Exploiting Linked Data in Financial Engineering

Vivian Lee, Masatomo Goto, Bo Hu, Aisha Naseer, Pierre-Yves Vandenbussche, Gofran Shakair, Eduarda Mendes Rodrigues

Exploiting Linked Data in Financial Engineering

Vivian Lee†, Masatomo Goto, Bo Hu, Aisha Naseer, Pierre-Yves Vandenbussche, Gofran Shakair, Eduarda Mendes Rodrigues

Fujitsu
vivian.lee@uk.fujitsu.com

Abstract. In this paper, we report on a recent initiative that exploiting Linked Data for financial data integration. Financial data present high heterogeneity. Linked Data helps to reveal the true data semantics and "hidden" connection, upon which meaningful mappings can be constructed. The work reported in this paper has been well-accepted at several public events and conferences, including the 26th XBRL conference, involving the realisation of the XBRL (eXtensible Business Reporting Language) prototype called HIKAKU, which means "comparison" in Japanese. It demonstrates our approach to exploit the power of Linked Data in enhancing flexibility for data integration in the financial domain.

Keywords: Financial Engineering, Financial Reporting, XBRL (eXtensible Business Reporting Language), Governmental Data, Linked Data, Data driven platform

1 Introduction

The Ki2NA platform is a data-centric platform whose applications are driven by interactions with Linked Data in a unified way. It is developed with shorten the development cycles and thus the time to the market as the main design principle. To achieve this goal, adopts a universal approach to treat data is the key. The Linked Data basis of the platform is designed to ensure the universal data treatment is applied. This is evident from the following design decisions: 1) the underlying data storage is heterogeneous and can be assembled for optimisation according to the specific needs of each application? 2) metadata services from LOD4all provide data discovery services? 3) modelling constructs such as the data-cube observation model are transferable across multiple application domains? 4) APIs and query language provide consistent access to data and associated services that drive applications.

The purpose of the HIKAKU application is twofold: i) the business value of the LOD4all service (a Linked Open Data Repository being developed in the scope of Ki2NA) through the development of a financial prototype featuring XBRL data and ii) the benefit of using Linked Data technologies to combine XBRL data with open data to enhance financial analysts' experience.

In order to motivate our research and development, we firstly outline current issues in the financial domain and how this impacts on corporate financial reporting. This is then followed by a review of related work and an exploration of the potential of Linked Data to address these issues. The current HIKAKU application is also discussed in detail, together with the key user benefits. In conclusion, we provide the future roadmap and proposed extensions to this prototype.

1.1 Financial Reporting

Financial reporting is the communication of financial information about an enterprise to the external world/public. Thus far, the corporate financial reporting practice has been questioned on two counts. On the one hand, the current financial reporting framework was largely shaped during and immediately after the first industrial revolution in response to the emergence of corporate form, stock market, and the regulation of accounting and auditing practices. Due to the intricate nature of financial instruments, the complexity of financial reporting is inevitable. Obscure legal terms designed to avoid stating responsibility in a black-white fashion have aggravated the magnitude of complexity. As a result, it becomes increasingly challenging for investors, who are not professionally educated/trained, to distil the messages conveyed in such reports [5]. Financial report tooling should, therefore, not only assist authoring, but also facilitate comprehension.

On the other hand, since the latest technological revolution, the corporate structure has undergone fundamental changes that start to render the conventional reporting approach less desirable for modern companies. Some of such fundamental changes include the difference between the market value and book value of company assets, the establishment of off-shore financial centres and off-shore financing channels, far-reaching globalisation, etc. Evidence of rigid constraints and a general lack of flexibility translate into financial reporting that is based on conventional auditing and accounting methodology, which starts to struggle in faithfully reflecting the performance of companies. This is particularly true for social media and e-commerce businesses, whose true value can only be revealed using data other than balance sheets, profit/loss statements, and cash flow statements.

However, there have been some efforts to address such problems. The recent advances in ICT, especially in the automation of data integration, has already been embraced by the business world and paved the way of for financial reporting mechanism that is well aligned with the emerging business practice.

1.2 Related Work

One important benefit of applying Linked Data principles to the financial domain is the increased data inter-operability across multiple financial systems and financial instruments [7]. The financial industry has long acknowledged the necessity of aligning different data providers[2]. This is evident in the Financial Reporting arena, where international collaboration is already firmly established. For instance, the US Securities and Exchange Commission (SEC) has mandated that by 2014 all financial entities should adopt the eXtensible Business Reporting Language (XBRL). XBRL is a family of XML-based global standards, enabling the automated exchange of business information through machine-interpretable tags. XBRL taxonomies are constantly revised to reflect the regulations/rules. Apart from the US SEC, major players of XBRL include US FFIEC, Japan Financial Supervisory Agency , Bank of Japan, Tokyo Stock Exchange, UK HMRC and many other European and Asian financial regulators.

Despite the benefits of adopting established standards, these can also become barriers, hindering the adoption of new technologies that are not fully compatible with existing ones. In recent years, the value of semantically enriched XBRL has been recognised [6]. The semantics of XBRL constructs are to benefit significantly

from the ongoing Financial Industry Business Ontology (FIBO) initiative, aiming to provide an industry-wide generic ontology [1]. To the best of our knowledge, however, there are not yet any deployed XBRL tools that take advantage of Linked Data offerings and release the full power of XBRL. Leaving non-technical issues aside, the lack of large scale adoption can be attributed to several reasons. Firstly, full-scale conversion from XML-based XBRL instance and taxonomy documents to genuine Linked Data format (i.e., a graph data model coded in the Resource Description Framework, RDF) is not straightforward. Naive conversion can lead to badly distorted RDF graphs, missing relations, and knowledge loss [4]. Secondly, without properly populated RDF models, the advantage of semantic inferences cannot be fully appreciated. Thirdly, XBRL leverages a large number of operational knowledge, defining how financial figures are computed. However, RDF, a knowledge representation paradigm that is rooted in Description Logic and specialises in non-numeric-conceptual modelling, may find difficulties in reasoning and reconstructing such knowledge. Finally, XBRL presents well-defined semantics for a well-defined purpose. Portraying complete pictures of companies requires data that are not annotated with XBRL. Using a single model for such a diverse mission may result in modifying or extending the XBRL models and thus raise operating costs and incur doubts among the established XBRL community.

2 HIKAKU-A Company Comparison Application

Better financial data integration allows both professional analysts and amateur individual investors to understand the performance of a particular company more efficiently. HIKAKU provides capabilities for seamlessly linking heterogeneous financial data and trace their provenance, thus enriching current financial reporting practice with the Linked Data computing paradigm [3]. It addresses the shortcomings of the current state of the art tools, in terms of data timeliness, data completeness, and data consumption. A key differentiator is that we compile data across multiple companies to offer performance comparison instead of isolated figures from individual ones. In the mean time, we hide the unnecessary financial reporting complexity and just present the data to ordinary investors/analysts in an easy-to-understand fashion. In addition, we expand the scope of where data is solicited from, linking not only conventional financial reporting information sources, but also Linked Open Data in order to construct performance summaries that go well beyond balance sheets.

Currently, the HIKAKU application utilises three main sources XBRL as a freely available and global mechanism for exchanging business information? Linked Open Data such as DBPedia and Crunchbase, which offer general information about companies? and finally, company sentiment extracted from news media (e.g. NY Times), which provides up-to-date information of a particular company that is attracting lots of media attention due to their performance or internal affair. The application aims to address the following inefficiencies in the current offering:

1. Semantic and syntactical discrepancies abound in individual reports, even with authoring support.

2. Lack of tooling for analysis across multi-sources reports.

3. Analysis is largely single faceted whereas financial requirements become increasingly multiple faceted.

4. Financial reports with release and audit latency fail to give timely results.

5. Failure to provide contextual information, helping a user to understand a company's performance.

2.1 Linked Data Driven Platform

The HIKAKU application is powered by the Ki2NA platform, which is a flexible, large-scale data processing platform that builds applications capable of delivering value to the user through knowledge-enabled networks (interconnected data systems supported and driven by Linked Data). The Linked Data model is used both at the data layer to deliver the end-user application, and also on the process-layer to model data flows between processes and define the interaction between all platform components.

The architecture and interaction of different processes are driven by the flow of information/data. Linked Data allows data, which is stored in the platform, to easily be further enriched and new knowledge to be produced by integrating and connecting existing information. Fig. 1 shows the Ki2NA architecture, consisting of three tiers and eight layers.

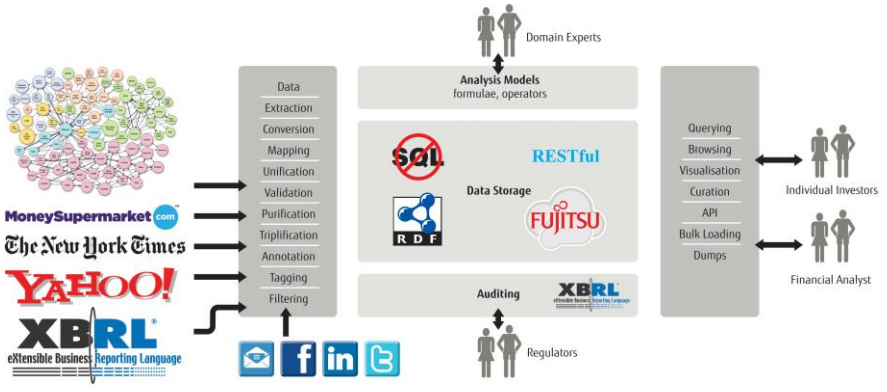

Fig. 1: Ki2NA Platform Architecture

The architecture can be divided into three tiers: a typical presentation tier, a (business) logic tier and a data tier. Whereas the presentation tier provides the communication interface between applications and clients on the one hand and the underlying system on the other hand, the logic tier executes system and application specific (business) logic. The data tier provides all necessary data management functionalities.

In the data tier, the platform deals with data heterogeneity and aims to handle a wide variety of data formats and storage solutions through a uniformed access

mechanism to the above logic tier. This flexibility is a prerequisite of the HIKAKU financial use case. HIKAKU mashes up data from a variety of sources, integrating them into financial reports and therefore greatly assisting the decision-making process.

2.2 Financial Regulation and Governmental Data

As already mentioned in section 1.2, XBRL is a standard that enables the communication and exchange of business information especially in the financial sector. Reporting information such as financial figures is usually provided in the form of an XBRL instance, and each semantic definition is defined in a taxonomy. XBRL includes a family of XML-based standards e.g. NewsML and MathML. Since each XBRL taxonomy definition gives additional semantics, it is considered as a specialised data type. In addition to XBRL documents, there are also three other types of identifiers that are considered crucial when reconciling financial information from different sources to a single individual or company entity.

These identifiers are:

- Legal Entity Identifier (LEI) is a global system of identifiers designated by the Financial Stability Board (FSB) in an effort to overcome the current fragmented systems of firm identifiers and to create a common identifier for financial institutions. The LEI code can be retrieved from https://www.ciciutility.org.
- Central Index Key (CIK) is a number given to an individual or a company by the United States Securities and Exchange Commission. The CIK code can be searched from http://www.sec.gov/edgar/searchedgar/cik.htm.
- Ticker Symbol is an abbreviation used uniquely to identify publicly traded shares of a particular stock on a particular stock market. Depending on the companies (in the HIKAKU case, major IT companies in the U.S.), the ticker symbol can be found under the two stock exchanges, namely New York Stock Exchange (NYSE) and NASDAQ.

The CIK code is normally embedded inside XBRL reports for identifying companies that have registered and filed disclosures with the U.S. Securities and Exchange Commission (SEC), and it does not cover corporations outside the scope of the SEC. In order to broaden company performance comparisons into a global view, it is important to use LEI as the ultimate global identifier for company entities. Moreover, mapping ticker symbols to LEI is also essential in order to extract accurate stock price data from stock markets for particular companies, enabling companies' stock prices to be brought into financial reports. The architecture of mash-up, extraction, and integration governmental public data can be shown in Fig. 2:

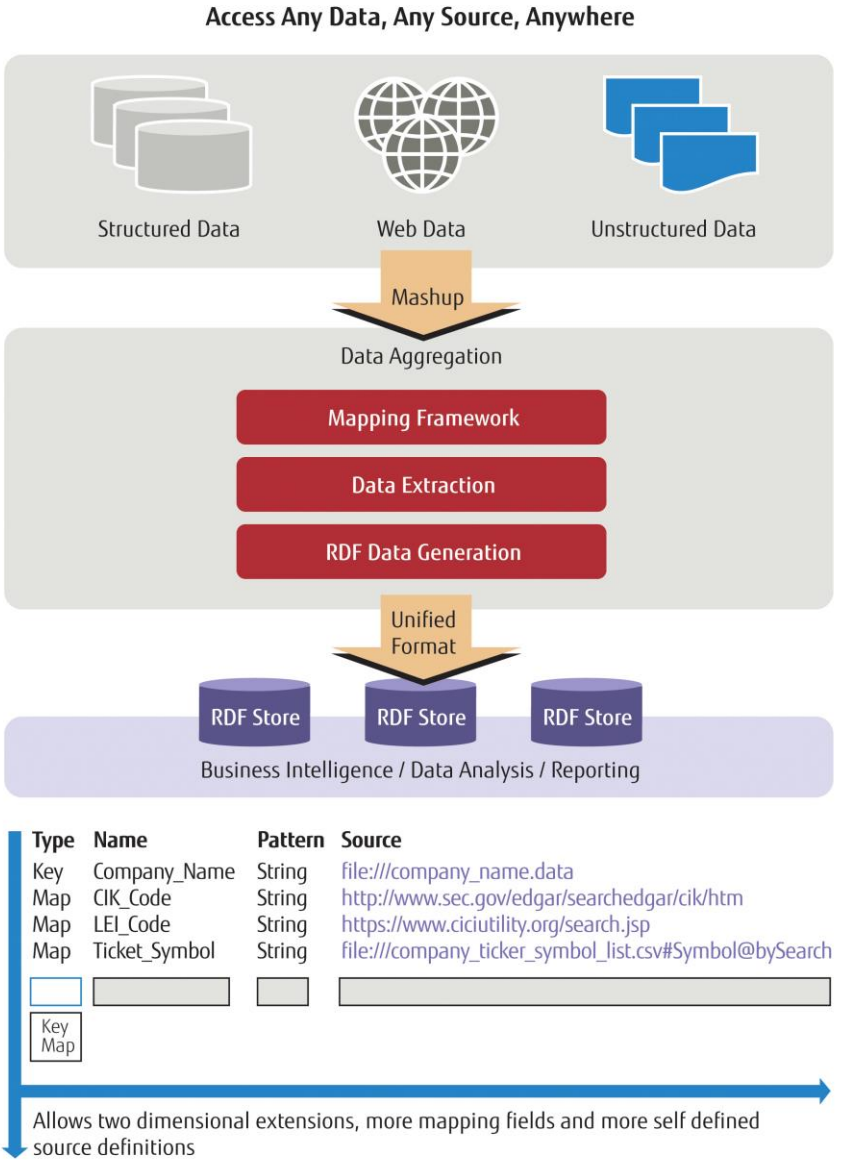

Fig. 2: Data Aggregation Architecture

Within HIKAKU, we elected to use company names as the key for searching and indexing other identifiers from the respective websites, enabling these different identifiers to be reconciled. This approach did raise some challenges. For example, MICROSOFT CORP can also be written as Microsoft Corporation while Yahoo can also be called YAHOO INC or Yahoo! Inc. etc., which ruled out exact string matching. In the initial prototype, the solution involves manually fixing the

problematic names into a consumable format for each website on top of string similarity based algorithms. This is based on the observation that identifier alignment has to be curated by human experts, due to the accuracy requirement of the application domain and the lack of tools to explicate full semantics of the name labels.

After extracting and mashing up data from heterogeneous sources, the pre-processed data are in temporary CSV format and are ready for converting to RDF. HIKAKU provides two conversion options, the W3C standard RDB to RDF Mapping Language (R2RML - see http://www.w3.org/TR/r2rml/) conversion and a SPARQL query construction that leverages a use-specified SV column to RDF type mapping. The second method treats the W3C RDF Data Cube Vocabulary (http://www.w3.org/TR/2013/CR-vocab-data-cube-20130625) as the basic ontology model, mixed with domain specific vocabularies. The HIKAKU data pre-processing is illustrated in the Figure 3:

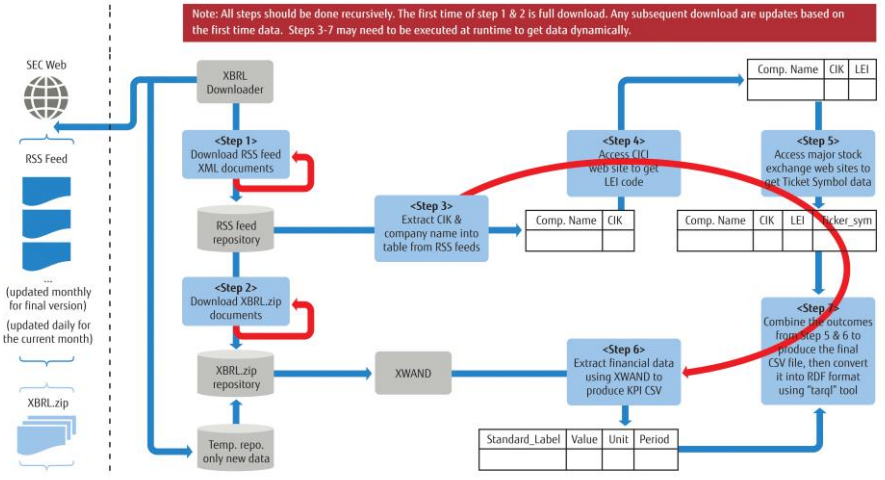

Fig. 3: Data Processing Flow

2.3 Linked Open Data

Numeric figures in financial reports can be unintentionally and/or deliberately manipulated to present a false and misleading image of a company (c.f. the recent scandalous acquisition of Autonomy by Hewlett-Packard [3]). Even though such incidents cannot be entirely avoided, incorporating other sources of data can promote informed decision making and minimise potential risks and mistakes due to a lack of transparency. Typical public data that can complement financial reports include stock price data, digitised mass media contents, mailing list/online bulletin board systems (BBS) and the emerging social media. The use of public data is based on the following observations: 1) Official financial reports are normally published quarterly (aka 10Q) and yearly (aka 10K). They normally lag behind media coverage of major events concerning the subject company; 2) Official financial reports tend to be

summarising over a long period of time and may not reflect the stock price fluctuation at given time points in that period; and 3) More and more users or customers are starting to share their opinions about a product, a service or a company in channels other than customer services. We witnessed customers' sentiments (boycott/promotional activities) campaigned through news articles that have strongly affected a company's performance in the real world.

Choosing the most appropriate data sets not only can impinge the scope and accuracy but also system performance in terms of query execution time and memory consumption. The data sets being considered by HIKAKU can be grouped into several categories.

LOD data sets. When comparing financial performance, one needs to cover a wide spectrum of aspects of corporate entities. Even though data sets published on LOD cloud may not explicitly bear a "finance" label in their titles, they can be of great assistance in discovering relevant information, which is otherwise hard to access. The following LOD data sets have been chosen at this stage.

- DBPedia is used for general company data such as logo and location, as well as KPIs such as the number of employees, revenue assets, equity, net income, etc. We also compose a company's subsidiaries out of DBPedia data. The quality of data varies. Hence, data collected from DBPedia are cross-validated and complimented with those from other sources.
- Linked CrunchBase is a free database of technology companies, people, and investors. From CrunchBase, we retrieve such data as funding, competitors, company acquisitions, main people in charge, and products. It allows us to identify similar and comparable company profiles. For instance, companies with similar size, products and competitors can be grouped together and recommended for performance comparison.
- Linked NewYork-Times, as of 13 January 2010, has published approximately 10,000 subject headings as linked open data8. It complements company profiles compiled from the above two sources.

Mass media. Mass media coverage provides more up-to-date information of a company and, on many occasions, actually leads/misleads the market on a wide and profound basis, e.g. causing stock briefly to plunge or rise. HIKAKU reflects this through the sentiment analysis of new articles. As an example, the NY Times is used to gather news articles and commentaries about a specific company, with Yahoo! Finance APIs used to get the stock price. Sentiment scores are then computed with off-the-shelf tools/services and accumulated to reflect a company's mass media image.

3 Implementation of the HIKAKU Prototype

The Ki2NA platform uses a triple store for data storage to meet the requirements of the HIKAKU application. This triple store contains all the heterogeneous data-sets represented according to the Linked Data guidelines (using HTTP URIs, RDF, and linking related resources together), making it easy to query for data from different sources using a single interface.

For end-users, the comparative results of the selected companies are centred on the web interface where users can decide whether to drill down into individual KPIs or roll-up to acquire an overview. The KPIs are extracted from the previously described heterogeneous data sources. The user interface uses a colour code to demonstrate the integration of such heterogeneous data to compose more complete information about companies, while at the same time serves as a legend for provenance information. For example, a company's description comes from DBPedia LOD data-source, whereas the CIK code is sourced from the U.S. Security Exchange Commission. These two KPIs will show up in the UI with distinct colours (DBpedia KPI in yellow and US SEC in blue).

There are three distinct types of visual analytics: bar chart view, time-line view and table view. Users can navigate between the bar and the time-line chart using the upper tabs while the table view is always visible. Users' interactions with the table view are reflected and synchronised on the bar chart and time-line views.

The current prototype also provide a feature for specifying new KPIs (i.e. new financial concepts) by arbitrarily combining pre-defined KPIs in mathematical formulae. The system will then calculate the formula on-the-fly and display the result as a normal concept. Such feature also demonstrates the strong data integration capability by enabling combine KPIs from different data sources.

4 Benefits

During HIKAKU research and development, we have acquired a better understanding of the socio-technical considerations, and our findings will be potentially valuable to researchers and practitioners planning similar initiatives. In particular, we highlight the following two aspects:

Early-adopter: In many application domains, the existing technologies create resistant old "habits". Finding the cutting point and early adopters becomes crucial. We decided to centre our Linked Financial Data application on financial reporting, and more specifically XBRL. The Financial Reporting community has already established a consensus on a common language for computer enabled data exchange? and has reached out to major technology vendors for help. Building our solution around established XBRL expertise therefore ensures a vast population of potential adopters. Moreover, as the beneficiaries of XBRL range from authorities/regulators, to financial institutes, and even to individual investors, the selected target group presents a wide diversity to allow us to implement different roll-out strategies.

Positioning: HIKAKU is designed as a value-added service on top of XBRL, consuming and unlocking the value of the latter. On the other hand, HIKAKU offers functionalities that are not available should XBRL be used in an isolated fashion. We acknowledge that XBRL and many other semantic-less XML-based languages will still serve as data exchange technologies in specialist domains. Linked Data technology is to assist rather than replace such technologies, offering better functionality and a better user experience. Meanwhile, Linked Data technology is not the answer to all the data integration challenges faced by the XBRL community (for instance, unambiguous alignment with universal identifier). It is always desirable to communicate any disadvantages fully at an early stage. During the conceptual design and development of the HIKAKU application, this strategy has helped us focus on true add-on values.

5 Conclusions and Future Plans

The public data enhancements, including Linked Open Data, social media, XBRL, LEI, and ticker symbol, allow us to extend conventional financial reports with: i) better comparison through semantic alignment, ii) support of unconventional, on-the-fly KPI definitions, and iii) timely access to external data other than the official financial reports.

The crux of our future work lies in improving the current prototype to reflect feedback from various public events and extended quantitative studies of employed technologies in the financial domain. More evaluations have already been scheduled. With the XBRL community already becoming the initial adopter, reach-out to other financial communities can be facilitated through the XBRL "channel".

Future plans also include an improved HIKAKU financial dashboard with features such as 1) time series analysis (e.g. "Fujitsu's performance since the latest tsunami disaster.", "is Fujitsu performing better this year comparing with other Japanese companies? "), 2) a data set explorer and quality-checker (e.g. "FT.com with a quality score of 0.8 and a trust score of 0.75"), and 3) user-defined KPI validation (e.g. "combining sentiment score and total number of employees does not make sense.").

References

1. Bennett, M.: Fibo: Best practice in big data. Journal of Banking Regulation (3-4), 255–268 (2013)
2. Burdick, D., Hernández, M.A., Ho, H., Koutrika, G., Krishnamurthy, R., Popa, L., Stanoi, I., Vaithyanathan, S., Das, S.R.: Extracting, linking and integrating data from public sources: A financial case study. IEEE Data Eng. Bull. 34(3), 60–67 (2011)
3. Florian, B., Martin, K.: Linked Open Data: The Essentials - A Quick Start Guide for Decision Makers. edition mono/monochrom, Vienna, Austria (2012)
4. García, R., Gil, R.: Triplificating and linking xbrl financial data. In: Proceedings of the 6th International Conference on Semantic Systems. pp. 3:1–3:8. ACM (2010)
5. Omberg, T., Sakr, S., Sethi, S., Murrell, B., Popken, A.: Key factors shaping financial reporting: The decade ahead. Deloitte (2011)
6. O'Riain, S., Curry, E., Harth, A.: Xbrl and open data for global financial ecosystems: A linked data approach. International Journal of Accounting Information Systems 13(2), 141 – 162 (2012)
7. O'Riain, S., Harth, A., Curry, E.: Linked Data Driven Information Systems as an Enabler for Integrating Financial Data. In: Yap, A. (ed.) Information Systems for Global Financial Markets: Emerging Developments and Effects, chap. 10, pp. 239–270. IGI Global (2012)