FLORA – Publishing Unstructured Financial Information in the Linked Open Data Cloud

Mateusz Radzimski, José Luis Sánchez-Cervantes, Alejandro Rodríguez-González,
Juan Miguel Gómez-Berbís, Ángel García-Crespo

[1] Departamento de Informática
Universidad Carlos III de Madrid, Spain
{mradzims, joseluis.sanchez, alejandro.rodriguez, juanmiguel.gomez,
angel.garcia}@uc3m.es

Abstract. In the world, where computers assist humans in information processing in almost every aspects of our lives, there are still huge gaps of unsurveyed areas, where data exists in an unstructured or unprocessable form limiting its usefulness and requiring extra human effort. Many times such data is extremely useful for many parties, as is the case of financial data. This paper describes an ongoing work of the FLORA system that aims at transforming unstructured financial data into Linked Data form and interlinking it with other relevant datasets of LOD initiative in order to provide a financial knowledgebase for financial data analysis framework.

Keywords: linked open data, financial data, data integration, data publishing

1 Introduction

With increasing number of financial data sources being published on the web, still doesn't come the easiness of analysis and retrieving relevant information. Documents containing financial statements appear to be structured, however many times only its content is structured, but not the data itself. Therefore any analysis or further processing of such datasets are limited by high cost of transformation in order to be fit into existing analytic models and tools [1]. This situation is also keeping the bar of multisource data integration very high. On the other hand we experience the blossoming development of Linked Open Data [2] cloud that offers best practices of sharing the data across the Web with great integration capabilities. Apart from bringing transparent access, Linked Open Data allows for easy combination of many information sources thus allowing for better data analysis. Financial information in such form could be of use by many entities, such as regulatory bodies detecting market anomalies, banks analyzing the risk of held assets or investors making better informed decisions.

Such semantic information integration starts to play crucial role in many domains such as bioinformatics, medical domain and other life sciences with growing amount of data gathered in repositories such as Bio2RDF [3] or Linked Life Data [5]. Integrating those dataset using Linked Data approach opens new possibilities for better data discovering, querying and visualization.

This paper presents a high level overview of an ongoing work in the FLORA project. FLORA aims at bringing the advantages of Linked Data (and Linked Open Data) to lower the integration obstacles [4] and to transform financial data into the form which can be used in automated environments in the interoperable way. Employing NLP techniques for information extraction will foster transforming unstructured data from public companies' statements dealing with accounting and financial perspective. Linked Data cloud of computation results is not only a suitable format for browsing and querying, but it also forms an input for further analysis services for decision support, multi-faceted data presentation [6] and visualization [7].

2 Related Work

There are several initiatives related with extracting unstructured information in the financial domain. Some of these works have obtained outstanding results through applying semantic technologies. In this section some of this works are described briefly.

The MONNET project [8] proposes a solution to the cross-language information access problem by using a novel combination of Machine Translation and Semantic Web Technology [9] for the public and financial sector. MONNET achieves this through semantically aware term translation based on a novel approach that integrates ontology-based domain semantics with linguistic information from the domain lexicon. MONNET project provides several benefits of scientific innovation and scientific impact as: ontology-lexicon model, multilingual ontology localization, cross-lingual ontology-based information extraction, cross-lingual knowledge access, presentation framework, formal model for multilingual, lexicalized knowledge representations, ontology localization services, methodology for developing cross-lingual information access applications, integrated approach to ontology localization, cross-lingual ontology-based information extraction, ontology-based language-independent information access, to mention a few.

The FIRST project [10] addresses the challenges of dealing with financial data in a near real-time with vast and constantly growing amounts of heterogeneous sources from financial markets. It aims at providing a large-scale information extraction and integration infrastructure for supporting financial decision-making process. The main result, Integrated Financial Market Information System, is based on a pluggable open architecture framework for non-ICT skilled end-users for on-demand information access and highly scalable execution of financial market analyses.

The XLite [11] project's main goal is to develop technology to monitor and aggregate knowledge that is currently spread across mainstream and social media, and to enable cross-lingual services for publishers, media monitoring and business intelligence. By combining modern computational linguistics, machine learning, text mining and semantic technologies with the purpose to deal with the following two key open research problems: the first is extract and integrate formal knowledge from multilingual texts with cross-lingual knowledge bases, and the second to adapt linguistic techniques and crowd-sourcing to deal with irregularities in informal language used primarily in social media.

LOD2 project [12] aims at developing technologies for scalable management of Linked Data collections in the many billions of triples and progress the state of the art of Semantic Web in data management, both commercial and open-source. It assumes RDF data representation as a viable choice for organizations worldwide and a premier data management format [13]. The LOD2 project contributes high-quality interlinked versions of public Semantic Web data sets, promoting their use in new cross-domain applications by developers across the globe. LOD2 also develops a suite of tools for data cleaning, linking and fusing that will help bootstrapping creation of datasets for new domains[1].

In [14] the lack of a well-documented software library for access and publication of data throughout the lifecycle of Linked Open Data and the problems related with the amount and quality sparse of the links among Linked Open Data Sources because of its growth were mentioned. The LATC project provides an alternative of solution to the problems previously mentioned. To support interested parties in Linked Data publication and consumption, LATC publishes and maintains a publication & consumption tools library along with screen-casts and tutorials. To provide an in-depth test-bed for data intensive applications, LATC publishes data produced by the European Commission, the European Parliament, and other European institutions as Linked Data.

In [15], the authors discuss how semantics can improve XBRL (Extensible Business Reporting Language) characteristics of expressiveness and interoperability beyond plain XML data representation. In a practical sense XBRL provides a potential platform for wide acceptance and adoption of Semantic Web. Finally the knowledge representation and Semantic Rules on the Web were mentioned.

These initiatives offer alternatives of solution to different problematic situations as: provide a solution for cross-language information access problem in public and financial sector, offer a large-scale information extraction and integration infrastructure for supporting financial decision-making process, develop technology to monitor and add knowledge that are spread across mainstream and social media, the develop a technology that allows the management of increased of Linked Data collections and the management of quality of the links among Linked Open Data Sources, to mention a few.

In comparative with the previously mentioned proposals, the main idea of our approach is based in obtaining large sets of unstructured financial information with the aim of use it in the creation of financial knowledge after the appliance of natural language processing techniques to filter out such information and get more accurate information which will be offered to stakeholders through after its publication in Linked Open Data cloud.

3 Conceptual Model

Financial data, both public and undisclosed, is traditionally represented in a vast number of different formats, ranging from textual descriptions to XML documents,

[1] LOD2 project objectives, as of February 2011: http://lod2.eu/WikiArticle/Project.html

such as XBRL format for public financial disclosures [16]. Although the terms and meaning is well known to those financial analysts that are dealing with it on a daily basis, it is still an enormous obstacle for machines that results in big integration efforts. Activities such as moving data between systems, getting data from quarterly reports for fundamental analysis or cross-domain data breakout need substantial manual effort of analyst. Even though there are numerous systems and formats dealing with financial data, they still need proper transformations for enabling interoperability. The idea of having data that is only usable within specific system is becoming out of fashion. On the other hand FLORA proposes data-driven approach for financial data integration, by following Linked Data principles, where the data instead of systems is the center of the overall process.

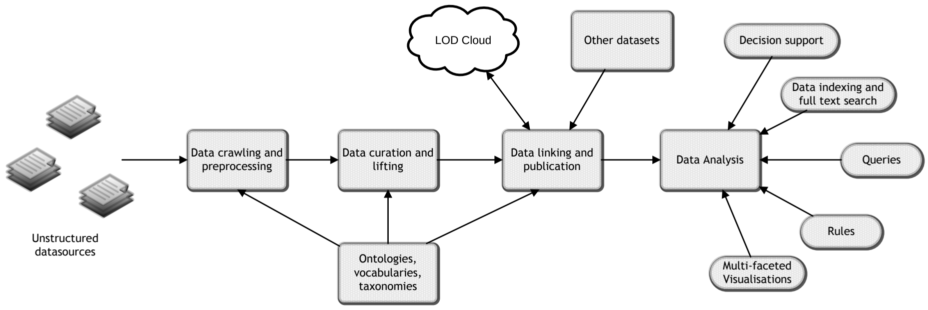

LOD Cloud
Other datasets
Decision support
Data indexing and full text search
Data crawling and preprocessing
Data curation and lifting
Data linking and publication
Data Analysis
Queries
Unstructured datasources
Rules
Ontologies, vocabularies, taxonomies
Multi-faceted Visualisations

Figure 1: Overview of the FLORA information extraction process

Transforming unstructured data into the structured form (as presented on Figure 1) comprise multiple steps that usually devise a similar high-level pattern that can be described as follows:

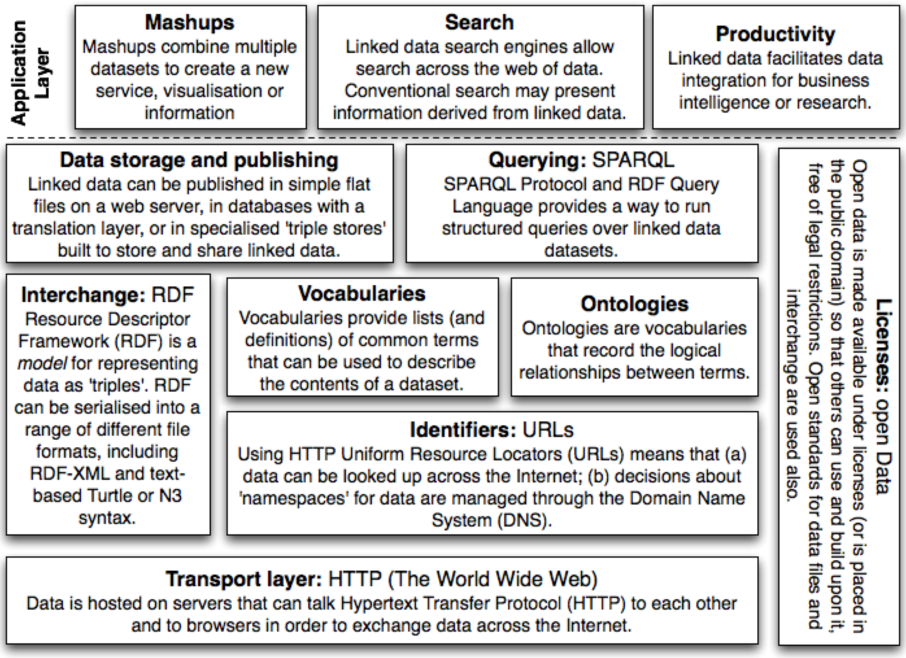
- Raw data is acquired and preprocessed in order to capture and extract the structure and isolate relevant data. Documents are cleaned and filtered according to input constraints. In case of documents having no defined structure, further NLP processes and ontology based information retrieval techniques, e.g. using GATE [17] for information retrieval.
- The data is lifted to the semantic form, using established ontologies and vocabularies that describe the dataset's domain. At this point, user-assisted data curation might be needed in order to improve the overall quality of data.
- Data can be published and interlinked with other datasets on the basis of using the same ontologies or describing same concepts. Linking and locating corresponding concepts across different datasets can be done automatically or semi-automatically, following approaches of LIMES [18] or SILK [19] projects.
- For improved data discovery, an indexing service is constantly traversing whole dataset facilitating full text search and results ranking.

Once published, the access to LOD dataset is realized by exposed SPARQL endpoint for data querying and HTTP access for data navigation. As it might be useful for lightweight data browsing and analysis, more sophisticated use cases might need the access to the whole RDF dataset. Such analysis services running on top of the dataset may provide further features, such as multi-faceted financial data visualizations (e.g. financial cockpit), decision support and data mining algorithms execution.

Adding reasoning capabilities on top of the RDF dataset will further improve of the data, as any data inconsistency can be easily detected and contradicting data removed or corrected.

4 Linked Data infrastructure for data-driven integration

Publishing results from previous steps in a form of Linked Data (and Linked Open Data in case of data with open license) forms an important part of the overall architecture. Most important building blocks of the Linked Data stack are show in the Figure 2. FLORA project covers all infrastructure blocks for making data available in the LOD cloud, following best practices in data publishing [20].



Mashups Mashups combine multiple datasets to create a new service, visualisation or information

Search Linked data search engines allow search across the web of data. Conventional search may present information derived from linked data.

Productivity Linked data facilitates data integration for business intelligence or research.

Data storage and publishing Linked data can be published in simple flat files on a web server, in databases with a translation layer, or in specialised 'triple stores' built to store and share linked data.

Querying: SPARQL SPARQL Protocol and RDF Query Language provides a way to run structured queries over linked data datasets.

Interchange: RDF Resource Descriptor Framework (RDF) is a *model* for representing data as 'triples'. RDF can be serialised into a range of different file formats, including RDF-XML and text-based Turtle or N3 syntax.

Vocabularies Vocabularies provide lists (and definitions) of common terms that can be used to describe the contents of a dataset.

Ontologies Ontologies are vocabularies that record the logical relationships between terms.

Identifiers: URLs Using HTTP Uniform Resource Locators (URLs) means that (a) data can be looked up across the Internet; (b) decisions about 'namespaces' for data are managed through the Domain Name System (DNS).

Transport layer: HTTP (The World Wide Web) Data is hosted on servers that can talk Hypertext Transfer Protocol (HTTP) to each other and to browsers in order to exchange data across the Internet.

Licenses: open Data Open data is made available under licenses (or is placed in the public domain) so that others can use and build upon it, free of legal restrictions. Open standards for data files and interchange are used also.

Elements of a Linked Open Data Stack - 2nd May 2011. CC BY-SA-NC
Draft sketch by Tim Davies (@timdavies / tim@practicalparticipation.co.uk) for IKM Working Paper on Linked Open Data for Development. Comments welcome. Search 'linked open data stack' on http://www.opendataimpacts.net for latest version.

Figure 2: Building blocks of the Linked Data stack[2]

[2] Elements of the Linked Open Data stack, rev 3, May 2011 http://www.practicalparticipation.co.uk/odi/2011/05/whats-in-the-linked-open-data-stack/

Combining the data from different datasets is possible by reusing the same vocabularies and ontologies for defining financial statements. Establishing links between corresponding concepts allows for augmenting financial reports with contextual information. In certain cases, such "data fusion" may result in the inference of the new knowledge, not explicitly stated before. Therefore the value of FLORA is higher that only the sum of its data.

For the data annotation, established financial ontologies are analyzed in order to be reused. Due to the broad scope of financial aspects covered by FLORA an upper ontology is considered to bridge different domains in the financial area. Further step is to provide necessary mappings to the concepts from to the core LOD dataset, DBpedia [21]. On top of the LOD stack, FLORA will provide services for data exploration and analysis, tailored to the financial domain.

7 Conclusions and Future Work

This article presents a complex, unified process of transforming unstructured financial data into an interlinked, navigable knowledge base for financial information management and information discovery. We described the system that is using ontology-based information extraction and data annotation for extracting relevant data that is further interlinked with appropriate LOD datasets. The whole underlying data infrastructure serves as a basis for providing services for data exploration, querying, discovery and visualizations.

In the broader sense this work will facilitate the financial data reuse and integration by following established data publication techniques. As presented in this paper, Linked Data-based approach adopted by FLORA is slowly becoming de-facto standard for structured data publication.

In the future work we aim at implementing envisaged system architecture and evaluate results based on such metrics as data quality and accuracy of the information extraction. We are also working on developing market surveillance and financial accounting use cases based on the FLORA results.

8 Acknowledgments

This work was supported by the Spanish Ministry of Science and Innovation under the project FLORA (TIN2011-27405).

References

1. Ciccotelloa, S.C. & Wood, R.E. An investigation of the consistency of financial advice offered by web-based sources, Information Systems, 10(1-4), pp. 5-18, 2001.
2. Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, 5(3), 1-22

3. Belleau, François; Nolin, Marc-Alexandre; Tourigny, Nicole; Rigault, Philippe & Morissette, Jean: Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics, Vol. 41 , Nr. 5 (2008) , S. 706-716.
4. O'Riain, S., Harth, A., & Curry, E.: Linked Data Driven Information Systems as an Enabler for Integrating Financial Data. (A. Yap, Ed.) Information Systems for Global, 239-270. 2011, IGI Global.
5. Momtchev, V., Peychev, D., Primov, T., & Georgiev, G. (2009). Expanding the Pathway and Interaction Knowledge in Linked Life Data. International Semantic Web Challenge, 2009.
6. David F. Huynh, David R. Karger, and Robert C. Miller. 2007. Exhibit: lightweight structured data publishing. In Proceedings of the 16th international conference on World Wide Web (WWW '07). ACM, New York, NY, USA, 737-746.
7. Oren, E. Delbru, R. Decker, S. Extending Faceted Navigation for RDF Data. Proceedings of the International Semantic Web Conference (ISWC06). Athens, Georgia. 2006.
8. Declerck, T., Krieger H. U., Thomas S. M., Buitelaar P., O'Riain S., Wunner T., Maguet G., McCrae J., Spohr D., & Montiel-Ponsoda E. Ontology-based Multilingual Access to Financial Reports for Sharing Business Knowledge across Europe, Internal Financial Control Assessment Applying Multilingual Ontology Framework, Chaper 4, 67-76, 2010.
9. Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., & McCrae, J. Challenges for the multilingual Web of Data, Web Semantics: Science, Services and Agents on the World Wide Web, 11(2), 63-71, 2012.
10. Grčar, M., Häusser, T., & Ressel, D. FIRST-Large scale information extraction and integration infrastructure for supporting financial decision making. September 201.
11. Grobelnik, M. Fact Sheet: XLike - Cross-lingual Knowledge Extraction. January 2012
12. Michael Hausenblas LOD2 Creating Knowledge out of Interlinked Data http://lod2.eu/WikiArticle/Project.html Retrieved February 20-2012
13. Klyne, G., & Carroll, J. (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation, Available at: http://www.w3.org/TR/rdf-concepts
14. Hausenblas, M.: Project Fact Sheet - The LOD Around-The-Clock (LATC) (2011).
15. Grosof, B. Opportunities for Semantic Web knowledge representation to help XBRL, Workshop on Improving Access to Financial Data on the Web. XBRL International and World Wide Web Consortium (W3C), 2009.
16. Roger Debreceny, Glen L. Gray, The production and use of semantically rich accounting reports on the Internet: XML and XBRL, International Journal of Accounting Information Systems, Volume 2, Issue 1, January 2001, Pages 47-74.
17. H. Cunningham, et al. Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science. 15 April 2011. ISBN 0956599311.
18. Axel-Cyrille Ngonga Ngomo and Sören Auer: LIMES – A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. Proceedings of IJCAI 2011.
19. Julius Volz, Christian Bizer, Martin Gaedke, Georgi Kobilarov: Silk – A Link Discovery Framework for the Web of Data . 2nd Workshop about Linked Data on the Web (LDOW2009), Madrid, Spain, April 2009.
20. Heath T., Bizer C., Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool.
21. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Zachary Ives: DBpedia: A Nucleus for a Web of Open Data, 6th Int'l Semantic Web Conference, Busan, Korea, 2007