

Challenges Ahead for Converging Financial Data

[Edward Curry](#)

[Digital Enterprise Research Institute \(DERI\) National University of Ireland, Galway](#)

[Andreas Harth](#)

[Institut für Angewandte Informatik und Formale Beschreibungsverfahren Karlsruher](#)

[Institut für Technologie](#)

[Sean O'Riain](#)

[Digital Enterprise Research Institute \(DERI\) National University of Ireland, Galway](#)

Introduction

Consumers of financial information come in many guises from personal investors looking for that value for money share, to government regulators investigating corporate fraud, to business executives seeking competitive advantage over their competition. While the particular analysis performed by each of these information consumers will vary, they all have to deal with the explosion of information available from multiple sources including, SEC filings, corporate press releases, market press coverage, and expert commentary. Recent economic events have begun to bring sharp focus on the activities and actions of financial markets, institutions and not least regulatory authorities. Calls for enhanced scrutiny will bring [increased regulation](#) and [information transparency](#)

While extracting information from individual filings is relatively easy to perform when a machine readable format is utilized (for example, using [XBRL, the eXtensible Business Reporting Language](#)), cross comparison of extracted financial information can be problematic as descriptions and accounting terms vary across companies and jurisdictions. Across multiple sources the problem becomes the classical data integration problem where a common data abstraction is necessary before functional data use can begin.

Within this paper we discuss the challenges in converging financial data from multiple sources. We concentrate on integrating data from multiple sources in terms of the abstraction, linking, and consolidation activities needed to consolidate data before more sophisticated analysis algorithms can examine the data for the objectives of particular information consumers (for e.g. competitive analysis, regulatory compliance, or investor analysis). We base our discussion on several years researching and deploying data integration systems in both the web and enterprise environments.

Financial Data Ecosystem

Financial Information Providers

Prominent providers of public domain financial information is the Securities and Exchange Commission (SEC) which through their EDGAR web site makes freely available a wide range of personal and company filings. Data from the SEC ranges from information about executives reporting the sale of equity in their companies (Form 4) to detailed annual reports (Form 10-K). Filings are in the older SGML format, free-text (HTML and PDF), or more

recently XBRL format. A wide range of other governmental or intergovernmental organisations publish data in various formats. For example, central banks use [RSS-CB](#) for publishing currency exchange rate data. The United Nations, World Bank, Eurostat, and the OECD are working towards a standard data format for publishing statistical information in [SDMX, Statistical Data and Metadata Exchange](#) format. Finally, there is considerable (financial) information, regarding companies and their executives available in Wikipedia which DBpedia publishes in the web standard [RDF \(Resource Description Format\)](#).

Financial Information Consumers

A large number of information consumers have varying degrees of interest in financial data. The integration and augmenting of financial information is of significant benefit for financial and business analysis as the following three use cases illustrate. The principle behind each is that XBRL information extracted from SEC filings receives a semantic metadata lift allowing the data to then be published in RDF format. The [Rhizomik](#) project, OpenLinks [sponger](#), and [Dave Raggerts](#) recent efforts are examples of first offerings for mapping XBRL to RDF. Once in RDF it can be linked and augmented with additional information from other financial data extracted from sources such as those previously mentioned as a consolidated financial 'mash-up'.

Competitive Analysis

An analyst looking at performing a competitive analysis with an appropriate source selection could work with a mash-up that associated both the financial figures and comments from an analyst summary call. Important comment from corporate officers or other external commentators could then be associated with financial facts allowing a more complete analysis which could otherwise have been easily overlooked from consideration in decision making.

Regulatory Compliance

The relatively new branch of [forensic economics](#) is an area which would benefit from the availability of financial, government and regulatory linked data. Interested in spotting patterns or conditions that suggest fraud, the economists look at the benefits of criminal activity as guides to spotting data footprint that suggest the activity is taking place. Identifying suspect activity sooner would be possible if regulation required companies to make relevant data available in a format that could easily be linked and aggregated for investigation. Financial regulators and fraud investigations could leverage such linked data within their forensic tools to improve their capacity to monitor regulator compliance or early fraud detection.

Investor Analysis

Individual and institutional investors alike spend considerable time looking at company and investment fund returns for investment potential. Financial results are often purposely presented in a convoluted manner, making direct comparisons difficult even for the financial professionals. Integrating the financial data in a common format with semantic markup would make similar financial instrument mapping and comparison easier. Overall it would increase the level of transparent and provide better information that would assist with the "What's the best performing fund?" or "What are the better values shared to buy?" type of financial analysis.

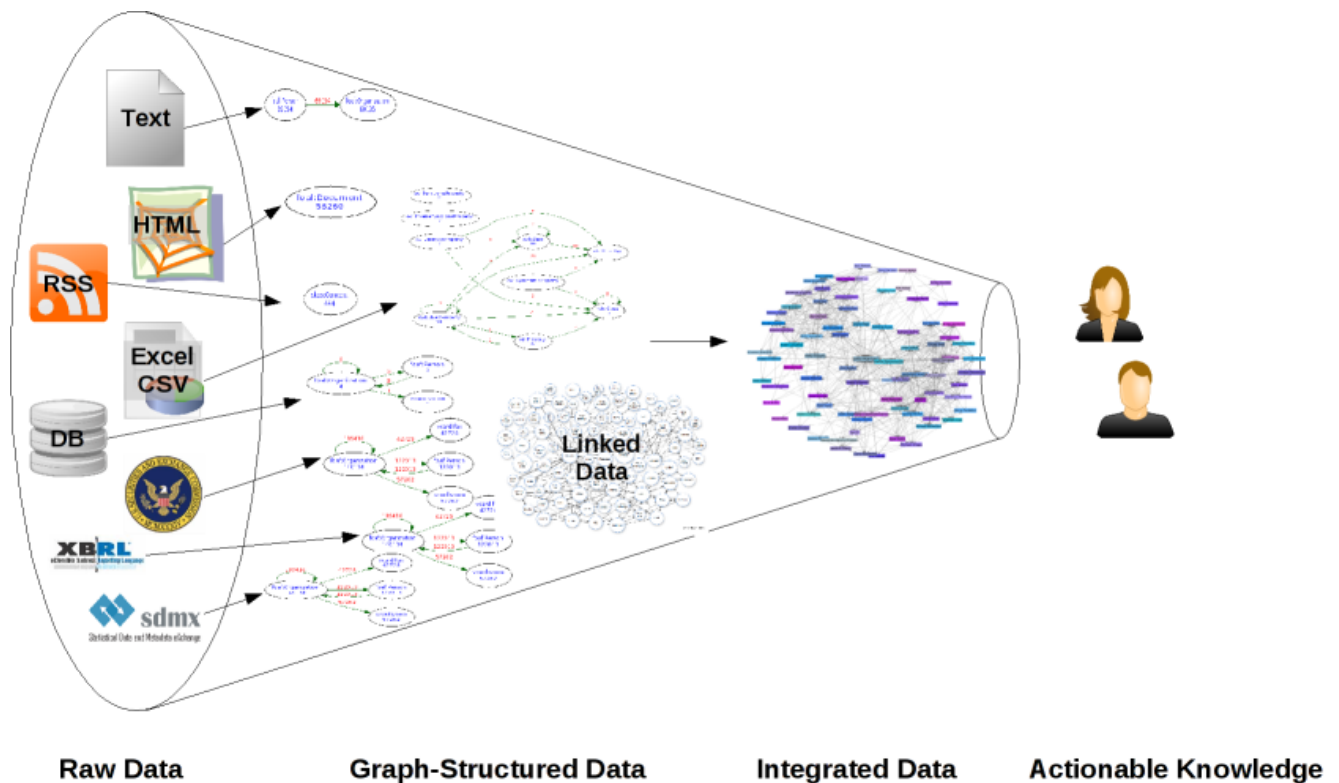
Converging Financial Data

Financial data in electronic form comes in various formats on a continuum of structure, ranging from unstructured text, to highly structured XML data, to graph-structured data in RDF. The goal is to allow for users to analyse the underlying datasets and derive actionable knowledge from the aggregated and integrated data.

Our data integration approach comprises several stages:

1. Lifting data sources to a common format, in our case RDF (Resource Description Format)
2. Integrating the disparate datasets (original sources plus reading existing linked data resources) into a holistic dataset by aligning entities and concepts from disparate sources
3. Optionally running analysis algorithms on the integrated data
4. Enabling interactive browsing and exploration of the integrated data or results of algorithmic analysis over the data

The different data processing steps from the raw data via graph-structured and integrated data to the end user are illustrated in the following figure.



Commonly Used Formats

In the following section we give a characterisation of the type of data we identified for addressing the outlined use case scenarios. We focus here on publicly available datasets; however, our approach is equally applicable to specialized in-house sources and formats.

Text: The traditional press is publishing financial news mainly in textual

format (e.g. Financial Times). In addition, sites such as seekingalpha.com make raw transcripts of investor calls available. Another source of text is certain quarterly SEC filings (e.g. form 10-K, Annual Report)

Hypertext: Websites carry data (in HTML); typically these sites are generated from relational database backends.

Spreadsheets: CSV files, Word documents, or PowerPoint presentations
Especially in corporate environments a large portion of data exchanged is encoded in MS Office documents

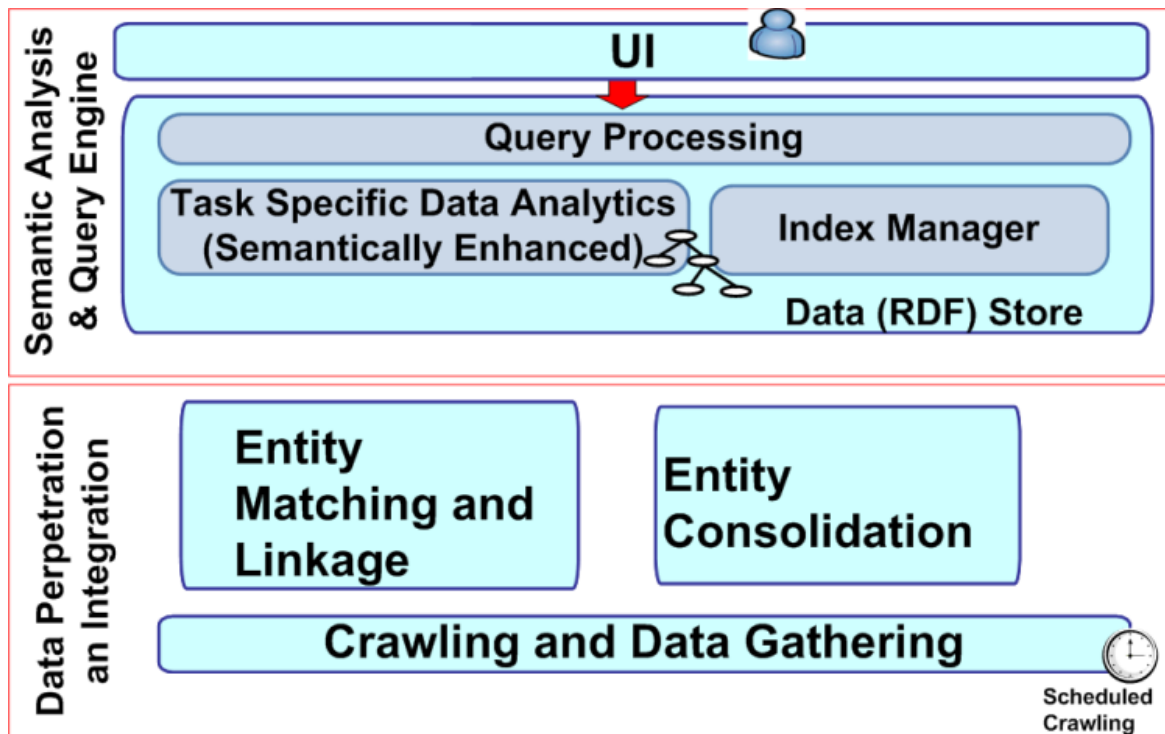
Structured Data: Typically structured data adhering to a fixed schema is encoded and published in XML. XBRL falls under this category as well as statistical information is being published either in CSV (comma-separated value) format (e.g. by Eurostat) or SDMX.

Graph-Structured Data: Sources such as DBpedia contain graph-structured RDF data describing companies, people, and the like. Financial information is typically not very structured - for example, there are manually maintained figures of assets or net income for the previous year. Other sources for financial information (especially about tech startups) include CrunchBase (for which a [version in RDF exists](#)).

The main obstacle preventing easy integration into a holistic dataset is that these types of data are in different formats, separate, and not interlinked. For a classification of data according to the types of information encoded (e.g. time series, taxonomic data) we refer the interested reader to Ben Shneidermans [paper](#).

Integration System Architecture

A typical system architecture for a data integration system is depicted in the following figure. The system consists of two components: a data preparation and integration phase to convert and bring data from different sources into a common format and a query and user interface module that operates over the integrated dataset. We will discuss each of the components later on.



Integrating data from multiple sources provides a common data platform from which search, browsing, analysis, and interactive visualisation can take place. Consolidation in semantic web terms leads to an aggregated source view or a coherent graph amalgamated, 'mashed up' from potentially thousands of sources, where an entity centric approach can provide a powerful single view point allowing information filtering and cross analysis. The key challenge for any information system operating in this space is the need to perform a semantic integration of structured and unstructured data from the open Web and monolithic data sources such as XML database dumps and large static datasets. This can be achieved using a hybrid data integration solution which amalgamates the data warehousing and on-demand approaches to integration.

From this integration emerges a large graph of RDF entities with inter-relations and structured descriptions of entities: archipelagos of information coalesce to form a coherent knowledge base. Entities are typed according to what they describe: people, locations, organizations, publications as well as documents; entities have specified relations to other entities: people can work for companies, people know other people, people author documents, organisations are based in locations, and so on.

Data Preparation and Integration

This step involves the process of collecting and integrating data from a plethora of sources in a multitude of formats such as native RDF, RSS streams, HTML, MS Office, PS, and PDF documents. This information can be located across multiple information systems such as databases, document repositories, government sites, company sites, and news sites in order to collect this information web crawlers or database wrappers can be employed.

In order to avoid having the knowledge contribution of entities split over numerous instances the system will need to connect sources that may describe the same data on a particular

entity. Within one of our case studies the results of analysing the connections between people and organizations from SEC filings (Form 4) identified 69,154 people connected to 80,135 organizations. The same analysis performed on database describing companies produced 122,013 people connected to 140,134 organizations. Once collected the base dataset needed to be enrich and interlinked using entity consolidation (a.k.a. object consolidation). In order to avoid having the knowledge contribution of entities split over numerous instances the system will need to connect sources that may describe the same data on a particular entity.

Semantic Analysis and Query Engine

Semantic analysis within these systems will be closely tied to the purpose of the system such as fraud detection, competitive analysis, profit projections, etc. While the exact algorithm used for analysis will vary, a number of common services will be needed to assist in the examination and query of the data including local index creation and management, distributed query processing, and runtime consolidation of results. Data based upon item consolidation rather than the XML document bases approach of XBRL provides not only insight but underpinned by linked data backbone (semantic web technology) allows a means of data querying that conventional tools do not. SPARQL, the semantic query language allows queries/questions such as the following to be asked:

- Show Texan start-ups with products in the same area that are funded by the same VC
- Project yearly GDP growth of China (only published once a year) by using quarterly company reports and show previous GDP growth and projected GDP growth in a time plot
- Who are competitors of HP?

Analysis of the consolidated data sources could also be performed by communities or groups of analysts who could be employed to annotate the data further to raise irregularities, for example. An example of this "crowd sourcing" approach to data analysis is the Guardian's site that asks readers to [tag and report UK MPs expense claims for further investigation](#).

Data Integration Challenges

There are a number of challenges to address when integrating data from different sources. We classify these challenges into four groups: text/data mismatch, object identifiers and schema mismatch, abstraction level mismatch, data accuracy.

Text/Data Mismatch

A large portion of financial data is described in text. Human language is often ambiguous - the same company might be referred to in several variations (e.g. IBM, International Business Machines, and Big Blue). The ambiguity makes cross-linking with structured data difficult. In addition, data expressed in human language is difficult to process via software programs. Since we aim for posing structured queries over an integrated, holistic view over the entirety of data, one of the functionality of a financial data integration system is to overcome the mismatch between documents and data.

Object Identity and Separate Schema

Structured data is available in a plethora of formats. Lifting the data to a common data format is thus the first step. But even if all data is available in a common format, in

practice sources differ in how they state what essentially the same fact is. The differences exist both on the level of individual objects and the schema level. As an example for a mismatch on the object level, consider the following: the SEC uses a so-called Central Index Key (CIK) to identify people (CEOs, CFOs), companies, and financial instruments while other sources, such as DBpedia (a structured data version of Wikipedia), use URIs to identify entities. In addition, each source typically uses its own schema and idiosyncrasies for stating what essentially the same fact is. Thus, Methods have to be in place for reconciling different representations of objects and schema.

Abstraction Levels

Financial data sources provide data at incompatible levels of abstraction or classify their data according to taxonomies pertinent to a certain sector. For example, the SEC provides a taxonomic classification of sectors which is incompatible with other commonly industry sector classification schemes. Since data is being published at different levels of abstraction (e.g. person, company, country, or sector), data aggregated for the individual viewpoint may not match data e.g. from statistical offices. Also, there are differences in geographic aggregation (e.g. region data from one source and country-level data from another). A related issue is the use of local currencies (USD vs. EUR) which have to be reconciled in order to make data from disparate sources comparable and amenable for analysis. Finally, countries have sometimes vast differences in legislation on book-keeping Euro region regulator indicators may not therefore be directly comparable with indicators from US-based regulators. The lack of agreement on even a common EU GAAP standard will hold up the publishing of XBRL in a standardized way and its multilingual accounting term translation. This will also be a problem for other legislators and jurisdictions when it comes to transcending national boundaries.

Data Quality

Data quality is a general challenge when automatically integrating data from autonomous sources. In an open environment the data aggregator has little to no influence on the data publisher. Data is often erroneous, and combining data often aggravates the problem. Especially when performing reasoning (automatically inferring new data from existing data), erroneous data has potentially devastating impact on the overall quality of the resulting dataset. Hence, a challenge is how data publishers can coordinate in order to fix problems in the data or blacklist sites which do not provide reliable data. Methods and techniques are needed to; check integrity, accuracy, highlight, identify and sanity check, corroborating evidence; asses the probability that a given statement is true, equate weight differences between market sectors or companies; act as clearing houses for raising and settling disputes between competing (and possibly conflicting) data providers and interact with messy erroneous web data of potentially dubious provenance and quality. In summary, errors in signage, amounts, labelling, and classification can [seriously impede the utility of systems operating over such data.](#)

Conclusion

The single largest barrier to developing sophisticated semantic analysis methods and algorithms for use in financial analysis, fraud or regulatory activities is the ability to integrate multiple financial data sources into a more holistic and transparent data set.

In this paper we have highlighted the data integration challenges facing the provision of transparent financial information and where semantic standards and approaches can be of direct benefit. An architectural approach is presented based upon previous case study experiences along with the remaining challenges in the areas. We feel that leveraging financial source data in the manner described will help distil actionable knowledge that can be used for driving business decisions and policy.

Acknowledgements

The work presented in this paper has been funded in part by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).