Discovering and Linking Financial Data on the Web

José Luis Sánchez-Cervantes
Universidad Carlos III de Madrid
Av. Universidad 30, Leganés, 28911,
Madrid, Spain.
+34 91 624 5936
joseluis.s.cervantes@alumnos.uc3m.es

Gandhi S. Hernández-Chan
Universidad Carlos III de Madrid
Av. Universidad 30, Leganés, 28911,
Madrid, Spain.
+34 91 624 5936
gandhi.hernandez@alumnos.uc3m.es

Mateusz Radzimski
Universidad Carlos III de Madrid
Av. Universidad 30, Leganés, 28911,
Madrid, Spain.
+34 91 624 5936
mradzims@pa.uc3m.es

Juan Miguel Gómez-Berbís
Universidad Carlos III de Madrid
Av. Universidad 30, Leganés, 28911,
Madrid, Spain.
+34 91 624 5936
juanmiguel.gomez@uc3m.es

Ángel García-Crespo
Universidad Carlos III de Madrid
Av. Universidad 30, Leganés, 28911,
Madrid, Spain.
+34 91 624 5936
acrespo@ia.uc3m.es

Abstract—**The constant publishing of large volumes of financial information by various business sector organizations through its financial statements is a fact that must be exploited by using semantic technologies. This paper describes the use of Linked Data to generate a financial dataset that is part of the ongoing work of the FLORA (Financial Linked Open Data Reasoning and Management for Web Science) project. Furthermore, we describe a process to discover other relevant links within the LOD cloud which can be related to FLORA dataset in order to provide a financial knowledgebase, which might be useful for: data analysis to support decision making, generating predictions or performing own financial discoveries to mention a few. However, the results of experiments performed, show that the coverage of the financial domain of public companies is rather small and ambiguous to link them to the FLORA dataset. Thus, is necessary involve more data as CIK (Central Index Key) or ticker symbol of the companies to objective of improve the results quality and implement techniques for disambiguation and manual verification.**

Keywords-Financial data; Semantics; Linked Open Data; FLORA; SILK framework.

I. INTRODUCTION

The current trend for companies to publish their financial statements under the Generally Accepted Accounting Principles (US-GAAP Financial model) and following the eXtensible Business Reporting Language (XBRL standard), increases the capacity for recovery and analysis of relevant data containing financial data semi-structured which is derived to facilitate the extending of financial datasets. Unlike datasets integrated by unstructured data and published on the Web, which represent a limitation by the high cost of transformation in order to be fit into existing analytic models and tools [1]. In our work, we have exploited the advantages of using semantic technologies [2], Linked Data (and Linked Open Data) [3] itself, which involves make the most of the best practices of sharing the data across the Web with great integration capabilities [4], basically we present a high level overview of an ongoing work in the FLORA project. FLORA fosters the transparent access to financial data through its dataset developed from extraction and triplification of the data contained in the financial statements of companies registered on U.S. Securities and Exchange Commission (SEC) [5] through fillings and forms stored on EDGAR System [6] furthermore, it allowing the easy combination of many information sources thus favoring the optimizing for data analysis. However, we consider important increasing the FLORA functionality through use of frameworks for discovering relationships between companies contained on FLORA dataset and data items within different Linked Data sources. This paper is structured in five sections, which are described briefly below. The introduction provides an overview of FLORA project, the second section corresponds to related work and it is divided into two subsections, the first subsection includes tools for discovery of links on the Web of data and second subsection describes some Linked Data projects. In section three, we describe the process for generation of financial dataset and discovery of related data items with companies stored on FLORA dataset. The fourth section, we describe the experiments for discovery of data elements in the LOD cloud and we present quantitative results obtained. Finally, we mention a conclusion and the future of our work.

II. RELATED WORK

In recent years, the systems for "triplification" and publication of datasets from multiple domains based on the Linked Data principles are becoming increasingly important. In this sense, there are several initiatives for the discovery of data items contained within the Linked Open Data cloud (LOD). We have classified these initiatives in two types, which are described briefly below:

A. Tools for discovery of links on the Web of data

Silk - Linking Framework [7], a toolkit for discovering and maintaining data links between Web data sources was presented in [8]. Silk is a tool that allows discovering

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-295-0

relationships between data items within different Linked Data sources. Data publishers can use Silk to set RDF (Resource Description Framework) [9] links from their data sources to other data sources on the Web. The presentation and evaluation of LIMES (Link Discovery Framework For Metric Spaces) [10] a novel time efficient approach for link discovery in metric spaces was described in [11]. The authors approach utilizes the mathematical characteristics of metric spaces to compute estimates of the similarity between instances. Thus, LIMES can reduce the number of comparisons needed during the mapping process by several orders of magnitude. In [12], LinQL framework was presented. It is a generic and extensible framework that works like an extension of SQL (Structured Query Language). This tool allows users to interleave declarative queries with interesting combinations of link discovery request. Its goal is to facilitate experimentation and help users find and combine the link discovery methods that will work best for their application domain. In [13], Silk Server was presented. It is an identity resolution component, which can be used within Linked Data application architectures to augment Web data with additional RDF links.

B. *Linked Data applications*

In [14], current efforts interlinking music-related datasets on the Web were addressed. The authors detail the application of an algorithm in two contexts: a) to link the Creative Commons music dataset to an editorial, and b) to link a personal music collection to corresponding Web identifiers. One of the most important features of this algorithm is that it was developed, implemented and practically it deployed to interlink different music-related datasets facing the overlapping problem. Faviki is an example of linked data application. It is a social bookmarking tool that lets users tag Web pages with semantic tags stemming from Wikipedia [15]. In [16], K-Search was presented. K-Search is an implementation of Hybrid Search (HS) another manifestation of linked data. It combines the flexibility of keyword-based retrieval with the ability to query and reason on metadata typical of semantic search systems. HS is defined as: i) the application of semantic search for the parts of the user queries where metadata is available and ii) the application of keyword-based search (KS) for the parts not covered by metadata however, KS is often affected by two main issues, ambiguity and synonymy. The LDIF-Linked Data Integration Framework [17] can be used with Linked Data applications to translate heterogeneous data from the Web of Linked data into a clean local target representation mapping language for translating data from various vocabularies that are used on the Web into a consistent, local target vocabulary. It includes an identity resolution component, which discovers URI (Uniform Resource Identifier) based on user-provides matching heuristic. The goal of Linking Open Drug Data (LODD) project presented in [18] is to facilitate the integration of large amounts of biomedical data from many different sources by bringing these data sources onto the Web of Linked Data. The biomedical datasets selected allow strong connections to existing Linked Data resources, while providing novel data of interest to pharmaceutical industry and patients. FLORA allows search of financial information contained in its data set including the calculation of several additional financial ratios that help users in finding information relevant to them [19]. However, the development of this project is iterative to obtain a continuous improvement at every stage of research and development whereby which we intend to exploit the FLORA possibilities. In the following section, stages of FLORA process are described briefly.

III. FLORA FINANCIAL DATASET

The transformation and generation process of FLORA dataset consists in investigate appropriate data sources and rich in financial information among which include: balance sheet, cash flow and income statements from companies. So far, we have generated a dataset that stores RDF triples generated from the extraction of 409.374 financial statements, which have been published in XBRL [20] format by different U.S. companies through the EDGAR system fillings [21].

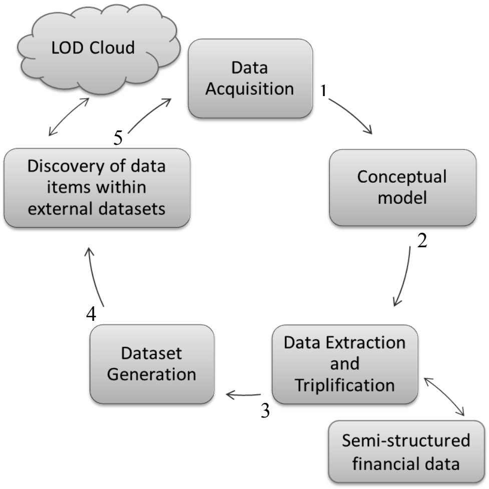
LOD Cloud
Data Acquisition
1
Conceptual model
2
Data Extraction and Triplification
Semi-structured financial data
3
Dataset Generation
4
Discovery of data items within external datasets
5

Figure 1. Stages of the FLORA process for generation of financial dataset and discovery of related data items

In Figure 1, each stage of the FLORA process has a defined function, which is briefly described below:

1. **Data Acquisition:** It is to research and get only the data sources of interest. The current focus of FLORA system is based on US-GAAP XBRL reports containing the Forms 10-Q, as published by the SEC EDGAR System. The published reports are crawled, downloaded and stored for further processing (Stage Data extraction and triplification). This stage needs to be repeated (at least quarterly) in order to retrieve newest reports and keep the system up-to-date.
2. **Conceptual model:** It is a meta-model that is the core of FLORA functionality because semantically represents the interaction between classes and subclasses that integrate it. This representation is

described in a high level of abstraction (see Figure 2) and includes an example of simplified taxonomy generated from published balance sheets under US-GAAP model including its general financial ratios, besides allowing the calculation of some additional ratios among which are: Total Asset Turnover, Non-Current Asset Turnover, Current Asset Turnover, Rotation of Rotation of Warehouse or Stocks, Working Capital, Cash Ratio, Debt Ratio and Ratio of Debt Quality.

3. **Data Extraction and Triplification:** this stage involves two simultaneous processes: the analysis (parsing) and extraction of data from financial statements made at the first stage and conversion to RDF triples extracted data in order to build the required dataset.
4. **Dataset generation:** this stage is closely related with the previous stage and consists in serialize the RDF triples to the semantic form. The dataset is following Linked Data principles and can be queried through SPARQL (Simple Protocol and RDF Query Language)-based queries.
5. **Discovery of data items within external datasets:** this is a stage that is currently underway and that we can consider the point to be discussed in this paper. As shown in Figure 2, the flow "*Discovery of related data items*" defines the search to find and create the links of all data stored in the LOD cloud that are related to companies registered in the FLORA dataset. Analyzing the state of the art (see Section 2) using frameworks like LIMES or Silk can be useful to achieve this goal.

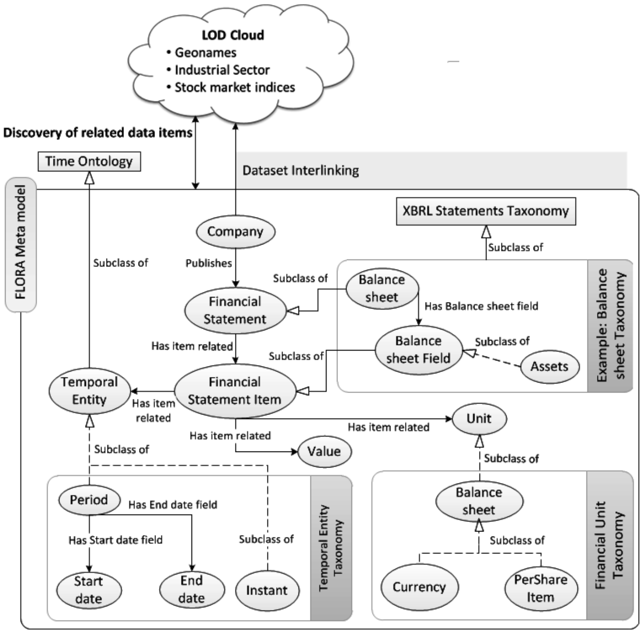

Figure 2. FLORA conceptual model

Discovering and linking additional data with data existing in FLORA dataset, will improve the search criteria and its reasoning capabilities. On other hand, through conceptual model we can identify the FLORA's behavior as well as detected and improve any data inconsistency with the possibility to correct or eliminate them.

IV. DISCOVERING DATA ITEMS IN THE LOD CLOUD

The volume of data stored in the FLORA system is increasing frequently to keep it updated. However, we find it interesting to investigate to what extent the LOD cloud covers the financial domain and what other data items may be linked with our dataset. To find this information, we have performed a set of experiments that allow linking external dataset with FLORA dataset through the SILK framework which provides the necessary mechanisms for discovering relationships between data items corresponding to FLORA dataset and different Linked Data sources within the LOD cloud.

As external datasets we have chosen DBpedia, Semantic XBRL and SEC triplified dataset. The latter two datasets are linking company concepts to DBpedia, but due to the outdated version (SEC triplified dataset discontinued as of 2010) and scarce mappings (Semantic XBRL) the only dataset is DBpedia.

The FLORA dataset that we have created is composed of 8915 public US companies obtained from SEC EDGAR fillings. In order to create mappings between concepts we considered comparing the following properties: (i) label, (ii) CIK (Central Index Key), (iii) ticker symbol. While the CIK and ticker symbol would clearly lead to the best results, most of the companies in DBpedia are lacking those data, leaving us with only label as a viable property for mappings.

To facilitate mappings, we performed 2 necessary steps: we created a list of stopwords for company names, including such terms as “inc”, “co”, “ltd” etc. and created a list of synonyms with stopwords filtered. In that example, a company “Apple Inc.” would have a synonym “Apple”.

For comparing strings, we used Levenshtein distance and a metric that measures the difference between two sequences. The Levenshtein distance between two strings is defined as the minimum number of edits needed to transform one string into the other, with the allowable edit operations being insertion, deletion, or substitution of a single character.

After that, we launched experiments with FLORA dataset against DBpedia SPARQL endpoint, with following parameters:

- Experiment 1 presents a Levenshtein distance with value 0 and the use of Stopwords.
- Experiment 2 includes a Levenshtein distance with value 0 and no Stopwords (no synonyms, only the official company name).

- Experiment 3 contains a Levenshtein distance with value 1 and use of Stopwords.
- Experiment 4 involves a Levenshtein distance with value 1 and no use of Stopwords (as in Experiment 2).

The summary of the experiments performed and the results obtained are shown in Table 1.

TABLE 1. SUMMARY OF THE EXPERIMENTS PERFORMED AND THE RESULTS OBTAINED

Linking external dataset experiment			
Number of companies in FLORA dataset: 8915			
Experiment	***Levenshtein distance value***	***Stopwords***	***Discovered Links***
1	0	No	96
2	0	Yes	1031
3	1	No	437
4	1	Yes	3652

The results presented in this section are experiments for development of our work and can be considered as partial results for the stage of "Discovery of data items within external datasets" (see Section 3). While for a simple string comparison (experiment 1) we obtain quite a few mappings, this is what previously was observed by García and Gil [22]. However the striking difference is the increase of number of links created when a list of synonyms is generated. The distance between strings might include many false links, especially for short company names, but provides user with more possible alternatives in case of longer company names (with typically non-letter or a white space character difference).

V. CONCLUSION AND FUTURE WORK

This article presented a complex, unified process of transforming unstructured financial data into an interlinked, navigable knowledge base for financial information management and information discovery within the Web of data through the use of frameworks developed for this purpose. In the future work we aim at implementing LIMES framework, applying of inference using SPIN rules and identify and use other financial data sources to perform more complex experiments.

The experiment, however, shows that the coverage of the financial domain (in case of this paper: public companies) is rather small. The lack of data that could be used for univocally identify company concepts (such as CIK or ticker symbol) makes dataset interlinking still a difficult task requiring various techniques for disambiguation and manual verification.

In the future work, we are focusing on generation of rich gazetteer for each company in order to increase the number of possible matches based on the company name. Also, other string comparing metrics are considered; that could use other advanced features for comparing company names. After that, we will perform a manual evaluation of created links in order to assess the gazetteer and synonym list for company mappings.

ACKNOWLEDGMENTS

This work was supported by the Spanish Ministry of Industry, Tourism, and Commerce under the projects: FLORA (TIN2011-27405) and (SEMOSA TSI-020400-2011-51). It was also supported by The Spanish Ministry of Science and Innovation under the project GECALLIA (TSI-020100-2011-244). Additionally, this work was sponsored by The National Council of Science and Technology (CONACYT).

REFERENCES

[1] Ciccotello, C. S. and Wood, R. E. An investigation of the consistency of financial advice offered by web-based sources. *Financial Services Review*, 10, 1-4 (2001), 5-18. DOI=http://dx.doi.org/10.1016/S1057-0810(01)00078-6

[2] Allemang, D. and Hendler, J. *Semantic Web for the working ontologist: effective modeling in RDFS and OWL*. Morgan Kaufmann, 2011. ISBN-10: 0123859654.

[3] Bizer, C., Heath, T. and Berners-Lee, T. Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5, 3 (2009), 1-22. DOI= 10.4018/jswis.2009081901

[4] O'Riain, S., Harth, A. and Curry, E. Linked data driven information systems as an enabler for integrating financial data. *Information Systems for Global Financial Markets*, Emerging Developments and Effects, (2011), 239-270. DOI=10.4018/978-1-61350-162-7.ch 010

[5] U.S. Securities and Exchange Commission, SEC. Retrieved June 16, 2013, from http://www.sec.gov/index.htm

[6] EDGAR System. Retrieved June 17, 2013, from http://www.sec.gov/edgar.shtml

[7] Silk - A Link Discovery Framework for the Web of Data. Retrieved June 17, 2013, from http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/

[8] Volz, J., Bizer, h., Gaedke, M. and Kobilarov, G. 2009. Silk– a link discovery framework for the web of data. *In Proceedings of the 2nd Linked Data on the Web Workshop LDOW2009*, (Madrid, Spain, April 20, 2009), 559-572.

[9] RDF, Resource Description Framework. Retrieved June 17, 2013, from http://www.w3.org/RDF/

[10] LIMES, Link Discovery Framework for Metric Spaces. Retrieved June 17, 2013, from http://aksw.org/Projects/LIMES.html

[11] Ngomo, A. N. and Auer, S. LIMES: a time-efficient approach for large-scale link discovery on the web of data. *In Proceedings of the Twenty-Second international joint conference on Artificial Intelligence* Volume Three. AAAI Press, 2011, 2312-2317.

[12] Hassanzadeh, O., Lim, L., Kementsietsidis, A., and Wang, M. A declarative framework for semantic link discovery over relational data. *In Proceedings of the 18th international*

conference on World Wide Web. (ACM New York, NY, USA, 2009), 1101–1102 DOI=10.1145/1526709.1526876

[13] Isele, R., Jentzsch, A., and Bizer, C. Silk Server-adding missing links while consuming linked data. *In 1st International Workshop on Consuming Linked Data (COLD 2010)*. (Shanghai, China, November 8, 2010)

[14] Raimond, Y., Sutton, C., and Sandler, M. Automatic Interlinking of Music Datasets on the Semantic Web. *In Proceedings of the 1st WorkShop about Linked Data on the Web LDOW2008*. (Beijing, China, April 22, 2008)

[15] Hausenblas, M. Exploiting Linked Data to Build Web Applications. *IEEE Internet Computing*, *13*(4), 68-73. DOI= http://doi.ieeecomputersociety.org/10.1109/MIC.2009.79

[16] Bhagdev, R., Chapman, S., Ciravegna, F., Lanfranchi, V. and Petrelli, D. Hybrid search: Effectively combining keywords and semantic searches. *The Semantic Web: Research and Applications*, 5021 (2008), 554-568. DOI=10.1007/978-3-540-68234-9_41

[17] Achultz, A., Matteini, A., Isele, R., Bizer, C., and Becker, C. LDIF-Linked Data Integration Framework. *In 2nd International Workshop in Consuming Linked Data*. (Bonn, Germany, October 2011)

[18] Jentzsch, A., Cheung, K., Zhao, J., Samwald, M., Hassanzadeh, O., and Andersson, B. Linking Open Drug Data. *Presented at the Triplification Challenge of the International Conference on Semantic Systems*. (2009), 3-6.

[19] Radzimski, M., Sánchez-Cervantes, J. L., Rodríguez-González, A., Gómez-Berbís, J. M., and García-Crespo, Á. FLORA–Publishing Unstructured Financial Information in the Linked Open Data Cloud. *In International Workshop on Finance and Economics on the Semantic Web (FEOSW 2012)*, (Heraklion, Greece. May 27-28, 2012), 31.

[20] XBRL - Extensible Reporting Business Report Language. Retrieved June 19, 2013, from http://www.xbrl.org/

[21] EDGAR System Fillings. Retrieved June 20, 2013 from http://www.sec.gov/cgi-bin/browse-edgar?action=getcurrent

[22] García, R., and Gil, R. Linking XBRL financial data. *In D. Wood (Ed.), Linking enterprise data*. *Springer US.* (2010), 103-125. DOI=10.1007/978-1-4419-7665-9_6