Accepting the XBRL Challenge with Linked Data for Financial Data Integration

Benedikt Kämpgen[1], Tobias Weller[1], Sean O'Riain[2], Craig Weber[3], and Andreas Harth[1]

[1] Institute AIFB, Karlsruhe Institute of Technology, Karlsruhe, Germany
kaempgen@kit.edu,tobias.weller@student.kit.edu,harth@kit.edu
[2] Digital Enterprise Research Institute, National University of Ireland, Galway
sean.oriain@deri.org
[3] Financial Intelligence, LLC, Los Altos, CA
cpatax@sbcglobal.net

Abstract. Analysts spend a disproportionate amount of time with financial data curation before they are able to compare company performances in an analysis. The Extensible Business Reporting Language (XBRL) for annotating financial facts is suited for automatic processing to increase information quality in financial analytics. Still, XBRL does not solve the problem of data integration as required for a holistic view on companies. Semantic Web technologies promise benefits for financial data integration, yet, existing literature lacks concrete case studies. In this paper, we present the Financial Information Observation System (FIOS) that uses Linked Data and multidimensional modelling based on the RDF Data Cube Vocabulary for accessing and representing relevant financial data. FIOS fulfils the information seeking mantra of "overview first, zoom and filter, then details on demand", integrates yearly and quarterly balance sheets, daily stock quotes as well as company and industry background information and helps analysts creating their own analyses with Excel-like functionality.

Keywords: #eswc2014Kampgen

1 Introduction

Analysts play a crucial role in the functioning of equity markets. Besides the actual analysis, e.g., comparing key performance indicators (KPIs) such as the Gross Profit Margin between companies, analysts spend a disproportionate amount of time with data curation, i.e., identifying, gathering and preparing data [4] and pursue to minimise time spent on tedious curation tasks. The Extensible Business Reporting Language (XBRL)[4] is an XML format for financial information that is more amenable to automatic processing than traditional financial information representations such as PDF, HTML and text documents.

[4] http://www.xbrl.org/Specification/XBRL-2.1/REC-2003-12-31/XBRL-2.1-REC-2003-12-31+corrected-errata-2013-02-20.html

Still, XBRL does not solve the problem of data integration – e.g., of company background information, balance sheets, stock quotes – for a holistic view on companies [8]:

- XBRL uses XML that is difficult to understand and process, e.g., due to an extension with link bases for referencing across documents [3].
- Automatically deriving information from XBRL is difficult since formal semantics are limited [12, 10]. Relationships between financial concepts, such as "SalesRevenueNet" and "Revenues" in the U.S. Generally Accepted Accounting Principles (US-GAAP), are only textually described.
- Financial information from different XBRL documents often cannot be compared since accounting and regulatory organisations do not align their taxonomies of financial concepts; new versions, e.g., of US-GAAP, lack backward compatibility; and XBRL allows publishers to define their own concepts.
- Gathering information about a company is difficult since there are no unique company identifiers across different reporting sources[5] and relationships between companies are obscure.
- Other finance-related Open Data such as stock quotes and background information are published using different data models.

Literature has proposed the use of Semantic Web technologies, but has not evaluated the benefit in financial case studies [12, 5, 1]. In this In-Use paper, after we describe a concrete XBRL scenario (Section 2), we present the Financial Information Observation System (FIOS) with the following contributions (Section 3):

1. For standardised data access, FIOS models XBRL and non-XBRL as Linked Data using the RDF Data Cube Vocabulary and other standard vocabularies.
2. FIOS integrates financial data using entity consolidation for background information, multi-company KPI, and cross-data-sources KPI analysis.
3. For intuitive and explorative analyses, FIOS provides SPARQL templates with visualisations, a Linked Data browser and a self-serve OLAP interface on top of a triple store.

For evaluation, we describe a case study implementing and applying FIOS for financial analysis (Section 4) and derive lessons learned (Section 5). We describe related work in Section 6 and conclude in Section 7.

2 Scenario: Integrating XBRL Data for Company Performance Analysis

In this section, we present a financial data analysis scenario inspired by the Annual XBRL Challenge organised by XBRL US: an investor wants to assess companies based on corporate XBRL data from the U.S. Securities and Exchange Commission (SEC) that since 2009 requires more than 8,000 U.S. companies

[5] `http://sunlightfoundation.com/sixdegrees/`

traded on the stock market to provide financial statement information such as quarterly and yearly balance sheets in the XBRL format to the SEC Edgar Database. The investor would find useful several analyses:

Background information analysis, e.g., looking at company information from different sources such as the address, the founding date and the industry.

Multi-company KPI analysis, e.g., comparing KPIs over time for several companies such as the stock market price for companies from the same industry.

Cross-data-sources KPI analysis, e.g., comparing values from heterogeneous datasets such as the Earnings per Share from yearly balance sheets with prices per share from electronic stock quotes as well as Total Assets published using the US-GAAP version 2009 and version 2011.

We can derive the following requirements: Answering above queries requires integration of different entities such as yearly and quarterly balance sheets using different taxonomy versions of US-GAAP, company and industry background information from Wikipedia/DBpedia and daily stock quotes. Since there is no standard way to model and publish finance data, data from the the SEC Edgar Database, from Wikipedia/DBpedia and from the Yahoo! Finance Web API need to be published as Linked (Open) Data and continuously extracted and stored (**Requirement 1**). To make the analyst understand and trust data that the system presents, the query interface needs to fulfil Shneiderman's information seeking mantra "overview first, zoom-in, details on demand" (**Requirement 2**). Also, since analysts can not use complex query languages, the analysis system needs to help them creating their own analyses, if possible with Excel-like functionality (**Requirement 3**).

3 Financial Information Observation System (FIOS)

We now describe our approach using Linked Data. As illustrated in Figure 1, FIOS' architecture is separated into two types of components, the *offline ETL components* and *online analysis components*, described in the following.

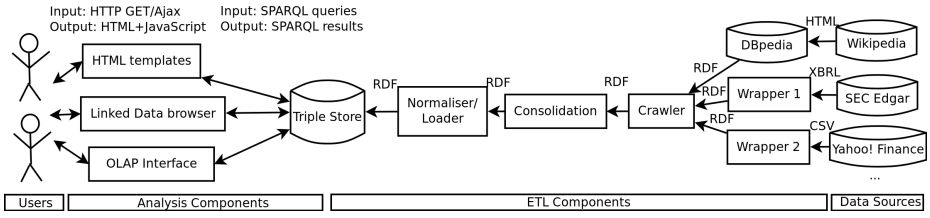

Fig. 1. Flow diagram illustrating architecture of Financial Information Observation System (FIOS)

3.1 Identification and Acquisition of Distributed Data

FIOS uses the Linked Data principles to identify and retrieve relevant information spread across different web servers.

Identification: We uniquely name things/entities with URIs: XBRL balance sheets from the SEC Edgar Database, including their taxonomies, and daily stock quotes from the Yahoo! Finance Web API; companies and industries listed by the SEC, Yahoo! Finance and Wikipedia/DBpedia. The SEC uniquely identifies the companies using a Central Index Key (CIK, e.g., Mastercard has "1141391"), Yahoo! Finance uses Ticker symbols (e.g., "MA" for Mastercard). Wikipedia uses their own non-standardised identifiers that are typically based on the name of the company, e.g., "Mastercard".

If looked up, URIs provide useful information in RDF either by originating from Linked Data providers or created by wrappers around data sources not publishing Linked Data. Wrappers mint new URIs and internally transform available information about such entities in the data source to RDF. Since the actual URIs are application-specific, in the following, we simply abbreviate URIs from our selected data sources using intuitive namespaces (abusing CURIE syntax): `edgar` for entities from the SEC Edgar Database, `yahoo` for Yahoo! Finance Web API and `dbpedia` for Wikipedia. Table 1 shows example mappings between things/entities, data sources with useful information about these entities and URI identifying those entities in Linked Data.

Table 1. Example mappings between things/entities, data sources and URIs

Entity	Original data source	URI
Company Mastercard	Mastercard from DBpedia	`dbpedia:Mastercard`
Company Mastercard	SEC Edgar company Mastercard with CIK 1141391	`edgar:cik/1141391#id`
Company Mastercard	Yahoo! Finance company Mastercard with Ticker MA	`yahoo:ticker/MA#id`
Balance sheet	XBRL document from SEC Edgar[6]	`edgar:archive/1141391/0001193125-11-207804#ds`
Stock Quotes table	Stock Quotes table from Yahoo! Finance Web API[7]	`yahoo:archive/MA/2010-12-01#ds`

Acquisition: For a holistic view on selected companies from the SEC Edgar Database, FIOS regularly looks up their URIs and checks the RDF (as well as the RDF of linked entities) for new data.

3.2 Modelling and Linking of Finance Data

To allow FIOS to use the retrieved information, we model financial data reusing existing Linked Data vocabularies and link entities from different sources.

[6] `http://www.sec.gov/Archives/edgar/data/1141391/000119312511207804/0001193125-11-207804-xbrl.zip`

[7] `http://ichart.yahoo.com/table.csv?s=MA&a=11&b=01&c=2010&d=11&e=01&f=2010&g=d&ignore=.csv`

Modelling: Whereas there are well-adopted vocabularies for all kinds of metadata, e.g., SKOS, FOAF and the DBpedia ontology, there is no standard way to represent XBRL data as Linked Data [5, 1, 12]. XBRL distinguishes instance and taxonomy documents. An *XBRL instance document* (also called "filing") contains financial facts with a numeric value and a unit such as USD. A fact has a context, e.g., describing the issuing company such as Mastercard, the time period of a financial fact (often, a quarter of a year or full fiscal year) and so-called segment information, e.g., allowing to specify subgroups of financial facts, e.g., that facts are published for subsidiary members. Most importantly, a fact specifies a certain disclosed financial concept such as "Total Assets". Financial concepts are taken from *XBRL taxonomy documents*. XBRL taxonomies can be standardised, e.g., the US-GAAP, and their concepts used across many instance documents. Also, companies may create their own taxonomies and financial concepts. Within taxonomies, concepts may be given additional information, e.g., labels, and may have relations to other concepts, e.g., "part of" relationships.

We model every XBRL instance and taxonomy as a multidimensional dataset, i.e., collection of facts with independent dimension variables and dependent measure variables, using a well-adopted Linked Data vocabulary, the RDF Data Cube Vocabulary (QB)[8] as follows: for any XBRL instance with taxonomy a multidimensional dataset (`qb:DataSet`) and data structure definition (`qb:DataStructureDefinition`) are created. For any single financial fact within an XBRL instance an observation (`qb:Observation`) is created with dimensions issuer, time period (`edgar:dtstart`, `edgar:dtend`), the financial concept (`edgar:subject`), segment and one decimal measure with a unit.

Similarly, stock quotes from Yahoo! Finance can be modelled using QB: every daily collection of values is a dataset, every stock quote contains as dimensions the company (`yahoo:issuer`), the date the value is valid and the stock quote type such as price at stock market opening (Open).

Linking: We use QB for the following reason: Given datasets contains observations with certain companies, certain financial concepts and certain periods in time, financial data integration boils down to identifying and consolidating equivalent dimensions and dimension values in multidimensional datasets.

See Figure 2 for an illustration of the linking between different entities or properties. Here, the fact of an XBRL instance document disclosing Total Assets (`edgar:vocab/us-gapp-2009-01-31#Assets`) and a Opening stock quote are linked via equivalent dimensions, e.g., `dcterms:date` / `ical:dtstart` and `edgar:issuer` / `yahoo:issuer`, and via equivalent dimension members, e.g., Mastercard `edgar:cik/1141391#id` / `yahoo:ticker/MA#id`.

Whereas time periods can easily be matched by comparing canonical representations of time, for linking between different URI for companies and financial concepts across data sources mappings need to be available. Entities or properties in RDF can explicitly be stated as equivalent via `owl:sameAs` or `owl:equivalentProperty` relationships between their URIs.

[8] `http://www.w3.org/TR/vocab-data-cube/`

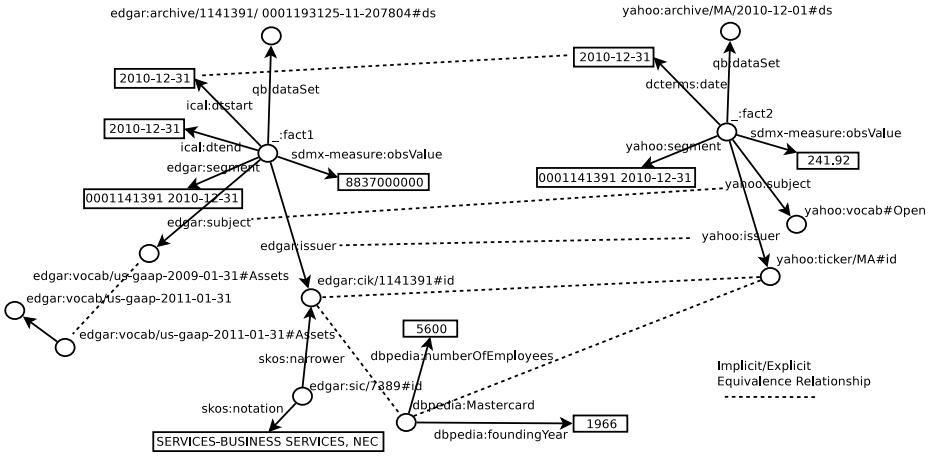

Fig. 2. Illustration of linking between Total Asset fact (top left), opening stock quote (top right), Mastercard in DBpedia (bottom center) in FIOS

To model finance related metadata, e.g., about companies and industries, we use widely-adopted Linked Data vocabularies, e.g., FOAF and the DBpedia ontology. The industry of a company can be represented using SKOS classification hierarchies, e.g., the SEC provides for companies the Standard Industrial Classification (SIC) hierarchy, e.g., SIC concept "SERVICES-BUSINESS SERVICES, NEC" `skos:narrower` "MASTERCARD INC".

3.3 Consolidation, Normalisation, Loading and Validation of Data

FIOS allows to pre-process and store acquired data for fast access as well as to check its quality.

Consolidation: Entity consolidation in FIOS – making explicit and merging all available information about an entity, so as access to that information is available independently from a specific distribution across sources – results in simpler queries and is only different from Hogan et al. [6] in that we also consider equivalent relationships between predicates such as dimensions of datasets.

Normalisation: Queries on our consolidated data are complicated by the fact that all entities described by FIOS use distributed namespaces. Resources from FIOS thus link to external servers with non-preprocessed data, confusing the user or complicating the application. Therefore, we mint dereferenceable URIs for all entities, including URIs in the predicate position, in an own FIOS namespace `fios`. For provenance reasons, we create `owl:sameAs` and `owl:equivalentProperty` links from FIOS entities to the original entities.

Loading: After pre-processing, data is loaded into a triple store that supports SPARQL 1.1 for analytical aggregate queries and is indexed for performance.

Validation: We use SPARQL queries for quality checks, e.g., validating QB integrity constraints (as described in the specification) or XBRL-specific integrity

constraints such as defined between financial concepts in XBRL calculation relationships.

3.4 Analysis of Integrated Financial Linked Data

Semantic Search engines (e.g., [6]) are too general for financial analysis and data analysis tools such as van Hage's and Kauppinen's SPARQL Package for R are too complicated for domain experts. FIOS uses three different kinds of interfaces for views on financial Linked Data.

SPARQL Templates with Visualisations, i.e., webpages that show results of SPARQL 1.1 queries on the triple store in visualisations, give a general overview of data in FIOS, e.g., number of datasets. Also, we create domain-specific reports about companies that require data integration. Templates can be parameterised with input by the analyst, e.g., a company identifier.

A **Linked Data Browser**, i.e., webpages of things/entities in RDF that show all ingoing and outgoing triples of a resource and allow follow-your-nose browsing from resource to resource, provides a more detailed view on any RDF data in FIOS.

An **OLAP Interface**, i.e., an intuitive and explorative data analysis method allows analysts to create own visualisations on multidimensional datasets. Since (parameterised) SPARQL templates have a fixed structure and Linked Data browsing does not aggregate triples, we use our approach [7] to evaluate OLAP operations using SPARQL on RDF reusing QB.

4 Implementing and Applying FIOS in a Case Study

We successfully submitted implementations of FIOS to the XBRL Challenge 2012[9] and 2013[10]. For evaluation, we now first describe the most current implementation of FIOS, then a case study applying FIOS to our scenario.

From the FIOS start page[11], we give information about the ETL process and an overview of available entities: publishing companies (`fios:issuer`/64 different values), valid time periods (`ical:dtstart`/234, `ical:dtend`/223, and `dcterms:date`/5,937) financial concepts (`subject`/3,781) and specific information (`segment`/58,395). From linked histograms, we see that most observations are from the time period between 2008 and 2013. Also, we see that we have quite evenly spread a number of observations for each company. Also, we get a good understanding of what financial concepts are published very often, e.g., `us-gaap-2009:Revenues`. Both FIOS ETL and analysis components run on a Virtual Machine with QEMU Virtual CPU version 0.12.3 with 2673.330 CPU MHz and 1GB memory and are described in the following:

[9] `http://xbrl.us/research/appdev/Pages/275.aspx`
[10] `http://xbrl.us/research/appdev/Pages/423.aspx`
[11] `http://fios.linked-data-cubes.org/FIOS_2_0/Queries/`

ETL components: For the `edgar` and `yahoo` namespaces, we have developed the SEC Edgar Wrapper[12] and the Yahoo! Finance Wrapper[13] using Google App Engines. Some information, e.g., XBRL calculation linkbases and footnotes currently are not considered, however, could be extracted and published as Linked Data to provide additional interesting information [9, 2].

Yahoo companies link to Edgar companies using a Ticker-to-CIK mapping provided by the Yahoo! Finance API. Edgar companies link to DBpedia companies via Freebase. Datasets from SEC and Yahoo! Finance are linked by manually stating the equivalence of dimensions, such as the company, the valid time period and the financial concept. In cases where structures of datasets are less similar, approaches for data warehouse integration could be applied [11].

We created a Java program fios-etl[14] containing separate components for crawling data, applying consolidation and normalisation algorithms to the collected data and loading the data into a triple store. As crawler, we used the Open Source software LDSpider (Stable Version 1.1e).

For each run, fios-etl automatically fills a seed list with selected companies and new balance sheets from where LDSpider starts to crawl. We selected company URIs from several industries, e.g., "finance, insurance and real estate" companies such as Visa and Mastercard. New balance sheet URIs are taken from an SEC RSS feed. For example, LDSpider would start crawling at the URI of Mastercard in Yahoo! Finance Wrapper that provides links to stock quote datasets from 1990-01-01 to today and `owl:sameAs` links to Mastercard in the Edgar Linked Data Wrapper. From Edgar Linked Data Wrapper, further `owl:sameAs` links to Mastercard in DBpedia and links to SEC balance sheets would be followed. We setup LDSpider to crawl with breadth-first strategy and a depth of the traversal of 3, with a maximum number of 10 URIs crawled per round per pay-level domain. Consolidation and normalisation algorithms we implemented as described for FIOS. Experiments with differently-sized datasets show that consolidation time increases exponentially with the number of equivalence statements, normalisation time increases linearly with the number of triples. Data was then bulk-loaded to an OpenLink Virtuoso Server v06.01.3127 running in Apache/2.2.14. For our case study, we run fios-etl daily during the XBRL Challenge 2013 submission time from 15 Feb 2013 to 27 Feb 2013 GMT. On average crawling, pre-processing and loading took 25min; loading can be done offline and could further be accelerated using differential loading. In total, we crawled 1,238,041 triples.

We created integrity constraints using SPARQL ASK queries that can be manually run, e.g., evaluating whether Earnings per Share for a company in fact is computed by the ratio of net income and outstanding shares. Since we have not found an automatic way of retrieving and validating integrity constraints, we have only implemented few checks.

[12] `http://edgarwrap.ontologycentral.com/`
[13] `http://yahoofinancewrap.appspot.com/`
[14] `https://code.google.com/p/fios-etl/`

Analysis components: The SPARQL Templates with Visualisations for overviews and domain-specific reports we implemented using the JavaScript library SPARK[15]. For some templates, especially the company template, users may need to wait several minutes before all results are displayed, due to large number of separately issued SPARQL queries. As Linked Data Browser we deployed the Open Source software Pubby. For the OLAP Interface we use the Open Source OLAP client Saiku and OLAP engine olap4ld[16]. The OLAP Interface shows long loading times due to large number of multidimensional elements such as financial concepts (3781) that need to be loaded in memory. In the remainder of this section, we show how FIOS fulfills the three requirements of our scenario.

4.1 Integrating Data Across Sources (Requirement 1)

We now describe four exemplary analyses integrating entities across data sources.

1) Background information analysis: The FIOS start page provides a link to analyse companies in a *SPARK company template*. After inserting the CIK for a company in the parameterised template, e.g., "1141391" for MASTERCARD INC, the user is presented with information from various sources, e.g., address and number of employees from Wikipedia, and various overviews of available KPIs from SEC Edgar Database and Yahoo! Finance Web API. Note, since companies from SEC, Yahoo! Finance and DBpedia are explicitly stated as equivalent in RDF and consolidated, we have one identifier for MASTERCARD INC that summarises all information from those data sources; queries do not need to consider equivalent links and thus are easier to write.

2) Multi-company KPI analysis: On the SPARK company template for a company, an overview of "Adjusted Closing Price" over time is given that interactively can be extended with companies from the same industry via the SIC classification as provided by SEC Edgar.

See Figure 3 for adjusted closing price for MASTERCARD INC and other companies in SERVICES-BUSINESS SERVICES, NEC (SIC) industry. We see that MASTERCARD INC stock quotes always have been higher than VISA INC and COMSCORE INC stock quotes and at the beginning of 2013 were at an all-time-high with over 500 USD per share.

3) Cross-data-sources KPI analysis: On the SPARK company template we also show an analysis taking into account "Earnings per Share" from SEC balance sheet and the "Opening Price per Share" from Yahoo! Finance stock market data. Earnings per Share is considered the single most important variable in determining a share's price, thus an analyst may be interested to check for an obvious correlation for a company.

In Figure 4, we return for each reporting end date the maximum Earnings per Share as published in quarterly or yearly balance sheets together with the maximum opening stock market price of values between the valid start and end data of the Earnings per Share financial ratio for MASTERCARD INC. Since

[15] `http://km.aifb.kit.edu/sites/spark/`
[16] `http://olap4ld.googlecode.com/`

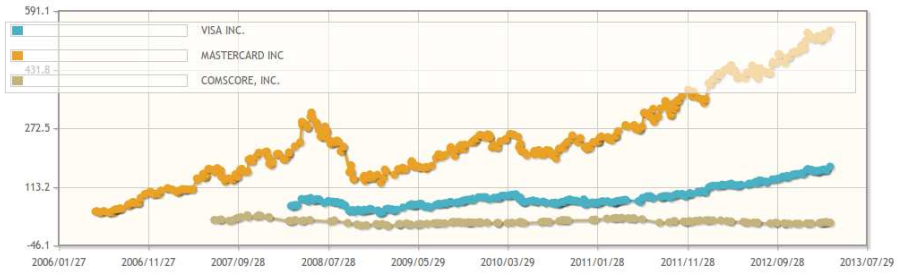

Fig. 3. Example multi-company KPI analysis of adjusted closing price for companies in industry SERVICES-BUSINESS SERVICES, NEC (SIC)

numbers are not normalised and SPARK visualisations would not allow several separate y-axes, it is difficult to see correlations in the figure. Another interesting analysis is the % rate of increase (decrease) for comparable periods, however, that was not easily doable since Edgar did not explicitly represent the sequence of balance sheets.

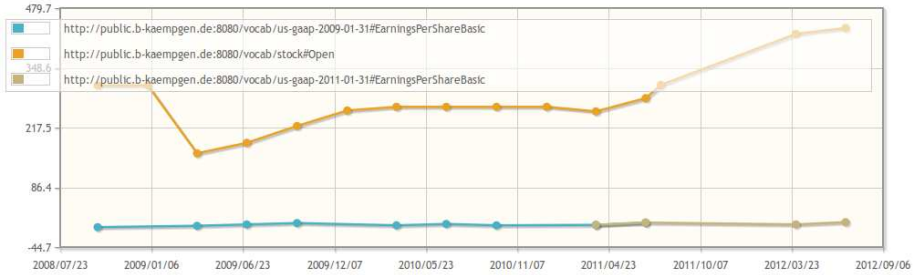

Fig. 4. Example cross-data-sources KPI analysis of Earnings per Share versus price per share for MASTERCARD INC

Note, since balance sheets and stock quote tables are integrated, it suffices to ask for specific values for the financial concept dimension `fios:subject` to query for financial concepts across the SEC Edgar Database and Yahoo! Finance.

As illustrated in Figure 5 for MASTERCARD INC, we can also query across different taxonomy versions: If we browse from the SPARK company template to a "Balance Sheet" template and click on "Total Assets", the company's Total Assets KPI over time is shown in a diagram from 2009 to 2012 although balance sheets from 2011 use a different US-GAAP taxonomy version. For that, Total Assets from US-GAAP-2009 and US-GAAP-2011 are stated as equivalent (either by consolidation or by adding UNION graph patterns to SPARQL queries). We see that MASTERCARD INC only twice has reduced its number of assets, at the end of 2009 (from 7.4B USD to 7.3B USD) and the end of 2010 (from 8.8B USD to 8.5B USD), for instance indicating reduced profits and a re-organisation.

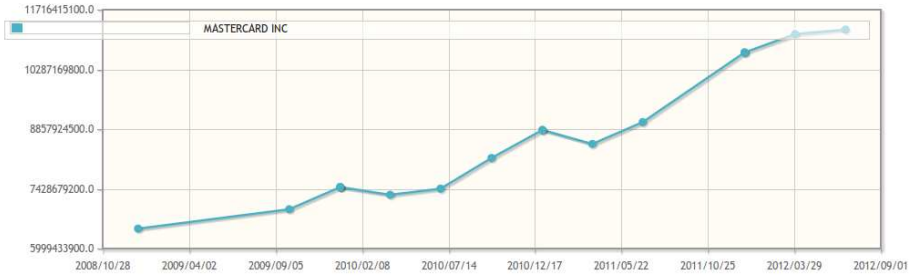

Fig. 5. Example cross-taxonomy analysis of Total Assets for MASTERCARD INC

4.2 Overview First, Zoom, Details on Demand (Requirement 2)

We now demonstrate the capability of FIOS to show the same data using our three different interfaces. As an example, we again visit from the SPARK company template the assets over time for MASTERCARD INC as displayed in Figure 5.

From the top of the SPARK company template, via "Pubby Link to Data", we can then start browsing all information related to MASTERCARD INC in the Linked Data Browser Pubby. For instance, we can browse to the balance sheets and there find the single observations visualised in the line chart. For instance, see Figure 6 for a screenshot of an observation found in Pubby describing the total asset on 2011-12-31.

Anonymous Resource #1

Property	Value
qb:dataSet	<http://public.b-kaempgen.de:8080/pubby/archive/1141391/0001141391-12-000006%23ds>
cal:dtend	2012-03-31
cal:dtstart	2012-01-01
?:issuer	<http://public.b-kaempgen.de:8080/pubby/cik/1141391%23id>
sdmx-measure:obsValue	318000000 (xsd:double)
is rdfs:seeAlso of	<http://public.b-kaempgen.de:8080/pubby/archive/1141391/0001141391-12-000006%23ds>
?:segment	0001141391 2012-01-01 2012-03-31
?:subject	<http://public.b-kaempgen.de:8080/pubby/vocab/us-gaap-2011-01-31%23IncomeTaxExpenseBenefit>
rdf:type	qb:Observation

Fig. 6. Example Linked Data browser view on total asset in 2011 for MASTERCARD INC.

From the SPARK company template, we can also visit the OLAP Interface, Saiku, to create the same report as shown in the total asset line chart. For that, we create a pivot table with the issuer dimension filtered by Mastercard on columns, date dimension on rows and filtered on subject dimension with `us-gaap-2009:Assets` and `us-gaap-2011:Assets` on columns, as can be seen in Figure 7.

Note, though connected through their underlying data, a better interlinking between the three provided interfaces was difficult due to technical problems: SPARK tables did not allow to show browseable links; SPARK diagrams often contained aggregated or densely-displayed facts that are difficult to select for

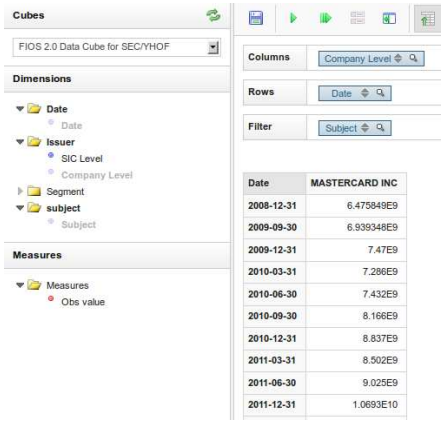

Fig. 7. Example OLAP Interface query on Total Assets over time for MASTERCARD INC

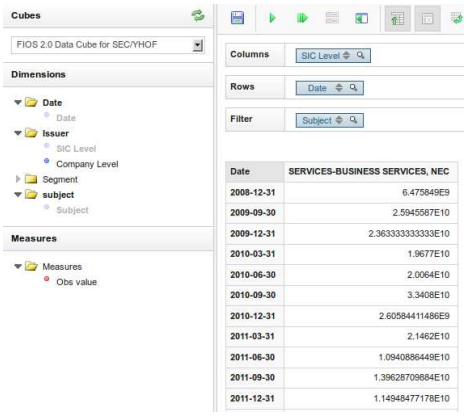

Fig. 8. Example OLAP Interface query on Total Assets over time for SERVICES-BUSINESS SERVICES (SIC)

browsing; single facts often were modelled as blank nodes in QB and thus are not directly referenceable; and to browse a URI from FIOS, Pubby required adding "pubby" and converting "#" to "%23".

4.3 Intuitively Create Own Reports and Analyses (Requirement 3)

We now show that a user can create a typical report on our integrated financial data with intuitive OLAP operations: requesting a pivot table showing the Total Assets over time, similarly as for the total asset line and pivot charts, but this time aggregated to the industry level of Mastercard as visible in Figure 8.

Projection: By drag & drop of a measure to Columns, Rows or Filter fields in the pivot table, a user can select a certain measure. Since our data cube only contains one measure, projection is not necessary.

Dice: A user can filter for certain facts by clicking on the magnifier symbol of a dimension on the Columns or Rows fields. In our case, the user filters for certain subjects (`us-gaap-2009:Assets` and `us-gaap-2011:Assets`) as well as certain companies (Mastercard).

Slice: Any dimension that a user does not drag & drop to either Columns or Row fields gets sliced, i.e., removed and aggregated over. Since QB does not provide means to describe aggregation, FIOS uses the AVERAGE function as default; for numeric values the average returns an easy-to-understand measurement.

Roll-up: Any dimension listed on the left side can exhibit a hierarchy of several levels. For instance, for the issuer dimension, either Company or SIC Level can be selected. SIC Level groups companies by their SIC industry classification. In our example, we rolled-up to SIC Level and filtered for the SIC of Mastercard, "SERVICES-BUSINESS SERVICES, NEC".

Drill-across: Although not required in our example, an analyst may request a pivot table containing both observations from balance sheets and stock quote

tables. Since the Saiku interface only allows to select one dataset per pivot table, we extend the ETL components with one extra SPARQL CONSTRUCT query linking observations from several integrated datasets (datasets with the same structure) to a new integrated multidimensional dataset "FIOS 2.0 Data Cube for SEC/YHOF".

5 Discussions and Lessons Learned

FIOS benefits from Semantic Web technologies, e.g., in modelling and integrating balance sheets from the SEC Edgar Database, stock quotes from the Yahoo! Finance Web API as well as company metadata from Wikipedia/DBpedia using existing vocabularies; the Linked Data principles ensure access to data in a standard and modular way. Since the schema of RDF is flexible, new data can easily be added by allowing the crawler to reach further entities. SPARQL allows quality checks and is sufficiently expressive to implement background information, multi-company and cross-data-sources analyses. Formal semantics such as explicit equivalent statements simplify access via entity consolidation. Three interfaces with different purposes use the same backend: any data that is added to the triple store can directly be visualised in SPARQL templates, browsed using the Linked Data Browser and queried using the OLAP Interface. Consequently, we argue that Semantic Web technologies allow a continuous integration of new data. With more heterogeneous datasets and frequent addition and updates of data sources, FIOS will develop its full potential if research resolves the following challenges:

Develop interfaces and visualisations sufficiently specific to provide added value and generic to have new data immediately considered. If new information such as from text or structured databases, e.g., subsidiary relationships, product classifications or organisational structures, are continuously added to FIOS, specialists are needed to adapt or create new SPARQL templates; the Linked Data Browser provides data only on the triple level; and the OLAP Interface requires integrated QB datasets. Ideally, new data sources seamlessly and without much effort will result in extended visualisations, e.g., providing more detailed provenance information, adding new data points or allowing additional interaction capabilities such as roll-up and drill-across.

Increase coverage and quality of information by continuously integrating data sources. Integrity constraint checks need to be manually extracted and run. There may still be errors in the data, e.g., companies that share CIKs or ticker symbols because of a merger. Debreceny et al. [4] have shown that some information may be derived only in a best-guess fashion. New data sources promise to reduce data quality issues if integrated to one well-interlinked model. Then, the same KPIs can be calculated in different ways to identify differences between data sources, e.g., DBpedia "Operating Income" and the last yearly balance sheet net income loss. FIOS would need to consider uncertainty, to draw declarative knowledge from experts or other data sources, and to describe both static and dynamic relationships between financial data.

Improve query processing performance. Although currently no issue in FIOS, pre-processing and integration will take too long for continuously updated and larger data sources. FIOS' current performance bottlenecks are large numbers of separately issued SPARQL queries and large numbers of multidimensional elements to load into the OLAP user interface. In more complex data integration and analysis scenarios optimisations such as parallelisation will be required, e.g., analytical queries that scan a large number of observations, contain filters of varying selectivity and compute aggregation functions on schema-flexible and heterogeneous data require specific data processing optimisations [7].

6 Related Work

We distinguish other financial data integration and analysis applications and related work about modelling XBRL data using Semantic Web technologies.

The Rhizomik Semantic XBRL demo [5] ties RDF representations of XBRL close to the original XML data which make mixing with other data sources difficult. The Business Intelligence Cross-lingual XBRL (BIXL) demonstrator [9] focuses on retrieving facts from unstructured text in filings as well as a multilingual interface, however does not consider data integration of XBRL balance sheets with stock quotes. Midas [2] implements a pipeline similar to FIOS without using Semantic Web technologies. Their main focus lies in extracting and linking of information about entities such as company and key people from semi-structured XML documents. However, it is unclear to what extend information from SEC and FDIC sources were integrated and what efforts would be needed to add new data sources such as Wikipedia.

Although judges saw potential, FIOS did not win the XBRL Challenge. Other submissions, in particular the winners – Calcbench and Sector3 – were more robust (e.g., FIOS still is limited to certain browsers), provide keyword search or filtering for companies (e.g., "revenue higher than"), include a larger number of companies and filings (also non-balance-sheets), exhibit short update intervals with new filings (10-15min) and often provide MS Excel exports for further processing and analysis (Saiku also provides that).

In summary, although important for a holistic view on companies, current systems do not focus on integration of different data sources: whereas multi-company KPI analysis with an Excel export often is possible, background information, such as from Wikipedia, rarely is embedded in the interfaces. Calcbench shows the actual stock quote of a company, yet, no other system allows for comparison of balance sheet KPIs with other numbers such as stock quotes over time. If systems find correspondences between companies or financial concepts, it is unclear whether the matching is hard-coded or flexibly represented with a formalism such as equivalence statements.

Several recent papers have proposed Semantic Web technologies as a suitable way to manage and model XBRL data. Wenger et al. [12] consider the interoperability problems of different taxonomy versions, but apart from proposing the criteria they do not evaluate their approach. Bao et al. [1] tries to fully keep the semantics of XBRL in an RDF/OWL representation; however, the authors

do not describe the benefits of their representation in case studies. In comparison, we show that XBRL filings and taxonomies can be efficiently represented as multidimensional datasets using the RDF Data Cube vocabulary.

7 Conclusions

In this In-Use paper, we have described the Financial Information Observation System (FIOS) that models XBRL data using the RDF Data Cube Vocabulary; consolidates financial data for background, multi-company, and cross-data-sources KPI analysis; and provides intuitive and explorative analysis interfaces. The benefit of Semantic Web technologies are a flexible schema, standard access, expressive queries and formal semantics. Main challenges to scaling-up those benefits in continuous integration scenarios are interfaces sufficiently specific to provide added value and generic to have new data immediately considered; to increase coverage and data quality with added data sources; and to optimise analytical operations on flexible schemas and heterogeneous data. In future work we intend to evaluate and extend the FIOS approach to other domains.

Acknowledgements. This work was carried out with the support of the German Research Foundation (DFG) within project I01, SFB/TRR 125 "Cognition-Guided Surgery" and the German Federal Ministry of Education and Research (BMBF) within Software-Campus project "LD-Cubes" (01IS12051).

References

1. Bao, J., Rong, G., Li, X., Ding, L.: Representing Financial Reports on the Semantic Web: a Faithful Translation from XBRL to OWL. In: Proceedings of the 2010 international conference on Semantic web rules (2010)
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. (2011)
3. Carretié, H., Torvisco, B., García, R.: Using Semantic Web Technologies to Facilitate XBRL-based Financial Data Comparability. In: International Workshop on Finance and Economics on the Semantic Web (2012)
4. Debreceny, R.: Feeding the information value chain: Deriving analytical ratios from XBRL filings to the SEC. Tech. rep. (2010)
5. García, R., Gil, R.: Triplificating and linking XBRL financial data. Information Storage and Retrieval (2010)
6. Hogan, A., Harth, A., Umbrich, J., Kinsella, S., Polleres, A., Decker, S.: Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine. Web Semantics: Science, Services and Agents on the World Wide Web 9, 365–401 (2011)
7. Kämpgen, B., Harth, A.: No Size Fits All - Running the Star Schema Benchmark with SPARQL and RDF Aggregate Views. In: ESWC (2013)
8. O'Riain, S., Curry, E., Harth, A.: XBRL and open data for global financial ecosystems: A linked data approach. Int. J. of Accounting Information Systems 13 (2012)
9. O'Riain, S., Coughlan, B., Buitelaar, P., Declerk, T., Krieger, U., Marie-Thomas, S.: Cross-Lingual Querying and Comparison of Linked Financial and Business Data. In: The Semantic Web: ESWC 2013 Satellite Events (2013)

10. Spies, M.: An ontology modelling perspective on business reporting. Information Systems 35 (2010)
11. Tseng, F.S.: Integrating heterogeneous data warehouses using XML technologies. Journal of Information Science 31 (2005)
12. Wenger, M., Thomas, M.A., Jr., J.S.B.: An Ontological Approach to XBRL Financial Statement Reporting An Ontological Approach to XBRL Financial Statement Reporting. In: AMCIS 2011 Proceedings (2011)