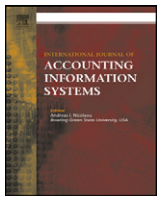
International Journal of Accounting Information Systems 13 (2012) 141–162

Contents lists available at SciVerse ScienceDirect

International Journal of Accounting Information Systems



XBRL and open data for global financial ecosystems: A linked data approach

Seán O'Riain [a,b,*], Edward Curry [a,b], Andreas Harth [c]

[a] Digital Enterprise Research Institute (DERI), NUI Galway, Ireland
[b] IDA Business Park, Lower Dangan, Galway, Ireland
[c] Institut für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB), Karlsruher Institut für Technologie (KIT), Germany

ARTICLE INFO

Article history:
Received 15 November 2010
Received in revised form 9 February 2012
Accepted 15 February 2012

Keywords:
Internet
World Wide Web
Metadata
Financial ecosystem
eXtensible Business Markup Language, XBRL
Resource Description Framework RDF, Open Data, Linked Data, Linked Open Data
LOD Financial Mashup

ABSTRACT

Information professionals performing business activity related investigative analysis must routinely associate data from a diverse range of Web based general-interest business and financial information sources. XBRL has become an integral part of the financial data landscape. At the same time, Open Data initiatives have contributed relevant financial, economic, and business data to the pool of publicly available information on the Web but the use of XBRL in combination with Open Data remains at an early state of realisation. In this paper we argue that Linked Data technology, created for Web scale information integration, can accommodate XBRL data and make it easier to combine it with open datasets. This can provide the foundations for a global data ecosystem of interlinked and interoperable financial and business information with the potential to leverage XBRL beyond its current regulatory and disclosure role. We outline the uses of Linked Data technologies to facilitate XBRL consumption in conjunction with non-XBRL Open Data, report on current activities and highlight remaining challenges in terms of information consolidation faced by both XBRL and Web technologies.

1. Introduction

In the 2008/2009 global financial crisis many banks had to quickly try and understanding their exposure to the changing market conditions. In an examination of the role of IT in the crisis, bank employees were classified as being involved in "detective work" having to piece together financial information distributed across multiple silos (Economist, 2009). Whether internal to an organisation or across its supply

* Corresponding author. Tel.: +353 91 495080.
E-mail addresses: sean.oriain@deri.org (S. O'Riain), ed.curry@deri.org (E. Curry), harth@kit.edu (A. Harth)

1467-0895/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.accinf.2012.02.002

chain, integrating financial information and data remains a fundamental challenge. Addressing the challenge requires a flexible approach to financial and business information integration with an ability to connect and consume large quantities of data sources.

The eXtensible Business Reporting Language (XBRL) standardises financial reporting and with a machine-interpretable format that makes corporate reports easier to consume and integrate. However, XBRL-based information sources only provide part of the picture, and many other data sources are used in conjunction with XBRL. As Hatsu Kim, VP Global Fundamentals Thomson Reuters, noted "when it comes to valuing companies, quarterly and yearly filings are insufficient in information terms and have to be considered together with information on markets and exchange rates (W3C, 2009).

Analysts and investors constructing detailed insight into an organisation develop their understanding through examination of a diverse range of business and financial information. Performing an analysis can require information varying from operational figures, new product announcements, risk exposure, sector spend, independent analysis and customer sentiment. The source of this information includes internal reports, social platforms, marketing briefs, regulatory filings, analyst reports, press releases, government statistics and third party information providers. Information professionals face the difficulty of how to achieve faster and more accurate analysis across these disparate financial information sources that present and behave as islands of information.

The XBRL International Standards Board (XSB) also recognises the potential of generating an integrated financial information environment, specifically noting the early stages of "*building an ecosystem in which XBRL information is generated, reported, reused, combined and analysed throughout the business community and all along the business reporting supply chain*" as also facilitating XBRL consumption (XSB, 2010).

Within the wider global financial and business information ecosystem of corporate press releases, government statistics, market press coverage and third party information providers,[1] XBRL data repositories represents yet another data silo in a global landscape of disconnected silos. Combining XBRL with the plethora of non-XBRL financial information remains at an early stage of realisation. Significant efforts are required on the part of financial information consumers to integrate XBRL data with other data expressed in a wide variety of data formats.

The last few years has seen the emergence of a "Web of Data" fuelled by Open Government transparency initiatives that have made significant amounts of public sector information freely available for use and redistribution without restriction. Notable examples within this Open Data[2] movement are data.gov, recovery.org (US), data.gov.uk (UK), Eurostat[3] (EU), the World Bank[4] and International Monetary Fund.[5] The EU has also mandated that collected financial, economic and legal data sets be made available as Open Data (EC, 2003) for integration and innovative reuse with new products and services (EC, 2006). Open government data is a significant player directly supporting *data innovation,* an approach where companies analyse raw government data to better inform their own business circumstance, those of stakeholders, or the development of new service opportunity (IDG, 2009).

Semantic Web technologies and standards play an important role in sharing of large quantities of data via the Web. The resulting Web of Data enables machine interpretation of the meaning of information.[6] Linked Data[7] (detailed in Section 3) is a best practice approach used to expose, share and connect data on the Web based on World Wide Web Consortium (W3C) standards.

XBRL consumption is typically restricted to the transformation of entity specific reporting concepts extracted from financial statement to financial statement ratios (Debreceny et al., 2009) that then drive analysis and insight generation. XBRL's lack of association with data external to financial statements restricts the information consumers opportunity for holistic investigative analysis of other tangible sources (e.g. company Web sites and government data, financial news, financial discussion forums). Linked Data technologies can be used to increase the association between XBRL and Open Data silos and contribute towards XBRLs consumption and exploitation potential. There have been initial efforts to combine XBRL

[1] Further examples are given in the Appendix on Web Based Financial & Economic Open Data Sets
[2] Refer to Appendix on Web Based Financial & Economic Data Sets for a broader list of examples
[3] http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/
[4] http://data.worldbank.org/data-catalog
[5] http://www.imf.org/external/np/fin/tad/exfin1.aspx
[6] http://www.w3.org/2001/sw/SW-FAQ
[7] http://linkeddata.org/

and Semantic Web as part of a broader information ecosystem. The W3C and XBRL joint workshop on "Improving Access to Financial Information on the Web" discussed complimentary approaches to bridging the two communities and identified significant remaining challenges (W3C, 2009).

This paper investigates the feasibility of leveraging XBRL data with Open Data as part of an integrated information ecosystem using a Linked Data approach by discussing the following research questions:

- First, *what role has Open Data to play in the information value chain?* What types of Open Data sources are accessible? How can these data sources be leveraged with XBRL where connections and relationships between diverse data sets can be established?
- Second, *how can XBRL interactive data integrate with Open Data?* What are the Semantic Web technology fundamentals required? What role does Semantic Web technology play in simplifying the integration of XBRL data with other relevant Open Data sources? Are there any successes to date?
- Third, *what barriers are there to information accessibility?* What reduces usability of the data? Relating to information accessibility and use, what are the remaining open challenges for a global financial data ecosystem?

Insight resulting from this study provides the XBRL community, regulators, financial information providers and analysts with a fundamental grounding in Open Data and Semantic Web technologies. The remainder of the paper is structured to address each research question as follows.

Section 2 introduces Open Data, discusses its incorporation into the Information Value Chain, and outlines how Open Data can be leveraged. Section 3 introduces the technology fundamentals of Semantic Web and Linked Data. Section 4 discusses Linked Data standards and technologies and shows how they can be applied to the integration of XBRL and Open Data. Section 5 outlines areas of current synergy between XBRL and Linked Data that could be further developed into mainstream activities. Section 6 addresses targeted open challenges not discussed in other areas of the paper regarding global financial data ecosystem evolution. Finally, Section 7 summaries findings and concludes the paper. Throughout the paper comment and comparison on XBRL and Semantic Web technologies are incorporated into the discussion as it develops.

2. Open data in the information value chain

An information ecosystem can be considered in broader terms as interconnected structures of organisations, technologies, consumers and products (Gundlach, 2006) or in corporate reporting terms as information generation, reporting, reuse, combination and analysis throughout the business community (XSB, 2010).

Adhering to an information focus we first introduce Open Data. We motivate the leveraging of XBRL with non-XBRL Open Data using a business analyst usage scenario driven by an information requirement that involves multiple distributed sources. We further highlight the potential for leveraging data when relationships between data sets can be established.

2.1. Open data

Demands for greater levels of transparency have resulted in Open Government initiatives that have made available large numbers of sector, statistical, financial, and economic data sets for public consumption in varying formats (e.g. XBRL, XLS, CSV, PDF, RDF, text). The Security and Exchange Commission's (SEC) EDGAR repository,[8] providing public access to corporate filings in multiple formats, is one of the more widely known examples of business and financial Open Data. Open Data has considerable potential for re-use, evident in the fact that EDGAR has been heavily exploited by third party information providers such as Hoovers, Morning Star, Yahoo! Financial and EDGAR Online to provide value add service offerings.

Open Data can also include generalised business news, marketing information, and competitor data that is available from an assortment of Web sites. Two popular datasets are DBpedia and Freebase. DBpedia[9] is a community initiative that extracts general-purpose information from Wikipedia and

[8] http://edgar.sec.gov/
[9] http://wiki.dbpedia.org/Datasets

makes it freely available. Freebase[10] harvests information from multiple sources to promote data about people, places and organisation. Combining Open Data allows companies to "data innovative" and enhance understanding of their own business circumstance, develop new insights and identify new service opportunities. A novel example demonstrating Open Data utility was the crowd-sourcing approach used by the UK Guardian newspaper for community tagging of Parliament members expense irregularities that reporters then used to guide investigation (Guardian Newspaper, 2010). The approach is an example of community based data curation (Curry et al., 2010) that is equally applicable to an analyst peer group tagging financial statements to share insight, comment and opinion.

Using the Web to publish Open Data has made information more accessible, but varying information formats can make consumption difficult. Linked Data[11] (cf. Section 3) based on the W3C Resource Description Framework (RDF) standard caters for multiple formats in providing a common interoperable format and model for data linking and sharing on the Web. Fig. 1 shows a section of the larger Linking Open Data (LOD) Cloud, which comprises 464 interlinked RDF datasets and 205 vocabularies[12] within the wider ecosystem that are being actively used by industry, government and scientific communities. Examples of standardisation attempts to enhance quality and reduce consumption costs of Open Data Sets[13] are the use of RSS-CB[14] by central banks to publish exchange rate data and the Statistical Data and Metadata Exchange format[15] (SDMX) used by Eurostat, and under consideration for use by the OECD, World Bank and United Nations. The World Bank[16] also makes its statistical data sets available in RDF.

2.2. XBRL and Open Data usage scenario

The information value chain (Fig. 2) provides a framework for determining how to leverage information resources within environments that spans organisations and their external third party connections. Its core stages of 1) information acquisition, 2) information processing and 3) information distribution represent between 15 and 35% of the total resourcing effort and cost (Feldman, 2003). XBRLs potential to improve the information value chain through regulatory filings and business information processing automation was an important consideration in the US SEC adoption of XBRL as interactive data (Debreceny et al., 2009).

The information value chain has three main lifecycle stages of acquisition, processing and distribution.

- *Information acquisition* concentrates on the sourcing and gathering of data from legacy databases, websites, or documents.
- *Information processing* concerns activities that support integration of the collected data. Integration also augments data understanding through consolidation.
- *Information distribution* makes the integrated information available to applications and end users via APIs, reports and interactive tools such as portals and dashboards.

Having adequate amounts of relevant information to work with remains a contentious issue for information professionals. Not satisfied with the information functionality offered by Business Intelligence tools, financial information consumers are looking to include ancillary and peripheral information sources in their analysis. Where the established information value chain does not cater for an analyst's information requirement, the analyst will intuitively look to develop their own, sourcing relevant information and performing elements of information processing as necessary.

Fig. 2 considers a business analyst investigating the risk profile of a company under investment consideration. The analyst has the following queries (also listed in Table 1) but lacks sufficient information to attempt an informed answer.

- *What was the first quarter revenue of MetLife in 2010?*
- *What is the media coverage on Metlife's recent SEC filing?*

[10] http://www.eba.europa.eu/Supervisory-Reporting/COREP.aspx
[11] http://linkeddata.org/
[12] http://stats.lod2.eu/
[13] Samples provided in Appendix Web Based Financial & Economic Open Data Sets
[14] http://cbwiki.net/wiki/index.php/rss-cbmain
[15] http://sdmx.org/
[16] http://data.worldbank.org/data-catalog

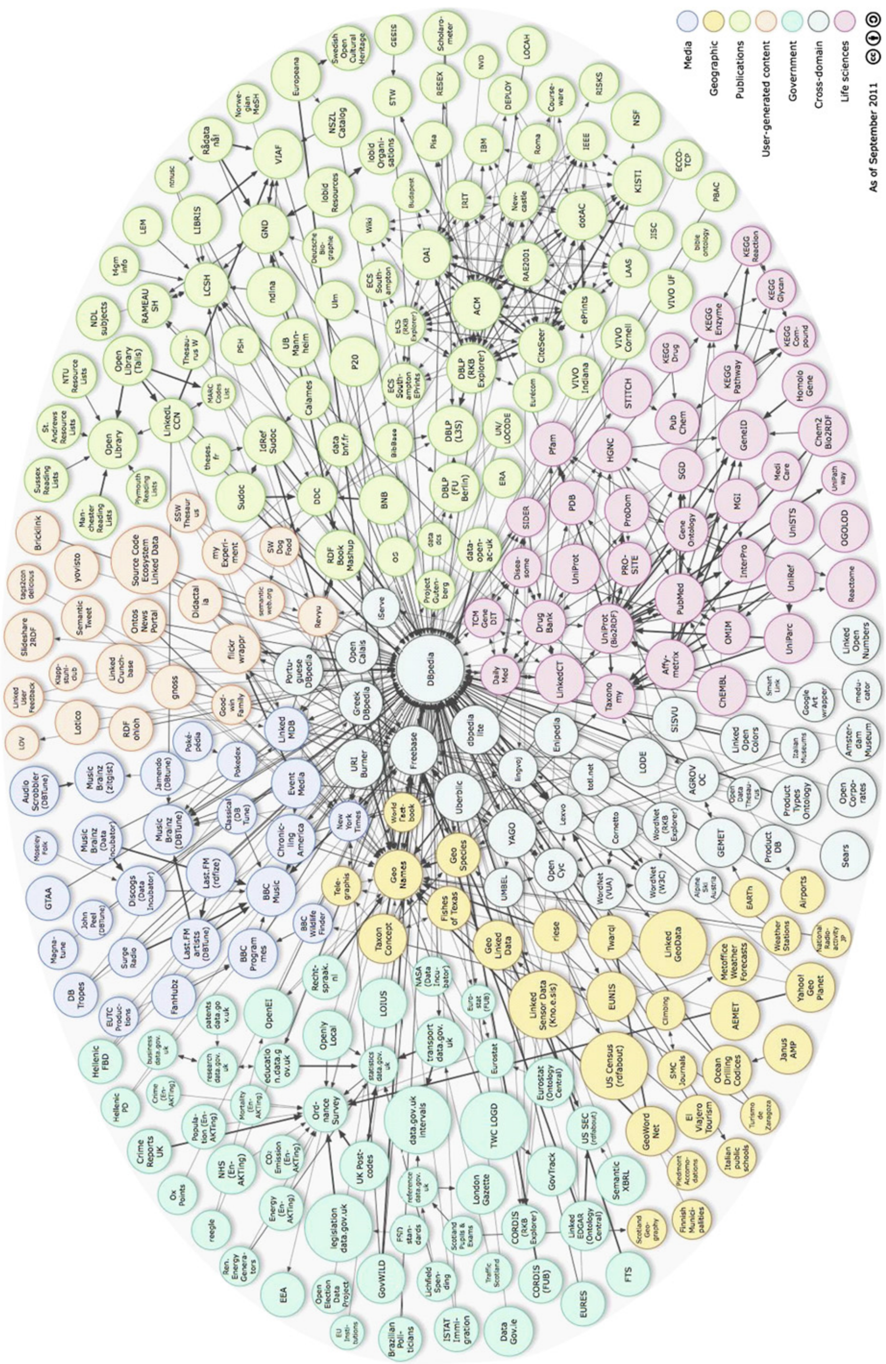


Fig. 1. Linked Open Data cloud, http://lod-cloud.net/.

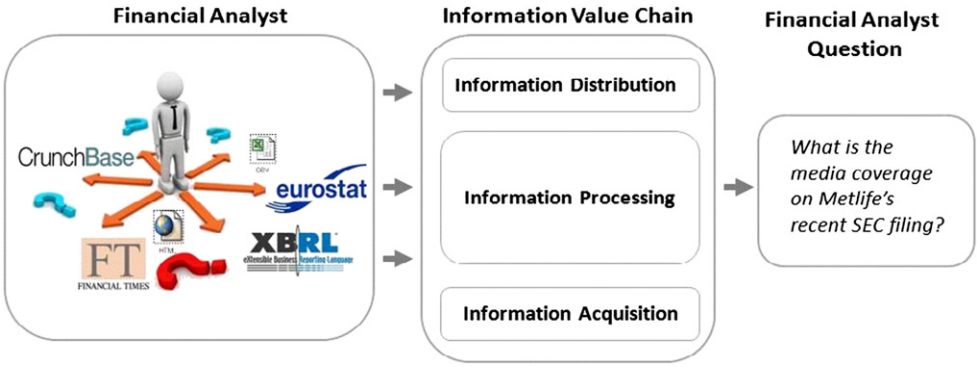


Fig. 2. Analyst information value chain.

- *Who are MetLife's suppliers?*
- *Which companies offer Life Insurance with revenue > USD 1 M?*
- *What companies with German clients have revenue > USD 1 M?*

To satisfy the information requirement for a query such as '*Which companies offers Life Insurance with revenue > USD 1 M?*' the analyst sets about developing their own information value chain by performing:

Information acquisition which requires the analyst to identify the relevant information sources, along with the type of data formats to accommodate the query. Looking at the query '*Which companies offers Life Insurance with revenue > USD 1 M?*' in Table 1:

- The type of information–financial data, company data and industry sector is first established.
- Next the sources of these information types are defined. Here financial data can be retrieved from XBRL EDGAR filings, company data from CrunchBase (a collaboratively-edited database about companies and financial organisations) and sector classifications from DBpedia.
- The type of information can be then be refined to the exact type required from each source. From company data the company name, location and address can be extracted and from sector the industry classification.

Table 1
Information requirement and applicable information sources.

Query (question)	Source information		Extracted information	
	Type	Source	Type	Format
What was the first quarter revenue of MetLife in 2010?	Company data (financial)	http://edgar.sec.gov/	Company Name CIK, Ticker Symbol	HTML, XBRL
What is the media coverage on Metlife's recent SEC filing?	Company data, market coverage, financial news	http://edgar.sec.gov/, http://bloomberg.com/, http://nytimes.com/	Company name, news, events	XML, XBRL, RSS
Who are MetLife's suppliers?	Internal company information,external company data	Corporate database,http://edgar.sec.gov/	Supplier name,Company name	XML, XBRL, RDB
Which companies offer Life Insurance with revenue > USD 1 M?	Company data (financial), sector information	http://edgar.sec.gov/, http://dbpedia.org/, http://crunchbase.com/	Company name, location, address, products, industry classification, competitors	XML, XBRL, RDF
What companies with German clients have revenue > USD 1 M?	Clients information, company data (financial), geography information	Corporate Database, http://edgar.sec.gov/, http://dbpedia.org/, http://crunchbase.com/, http://geonames.org/	Company name, location, address, products, industry classification, customers, geographical location	XML, XBRL, RDF, RDB

- Finally the format of the data target for extraction is noted. To service the query *'What is the media coverage on MetLife's recent SEC filing?'* additional company data, market data and financial news will have to be sourced. Market data can come from Bloomberg and financial news from both the New York Times and Bloomberg news feeds or Web site. The format that have to be dealt with XBRL from EDGAR, XML from Bloomberg Websites and RSS feeds (text) from the NY Times.

Information processing requires that the analyst performs the task of consolidating and merging the retrieved information. In order to make sense of the data it can help to create a model around a central entity such as company and then to establish relationships to related data items. For example, a query by company sector and location will require the analysis to extract XBRL financial statements linking the particular company with its sector type and company address within a particular geographical location.

Information distribution allows the analyst to interrogate the integrated data sources. Viewing information for ad-doc sources will not fit well with established tools and reporting output mediums. XBRL filings use specific interactive viewers (such as Fijitsu Interstage XWand[17]), SEC or proprietary tools that accommodate filing instance display. Hybrid approaches such as inline Extensible Business Reporting Language (iXBRL[18]) allow XBRL tagging combined with HTML to produce more flexible Web browser viewable reports. Moving to a situation of querying and analysis across XBRL with other source requires a more flexible representation, navigation and interaction user experience.

We next present the technology fundamentals that Semantic Web uses to address these issues.

3. Semantic Web technology foundations

The evolution of the Web from being a document-based Web to one that also supports data has provided information consumers easier access to greater amounts of content. Semantic Web technologies offer flexible means for data publishing, integration and interpretation. The following section introduces several fundamental Semantic Web building blocks that provide familiarity with the technology principles, data model and representational format required to consider their application for Open Data usage within a Linked Data supported financial ecosystem.

3.1. Universal Resource Identifiers

Universal Resource Identifiers (URIs) unambiguously identify arbitrary resources on the Web. HTTP URIs are based on the Webs Domain Name System (DNS) which allows organisations and individuals to register global domain names, which themselves can be used to construct entity identifiers. Resources can be files (e.g. XBRL documents), real world entities (e.g. product, company) or an abstract concept (e.g. supplier relationship, profit before tax). URIs can be retrieved (dereferenced) via HTTP in a Web browser or application and provide the basic mechanism to associate distributed pieces of data. Based on the Web's HTTP infrastructure URIs can be used to access and integrate distributed data both intra- and inter-organisational without impacting the existing information infrastructure. Analysts can therefore 'pull' information from multiple sources or traverse an information thread moving from one data set to another via the established interlinks.

3.2. Resource Description Framework

The Resource Description Framework (RDF) is the basic machine-interpretable information representation format on the Semantic Web. RDF provides a common (graph-based) data format and an identifier scheme that can serve as foundation to unify data from a large number of sources. Data and facts are specified as RDF statements with atomic constructs of a subject, predicate and object, also referred to as triples (example in Fig. 5). RDF statements from sources such as DBpedia and data converted to RDF, for example from SEC's EDGAR, can be merged into a single representative graph.

[17] http://fijitsu.com/global/services/software/interstage/xbrltools
[18] Used by the UK Revenue to inspect company filings.

RDF should be considered for use in situations where:

- Multiple source data integration is required without the overhead of a large development effort.
- Data will be made available for re-use by stakeholders.
- Data is available in a decentralised manner, that is, no single stakeholder has responsibility for the entirety of data.
- Enhanced use of large amounts of structured data is required (browse, query, match, extract, input).

Data locked in organisational data silos can also benefit by having their data exposed as RDF without modification (Auer et al, 2009) allowing existing information architecture and original data representation remain unaltered. Once the data is exposed several frameworks (Volz et al., 2009) can be used to establish and discover relationships between and with other Web accessible Linked Data. Legacy databases can then be queried using the Semantic Web query language SPARQL (Bizer and Cyganiak, 2006). The key point is that RDF use is for establishing a common representation of information contained in other formats to assist combined pre-processing rather than as a replacement of the originating format. The Semantic Web community refers to the activity of converting Open Data to RDF as *RDFication* or *RDFizing*.

3.3. Vocabularies

Most business domains have their own particular vocabulary which provides distinct descriptions and definitions of concepts used their schema and recommendations on how to model the domain. XBRL allows easy access to attributes and values in instance documents but comparing information between alternate XBRL taxonomies requires extension or alternate representation. Semantic Web vocabularies support shared understanding through related vocabulary identifiers and can accommodate multiple levels of abstraction. XBRL concepts model financial statements with dimensions such as reporting period and reporting entity (company or filer). Semantic Web vocabularies can define higher abstractions placing real-world entities such as company as primary concepts with financial statements as properties representing a movement from a statement-centric view to an entity-centric view.

Vocabularies can also be included in other vocabularies (also allowed by XBRL) and their definitions re-used. Semantic Web examples of alternate taxonomies used to enhance integration and query capability are W3C's Geo vocabulary, a geospatial schema, SKOS, the Simple Knowledge Organization System vocabulary, used to encode classification schemes and Good Relations for retailer[19] product offerings and description tagging (Hepp, 2008).

3.4. Linked Data

The Semantic Web community promotes Linked Data as the main deployment practice for data publishing. Linked Data can be classified as a set of best practice principles[20] which recommends to:

1. Use URIs to name things.
2. Use HTTP URIs so that one can look up those names.
3. When someone looks up a URI, provide useful information, using standards (e.g. RDF, SPARQL)
4. Include links to other URIs so that one can discover more things.

Publishing data which adheres to these basic principles provides a common standard based model for data access and inter-linkage. Where Open Data is made available in non-proprietary formats, wrappers and converters that map information from their source format into Linked Data can be employed on an as-needed basis. Examples are detailed in the next section.

[19] Best Buy reported a 30% increase in traffic across their online product catalogues containing the annotated structured information (NYT, 2010)

[20] http://www.w3.org/DesignIssues/LinkedData.html

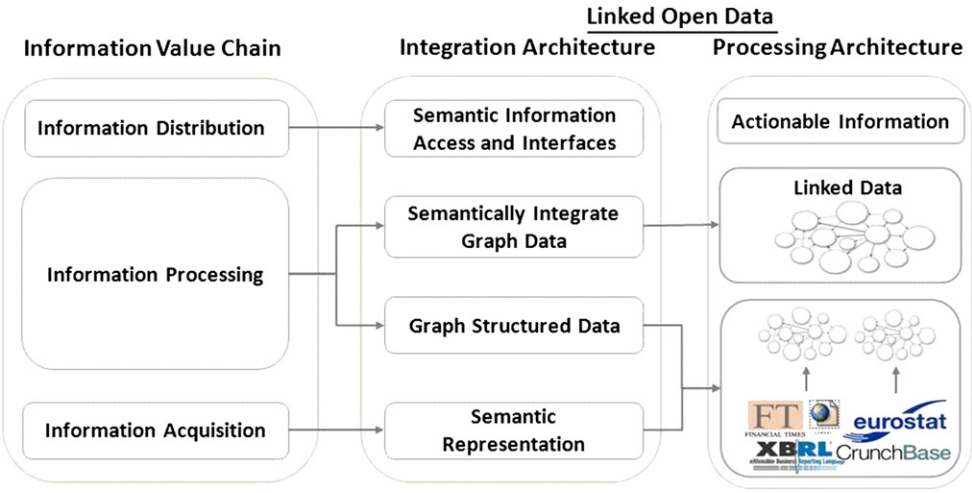


Fig. 3. Open Data source integration using Linked Data.

4. Combing XBRL and Open Data

Previous research on Web-based financial information identified the lack of standardised reporting as an area that stood to benefit from the application of an XBRL–RDF approach (Debreceny and Gray, 2001). Since then transparency and regulatory initiatives have enhanced both accessibility and availability of standardised financial reporting through the adoption of XBRL. However synergies based on combined use of XBRL and Open Data remains at an early stage of realisation mainly demonstrable through academic initiatives. With the advent of Enterprise 2.0 (McAfee, 2006) technology mashups[21] based on Web standards began to first emerge in the social space allowing data interconnection through 'mashing' multiple competitor API's and Open Data to generate a more holistic view over data. When faced with integrating and providing timely access to information for scenarios such as those previously outlined the mashup represents a relatively new but powerful approach to multiple data sources integration. Gartner has identified the mashup as one of the top 10 strategic enterprise technologies due to its transformation potential (Gartner, 2009).

The emergence of Linked Data from Semantic Web efforts has promoted the availability of semi-structured and structured data on the web in a format that provides standard access and interoperability between and among Open Data sets. Linked Data directly supports data integration, offering a path for XBRL and Open Data linkage as part of a wider ecosystem. To illustrate how synergies can be established between the two technologies of XBRL and RDF we introduce the Linked Data integration architectures and discuss activities associated with adoption and usage. As the RDF representation format is a key enabler for interoperability and linkage we discuss the transformation of Open Data and specifically XBRL to RDF.

4.1. Linked data enabled financial data integration

Fig. 3 illustrates the intrinsic relationship between the information value chain and corresponding stages of the Linked Open Data (LOD) Integration Architecture (centre of figure), and also provides a line of sight from raw data acquisition to end data usage. It also illustrates from a practical view point what each of these integration processing stages actually involve (right portion of figure).

[21] A mashup may be considered as an application or Web page that combines data from multiple resources (applications, services) to create something new.

The integration architecture has the main activity stages of:

Semantic Representation involving the conversion of multiple financial data format types and structure that range from unstructured text, structured XML, CSV Open Data etc., to RDF adhering to the Linked Data principles. Data is extracted and serialised in RDF based upon HTTP lookup on the entities URI. In effect the source data.

Semantically Integrate Graph Data: The individual graphs are integrated into a holistic dataset using entity alignment and linking. Available Linked Data can additionally be harvested and added.

Semantic Information Access: Users can analyse and navigate the integrated Linked Data through a single entry point that supports interactive exploration and semantic querying based interfaces.

The key to the Linked Open Data processing architecture is the ability to map different data structures and taxonomies to a common interoperable format that captures the necessary semantic interpretation. Each of the integration architectural stages is next discussed with emphasis on the conversion of XBRL and its integration into the ecosystem. For further detailed discussion on component architecture of a Semantic Web mashup we refer the reader to (Curry et al., 2009; García et al., 2010) or for commercial mashup development environments to (Grammel and Storey, 2010.

4.2. Semantic Representation of XBRL

Merging information revolves around transforming source information into a proprietary or standards based common format that facilitates information interoperability, merging, and integration. The ability to successfully model between source and destination formats will determine whether the transformation can be wholly or partially automated. XBRL, similar to any other information sources, requires that its semantic representation be explicitly mapped, before financial statement values can be processed. XBRL semantics are distributed across its label, presentation and calculation linkbases.

The label linkbase provides human-readable definitions and descriptions of concepts. Parent–child relationship hierarchies are found in the presentation link base and the calculation link base relates concepts through application of basic financial calculation rules. Extracting the semantic hierarchy requires that:1) semantic context be established and 2) variations between the instance document and taxonomy be addressed.

4.2.1. Semantic context

Fig. 4 illustrates how the semantic hierarchy for the element *us-gaap:FeesAndCommissions* requires establishing semantic context from both the presentation and calculation linkbases. The presentation linkbase details the parent–child relationships of *OperatingIncomeLossAbstract*, *GrossProfitAbstract*, *RevenuesAbstract* and *FeesAndCommissionsAbstract*. The *FeesAndCommissionsAbstract* in turn has a parent–child relationship with elements *FeesAndCommissions*, *FeesAndCommissionsFiduciaryAndTrustActivities*, *InsuranceCommissionsAndFees* and

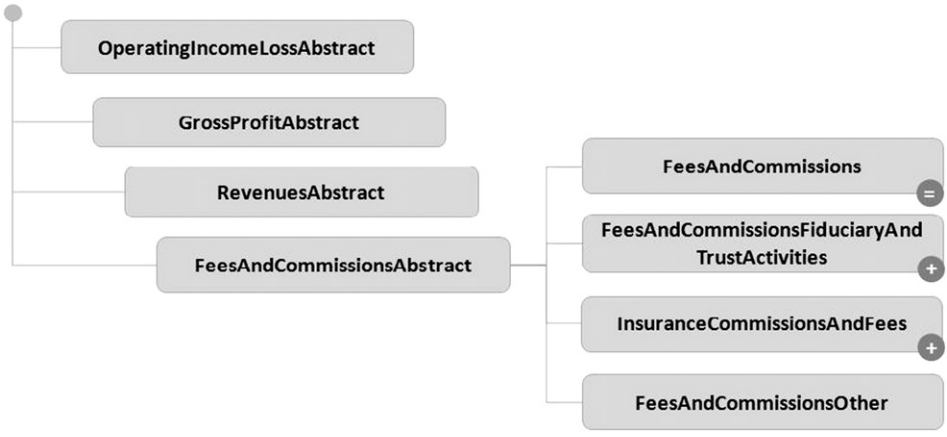


Fig. 4. U.S. GAAP presentation hierarchy extract.

FeesAndCommissionsOther. Querying the calculation linkbase establishes the explicit semantic hierarchy of *FeesAndCommissions* using its validation rule, overlaid on the element names in Fig. 4 of: *FeesAndCommissions* = *FeesAndCommissionsFiduciaryAndTrustActivities* + *InsuranceCommissionsAndFees* + *FeesAndCommissionsOther*.

XBRL also exhibits variation between what an instance document discloses and the taxonomy provides as hierarchy. For example even though the instance document may disclose *NonInterestIncome* as being derived from *AssetManagementFees*, the taxonomy denotes *NonInterestIncome* as being calculated from *InvestmentBankingAdvisoryBrokerageAndUnderwritingFeesAndCommissions* which in turn is calculated from *FeesAndCommissions*. The instance document would add to it that *FeesAndCommissions* comprise *AssetManagementFees*. Therefore when extracting values: first, relationship information in both the instance document and the taxonomy need to be considered; second, the transitive closure of allowable calculations needs to be determined from the taxonomy to ensure that the same information is being retrieved. The transitive closure would contain calculation relationships to its direct children, as well as the children's' children, For *NonInterestIncome*, this means relationships to *AssetManagementFees*, as well as *InvestmentBankingAdvisoryBrokerageAndUnderwritingFeesAndCommissions* and *FeesAndCommissions*. Once the transitive closure is established, determining what pair exists, and answering queries such as whether element A contain element B, is straightforward.

XBRL networks allow different views on concept and relationship, termed roles, and use the extended link roles to define and differentiate between these networks. The semantics of these views can be drawn from the label, presentation and calculation linkbases or added as extensions. Modelling extensions that deviate from the core taxonomy specification remain problematic as element interpretation relying upon concept similarly will depend on the particular view taken. The XBRL extended link roles act as data containers and provide unique identifiers accessible through a URI, although they are not resolvable or globally valid. Relevant recent work by (Debreceny et al., 2009) investigated the ability of SEC filings to automatically populate financial ratios using financial concepts. A methodology was developed to identify, match and determine alternate elements suitable for substitutions where concepts were not supported based on a set of six generic patterns. Both methodology and patterns represent a useful reference for future work on semantic context establishment and transitive closure checking.

4.2.2. Converting XBRL to RDF

Converter tools are available to automate the conversion of XBRL taxonomies and instances to their RDF (Raggett, 2009) or OWL (Declerck and Krieger, 2006; García and Gil, 2009; Bao et al., 2010) equivalent representation. The EDGAR wrapper[22] generates Linked Data facts from XBRL forms 10-Q, 10-K, 8-K, S4-A, and 6-K. Targeted XBRL mapping to RDF to facilitate integration with and querying across SEC, DBpedia and US Census data has also been attempted (Openlink, 2009). Business uses cases targeting the use of XBRL as a background semantic resource to make financial risk facts extracted from business reports semantically explicit (Declerck et al., 2008) and transforming XBRL taxonomies and instances into RDF to provide business users transparent access to company information across national and linguistic boundaries have been demonstrated (Wunner et al., 2010). OWL has also been used to support fundamental integrative approaches for business performance management standards compliant reporting and business analytics for XBRL (Spies, 2010).

Fig. 5 provides an RDF representation of a basic XBRL financial statement fact. The upper portion of the figure provides an XBRL document instances excerpt of the taxonomy element *us-gaap:AccruedIncomeTaxes.* ContextRef denotes the monetary values for multiple reporting periods. The first financial statement fact states "accrued income taxes in 2010 were 2.97 billion USD" and the second that "accrued income taxes for the first quarter of 2011 as 1.13 billion USD". The representation of the first fact is also given in its equivalent RDF turtle format representation generated using an XBRL to RDF convertor tool.[23]

[22] http://edgarwrap.ontologycentral.com/
[23] Monnet project, http://monnet-project.eu/

```
<us-gaap:AccruedIncomeTaxes contextRef="BalanceAsOf_31Dec2010"
    unitRef="USD" decimals="-6">297000000</us-gaap:AccruedIncomeTaxes>
<us-gaap:AccruedIncomeTaxes contextRef="BalanceAsOf_31Mar2011"
    unitRef="USD" decimals="-6">113000000</us-gaap:AccruedIncomeTaxes>

@prefix xbrl-us: <http://xbr.us/us-gaap/2009-01-31/> .
@prefix xbrl: <http://www.xbrl.org/2003/instance/> .
@prefix monnet: <http://monnetproject.eu/xbrl/> .
@prefix ex :<www.deri.ie/about/team/member/sean_o'riain/example>

ex:Assets1 rdf:type xbrl-us:Assets ; monnet:value "6782295000"xml:double ;
          xbrl:unitRef iso4217:USD ; xbrl:contextRef xbrl:Context1 .
ex:Context rdf:type time:Instant ; rdf:type xbrl:Context ;
           time:inDateTime [ rdf:type time:DateTimeDescription ;
                time:day "31" ; time:month "12" ; time:year "2010" ; ]
```

Fig. 5. XBRL instance from Metlife Form 10-Q, 31.03.2011 and its RDF representation.

4.3. Semantically integrated graph structured data

The XBRL abstract model defines the semantic level conceptualisation of the underlying structures and relationships (XBRL, 2011). Having clear semantics on XBRL data will help automated pre-processing and transformation into technologies such as RDF and OWL, unification with existing XBRL taxonomies and investigation into XBRL behaviour in the larger financial ecosystem. Understanding XBRL at the abstract level should also help predict and address issues with technology combinations that increase comparability, enhance XBRL ease of use and reduce the barrier for software developers. The XBRL abstract model fosters entity-centricity in XBRL data and as such improves integration of XBRL data with other data sources. The abstract model does however remain unclear as to how to actually link instances and concepts with other data sources. RDF with its well defined and understood abstract model assists application developers adopt a data level mashup approach to integrate data sources. DERI Pipes for example (Le Phuoc et al., 2009) allows for data feeds to be merged. Google Refine[24] (a tool for working with messy data) allows for linking to other databases such as Freebase and Google Refine with RDF Extension[25] can export data as interlinked RDF.

Essential to integration is the establishment of relationships between data items and identifiers that uniquely identify entities. Extracting and integrating financial information from XBRL instances and other data sources relies upon establishing equivalence to a central entity (e.g., company). Fig. 6 depicts the unified modelled distributed data environment with the interconnections in place that would support the previous scenario and queries from Table 1. For clarity we take an idealised view that instances of each entity are identified via a URI. Solid lines denote relations that already exist in the data; dotted lines are inferred, either via natural language processing techniques (for example, identifying organisations in news items via named entity recognition) or logical inference (for example, establishing equality between companies in SEC's EDGAR and Freebase via the shared CIK attribute).

The *owl:sameAs* relation establishes equality between data instance or schema-level URIs allowing the connecting and equating of data from different sources together. The relation is also transitive, so if Company A mentioned by Bloomberg is the same as organisation B from DBpedia, and organisation B from DBpedia in turn is the same as Company C from Freebase, then Company A is *the-same-as* Company C. Equality is an important and useful assist in integrating and linking entities such as organisations whose name may change over time or across jurisdictional boundaries that Semantic Web standards accommodate. To avoid having descriptions of entities split over numerous instances the same data is connected with a particular entity allowing the removal of duplicates and the central entity be incrementally augmented with additional attributes and data such as competitor relations, CEOs or share price as they become available.

[24] http://code.google.com/p/google-refine/
[25] http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/

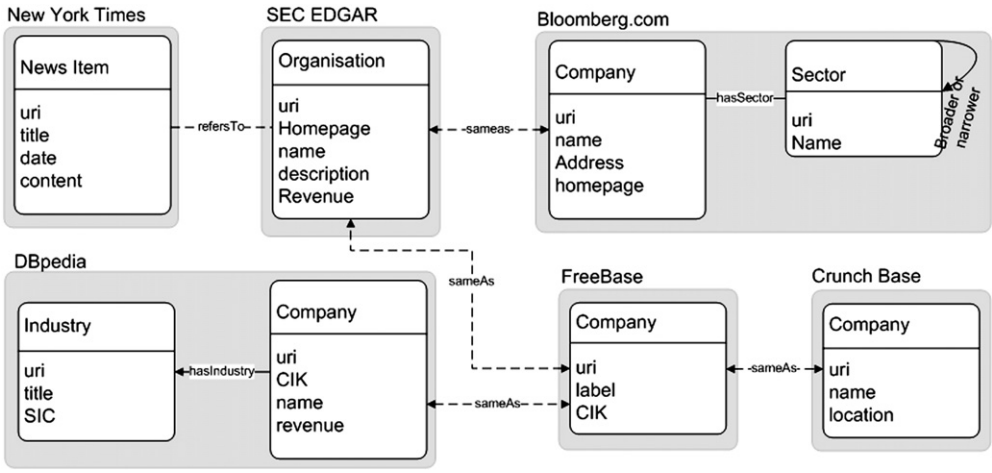


Fig. 6. Source, class and attribute integration modelling.

4.4. Semantic information access

The mashup with sources drawn from the wider ecosystems represents its own localised information value chain providing a search and interaction platform that allows information professionals freely navigate the information space of entities, their defined relationships and query data beyond the limited expressiveness of keyword searches. Semantic Web integrated sources provide centralised Web access that facilitates complex question answering, information search and navigation for both application developers and information professionals. The environment also allows for keyword search, faceted search, semantically enhanced entity centric search (people, company, financial instruments) and traversal through entity relationships that allow alternate views and discovery possibilities of the data be made (O'Riain et al., 2011).

Linked Data publishing tools render content directly from RDF stores or can provide a Linked Data view over non-RDF legacy data. RDF stores are the Semantic Web version of the relational database and there

Table 2
Selected sample of RDF data stores.

Repository	Details	Provider
SESAME	Open source framework for storage inferencing and query over RDF. Open Source/ BSD license.	Originally developed by Aduna, *http://www.openrdf.org/*
BigOWLIM	Layer over the SESAME RDF repository, academic, free for research purposes	Ontotext, *http://www.ontotext.com/owlim/big*
Talis Platform	Provides native storage for RDF/Linked Data. Platform store accessible via a SPARQL endpoint and REST APIs. Developer trial licenses, commercially licensable	Talis, *http://www.talis.com/platform/*
Virtuoso	DBMS systems which combines database and RDF triple store functionalities. RDF data can be stored directly in Virtuoso or created on the fly from non-RDF relational databases. Open sourced, commercial	OpenLink Software, *http://virtuoso.openlinksw.com*
BigData	Supports SPARQ, RDFS and limited OWL inference. Open-source license (GPL v2)	BigData, *http://www.systap.com/bigdata.htm*
AllegroGraph	Supports SPARQL, RDFS++, and Prolog reasoning. Commercial	Franz, Inc., *http://www.franz.com/agraph/allegrograph/*
Oracle Database 11 g Semantic Technologies	Open standards-based, RDF management platform using the relational database management system. RDF data (triples) are persisted like other object-relational data types. Commercial	Oracle, *http://www.oracle.com/technetwork/database/options/semantic-tech/index.html*

are a variety of academic and commercial stores. Table 2 provides a sampling of the more commonly used stores.

Loaded into an RDF repository, the data set is open to Web-based interactive exploration, and complex queries using the RDF query language SPARQL. The SPARQL representation for the previous query '*Which companies offers Life Insurance with revenue > USD 1 M?*' is:

```
SELECT DISTINCT ?companyName ?revenue
WHERE { ?company dcterms:subject category:Life_insurance_companies.
    ?company rdfs:label ?companyName .
    ?company dbpedia-owl:revenue ?revenue .
    FILTER ( datatype(?revenue) = http://dbpedia.org/datatype/usDollar
    && xsd:float(str(?revenue)) > 1000000)} }
```

Returning raw triple format output of:

"Protective Life"@en	"3.06E9"	<http://dbpedia.org/datatype/usDollar>
"Assumption Life"@en	"1.057E8"	<http://dbpedia.org/datatype/usDollar>

The example demonstrates the distributed nature of the information sources, definitions based on multiple vocabularies and the ability of the language to query cross vocabularies. Previous examples have included the US-GAAP taxonomies to describe the entities in use. Here Dublin Core (dcterms) is used to define the company category, and DBpedia (dbpedia-owl), to define revenue and monetary type US Dollar. Challenges associate with querying Web data due to users lack of an a-priori understanding of the available datasets, have also resulted in search approaches that looks to entity-centric, structured, question answering and best-effort natural language methods (Freitas et al., 2012).

5. Current XBRL and Semantic Web technology interactions

Interested in activities that would assist data innovation enabled by information integration we categorised activities found as applicable to: 1) Semantic approaches to fact extraction from financial text that linked Data would assist consumption of; 2) Semantic Web approaches to XBRL taxonomies leveraging for enhanced understanding; and 3) Semantic Web approaches to the use of ontologies as architectures for vocabulary reconciliation, next discussed.

5.1. Fact extraction from financial text

On-going efforts to leverage unstructured text include works to develop new insight from Web based financial news articles and feeds (Bovee et al., 2005; Saggion et al., 2007), extract corporate earning facts (Conlon et al., 2007) and develop stock market prediction (Schumaker and Hsinchun, 2009) using text analytics. SEC and FDIC filings have also been used to identify critical banking hubs as part of systemic risk analysis based on connections between financial firms, derived from fact aggregation (Hernandez et al., 2010).

Prior to the availability of XBRL, financial information sourcing looked to various text-based extraction approaches to extract from corporate filings (Leinnemann, et al., 2001; Gerdes, 2003; Grant and Conlon, 2006; O'Riain and Spyns, 2006). XBRL has made the extraction of financial facts through taxonomy labels easier but the disclosure notes remain problematic due to their natural language content (Grant and Conlon, 2006). Characteristic of these efforts is their handcrafted approach that lacks a common format that would facilitate newly generated information to be included, combined and reused, externally to their proprietary environment.

Fig. 7 illustrates how a sentence within disclosure notes could be classified, have their terms defined and transformed into financial facts making inherent semantics explicit, for mashup and inclusion with existing financial data as part of a process to build a holistic picture of related information. The vocabularies are used to enhance elements of the projects abstract model not present in the core vocabulary. The IFRS is used to defines *revenue*, the XBRL formatted European Business Registry taxonomy (xebr namespace) defines *net income*, the XBRL US GAAP *reporting period* and ISO4217 *monetary type* definition. Without further context, monetary units cannot be established for the second statement resulting in its eliminating as a comparative financial fact candidate.

"The company generated sales revenues of 54 million euros in the 2007 fiscal year"

```
:Revenue_1
    a ifrs::Revenue ;
    monnet:contextRef xbrli:FROM_01_2007_TO_31_2007 ;
    monnet:value "54000000" ; xbrli:unitRef iso4217:EUR .

xbrli:FROM_01_2007_TO_31_2007 a monnet:Context , time:Interval ;
    time:hasBeginning [time:inDateTime [time:day "01" ; time:month "01" ; time:year "2007"]];
    time:hasEnd [ time:inDateTime[time:day "31" ; time:month "12" ;time:year "2007"]] .
```

"Siemens again showed outstanding profitability in the second quarter of fiscal 2010, increasing net income 48% year-on-year, to 1.5 billion." into

```
:NetIncome_1
    a xebr:NetIncome ;
    monnet:contextRef xbrli:FROM_01_2010_TO_06_2010 ;
    monnet:value "1500000000" .

xbrli:FROM_01_2010_TO_06_2010 a monnet:Context , time:Interval ;
    time:hasBeginning [time:inDateTime [time:day "01" ;time:month "01" ; time:year "2010"]];
    time:hasEnd [ time:inDateTime [ time:day "30" ;time:month "06" ; time:year "2010"]].
```

Fig. 7. XBRL financial fact example (in Turtle syntax).

5.2. Taxonomies as lexical resources

Existing XBRL, IFSB, IASB financial and regulatory taxonomies standards and specifications can be used as domain lexical resources to drive term identification and make semantics explicit. The degree of fit on a terminological level, where taxonomy elements were compared to terms found in the financial statements has been applied to the Commercial & Industrial XBRL taxonomy to target accounting terms for extraction from semi-structured SEC filings (Bovee et al., 2005). XBRL taxonomies have also been applied as domain knowledge sources for multilingual fact interpretation and extraction (Wunner et al., 2010). SFAS 87 and 158 guidelines were used to classify, and semi-automatically create a taxonomy structure for the pension related footnotes of Form 10-K (Vasundhara and Miklos, 2010). To assist hierarchical formulation historical data taxonomy structures were compared to the XBRL US-GAAP and found to contain additional terms in different hierarchical locations. Taxonomy differences were attributed to company reporting trends adding greater levels of aggregated information, new terms and sections to the footnotes as opposed to the disaggregated structure followed by XBRL. Mapping pension disclosure data to taxonomy tags was also used to evaluate taxonomy comprehensiveness (Vasundhara and Miklos, 2010).

The use of existing lexical resources to identify domain terms and their variations, assist with taxonomy creation or suggest extensions and can contribute towards further domain conceptualisation. Both are useful for instructing information extraction techniques and enhancing meta-tagging that will directly enhance indexing, search and processing ability. From here an additional research topic is the rigorous evaluation of XBRL taxonomies for target financial sub domains applicability as a generally repeatable approach to dealing with unstructured and semi-structured text. The research question could then advance to determining whether: 1) additional semantic tagging enhances search and look up and; 2) whether multiple fact type integration does actually deliver on enhanced analysis.

5.3. Taxonomy alignment

XBRL formatted financial reports are exploited by stakeholders across a wide range of organisations and jurisdictions. As XBRL taxonomies model domain standard reporting frameworks, comparisons of reported facts across frameworks and jurisdictions remains an unresolved area (Bonsón et al., 2009). Lack of resolution at the schema level, varying accounting standards and practices make comparison based on concept equivalence difficult to conduct (Curry et al., 2009). For example the element *us-gaap:GoodwillImpairmentLoss* equates to *ifrs:ImpairmentLossRecognisedInProfitOrLoss*.

The IFRS and the European Business Registry taxonomy (xEBR) have aligned core accounting concept across EU jurisdictional XBRL GAAPs to support multi-lingual reporting and querying of country specific

GAAPs (Wunner et al., 2010). An ontology architecture is used for taxonomy alignment based on xEBR core terms and concept equivalence established using OWL language constructs. XBRL lacking such constructs can either look to formally adding semantics through the specification or adopt a similar approach toward ontology use for semantic inter-operability and translation between XBRL formats as proposed by (Wenger et al., 2011). Although only at the conceptual level, a financial statement ontology was outlined to reconcile between industry and country taxonomies. The conversion of financial statements (foot notes, management discussion or non-financial disclosures) were left as pragmatic modelling decisions due to language translation difficulties inherent in the conversion process.

Further incremental evaluation of the XBRL ontological architecture on defined taxonomies to establish validity on the approach should be undertaken. A first step would be to establish taxonomy pairs such as the US GAAP and XBRL US Management's Discussion and Analysis Taxonomy or US GAAP and XBRL US Schedule of Investments Taxonomy. Next steps would be to perform reconciliation based on some defined subset and filings processed to extract facts based on taxonomy elements. Once instantiated with financial statements the ontology usefulness to assist Open Data integration could be evaluated for success based on the criteria that sought to establish conversion utility such as what are the benefits, how repeatable and maintainable is the approach? Analysts could then undertake analysis based on the integrated data and evaluate its contribution to analysis outcome.

6. Challenges for Semantic Web and XBRL global financial data ecosystems

A report from the XBRL International Standards Board (XSB) outlines future business goals and areas of technical developments for XBRL and addresses the broader goals of easier access for developers, making XBRL information more compatible across taxonomies and simplifying XBRL consumption to increase its momentum and enhance business information processes (XSB, 2010). For an ecosystem where business information flow is continuous, the goals will be realised through the ability of the environment to accommodate information access, facilitate information integration and provide methods for ease of reuse.

To identify where research on XBRL and Linked Data technologies can make a contribution to ecosystem evolution, we decomposed the broader XSB goals into a series of topic areas that support information integration either directly or indirectly. The result listed in Table 3 provides a summary of research discussion made throughout the paper and literature review. The table should be interpreted starting from an XSB goal, moving to a sub consideration and then to what contribution can be expected from XBRL and then of Linked Data/Semantic Web. The XSB goal of easier access decomposes into three considerations of availability, accessibility and representational format. Accessibility further decomposes into taxonomy and data set. XBRL contribute to data sets through filings availability across the jurisdictions in XBRL format and Linked Data from multiple sources in multiple structured formats.

As the majority of topics within the table have been previously discussed we select two for further research discussion due to their contribution to information integration and business impact for Open Data consumption. The first recognised as a key problem by both XBRL and W3C (W3C, 2009) is the ability to uniquely identifying higher order entities to guide consolidation of multiple information types. Data source addition is referred to as information *type* to reinforce that varied types of information such as event facts from financial news bulletins or statement facts from analyst reports can be added once they are extracted and converted to RDF structured format. The other associated with pervasive Web data usage highlights concerns with source trustworthiness, data provenance and legal aspects of data consumption.

6.1. Unique identification of financial entities

All data integration typically depends on a conceptual model that details central consolidating entities and a physical model that specifies unique identifiers to drive data integration efforts. Open Data available from multiple Web locations exacerbates the problem as data sets are accompanies by their own unique identifiers issued by some originating authority. An important aspect for Semantic Web technologies is the issue of uniquely identifying resources essential for integrating data across sources. Assignment of URIs to entities is optional within RDF and currently there is limited agreement on the

Table 3
High level XBRL and W3C research area compatibility matrix.

	Consideration/ topic area	XBRL contribution	Linked data contribution
Easier access	Availability	Government, largely confined to consolidated reporting	Government, private, organisational financial, economic, business, general
	Accessibility	Publically available on Web	Publically available on Web
	Taxonomy	Financial domain specific	Generally non-financial domain specific vocabularies and ontologies.
	Data Set	*Jurisdiction limitation on availability apart from SEC EDGAR. XBRL format.*	*Multiple topic (e.g. government, business, financial, scientific) freely available on multiple structured formats*
	Representational format	Jurisdictional variants	RDF, OWL
	Mark-up method	Commercial, open source tools	Commercial, open source convertors
Compatibility across taxonomies	Multiple information type inclusion	Confined to structured financial statements	For any RDF/OWL serialisation
	Lexical Resources	Domain specific taxonomies, specifications	Few domain specific vocabularies available
	Availability		
	Tool Support	Proprietary, third party	Proprietary, open source, ontology based information extraction
	Interoperability with other technologies	Not yet addressed, compatibility committee in place.	Multiple format convertors available
	Unique Identification Scheme	*Filing URI, filing authority unique id's.*	*Web based (global) URIs*
	Link to 'other' information	Not catered for in representational format	Catered for in representational format
	Concept equivalence	Versioning, across standards not catered for	Cater for in representation language
	Shared understanding	Difficult to establish, handcrafted	Vocabulary and ontology alignment
Simplifying consumption	Abstract Model	Initial	Established
	Query	Financial statement labels, key word, financial comparison	Keyword, object
	Navigation	Faceted, concept hierarchy drill down	Faceted. path traversal through and across data sets

use of common URIs for the same instances across sources. Lack of agreement on URIs for resources can result in missing associations between resources during integration and entity consolidation difficulty. Linked Data implementations have found that infrequent and inconsistent reuse of instance identifiers across sources can lead to problematic data integration and data about the same entity remaining fragmented across multiple instances. As the desired outcome of a Linked Data effort is an integrated well connected data space, agreement on identifier naming is crucial for entity association. Fusing identifiers is also important for entity-centric applications as it underpins search and query engines and interactive browsing tools capability. Lack of formal specification for equivalences determination confines most approaches to probabilistic methods whereas formal approaches can rely on properties unique to the entity to determine its identity (e.g. the SEC CIK for company identification, email addresses, URI of an XBRL filing). XBRL would benefit from greater re-use of URIs that are resolvable and globally valid.

Data elements tied to companies could be globally referenced (and retrieved) rather than referenced just within the filing document. Evaluating the potential use to XBRL of authoritative identifiers for schema elements and data instances to assist wider data type integration would make an interesting research area. Schema level mismatches such as the SEC CIK to identify corporate officers or financial instruments with DBpedia use of URI to define the same, might then be resolvable. Criteria for evaluation would target data quality dimensions such as completeness of consolidated items and level of fact duplication.

A longer term research goal would be establishing a naming convention and description as to what a semantically rich organisational entity would be. Business relevant solution characteristics for entities should include uniqueness, international validity, authoritativeness and hierarchical. Suggested elements

for incorporation would be a unique identifier, attributes of name and business parameters such as business type, jurisdiction, business segment and next filing date (W3C, 2009).

6.2. Use of Web data

In an open environment, data aggregators have little influence on data publishers. Financial data is provided for consumption at with varying levels of abstraction, classification and aggregation (Curry et al., 2009). Accuracy of corroborating facts and statements and settling disputes between conflicting data and providers can impede the utility of applications operating over such data (Bartley et al., 2009).

Understanding the origins and trustworthiness of data presented as part of any analysis is often the first point that requires clarification. In information processing, heuristics can play a part but where data sets have not previously been encountered there is little context to establish data quality. For automated integration of non-XBRL sources, link-based algorithms taking into account the sources of information similar to PageRank can provide some coarse measure but they need to be adapted to Web linkage patterns (Harth et al., 2009). Data provenance can play a fundamental part in accessing data trustworthiness. Tracking data origins can serve as basis for assessing trustworthiness of aggregated data particularly where a work flow (e.g. report creation, comparative analysis) has altered the data from its original state. In such instances, understanding the context and subtleties of the data will be lost without access to the original financial statements (Debreceny and Gray, 2001). Provenance of a resource represents an information record that describes entities and processes involved in production, delivery or influences on that resource (W3C, 2012). Semantic Web provenance models propose a common approach for data quality assessment (Hartig, 2009) and coverage to accommodate information such as location, timeliness, origin, owner organisation and operating conditions (Freitas et al., 2011).

Emergent future research topics for the consumption of Open Data and XBRL data is the provision of machine-readable licensing information with data. Even though XBRL filings are made available from authoritative sources that are normally re-usable without charge, information consumers will require an understanding of the provenance of financial statements extracted and combined with other statements. Open Data catalogues are made available for use adhering to open licenses. Licensing of data should be explicit to avoid misinterpretation on usage terms by consumers. Common licenses in use are Creative Commons,[26] Public Domain Dedication and License[27] and GNU Free Documentation License.[28] Implication for data aggregation, re-use or recommender type financial service provision require investigation. While there is active Web community discussion on Linked Data Business Models e.g. (Brinker, 2010), the licensing issue first requires consensus. Protecting privacy on a Web that operates as a single global ecosystem will require a combination of technical and legal awareness as to what data to provide, in which context and how it can be consumed. Recent research efforts in the domain are the Transparent Accountable Data Mining Initiative on Information Accountability (Weitzner et al., 2008) and privacy on the data Web (O'Hara and Shadbolt, 2010).

7. Conclusion

We investigated the feasibility of leveraging XBRL data with Open Data as part of a wider integrated information environment and developed a series of research questions. The first considers the role of Open Data within the information value chain, detailing the use of Open Data and how it could be leveraged with XBRL. The second investigates how XBRL can be integrated with Open Data using a Linked Data approach, providing examples of existing synergies between the two technologies. The final research question is concerned with remaining challenges not addressed

[26] http://creativecommons.org/
[27] http://www.opendatacommons.org/licenses/pddl/
[28] http://www.gnu.org/copyleft/fdl.html

elsewhere in the investigation that would assist information accessibility and contribute towards financial ecosystem formulation.

Findings from the overall analysis highlight that using the RDF representational format and adhering to Linked Data principles exposes the semantics of information making it easier to establish relationships between data items. Linked Data was also found to be a useful mechanism with which to integrate financial facts extracted from unstructured financial sources with existing structured information. Key findings are: 1) RDFs usefulness as a common data representation for data combination purposes and 2) the ability to introduce Linked Data in a minimally disruptive fashion into existing information infrastructure by deploying it in a layered fashion that supports the exposure of existing data without requiring modification of the original format.

Despite the fact that Semantic Web formalisms have been identified as a natural selection for data integration from different sources (Bao et al., 2010), current evidences of leveraging XBRL with Open Data are largely confined to academic use cases. To progress further several non-trivial issues relating to representation differences remain. Particularly the ability of Semantic Web vocabularies to semantically model mathematical relations contained in the XBRL calculation and formula has to be conclusively addressed (Lara et al., 2006). Relationship modelling is important as it can be used directly for data validation and hierarchical modelling. Rule based languages such as SPIN[29] or RIF[30] have been proposed. Semantic Web Rule Language (SWRL) has also been considered but only as part of a conceptual ontological architecture (Wenger et al., 2011).

Both XBRL and Linked Data can be treated as 'customisable architectures' to assist with information interpretation and integration. XBRL taxonomies can be used as domain lexical resources to help interpret unstructured content sections, and suggest extensions to base taxonomies. Both have been proven through industrial and academically use cases. Advancing this capability to the level of generalised functionality would benefit taxonomy and software developers and enhance XBRL consumption. Learning's and approaches from the established practice of using ontologies for vocabulary translation within Semantic Web can also be directly applied to further develop, instantiate and evaluate the proposed ontology reconciliation approach for XBRL taxonomy formats (Wenger et al., 2011). As part of a wider ecosystem XBRL should also consider the usefulness of URIs for financial resource identification on the Web.

There is a growing trend toward wider data integration using enhanced semantics that also looks to inclusion of unstructured information. Whether part of an ecosystem, mashup or semantic extract-transform-and-load process, the approaches reflect the use of Semantic Web technologies to help handle data heterogeneity (Niinim, 2009; Nebot and Berlanga, 2010) and fact inclusion using natural language processing (Simitsis et al., 2010). Academic research projects have demonstrated the possibility of using Semantic Web technology to subsume XBRL information and make it available as Web resources. Linked Data technology offers assistance in overcoming Open Data interoperability issues between financial and non-financial information. Linked Data can also serve both as a technology framework and model to support the delivery of these information resources. XBRL interoperable with Linked Data allows ecosystem evolution that will contribute to wider use and consumption of XBRL.

Acknowledgements

The work presented in this paper has been funded in part by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2), the EU FP7 Activity ICT-4-2.2 under Grant Agreement No. 248458, Multilingual Ontologies for Networked Knowledge (MONNET) and the EU Network of Excellence PlanetData (ICT-NoE-257641) project. We would also like to thank Tobias Wunner and Benedikt Kämpgen whose research contributed to the XBRL examples and figures used.

[29] SPARQL Inferencing Notation modelling vocabulary of RDF used to specify rules and logical constraints
[30] Rules Interchange Format, a standard rule language for the Semantic Web infrastructure

Appendix Web based financial & economic Open Data sets

Open Data set type	Web Site (Verified 01 March 2011)
Historical FX rates	http://oanda.com/convert/fxhistory
Historical stock prices	http://finance.yahoo.com/q/hp?s=yhoo
Mortgage borrowing, unsecured credit, credit card use and deposits	http://www.bba.org.uk/statistics
Historical and current market data analysis, implied and realized (historical) volatility	http://www.ivolatility.com
Commodity derivatives	http://www.liffe-commodities.com/
US Fundamentals	http://www.sec.gov
UK government statistics	http://www.statistics.gov.uk
US government statistics	http://www.census.gov
Bureau of labour statistic	http://www.bls.gov
Bureau of economic analysis	http://www.bea.gov/
Economic research service	http://www.ers.usda.gov/
Board of Governors of the Federal Research system	http://www.federalreserve.gov http://www.federalreserve.gov/econresdata/default.htm
US Treasury	http://www.treasury.gov
Historical Data for S&P 500 Stocks	http://kumo.swcp.com/stocks/
NYSE Euronext's stock markets five year database	http://www.bourse-de-paris.fr/en/index_fs.htm?nc=2&ni=11&nom=marche
Market Watch	http://bigcharts.marketwatch.com/historical/
S&P 100 index, Yahoo!finance	http://finance.yahoo.com/q/cp?s=%5EOEX
Open Government at the Department of Defence	http://open.dodlive.mil/link-library/financial-data/
Department of Defence Information Related to the American Recovery and Reinvestment Act of 2009 (Recovery Act)	http://comptroller.defense.gov/Budget2011.html
Eurostat, European Commission, Economy & Finance	http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/themes
Commonwealth of Massachusetts annual budget including current and archived budget documents, revenue projections	https://wiki.state.ma.us/confluence/display/data/Data+Catalog https://wiki.state.ma.us/confluence/display/data/Financial+Data

References

Auer S, Dietzold S, Lehmann J, Hellmann S, Aumueller D. Triplify – Light-Weight Linked Data Publication from Relational Databases. Proceedings of the 18th World Wide Web Conference; 2009.

Bao J, Rong G, Li X, Ding L. Representing Financial Reports on the Semantic Web: A Faithful Translation from XBRL to OWL. Rule ML. 4th International Web Rule Symposium: Research Based and Industry Focused., VA, USA; 2010.

Bartley JW, Chen YA, Taylor E, Zalkin A. Comparison of XBRL Filings to Corporate 10-Ks–Evidence from the Voluntary Filing Program. Available at SSRN:; 2009.

Bizer C, Cyganiak R. D2R Server–Publishing Relational Databases on the Semantic Web. Poster at the 5th International Semantic Web Conference (ISWC2006); 2006.

Bonsón E, Cortijo V, Escobar T. Towards the global adoption of XBRL using international Financial Reporting Standards (IFRS). Int J Account Inform Syst 2009;10:46–60.

Bovee M, Kogan A, Nelson K, Srivastave R. Financial Reporting and auditing agent with net knowledge (FRAANK) and eXtensible business reporting language (XBRL). J Inform Syst 2005;19(1):19–41.

Brinker S. Chief Marketing Technologist, Business models for linked data and Web 3.0. http://www.chiefmartec.com/2010/03/business-models-for-linked-data-and-Web-30.html2010.

Grammel L, Storey MA. A Survey of Mashup Development Environments; 2010.

Hernandez AM, Ho H, Koutrika G, Krishnamurthy R, Popa, Stanoi IR, et al. Unleasing the Power of Public Data for Financial Risk, Measurement, Regulation and Governance in World Wide Web International Conference. Raleigh, North Carolina; 2010.

Conlon S, Lukose S, Jason GH, Vinjamur A. Automatically Extracting and Tagging Business Information for E-Business Systems Using Linguistic Analysis. In: Salam AF, Stevens J, editors. Semantic Web Technologies and E-Business: Toward the Integrated Virtual Organization and Business Process Automation. Hershey: IGI Global; 2007.

Curry E, Harth A, O'Riain S. Challenges Ahead for Converging Financial Data. Workshop on Improving Access to Financial Data on the Web, Co-organized by W3C and XBRL International, FDIC, Arlington, Virginia USA, 5-6 October; 2009.

Curry E, Freitas A, O'Riain S. The Role of Community-Driven Data Curation for Enterprises. In: Wood D, editor. Linking Enterprise Data. Springer978-1-4419-7664-2; 2010.

Declerck T, Krieger H. Translating XBRL Into Description Logic. An approach using Protege, Sesame & OWL. Proceedings of Business Information Systems (BIS); 2006. p. 455–67.

Declerck T, Kreiger H, Saggion H, Spies M. Ontology-Driven Human Language Technology for Semantic-Based Business Intelligence. Proceedings of the 18th European Conference on Artificial Intelligence (ECAI); 2008.

Debreceny R, Gray GL. The production and use of semantically rich accounting reports on the Internet: XML and XBRL. Int J Account Inform Syst 2001;2(1):47–74.
Debreceny R, Farewell S, Felden C, d'Eri A, Piechocki M. Feeding the Information Value Chain: Deriving Analytical Ratios from XBRL filings to the SEC. Presented at the 20th XBRL International Conference, Rome, April; 2009http://20thconference.xbrl.org/.
The Economist, 2009. Banks and information technology: "Silo but deadly, Messy IT systems are a neglected aspect of the financial crisis". Print edition, http://www.economist.com/node/15016132 (Dec 3rd).
European Commission. European Information Society, Public Sector Information–Raw Data for New Services and Products. http://ec.europa.eu/information_society/policy/psi/index_en.htm2003.
European Comission. European Information Society, MESPIR Study. http://ec.europa.eu/information_society/policy/psi/mepsir/index_en.htm2006.
Freitas A, Knap T, O'Riain S, Curry E. W3P: Building an OPM based Provenance Model for the Web. J Future Generat Comput Syst 2011;27(6):775–80.
Freitas A, Curry E, Oliveira JG, O'Riain S.,: Querying Heterogeneous Datasets on the Linked Data Web: Challenges, Approaches, and Trends. IEEE Internet Computing. 16, 24–33 (2012).
Feldman S. The High Cost of Not Finding Information. International Data Corporation; 2003.
García R, Gimeno JM, Perdrix F, Gil R, Oliva M, López JM, et al. Building a Usable and Accessible Semantic Web Interaction Platform. World Wide Web, Vol. 13, No. 1-2. Springer; 2010. p. 143–67.
García R, Gil R. Publishing XBRL as linked open data. World Wide Web Workshop: Linked Data on the Web (LDOW2009); 2009.
Gartner. Identifies the Top 10 Strategic Technologies for 2009. Orlando, Florida: Gartner Symposium/Itxpo; 2009.
Gerdes J. EDGAR-Analyzer: automating the analysis of corporate data contained in the SEC's EDGAR database. Decision Support Systems; 2003.
Grant GH, Conlon SJ. EDGAR extraction system: an automated approach to analyze employee stock option disclosures. J Inform Syst 2006;20(2):119–42.
Guardian Newspaper. Investigate your MP's Expenses. http://mps-expenses.guardian.co.uk/.
Gundlach GT. Complexity science and antitrust? Antitrust Bulletin 2006;51(1):17.
Harth A, Kinsella S, Decker S. Using Naming Authority to Rank Data and Ontologies for Web Search. Proceedings of the 8th International Semantic Web Conference, Washington DC; 2009.
Hartig O. Provenance Information in the Web of Data. Proc. of the Linked Data on the Web Workshop at World Wide Web Conference (WWW); 2009.
Hepp M. GoodRelations, An Ontology for Describing Products and Services Offers on the Web. Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management (EKAW2008), vol. 5268. ; 2008. p. 332–47.
IDG News. Grant Gross, Companies Offer Services to Crunch Gov't Raw Data. http://www.pcworld.com/article/170105/companies_offer_services_to_crunch_govt_raw_data.html. Aug 12.
Lara R, Cantador I, Castells P. XBRL Taxonomies and OWL Ontologies for Investment Funds. 1st International Workshop on Ontologizing Industrial Standards at the 25th International Conference on Conceptual Modeling. Tucson, Arizona, USA; 2006. p. 14.
Le Phuoc D, Polleres A, Morbidoni C, Hauswirth M, Tummarello G. Rapid semantic Web mashup development through semantic Web pipes. Proceedings of the 18th World Wide Web Conference (WWW2009); 2009.
Leinnemann C, Schlottmann F, Seese D, Stuempert T. Automatic extraction and analysis of financial data from the EDGAR database. S Afr J Inform Manag 2001;3(2).
McAfee A. Enterprise 2.0: The Dawn of Emergent Collaboration. MIT Sloan Manage Rev 2006;47(No. 3):21–8.
Openlink Software Integrating Open Sources and Relational Data. http://www.openlinksw.com/dataspace/dav/wiki/Main/VOSArticleRDFandMappedBI.
Nebot V, Berlanga R. Building data warehouses with semantic data. Proceedings of the 1st International Workshop on Data Semantics–DataSem '10. New York, New York, USA: ACM Press; 2010.
New York Times. How Best Buy is using the Semantic Web. http://www.nytimes.com/external/readwriteWeb/2010/07/01/01readwriteWeb-how-best-buy-is-using-the-semantic-Web-23031.html.
Niinim M. An ETL Process for OLAP Using RDF/OWL Ontologies. In: Spaccapietra S, Zimányi E, Song I-Y, editors. Journal on Data Semantics, XIII. Heidelberg: Springer Berlin; 2009. p. 97-119.
O'Riain S, Spyns P. Enhancing the Business Analysis Function with Semantics. On the Move to Meaningful Internet Systems, Ontologies Databases and Applied Semantics (ODBASE); 2006.
O'Riain S, Harth A, Curry E. Linked Data Driven Information Systems as an enabler for Integrating Financial Data in Information Systems for Global Financial Markets: Emerging Developments and Effects. In: Yap A, editor. IGI; 2011.
Raggett D. XBRL Import An XBRL to RDF translator. http://sourceforge.net/projects/xbrlimport.
Saggion H, Funk A, Maynard D, Bontcheva K. Ontology-based Information Extraction for Business Applications. Proceedings of the 6th International Semantic Web Conference (ISWC 2007). Busan, Korea; 2007.
Schumaker R, Hsinchun C. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Trans Inform Syst 2009;27(2). Article 12.
Simitsis A, Skoutas D, Castellanos M. Representation of conceptual ETL designs in natural language using Semantic Web technology. Data Knowl Eng 2010;69:96-115.
Spies M. An ontology modelling perspective on business reporting. Information Systems, Volume 35, Issue 4, Vocabularies, Ontologies and Rules for Enterprise and Business Process Modelling and Management; 2010. p. 404–16.
Volz J, Bizer C, Gaedke M, Kobilarov G. Silk – A Link Discovery Framework for the Web of Data. Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009); 2009.
Weitzner D, Abelson H, Berners-Lee T, Feigenbaum J, Hendler J, Sussman G. Information Accountability. Comm ACM 2008;51(6):82–7.
O'Hara K, Shadbolt N. Privacy on the data Web. Comm ACM 2010;53(3):39–41.
Vasundhara C, Miklos V. Automating the process of taxonomy creation and comparison of taxonomy structures. 19th Annual Research Workshop on Strategic and Emerging Technologies American Accounting Association. San Francisco, CA, USA; 2010.
World Wide Web Consortium (W3C), Provenance Working Group, Working Draft, "The PROV Data Model and Abstract Syntax Notation", 02 February 2012 http://www.w3.org/TR/2012/WD-prov-dm-20120202/

W3C. Report for the Workshop on Improving Access to Financial Data on the Web, Co-organized by W3C and XBRL International, and hosted by FDIC, Arlington, Virginia USA, 5-6 October; 2009http://www.w3.org/2009/03/xbrl/report.html.
Wenger M, Thomas MA, Jeffery SB. An Ontological Approach to XBRL Financial Statement Reporting, AMCIS, Paper 448; 2011.
Wunner T, Buitelaar P, O'Riain S. Semantic, Terminological and Linguistic Interpretation of XBRL. Reuse and Adaptation of Ontologies and Terminologies Workshop at 17th International Conference on Knowledge Engineering and Knowledge Management (EKAW); 2010.
XSB. XBRL International Standards Board, "XBRL: Towards a Diverse Ecosystem", Discussion Document. http://www.xbrl.org/2010TechDiscussion/2010TechDiscussion.pdf2010.
XBRL. International Standards Board, XBRL Abstract Model 1.0 Public Working Draft 19 October. Specification available fromhttp://xbrl.org/Specification/abstractmodel-primary/PWD-2011-10-19/abstractmodel-primary-PWD-2011-10-19.html.