Social Semantics and Linked Data for improving access to financial data on the Web

Alexandre Passant [1], Michael Hausenblas [1], Sean O'Riain [1], Kingsley Idehen [3], John G. Breslin [1,2], Stefan Decker [1]

[1] Digital Enterprise Research Institute, National University of Ireland, Galway, Galway, Ireland

[2] School of Engineering and Informatics, National University of Ireland, Galway, Galway, Ireland

[3] OpenLink Software, Burlington, Massachusetts, USA

A need for transparency

Financial data extracted from financial instruments represents the main information driver for financial analysis and business decision making. However, what decision makers look for but rarely get are the benefits of insight and opinion from their peers, from industrial commentators, and from the general discussions of customers and citizens who may have a multitude of opinions on corporate financial results, market position and actions. In addition, financial data is generally locked into closed-world data silos, due to a lack of interoperability between applications.

Combined together, Social Semantics and Linked Data can enable the desired features for financial data, making it more easily accessible and understandable. On one hand, Linked Data provides the interoperability framework that allows interoperability between sources that also publish their data using the same principles, wherever they come from. On the other hand, SIOC (Semantically-Interlinked Online Communities) allows social media applications to publish related discussions in a format that also permits better aggregation and interoperability.

In this position paper and during the workshop, based on our expertise on the Social Semantic Web, Linked Data and eBusiness, via DERI's Social Software Unit, the Linked Data Research Center and the eBusiness Domain Unit respectively, we will discuss how these two layers can be combined and we will suggest some future steps for this topic within the W3C.

Enabling data interoperability

Providing Financial Linked Data (FLD) on the Web

In order to provide financial data as Linked Data on the Web, hence enabling Financial Linked Data (FLD), the four principles of Linked Data can be applied to this data as follows:

- **Use URIs as names for things**: Each financial record will have its own URI, hence being uniquely named and referenceable by any application, whether it be an application used for browsing that data or one that is used for advanced analytics and reports;
- **Use HTTP URIs so that people can look up those names**: In addition, each bit of financial data should be accessible from the Web, without requiring additional protocols or tools, rather than a browser (either HTML browser or RDF browser);
- **When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)**: This would help one to get information about a financial report by simply dereferencing its URI (i.e. retrieving it using a software agent). While current practices imply querying a database (often closed) and then translating this information to a generic

format, this method will provide a uniform way to get to the desired financial information;

* **Include links to other URIs so that they can discover more things**: In addition, this last step will not only allow financial data to live beyond a closed world, but to be interlinked on a global scale, as we will detail in the next section.

Data items from profit and loss statements, as well as balance sheet items or source and application of funds statements could be interlinked on the Web (for example, using the XBRL Ontology) from various institutions, agencies and countries. Current applications can already be deployed to represent existing financial data as Linked Data. For example, the OpenLink Virtuoso Sponger can provide translations from existing financial data items into FLD. Some examples can be browsed at http://delicious.com/kidehen/xbrl_linkeddata_demo, and the following screenshot gives an example of such a translation.

Figure 1: Example of RDF statements generated from financial data

Existing efforts from the microformats community can also be bridged to FLD, including: the proposed banking and currency microformats; representations for financial calendaring information; and microformats embedded within Excel spreadsheets. By using dedicated services or GRDDL transformations, these too can be linked on the level of RDF semantics, and this could be discussed within the W3C.

Interlinking with other datasets

Figure 2: The Linking Open Data cloud, July 2009 (Source: Richard Cyganiak and Anja Jentzsch)

As of July 2009

While providing FLD from existing sources is the first step towards achieving better interoperability for financial data, linking it to other datasets from the Linking Open Data effort would be of a great benefit. One use case would be to systematically reuse GeoNames.org URIs to represent geolocational aspects of such FLD. GeoNames not only provides geographic representation features (such as coordinates), but it includes facts about some places (such as their populations) and it also provides links to DBpedia. As an example, by linking financial data about Galway City to its GeoNames URI (<http://sws.geonames.org/2964180/>), one can identify that this financial data is related to a city in Ireland, with a population of about 70,000 inhabitants (since this information is available in the related dereferenced file <http://sws.geonames.org/2964180/about.rdf>), enabling new possibilities for financial statistics.

Advanced statistics then become possible, such as ordering financial data by population, country, etc. For example, expanding the previous use case, this could be used for comparing the public budgets of all Irish localities with between 50,000 and 100,000 inhabitants. Relevant datasets include statistical data from riese as well as various sources of governmental data, an effort currently being pursued in both the USA and in the UK.

Enabling better transparency

Achieving transparency for financial data

In various contexts (B2B, B2C, intra-organisational, etc.), enabling transparency of discussions is needed to augment one's understanding and to provide more insight into the underlying financial assets being discussed. Such transparency can be achieved:

- By letting people comment on and interact with the FLD provided on the Web (via comments related to each financial data snippet in grant proposals, budget reports, etc.); and
- By providing this information in a machine-readable way, using the aforementioned Linked Data principles as applied not just to the financial record but to the related discussions.

Enabling open access to related discussions

Whether they are internal to a corporate environment or available as part of a community of interest on the Web, aggregated social content and discussions can be used to augment one's understanding of financial data. So far, these discussions are being stored in dedicated applications that act as closed-world data silos: government databases, blogs, content-management systems, etc.

The SIOC (Semantically-Interlinked Online Communities) and FOAF (Friend Of A Friend) formats can be used to deliver aggregate sets of Social Semantic data for indexing by semantic query engines, or for further analysis of emergent trends and public sentiment (e.g. with respect to government budgetary reports, grant proposals etc.). SIOC provides a unified model for representing authors' contributions (blog posts, wiki pages), while FOAF is used to represent personal identity and some characteristics of social networks. By combining both, we can (for example) identify that a particular contribution comes from an employee of a certain funding agency, thereby providing a first level of trust with respect to related financial discussions.

In addition, these models have not just been theoretically defined, but are already widely deployed on the Web. Since this information is also used by search engines such as SearchMonkey to enrich the presentation of their results, it also provides a way, when searching for financial assets on the Web, to directly identify the number of related discussion threads and people involved in this thread, giving an idea of the "hot topics" in a particular domain.

Such discussions can also benefit from vocabularies for semantic tagging, such as Common Tag. This allows users to group together their conversations regarding the aforementioned financial items, enabling better interoperability between FLD and Social Semantics, and therefore producing a better understanding of both.

An example vocabulary

The following picture represents a graph of the kind of facts that could be represented by combining FDL (on the left side of the picture, in yellow, using a dedicated ontology) and Social Semantics (on the right, using SIOC and FOAF).

Figure 3: Interlinking social data and financial information
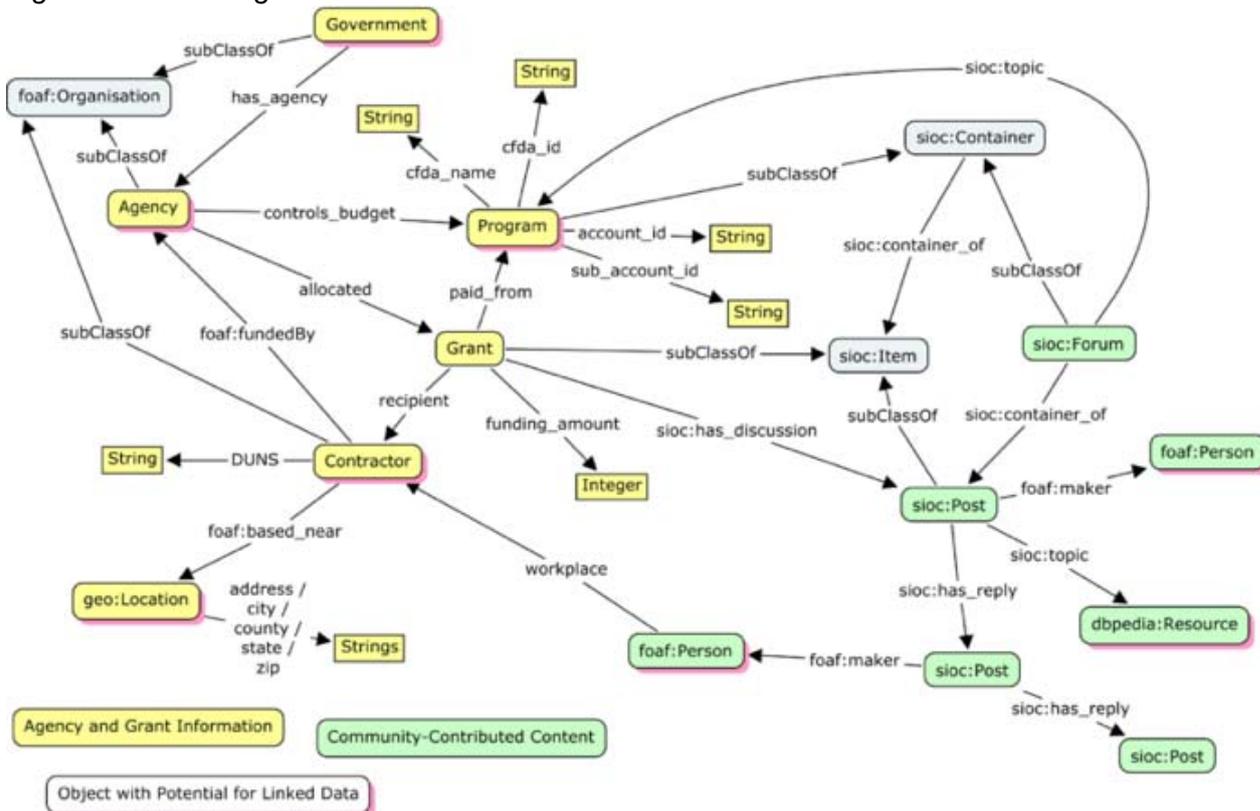
Financial data such as grants, project funds, etc. can be linked with user contributions in the form of associated threaded discussions on different topics, all to be exposed as Linked Open Data using custom financial schemas plus common schemas like SIOC and FOAF, from across a number of websites. The picture illustrates how the data within one financial site, that publishes funds information (in yellow) and allows citizen feedback (green), could be modeled and interlinked.

We can imagine how multiple financial amounts, account holders, fund recipients, funding topics etc. could be linked together by expanding this picture and by using this common representation to link the data together. Also, it is not just across financial data providers that financial recipients or funding areas could be linked, as many of the concepts shown (with shadows) could be linked to Linked Open Data sources for further background knowledge: company profiles, government departmental information, more data related to the location of particular funding recipients (for measuring impact, finding related projects, etc.).

This can also be achieved by augmenting the HTML pages describing financial information or allowing discussions with RDFa tags saying "this is a financial statement", "this comment is referring to this project", "this is a reply to this" or "this person said this about this financial balance" in a way that is reusable and can be linked to similar pieces of data on other pages or sites.

Next steps within the W3C

During the workshop, and in conjunction with other proposals, we would like to discuss some of the next steps that the W3C could lead in this context. In particular, we suggest that the following topics be addressed in an upcoming XG or WG:

- Carry out a state-of-the-art review of existing formats that are used to describe financial information on the Web, investigating possible mappings between these formats on the level of semantics;
- Determine the relevance of existing datasets from the LOD cloud in the context of linking them to financial information;
- Create a set of outreach documents regarding the potential of FLD on the Web; and
- Discuss issues relevant to FLD, such as privacy, security and trust.

Acknowledgments