN 0001-0782.
[17] Bizer, C.; Heath, T.; Berners-Lee, T. Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS), v. 5, n. 3, p. 1-22, 33// 2009. ISSN 1552-6283.
[18] Berners-Lee, T. Linked Data: Design Issues. Online, 29/03/2013 2006. Disponível em: < http://www.w3.org/DesignIssues/LinkedData.html >. Acesso em: 02/02/2015.
[19] Alexander, K. et al. Describing Linked Datasets On the Design and Usage of voiD, the "Vocabulary Of Interlinked Datasets". Linked Data Workshop at WWW09 2009.
[20] BEER, J.; WORRELL, BLOK, K. Future technologies for energy efficient iron and steel making. Annual Review of Energy and Environment, v. 23, p. 123- 205, 1998. Disponível em: https://ies.lbl.gov/iespubs/42774.pdf
[21]Tolmasquim, Mauricio T., Guerreiro, Amilcar, & Gorini, Ricardo. (2007). Matriz energética brasileira: uma prospectiva. *Novos Estudos - CEBRAP*, (79), 47-69.