Chapter 10
Linked Data Driven Information Systems as an Enabler for Integrating Financial Data

Seán O'Riain
National University of Ireland Galway, Ireland

Andreas Harth
Karlsruhe Institute of Technology (KIT), Germany

Edward Curry
National University of Ireland Galway, Ireland

ABSTRACT

With increased dependence on efficient use and inclusion of diverse corporate and Web based data sources for business information analysis, financial information providers will increasingly need agile information integration capabilities. Linked Data is a set of technologies and best practices that provide such a level of agility for information integration, access, and use. Current approaches struggle to cope with multiple data sources inclusion in near real-time, and have looked to Semantic Web technologies for assistance with infrastructure access, and dealing with multiple data formats and their vocabularies. This chapter discusses the challenges of financial data integration, provides the component architecture of Web enabled financial data integration and outlines the emergence of a financial ecosystem, based upon existing Web standards usage. Introductions to Semantic Web technologies are given, and the chapter supports this with insight and discussion gathered from multiple financial services use case implementations. Finally, best practice for integrating Web data based on the Linked Data principles and emergent areas are described.

INTRODUCTION

Consumers of financial information vary from personal investors looking for investment opportunity, business executives seeking competitive advantage over their competition, to government regulators investigating corporate fraud. While the particular analysis performed by each of these information consumers will vary, they invariably have to source, consider and evaluate information from multiple resources such as the US Security and Exchange Commission (SEC) filings, corporate press releases, market press coverage, third

party information providers, expert commentary and specialist communities of interest. Failing to consider information from alternate or complimentary data resources brings the risk of lacking adequate insight for investment decisions or, of making an uninformed judgement call. Recent economic events have begun to bring sharp focus on the activities and actions of financial markets, institutions and not least regulatory authorities. Enhanced scrutiny will bring increased regulation (Economist, 2009) and information transparency (Wired, 2009), further increasing the burden on investors, analysts and investigators.

The last five years has also seen a growing number of Open Government transparency initiatives to make such public sector information available. Notable economic and financial Boxs are EuroStat (http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/), data.gov.uk, sec.org, data.gov, recovery.org, the World Bank (http://data.worldbank.org/data-catalog) and IMF (http://www.imf.org). Intended to freely provide public sector information, statistics data and economic indicators for transparency purposes, this raw government data is also being ''crunched'' by companies looking to inform their own business function, those of their stakeholders or as opportunity for new service provision (IDG, 2009). The EU Public Sector Information directive has previously mandated that data sets produced and collected by the public sector containing legal, financial, and economic data be made available to explicitly target innovative use of the data. The resulting catalogues represent the largest source of raw information in Europe available for re-use and integration into new products and services targeting an estimated twenty seven billion euro market (EC, 2006). The sources represent rich repository sources of business related data, competitor data, market watch data and government spend; their data catalogues are freely available online for uptake and analysis by the financial community.

Consolidating similar and related information into a single view remains hugely problematic.

Originally the realm of data warehousing, data-related issues of format, duplication and differing schema definitions remain (e.g., different jurisdictions can have different understandings of the same accounting terms). Across multiple sources the problem remains the classical data integration problem, where a common interoperable data abstraction is necessary. However Open Data is published in formats such as CSV, PDF, XML or text making integration and reuse costly, acting in effect as a barrier to entry. Extracting information from individual filings and structured sources is relatively straight forward where a machine readable format is available e.g. data encoded in the eXtensible Business Reporting Language (XBRL, http://xbrl.org/).

Semantic Web technologies provide powerful integration capabilities based upon a standard representational format. Linked Data represents best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the web based upon those standards. Used to publish semi-structured and structured data on the web, and as a means to provide more tightly interlinked datasets for enhanced search and querying, its adoption and use represents an opportunity to achieve standard access and inter-operability between and among financial data sets both for data consumption and publishing.

The chapter focuses on the use of Semantic Web technology, in particular using Linked Data principles, as an enabler for financial data integration that spans the enterprise firewall to include web-based financial content as part of financial data ecosystem.

Introduction introduces the case study fundamentals and challenges associated with financial data integration.

Overview of Semantic Web Technologies provides an introduction into the fundamental principles and building blocks of Semantic Web technology, namely global Universal Resource Identifiers (URIs), the Resource Description Framework (RDF) and its conceptual data model,

Linked Data principles and vocabularies. This section is intended to prime the reader for the follow on section on Financial Data Integration.

Anatomy of Web enabled Financial Data Integration begins with an outline of the high level process steps necessary and components architecture required to implement web-enabled data integration. The important step of how to convert structured data into the Semantic Web usable RDF format is then outlined. The remainder of the section discusses in greater detail the architecture component anatomy of a financial ecosystem, namely i) data preparation and integration, and ii) semantic analysis, querying and browsing. Practicalities of web-enabled data integration based on our case study prototype deployment experiences are included in the discussion.

Best Practices for Integrating Web and enterprise Data outlines best practice considerations and take-away points on standards to employ, as well as discussion on alternative architectural approaches and process steps.

Financial Data Ecosystems introduces the idea of a financial ecosystems along with its use and value to the information professional. The relatively new area of web and enterprise data merging is touched upon.

Conclusion concludes the chapter with a summary discussion and indications of emergent areas.

CASE STUDY FUNDAMENTALS

Discussion in the chapter is informed by our experiences gained from the design and implementation of a financial data integration case study. Our prototype includes the data sources detailed in Table 1 gathered from multiple distributed information systems such as databases, document repositories, government data sets, company data sets and news sites.

Of particular interest was the analysis of connections (relationships) between people and organisations, their financial and accounting records, sectoral classifications, products and associated information such as news items, patents and so on. The data was linked, consolidated and later enriched with other sources, such as current events, using an object (also referred to as entity) consolidation approach. Using Citigroup as an Box, company is the entity and Citigroup an instance. To avoid having descriptions of entities split over numerous instances the system connected the same data on a particular entity, removing duplicates and augmenting the central entity with ad-

Table 1. Case study data sources and elements

Resource Type	Type Instance	Data Targeted for Integration	Format
Company data	http://dbpedia.org/, http://crunchbase.com/ http://finance.yahoo.com/, http://hoovers.com/	Company name, location, address, description/history, officers, investors, products, industry classification, competitors, stock price	HTML, XML,RDF
Financial statements	http://edgar.sec.gov/	Company name, officers,financial and accounting data	XML, CSV, XBRL
News and current events	http://bloomberg.com/, http://seekingalpha.com/http://ny-times.com/	Company name, news, events	RSS, RDF
Jobs	http://indeed.com/	Company name, job posting, location, position	XML
Patents	http://epo.org/	Company name, patent	CSV
Trademarks	http://uspto.gov/	Company name, trademark	HTML
Geospatial data	http://geonames.org/	Geographical location	XML, RDF

ditional new attributes and data was essential. Integrated into a repository, the data set was then open to web-based interactive exploration, allowing complex queries such as:

- *Q: Information about CitiGroup?*
- *Q: Competitors of CitiGroup?*
- *Q: Key people of competitors of CitiGroup?*
- *Q: Reported financial results of CitiGroup?*
- *Q: Funding pattern of Vulcan Inc.?*
- *Q: News about financial companies with offices in Boston?*
- *Q: Job offerings of competitors of CitiGroup?*
- *Q: IP portfolio of companies that Vulcan Inc. invested in?*

Using the data sources listed in Table 1 we designed, implemented and evaluated a financial prototype data integration system that first collected and integrated the data and then allowed user search, interactive browsing and further analytics. Central to our approach was the use of Semantic Web technologies and its graph-centric knowledge representation, the more important principles and concepts of which are covered in the Overview of Semantic Web Technologies. Insights and Boxs from the implementation are used throughout the chapter.

Financial Data Integration Challenges

The key challenges to address for a data integration system may be classified as those related to increasing data inter-dependencies, data quality, text/data mismatch, entity identifier and schema mismatch and abstraction level mismatch. The database community address the challenges using the extract, transform and load approach (ETL) for data warehousing and datamart construction. With Semantic Web technologies however the nature of the challenge and the means of addressing them can vary due to the modelling primitives involved.

Increasing Inter-Dependencies between Data Sources

Having adequate amounts of relevant information to work with, remains contentious for most information professionals. Not satisfied with the information functionality that Business Intelligence tools provide financial information consumers are looking to include ancillary and even peripheral information sources in their deliberations. For an information consumer, increased inter-dependence poses its own particular challenges as the consumers have to source and include data themselves, rather than have the system perform the activity. As an example consider a financial analyst researching a particular company to develop financial insight using a variety of sources such as: i) the company's SEC Form 10-K to provide numerical insight, ii) a fund specialist such as MorningStar (http://xbrl.org/) providing information on investment funds which invest in the company, iii) Hoovers (http://hoovers.com) providing additional insight on proprietary company profiles and industry information, iv) Crunchbase, with some degree of data overlap from previous resource, may also supply data about technology companies, people, start-up's, investors and competitor information, v) finally Eurostat, the EU Statistical Office, could provide detailed economic indicators and statistics for consideration.

Data Quality

Data quality is a general challenge when automatically integrating data from distributed and independent sources. In an open environment the data aggregator has little influence on the data publisher, data is often erroneous, and combining data often further aggravates the situation. When performing additional data processing, such as reasoning (automatically inferring new data from existing data), erroneous data can negatively impact on overall quality of the generated results. In

summary, errors in signage, amounts, labelling, and classification will impede the utility of systems operating over such data (Bartley et al., 2010). The challenge for data publishers or consumers is how to coordinate in order to fix problems in the data or blacklist sites which do not provide reliable data.

Methods and techniques are needed to:

- Corroborate evidence;
- Determine the probability of given statement being true;
- Check integrity, accuracy and identify sanity checks;
- Act as clearing houses for raising and settling disputes between competing (and possibly conflicting) data providers;
- Interact with erroneous web data of potentially dubious provenance and quality.

Text/Data Mismatch

A large portion of financial data is described in text. As human language is often ambiguous, the same thing might be referred to in several variations. For instance, IBM can stand for International Business Machines or Integrated Business Management. Such ambiguity makes cross-linking with structured data difficult and also makes free text data difficult to process and use in software programs. Since the process of data integration aims towards the ability to pose structured queries over an integrated, holistic view of the entire data set, one of the requirements of a financial data integration system should be to overcome the mismatch between documents and data.

Object Identity and Separate Schema

Structured data is available in multiple formats such as CSV, XML, XBRL, and the record sets returned from relational databases. To integrate structured data it is necessary to first lift (map) that data to a common format. Having equivalent formats however does not guarantee consistency as the originating sources may state what is essentially the same fact differently. These differences exist at both the data description (schema) and actual data (individual object) levels (cf. Overview of Semantic Web Technologies). A relevant example of an object level mismatch would be where the SEC provides Central Index Keys (CIK) to identify corporate officers/people (CEOs, CFOs), companies, and financial instruments while other sources, such as Dbpedia, a structured data version of Wikipedia, uses URIs. Both sources use their own schema and conventions for fact representation and definition. Consolidating information from these sources therefore requires methods for identifying and reconciling different representations at both the schema and object levels.

Data Abstraction Levels

Financial data sources provide data at varying levels of abstraction. Boxs of different types of abstraction are:

1. **Type***:* Different data sources cover different topic areas, e.g., corporate officers (people), companies or services.
2. **Detail:** Data from different countries uses local currency, e.g., USD vs. EUR.
3. **Aggregation:** Data might differ in regional resolution levels, e.g. state-level data provided from one source and country-level data from another.
4. **Topic classification:** Different data sources use different taxonomy classifications, e.g. the SEC uses a different taxonomic classification of sectors incompatible with other industrial classification systems such as Standard International Trade Classification (SITC, http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=14).
5. **Principles and practices:** Countries can have vast differences in accepted accounting

practices and legislation base, e.g. various European Union region regulatory indicators may not be directly comparable with US regulations based on the US GAAP (http://xbrl.us/taxonomies/Pages/US-GAAP2009.aspx).

Whether integrating financial data for a more complete data set or better analysis and insight, the levels of abstraction have to be understood and catered for as they have a direct bearing on the integration activity and data provisioning (cf. Data Collection and Integration).

OVERVIEW OF SEMANTIC WEB TECHNOLOGIES

The Web continues to move from a medium that provides content as documents to a medium that provides content as data. For information consumers this means easier access to even greater amounts of data and easier access to more fine grained data, but the challenge remains how to manage and make best use of the data. For the publisher, it introduces a requirement on how fine grained data should be made available, and for the consumer, how to map and integrate additional data sets, along with catering for further analysis. Both face these challenges on an on-going basis. Semantic Web technologies offer flexible means for data publishing, integration and interpretation. To consider their use and implementation in a Semantic Web supported Financial Ecosystem, familiarity with the technology, its principles, data model and representational format are required. To identify data items, data published on the Semantic Web uses Universal Resource Identifiers (URIs). RDF provides the means for representing data and Linked Data the set of core principles, which if adopted provide the foundation for enhanced publishing, linkage and data integration. These principles amended with vocabularies to establish and share understanding, provide the

basis for our case study and can form the core of any web-based financial ecosystem. Each of these topics is next discussed within the context of a financial ecosystem.

Universal Resource Identifiers

World-wide data creation requires a system to name things unambiguously. To avoid situations where multiple data sources create identical identifiers for denoting different things, coordination between participants is required. The internet with its established infrastructure is used to help create this naming scheme. URIs and more recently Internationalised Resource Identifiers (IRIs) are used to define identifiers for arbitrary resources names. A resource can be either some sort of file (such as a hypertext document, a digital image, or a digital video), a real-world entity (such as a person, a company, or a product), or an abstract concept (such as financial instrument or types of relationships such as competitor of). URIs provide a basic mechanism to globally identify resources and form the basic mechanism used to associate otherwise disjoint pieces of data. URIs can be retrieved (dereferenced) via HTTP (Web browser or application).

For example, multiple URIs from multiple data sources identifies Vulcan Inc, a venture capital firm, and its founder Paul Allen (see Table 2).

Resource Description Framework

The Resource Description Framework (RDF) (Manola and Miller, 2004) is the basic machine-readable representational format used on the Semantic Web to represent information. RDF is a general method for encoding graph-based data which does not follow a predictable structure. RDF is schema less and self-describing, meaning that the labels of the graph describe the data itself. Data and facts are specified as statements and are expressed as atomic constructs of a subject, predicate and objects, also known as a triple. Within

Table 2. Example URIs

Entity	URI
Vulcan Inc.	http://cbasewrap.ontologycentral.com/financial-organization/vulcan-capital#id
	http://jobswrap.ontologycentral.com/jobs?location=Seattle&name=Vulcan+Capital#id
	http://data.semanticweb.org/organization/vulcan-inc
	http://dbpedia.org/resource/Vulcan_Inc.
	http://edgarwrap.ontologycentral.com/cik/1014931
	http://www.rdfabout.com/rdf/usgov/sec/id/cik0001014931
Paul Allen	http://data.nytimes.com/73984727404383484133
	http://dbpedia.org/resource/Paul_allen
	http://rdf.freebase.com/ns/en.paul_allen
	http://mpii.de/yago/resource/Paul_Allen
	http://www.rdfabout.com/rdf/usgov/sec/id/cik0000904057
	http://sw.opencyc.org/concept/Mx4r3I-uNKmxsQdiQt_OmSIkZ7w

the graph, subjects and objects are nodes, while a predicate is an arc. Figure 1 provides an example

a graph segment which models a description of Vulcan Capital: its name, homepage, a competitor, and the founder of the company.

Previously we noted that http://cbasewrap.ontologycentral.com/financial-organization/vulcan-capital#id, denoted a financial organisation named Vulcan Capital. Financial organisation is the *subject*, Vulcan Capital the *object* and *name,* the predicate. RDF is designed for the purposed of web-scale decentralised use in alternate graph models. For this reason the statement parts need to be identified so that they can be readily and easily reused. For identification, RDF uses these URIs. Expressing the previous statement in triple format then becomes (Box 1).

Looked at in identifier terms, the identifier has a name Vulcan Capital. Similarly but using identifiers for both subject and object, the home page of the identifier which denotes Vulcan Capital is http://capital.vulcan.com/ (see Box 2).
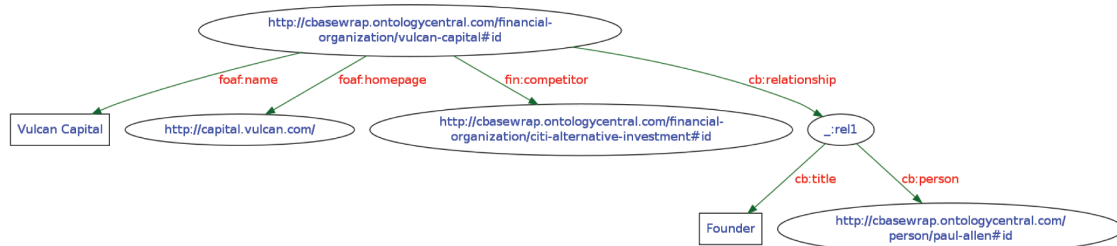
Or 'Vulcan Capital has a competitor Citi Alternative Investments' (Box 3).

Finally a more complex triple representation which introduces relationships between identifiers of (Box 5).

Box 1.

```
http://cbasewrap.ontologycentral.com/financial-organization/vulcan-capital#id
  foaf:name "Vulcan Capital" .
```

Figure 1. An RDF data graph describing Vulcan Capital

Box 2.

```
http://cbasewrap.ontologycentral.com/financial-organization/vulcan-capital#id
  foaf:homepage <http://capital.vulcan.com/> .
```

Box 3.

```
http://cbasewrap.ontologycentral.com/financial-organization/vulcan-capital#id
cb:competitor
<htttp://cbasewrap.ontologycentral.com/company/citi-alternative-
investments#id>.
```

Box 4.

```
http://cbasewrap.ontologycentral.com/financial-organization/vulcan-capital#id
cb:relationship _:rel1 .
_:rel1 cb:title "Founder" .
_:rel1 cb:person <http://cbasewrap.ontologycentral.com/person/paul-allen#id>
```

States that there is a relationship between Vulcan Capital and Paul Allen, where the type of relationship is founder. Paul Allen in turn can also be described in RDF.

In addition, we are able to represent numerical data in RDF, for Box the amount of cash and cash equivalents at carrying value for Microsoft (of which Paul Allen is a co-founder). For representing numerical data we used the RDF Data Cube vocabulary[1]. The triple in Box 5 relates the company to a dataset, which in turn contains the fact that the company had reported 8,823,000,000 USD in cash and cash equivalents at carrying value at the end of September 2009.

The take away points from the Boxs are that URIs are used to uniquely identify data and relationships, and the triples in RDF syntax are the fundamental building blocks that these link identifiers. The key observation is that since identifiers are unique they can be used to assist data integration by matching on the URIs and since they are HTTP based, the Web infrastructure can

Box 5.

```
http://edgarwrap.ontologycentral.com/cik/789019#idhttp://edgarwrap.ontology-
central.com/vocab/edgar#issued
<http://edgarwrap.ontologycentral.com/archive/789019/0001193125-10-239825#ds>
_:id1 http://purl.org/linked-data/cube#dataset
<http://edgarwrap.ontologycentral.com/archive/789019/0001193125-10-239825#ds>
_:id1 <http://edgarwrap.ontologycentral.com/vocab/us-
gaap#CashAndCashEquivalentsAtCarryingValue> "8823000000".
_:id1 <http://purl.org/dc/terms/date> "2009-09-30".
```

be leveraged to access and integrate data that is distributed, whether internal or external to an enterprise (see next section on Linked Data). A graph-based knowledge representation formalism, in conjunction with global resource identifiers, leads to a mechanism which facilitates data integration on an enterprise and even world-wide scale.

Implementing our financial ecosystem we used the RDF N-Triple format (http://www.w3.org/2001/sw/RDFCore/ntriples/) and extended the triple model (subject, predicate, object) to include context to describe the data. Context was used to encode the URL of the data source for the originating triple. While RDF N-Triples are called triples, we termed triples with context, "quadruples" or "quads" and the resulting data format, N-Quads (http://sw.deri.org/2008/07/n-quads). Context was used to track the provenance of data and to assist object consolidation in determining how data from different sources would be fused. Context can also be used to analyse which canonical URI to select for use.

Linked Data Principles

Linked Data has four basic principles that are required of data publishers (http://linkeddata.org/):

1. Use a Uniform Resource Identifiers (URIs) as names for things. A URI can identify any kind of object, concept or resource. e.g. Investment Fund, Risk, Person, or Company.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the Semantic Web standards RDF, SPARQL.
4. Include links to other URIs, so that more things can be discovered.

For Box, a data publisher has used the URI http://data.nytimes.com/73984727404383484133 to identify 'Paul Allen', adhering to principles 1 and 2. A lookup on the URI using an RDF-capable agent returns http://data.nytimes.com/73984727404383484133.rdf, the RDF document containing data about 'Paul Allen', adhering to principle 3. The RDF returned contains links to other data sources, for Box to http://dbpedia.org/resource/Paul_Allen, which, when dereferenced, provide their descriptions of Paul.

Data publishing adhering to these simple principles leads to a unified system for data access and interlinkage. Where data publishers use these standards, data consumers can also use those standards (as clients) to access the data and further process the data. The so-called Linked Data cloud of Semantic Web data has been growing considerably in the past years and provides a foundation upon which applications leveraging that data can be built on. A TED Talk by Berners-Lee provides a non-technical description of Linked Data (Tim Berners-Lee, 2009).

Vocabulary Descriptions

To establish a shared understanding between data providers and consumers of what certain schema-level identifiers constitute required a method to describe schema-level constructs. Consider Figure 2 which is a partial representation of the data model based upon the data sources listed in our Case Study Fundamentals Section. Please observe the abbreviated prefixes of {cb, fin, geo, foaf, skos, job}. Each denotes a particular domain vocabulary which provides distinct descriptions and definitions of their schema/namespace that they represents. cb refers to the Crunch bases schema; fin to the financial schema, modelled specifically for the case study from the financial, DBpedia, EPO and USPTO data sources; foaf to the Friend of a Friend vocabulary that describes people, geo to the geospatial schema, job to the job agency vocabulary and skos to the Simple Knowledge Organisation System Vocabulary (http://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html), for encoding classification schemes and thesauri, here used for sectoral classifications.

Vocabulary descriptions list these schema-level constructs of classes and properties and annotate them with certain data, such as human-readable labels and descriptions. Additionally there are certain knowledge representation constructs rooted in logic which allow for mathematical specification of meaning and allow for mechanical inference (such as subclass and sub-property relations).

Looking at the Financial Ecosystems ontology presented in Figure 2, foaf:Organisation stipulates that the definition for organisation is derived from the foaf namespace, http://xmlns.com/foaf/spec/Organization. When looked up (or dereferenced), it will established that foaf:Organization is seen as a class of which there can be instances such as Vulcan Capital or Citi Alternative Investments. Both of these instances can be stated in RDF in Box 6.

By dereferencing foaf:Organization a description of the class can be retrieved. The description can contain useful information such as: subclasses, useful for reasoning; text descriptions, useful for explaining the purpose of the class to any user or consumer of the class (foaf:Organization) and labels ("Organisation") which are used for results rendering purposes.

In addition to classes, properties can also be defined and dereferenced. From Box 3, the cb:competitor property can also be described. Firstly it has a defined domain, used to state that

any resource that has a given property is an instance of one or more classes, e.g. any instance which a value for property cb:competitor is an organisation. Secondly it also has a defined range, used to state that the values of the property cb:competitor are instances of one or more classes, e.g. organization. The domain and range of cb:competitor can be stated in RDF in Box 7.

Overall these modelling primitives allow for the reasoning which can provide a more complete picture of the information aggregated for multiple possibly incomplete sources. Class assists in this regard by grouping common instances into a single categorization. While different sources may use different class URIs for the same grouping, these can later be flexibly mapped on a schema level. Properties additionally allow for finding new instances of classes, for Box given the domain and range descriptions of cb:competitor in Box 7 and given the use of the property in Box 4 we can infer that Vulcan Capital and Citi Alternative Investments are organisations, even if not explicitly stated, as in Box 6. Beside new memberships of classes, new relationships between instances can also be inferred. If necessary axioms (rules) describing the meaning of properties such as cb:competitor rdf:type owl:SymmetricProperty, can be asserted. This symmetric property is useful as it allows inference to take place. For the Box above symmetric means that if Vulcan is a
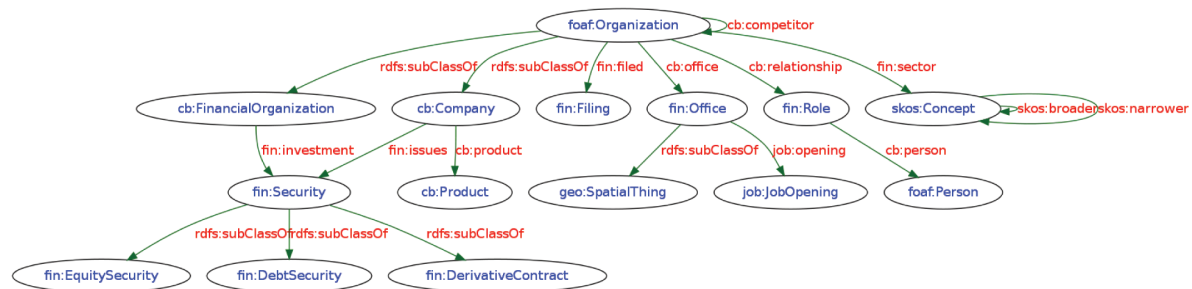
Box 6.

```
<http://.../vulcan-capital#id> rdf:type foaf:Organization.
<http://.../citi-alternative-investments#id> rdf:type foaf:Organization.
```

Box 7.

```
cb:competitor rdfs:domain foaf:Organization.
cb:competitor rdfs:range foaf:Organization.
```

Figure 2. Financial ecosystem ontology

competitor of Citi Alternative Investments, then Citi Alternative Investments is a competitor of Vulcan.

The main vocabularies used are RDF, RDF Schema (http://www.w3.org/TR/rdf-schema/) and the Web Ontology Language (http://www.w3.org/2004/OWL/). The latter are themselves encoded in RDF, amended with textual descriptions that specify the meaning of certain identifiers. Vocabulary descriptions provide a recommendation as to how things are modelled, that is they imply classes and types. Databases schemas on the other hand have to be enforced using integrity constraints. In contrast to relational database technology, representing the schema integrated with the data layer as opposed to the application layer, RDF allows for a more flexible approach to integration and follow on interoperability.

The important points to note are that classes, instances and properties are the basic building blocks of the vocabulary. Namespaces allow for dereferencing class and property definition, for re-use or extension by other vocabularies or data publishers, definitions that a community or domain would have previously agreed to. Relating vocabulary identifiers with one another (e.g. via a subclass relation) results in mappings between the vocabularies. Both together are quite powerful as you no longer need a single picture of an information landscape and can leverage multiple distinct pieces of the information puzzle, to facilitate either data integration or subsequent re-use.

THE ANATOMY OF WEB ENABLED FINANCIAL DATA INTEGRATION

Web enabled financial data integration like other data integration activity follows the established process steps of data acquisition, data conversion to some common format, the processing and integration of that data and making the resulting data available for consumption or further processing. Where data conversion and data integration are concerned, Web enabled data integration differs in the formats used to achieve a common interoperable format and how that data is consolidated.

To help illustrate the benefits that Linked Data usage brought to our financial data integration prototype and the manner in which its implementation was approached, we first introduce the process stages associated with its adoption and usage along with the component architecture necessary to support the activities involved in each of the stages. Then for each of the main architectural component areas a more thorough discussion on component composition and functionality is given. Discussion is based on our case study experiences.

Financial Integration Process and Architecture Overview

Figure 3 serves a three purpose function for Web based data source integration. It firstly illustrated the processing stages involved as a processing architecture funnel (left of figure) that provides a line of sight from raw data acquisition to end data

usage. It secondly provides a component block architecture (right of figure) to illustrate the main components necessary that provide the functionality to progress between process stages. Lastly it illustrated the intrinsic relationship between both.

Processing Architecture Funnel

The processing architecture comprised four stages:

- **Raw Data:** Useful data sources were first identified, inspected and individual data elements targeted for extraction (previously listed in Table 1).
- **Linked Data:** Raw data was then converted adhering to the Linked Data principles. Multiple electronic financial data format types and structure that range from unstructured text, to highly structured XML data were catered for. The data sources are

rendered to a common format and access protocol using wrappers. We developed wrappers that extract and serialise the data in RDF based upon HTTP lookup on the URI of the entity.This stage makes Linked Data available.

- **Integrated Data:** Next a crawler was used to harvest available Linked Data either via wrappers or as Linked Open Data on the Web. The system then iintegrated the collected data from a series of individual graphs into a holistic dataset by aligning and linking entities.
- **Actionable Information:** Finally, users can analyse the underlying datasets through interactive exploration, deriving further insight and understanding, leading to knowledge usable in decision support and decision making.

Figure 3. Integration of Web based data sources using linked data
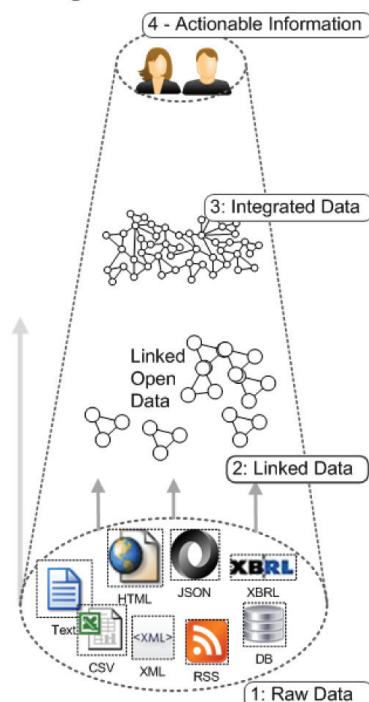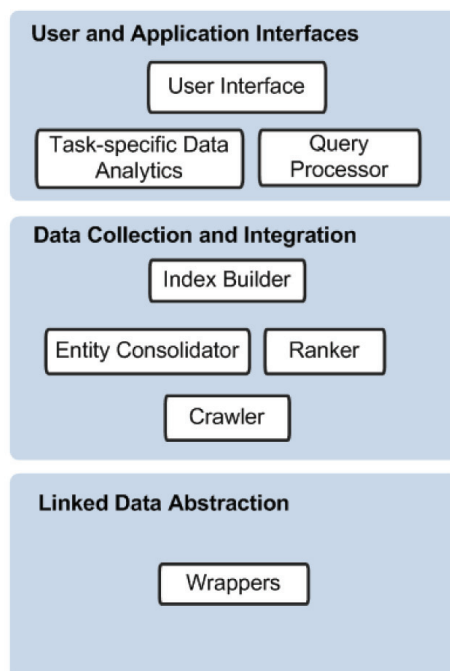
The process stages can be iterative for new source additions. Where a data refresh is required the identification of raw data can be excluded or if new steps are added to the data integration operation, only re-running the data integration step is required.

Component Integration Architecture

Each of the processing stages is complimented by its own component integration support activities. For Box, with reference to Figure 3, the process stage of obtaining raw data and rendering it as Linked Data is carried out by the Linked Data Abstraction layer, whose major component is a wrapper responsible for the data transformation activity. Moving from Linked Data to integrated data is performed by Data Collection and Integration, consisting of modules concerned with pre-processing data (crawler, reasoner, ranker, index builders). The integrated data is made accessible and usable to the financial community by the components of the User and Application Interfaces whose modules are concerned with query-time processing (index managers, query processor, user interface). The architecture follows the basic mode of operation that describes a web search engine, namely "crawl, index, serve" (Brewer, 1998). Our case studies navigation system extends this to include interactive data exploration based on the KOPF principles (cf. User and Application Interfaces Section). Having introduced the overall architecture, we now discuss the components constituents for each component integration architecture area and the data transformation activities performed to assist our financial integration prototype creation.

Linked Data Abstraction

Data integration begins with the collection of web-based and internal data and its representation in a common format (RDF) and access protocol (HTTP). An extractor transforms metadata from HTML documents (e.g. RDFa, GRDDL, or Microformats) and metadata embedded in various file formats (e.g. GIF, PNG, PDF, MS Office) into RDF. RSS 2.0 and Atom feeds are also translated into RDF. Wrappers allow for accessing legacy data as Linked Data. For Box, D2R (http://www4.wiwiss.fu-berlin.de/bizer/d2rmap/D2Rmap.htm) provides a mechanism for exposing data stored in relational databases as Linked Data. Custom-build wrapper software often provides Linked Data access to Web sources returning JSON or XML. The more commonly encountered formats for financial content on the web include:

- **Text:** The traditional press is publishing financial news mainly in textual format (e.g. Financial Times). Sites such as seekingalpha.com make raw transcripts of investor calls available and the SEC filings (e.g. Form 10-K, Annual Report) contain a large amount of free text financial comment.
- **Office documents:** CSV files, Word documents, or PowerPoint presentations, in corporate environments a large portion of data exchanged is encoded in MS Office documents.
- **Structured Data:** Typically structured data adhering to a fixed schema is encoded and published in XML. XBRL falls under this category as well as statistical information is being published either in CSV (comma-separated value) format (e.g. by Eurostat) or SDMX.
- **Linked Open Data:** Sources such as DBpedia contain graph-structured RDF data describing companies, people, and the like. Financial information is typically not very structured - for Box, there are manually maintained figures of assets or net income for the previous year. Other sources for financial information (especially about tech start-ups) include CrunchBase, for which a version in RDF exists.

Data Collection and Integration

Source and data ranking can additionally be used to assist the consolidation process. Indexing and storing the results in an RDF repository will then allow further analytics or direct access to the data. These activities are consistent with the data warehousing portion of the hybrid approach mentioned previously.

Crawling

A crawler (e.g., MultiCrawler (Harth et al., 2006) harvests data accessible via HTTP by following links. The result of the crawling step is a set of RDF graphs that can be processed further. Given that the wrappers lift all data sources to the common Linked Data abstraction, the crawling component can easily retrieve RDF data and follow links based on specified parameters (such as depth of the graph traversal or maximum number of data sources).

Ranking Data and Data Sources

Ranking provides a mechanism to prioritize data elements and assuage the noise inherent in datasets which have been aggregated from multiple sources or have been created in some decentralised way. We use a set of scalable algorithms for ranking over a general model of structured data collected from an open, distributed environment, based on the notion of naming authority. We adapted the general model to the case of RDF, taking the intricacies of RDF data from the Web into account. In comparison to using plain PageRank (http://www.google.com/corporate/tech.html) on a node-link graph representation of RDF, our methods exhibited similar runtime properties while improving on the quality of the calculated rankings. Contrary to methods which require manual input of a domain expert to specify schema weights, our method derives rankings for all identifiers in the dataset automatically. The approach has applications in

search, query processing, reasoning, and user interfaces over integrated datasets in an environment where assessing trust worthiness of sources and prioritising data items without a-priori schema knowledge is vital. Ranking the importance and relevance of results during interactive exploration is the responsibility of the Ranker. IdRank (Harth et al., 2009) is one such link analysis technique approach which simultaneously derives ranks for all RDF elements and data sources. Ranking is a vital component in search and query interfaces operating on web data and is used to prioritise presentation of the more relevant search results and prioritising alternate possibilities for user query construction.

Entity Consolidation

An important aspect of Semantic Web technologies is the issue of identity and uniquely identifying resources, which is essential for integrating data across sources. Currently, there is limited agreement on the use of common URIs for the same instances across sources. Since the assignment of URIs to instances is optional within the RDF, many entities are described anonymously without use of a URI. Where agreement on URIs for resources cannot be reached, multiple URIs may exist for the one resource resulting in an integrated dataset missing associations between resources. The problem of entity consolidation has received significant attention in the database community as record linkage, instance fusion, and duplicate identification. Due to the lack of formal specification for determining equivalences, these approaches are mostly concerned with probabilistic methods. In more formal approaches, properties unique to an entity can be used to determine its identity. Boxs are the SECs Company Identification Key (CIK), email addresses or the URI to an XBRL company. These inverse functional properties, specified in OWL descriptions are important as they allow data marked with such properties to indicate equivalence. Box 9 allows an applica-

tion conclude that the property CIK is unique to a company and the two instances sharing the same value for the properties must be the same real-world entity, i.e., http://edgarwrap.ontology-central.com/cik/789019 and http://dbpedia.org/resource/Microsoft are the same company.

There is much more agreement on values that are established in real-world areas or widely used applications such as email addresses or instant messaging usernames than on URIs as the identifying factor of entities. Thus, equivalence of identifiers based on inverse functional properties has to be taken into account when merging RDF graphs, since infrequent reuse of instance identifiers across sources can leads to problematic data integration and knowledge contribution and an entity remaining fragmented across multiple instances. As the desired outcome of a linked data effort is an integrated well connected graph, agreement on identifiers is crucial for entity association. Fusing identifiers is especially important for entity-centric applications relying on RDF data, such as search and query engines, interactive browsing tools and data mining systems. In implementation our own native algorithm (Hogan et al., 2007), the requirements were that it should be web scalable, robust to problematic data in terms of quality and completeness, and cater for URI reduction of URIs to a single canonical form when denoting the same entity. Faced with merging equivalent instances of RDF data sets and associated entities our approach first determined a list of inverse functional properties (e.g. hasCEO relation is inverse functional as for any company there is only one CEO) from vocabularies (ontologies) that describe instances in the dataset,

then determine instance equality based on the values of the inverse functional properties, store the transitive closure of the equivalences in a data structure for efficient lookup and finally generate a canonicalised URI for each equivalence chain, scan the dataset and rewrite identifiers. This canonical (consolidated) identifier is also referred to as the pivot identifier or element and is driver of instance merging. A pivot element must exist for each row in the equivalence list and the identifiers in the row as viewed as candidates. Lacking formal guidelines for pivot elements of RDF data and with multiple alternatives to selecting one from an equivalence row, we choose to use URIs as they can be recycled for future extensions of the resource description.

Object consolidation remains an active research area as evidenced by the literature. Chen et al. use a graph-based data model for representing data, and use inter-object relationships to detect clusters of similar items (Schumaker and Hsin-chun, 2009). (Michalowski et al., 2003) utilise secondary sources to determine equivalences. A recent survey summarising many of duplicate record detection methods can be found in (El-magamid et al., 2007)]. The SemTag effort as described by (Dill et al., 2003) assigns identifiers to web pages. (Bouquet et al., 2006) motivate the problem of (re)using common identifiers as one of the pillars of the Semantic Web, and provide a framework and fuzzy matching algorithms to fuse identifiers. We perform entity consolidation on a larger scale and avoid use of probabilistic methods. Brickley discusses in (Brickley 2002) implementation strategies for merging identifiers in RDF based on RDF reasoning engines. How-

Box 9.

```
dbp:secCik rdf:type owl:InverseFunctionalProperty.
<http://edgarwrap.ontologycentral.com/cik/789019> dbp:secCik "789019".
<http://dbpedia.org/resource/Microsoft> dbp:secCik "789019".
```

ever, the technique has not yet been applied to large datasets. Our approach handles data from a large number of sources obtained from the web; thus to cater for potentially long equality chains derived from many sources, efficient data structures for storing an equality table are required. In addition, we provide detailed statistics on the current usage of data on the Semantic Web. The Florid system (Frohn et al., 1997) is an F-logic inspired reasoning system that maintains an equality relation data structure to be able to deal with ground equalities. In the Hyperion project (Kementsietsidis et al., 2003), the issue of relating identifiers amongst peers is solved using so-called "mapping tables". (Park and Durusau, 2006) use the notion of subject-centric merging of ontologies based on a Topic Map approach to denote a fine-grained approach to subject identity. SILK (http://www4.wiwiss.fu-berlin.de/bizer/silk/) can be used to discover links between Linked Data from different sources. Using a declarative language, a developer specifies conditions that data from different sources has to fulfil to be merged, optionally using heuristics in case merging rules can lead to ambiguous results. In contrast, we used owl:InverseFunctionalProperty information to determine equivalences.

Indexing

The index builder (or indexer) provides a general framework for locally creating and managing inverted keyword indices and statement indices over the materialised dataset. We view these two index types as the fundamental building blocks of a more complex RDF index. The index structured used in the case study comprised a complete index on quadruples (Harth and Decker, 2005) with keyword search functionality based on a standard inverted index. The keyword index mapped keywords extracted from literals to subject nodes. Each statement index offers prefix lookups over a set of statements of fixed length in a particular order. Multiple statement indices provided (complete)

support for all possible lookups on statements of any length. They can also be used to develop optimised methods for problematic queries. Our framework, with combinations of keyword and statement indices, can be used to implement specialised systems for indexing RDF.

Once a general index is built over the materialised data set, additional analysis algorithms and approaches (not covered here but refer to (Schumaker and Chen, 2009) for further reading) to cater for a specific task or activity can be applied to generate a more specific index. Incorporating additional domain knowledge semantically enhances the resultant index and can make it to application catering for particular vertical markets, for Box investment fund search and lookup. The index manager manages and provides access to both the main and any task specific sub-indexes, offering atomic lookup functionality over them. The indices can include keyword indices on text and statement indices, such as quad indices on the graph structure, and join indices on recurring combinations of data values. Index managers if necessary can be distributed across a number of machines.

Discussion

Table 2 outlines the typical feature combination and differences between established large-scale information retrieval, relational database, data warehousing and Semantic Web systems architecture (Note that use case scenarios can introduce architectural variations). In general, information retrieval systems perform best-effort query processing over large document corpora. Precision (ability to retrieve only relevant items) and recall (ability to retrieve all relevant items) measures are used to determine quality and comprehensiveness of the results set. Database systems require exact queries but can deliver exact answers. For databases the ongoing process of data curation largely addresses data quality issues. Data warehousing systems aggregate analyse and provide

report generation capability of data collected from multiple relational databases but typically for a specific domain application or scenario. Semantic Web systems differ from these traditional database system as they have to cater for both structured data and documents, which has a knock on effect on the component architectures of semantic navigation systems. Early prototypes using the concepts of ontologies and semantics on the web include Ontobroker (Decker et al., 1998) and SHOE (Heflin et al., 1999). Semantic Web data is utilised by Swoogle (Ding et al., 2004) but is restricted to indexing native RDF data. Swoogle in looking at the Semantic Web still retains the document level view and provides keyword search facilities only over those documents. A semantic navigation system in contrast, should leverage data in both XML and RDF format and view the Semantic Web/Web of Data as a single massive directed graph with labels on both nodes and arcs. The main distinction between web search engines, Swoogle and semantic search engines is the ability to perform queries over structured data involving joins. The semantic system should therefore look to utilise database methods for index and query processing and consolidation techniques from

data warehousing and data integration. Entity centric data search is another approach, which is beginning to gain acceptance that can be used to address some of the limitations. The recently developed Falcon Search3 (http://www.search3.com/) also offers entity-centric searching over RDF instances. Sindice4 (http://sindice.com/) utilising entities as its knowledge representational core, offering a registry and lookup service for RDF over a Lucene index. Watson (http://kmi-web05.open.ac.uk/WatsonWUI/) approaches entity search across Semantic Web data using natural language search. Overall searching and navigating graph formatted structured data presents unique requirements for a system (as outlined in Table 3) and requires a system architecture comprising novel component combination and characteristics as argued by (Aberer et al., 2005).

User and Application Interfaces

The query processor has to support queries from both user and application programming interfaces. Query evaluation first requires the creation and optimisation of a logical plan (in principle the same as the query plan generated by an optimiser for

Table 3. Contrasting architecture characteristics of IR, RDB, data warehouse and semantic navigation systems

Dimension	Information Retrieval	Relational Database	Data Warehouse	Semantic (Web) Navigation Systems
Data Model	Documents	Relational	Relational	Graph
Data Provisioning	-	Transactional	Extract/Transform/Load	RDF, dereferenced URIs
Data Integration	-	Application Logic	Merge Rules	Entity matching, linking, consolidation
Index Structure	Inverted Index	B+-Tree	B+-Tree, Bitmap Index	Inverted index, Triple Index
Query Processing	Boolean Keyword	Complex Query	Complex Query	Complex Query
Ranking Scheme	Application Dependent	-	-	Application Dependent
User Interaction	Keyword-Result	Forms, Queries (parameter based)	Reports, Data Feeds (parameter based)	KOPF

an RDB) which is then executed against indexes through the interfaces provided by the index manager.

The user interface provides the search, browsing and interactive data exploration over the data. A semantic navigation system typically offers a combination of the following operations: keyword search (K), object focus (O), path traversal (P), facet selection (F) also termed a KOPF system. Similar to regular search environments, users can begin by specifying keyword searches but then incrementally build queries to browse and navigate the object graph, through relationships (paths), and retrieve information about objects.

Task-Specific Data Analytics

A consolidated data set represents a rich source of high quality data that can be exploited for uses such as risk monitoring, automated trading, market events, fraud detection, competitive analysis or profit projections. There are a large number of data mining and textual analysis techniques such as the predictive machine learning approach to choose from. (Schumaker and Chen, 2009) seeking to predict stock market prices used the predictive approach on different textual representations of named entities, bag of words and noun phrases. Unstructured textual financial new articles were used as sources. The same approaches could also be applied to RDF where structured data would provide more pattern signal for data mining algorithms to work with. More novel approaches can also be considered. The Guardian, a UK broadsheet employed a community crowd-sourcing approach to further analysis by requesting that readers tag, and report Members of Parliament expense claims to highlight irregularities for further investigation (Guardian, 2009). This is equally applicable to a community of analysts where tagging would share insight, comment and advice.

Querying Data

SPARQL Protocol and RDF Query Language (SPARQL, http://www.w3.org/TR/rdf-sparql-query/) is the W3C recommendation for querying RDF graph data. A SPARQL endpoint accepts queries and returns results via HTTP. Generic end points query against Web based RDF data and specific endpoints against particular datasets. A SPARQL query Box that models the question looks like (Box 10):

Which venture capitalist has invested in companies that are competitors of each other?

The PREFIX, also called a namespace provides a means of uniquely identifying elements of a particular vocabulary and defining them. It's the SPARQL equivalent of an XML namespace. Here the PREFIX statement defines the shorthand notation (or prefix) cb, which within the query will be expanded to the full URI <http://ontologycentral. com/2010/05/cb/vocab#> and the URI used to provide vocabulary definitions on investment and competitor. The SELECT clause specifies the data elements that the query should return, here the three variables, indicated by the prefix "?" of ?vc, ?co1 and ?co2. The FROM clause, specifies the location of the data set to query against, here the URI <http://cbasewrap.ontologycentral.com/ company> is dereferenced. Finally the WHERE looks for vc? that have cb:investment in competitors (that are companies). Each matching binding of the graph pattern's variables to the model's nodes becomes a query solution, and the values of the variables named in the SELECT clause. SPARQL query results can be rendered in a variety of formats: XML SPARQL specifies an XML vocabulary for returning tables of results; JSON, a JSON "port" of the XML vocabulary, particularly useful for Web applications; RDF, which can be serialised in a number of ways (RDF/XML, N-Triples, Turtle, etc.); HTML, when using an interactive form to work with SPARQL queries.

Box 10.

```
PREFIX cb: http://ontologycentral.com/2010/05/cb/vocab#
SELECT ?vc ?co1 ?co2
FROM <http://cbasewrap.ontologycentral.com/company>
WHERE {
      ?vc  cb:investment ?co1 .
       ?vc  cb:investment ?co2 .
       ?co1 cb:competitor ?co2 .}
```

Queries can be distributed to multiple SPARQL endpoints, computed in a distributed fashion, and results gathered, in a procedure known as federated query.

User Interface

A typical system architecture for any data integration system has two main phases: a data collection and integration phase to convert and map data from different sources into a common format and an application and user interface phase that operates over the integrated dataset. Integrating data from multiple sources provides a common data platform and source view from which search, browsing, analysis, and interactive visualization can take place.

Using entities differs from the standard interaction model for web search: the user specifies keywords which are matched with document contents, and may change the query based on spelling suggestions (Shneiderman et al., 1998). Norman (Norman, 2002) argues that the conceptual model of a system has to fit the user's own conceptual model about it, that is, what the users have to know about the system before interacting with it. With the entity model users perceive and act on entities/objects, in-line with early graphical user interfaces (Lipkia et al. 1982). In general, there is a one-to-one correspondence between the entities in the dataset and those displayed to the user, loosely following the "naked objects" approach (Constantine, 2002). Users are able to

search and navigate the entities in the dataset; a user query yields entities as a result. Users can choose to display the result set in detail, list, or table view; optionally, a timeline or map visualization is available if the result entity contain suitable information. In addition, users are able to export the results to application programs or services.

In order to leverage existing familiarity of users with search engines for our case study, the first step in our interaction model used keyword search to locate entities. We introduced a point-and-click extension to the traditional web search interaction model, which represents an incremental path towards higher quality search, while retaining the ease of use of traditional web search. In subsequent steps, users refine their query based on the navigation primitives; as such, the interaction model leads to an explorable system that can be learned through experimentation. Since the system computes further possible steps relative to the current result set, only valid choices were offered. Our search and navigation model extended the web search workflow with restriction and navigation operations on entities, rather than the standard web search engine interaction model (Shneiderman et al., 1998) of keyword search, list of documents, spelling correction and iteration. In summary our search and interaction model aimed to facilitate users to freely navigate the information space that spans all available entities, and has the ability to ask queries that go beyond the precision and expressiveness of simple keyword searches. Both are achievable with an entity centric data model.

Keyword search over hypertext documents is an established method used by majority of web users (Henzinger, 2007). Search engines operating over these millions of automatically collected documents offer the familiar but limited functionality of returning links to other sites and documents rather than attempting to directly the answer the questions asked or retrieve the data item sought. Keyword phrases used for search do not cater for complex information need (Henzinger, 2007) and unstructured natural language text as mentioned previously is ambiguous and hard to process automatically. Indeed adding financial numerical data serves to even further complicate the problem.

The Semantic Web Technologies Overview Section previously outlined the entity centric modelling structure behind our prototype which for the case study comprised financial/financial related textual and numerical Web Integrated data. Asking any question of or interrogating such data requires a flexible interaction model to present different types of visualization that alternative views of that data would require. Specifically we determined that for Web integrated data set the following requirements:

- **Intuitive use**, occasional users and subject-matter experts should be able to interact with the data as is. Results should be derivable quickly (a few clicks).
- **Universality**, users should be only minimally constrained in constructing queries while keeping the system easy to use.
- **Minimal transformation**, large levels of manual intervention on Web data in multiple formats and vocabularies would not be feasible.
- **Tolerance**: the user interface has to be tolerant enough to deal with diverse content, both in terms of schema and noise in addition to duplicates, malformed syntax and incorrect formatting gracefully.
- **Scalability**, with the Web as a data source the system has to scale competently, which has implications for system architecture and implementation.

Table 4. Comparison of query operations of systems operating over entity-structured data

System Keyword	Keyword Search	Object Focus	Path Traversal	Facet Selection	Results
Magnet (Sinha and Karger, 2005)	Yes	Yes	No	Yes	Set
Museum Finland (Hyvnen, et al., 2005)	Yes	Yes	No	Yes	Set
GRQL (Athanasis, et al., 2004)	No	Yes	Yes	rdf:type	Set
SWSE (Harth et al., 2007)	Yes	Yes	No	No	Set
Facet (Hildebrand et al., 2006)	Yes	Yes	No	Yes	Set
BrowseRDF (Oren et al., 2006)	Yes	Yes	No	Yes	Set
ESTER (Bast et al., 2007)	Yes	Yes	No	Yes	Set
TcruziKB (Mendes et al., 2008)	Yes	Yes	Yes	No	Set
Tabulator (Berners-Lee et al., 2006)	No	Yes	Yes	No	Tree
Falcons (Cheng, 2008)	Yes	Yes	No	rdf:type	Set
Humboldt (Kobilarov and Dickinson, 2008)	Yes	Yes	Yes	Yes	Set
Parallax (Huynh and Karger, 2009)	Yes	Yes	Yes	Yes	Set
ECSSE (Cyganiak et al., 2009)	Yes	Yes	No	No	Set
VisiNav (Harth, 2009)	Yes	Yes	Yes	Yes	Tree

- **User satisfaction**, users should be able to export and share results of their information seeking tasks with others and application programs for further analysis.

Our approach to mapping the requirements to a set of interaction features was to look at the feature sets found in existing Web based systems that interact with entity structured data sets. The interaction model that emerged uses the core set of query primitives (operations) common to a range of established browsing and navigation systems for semi-structured data. As outlined in Table 4 the feature sets condensed around the five operations of:

- **Keyword search**, where the user may specify keywords to pinpoint objects of interest e.g. an investment company, such as Citi.
- **Object focus,** similar to following a hypertext link in a web browser. From a set of results or a single result, the user selects an object which is used to create a new query and returns a result set containing the object rather than just a link to another document.
- **Path traversal**, rather than arriving at a single result by selecting a focus object, users can navigate a path along an object property to establish a new set of results. e.g. Start with Citi and using competitors, traverse to the competitor e.g Bank Of America.
- **Facet Selection**, is a means of assigning multiple classifications to the same object or property and literal value. Facets are calculated relative to the current result set and based on derived facets, the user can reformulate the query and obtain increasingly specific results.

VisiNav our prototypes interaction model binds keyword search, entity focus, and introduces the notion of result trees which extend single-set results to multiple result sets containing result paths for path traversal and additionally includes facet selection. Combining query operations and feature of existing Web based systems sets in response to our initial requirements, represents a degree of user conceptual familiarity and community consensus of sorts, which could be built upon. The initial step of our interaction model is a keyword search to locate entities, leveraging existing familiarity of users with search engines. In subsequent steps, users refine their query based on navigation primitives; as such, the interaction model leads to an explorable system that can be learned through experimentation.

Keyword search pinpoint entities of interest providing an initial set of results based on broad matching of string literals connected to those entities. We perform matching on keywords without manually extending the query for synonyms or other natural language processing techniques. Rather, we leverage the noise in Web data, i.e. the fact that the same resource might be annotated using different spellings or different languages. Keyword search has the useful property that the users do not need to be familiar with the data model, enabling users to pose queries without prior domain knowledge.

Figure 4 shows query result for companies in the Internet Industry, the screen shot shows a list of companies who are additionally defined as having the industry "Internet" (List View). At the top of the results page, you see that the query is formulated as two parts: (i) the first is "type Company" which initially restricts results to those give the type (loosely, category) Company in the data; (ii) the second is "Industry Internet" which further restricts the previous sets of all companies to those who are assigned the value "Internet" to the property "Industry". Both of these are called facets, and allow for iteratively refining the results set until the precise set of resources the user is interested in is achieved. Users are only offered facet selections which can lead to non-empty

result sets, i.e. based on the underlying data and will provide results. A facet is a combination of a property and a literal value or an object (distinguishing between datatype and object properties). Facets are calculated relative to the current result set. Based on derived facets, the user can reformulate the query and obtain increasingly specific result sets. The results themselves are headed by the primary label (name) for the entity, the full URI by which the entity is identified, the types found for the entity, and also images and textual snippets if found. Each result is prioritised according to a ranking, and results can be clicked on to display a more detailed view of the entity, showing all information associated with it (Detail View), 'Nokia', the company, in Figure 5 is a primary label. Entity focus as an operation is similar to following a hypertext link in a web browser. From a set of results or a single result, the user can selects an entity which is used to create a new query and have returned a result set which also containing the entity. Rather than arriving at a single result by selecting a focus entity, users are also able to navigate a path along an entity

property to establish a new set of results. Users can select an entity property which allows them to perform a set-based focus change i.e., they follow a certain path – either from a single result or a set of results. Aside from the possibility of domain-specific mashups and renderings of the data, VisiNav offers out-of-the-box visualisations which exploit the underlying structure of the data. For Box, date-time values in the data can be automatically detected, parsed and rendered in a Timeline View. Similarly, the Table View allows for displaying entity-information in rows, with columns referring to properties, and cells containing values for those properties (e.g., render all company names, stock tickers, and recent acquisitions in a table for each company in the current result set): subsequently, this data can be exported to Excel for offline processing.

Result sets comprising numerical data benefit from the right visualization. Hence, the user interface can optionally display numerical data using charts. For Box, the "Scatterplot View", in Figure 6, shows value pairs (in some instances multiple) attached to an entity plotted against each

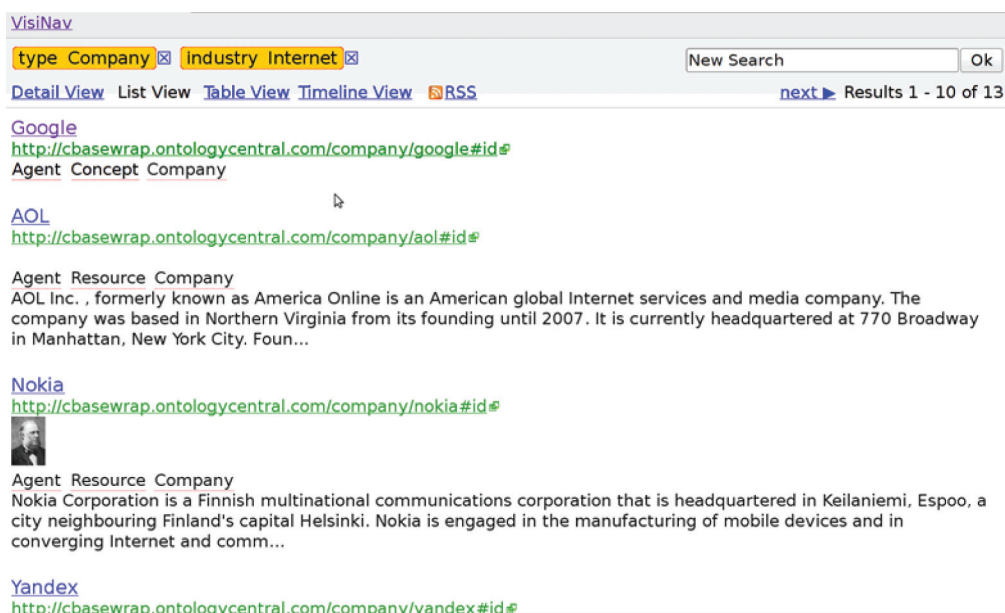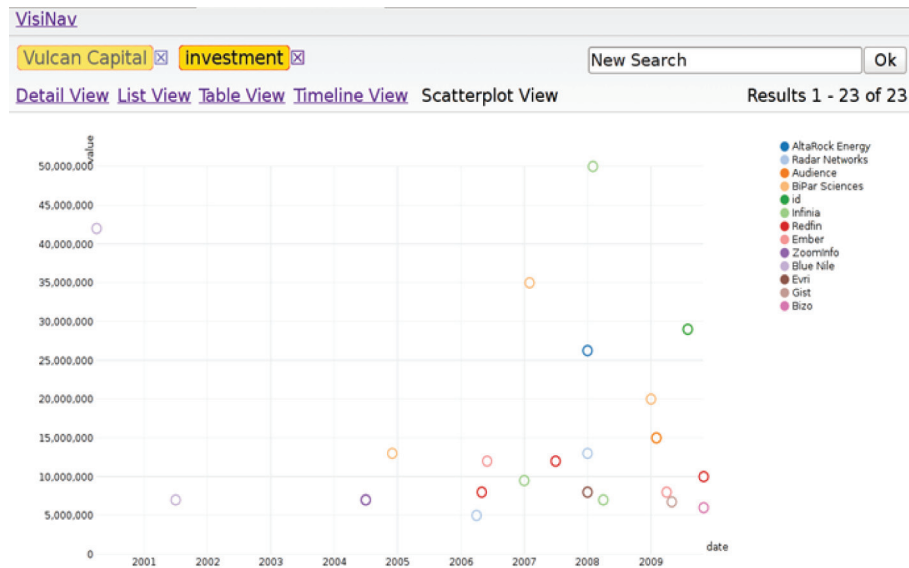*Figure 4. VisiNav user interface showing companies in the Internet industry*

Figure 5. VisiNav scatter plot of Vulcan Capital funding rounds

other. The x-axis represents time in years and y-axis the funding value in USD. The scatterplot graphs the progression of Vulcan Capitals investment funding rounds with 13 companies over the period 2001-2009. Audience, (mobile voice communication) is an Box of a company in receipt of multiple investment rounds.

Discussion

NLMenu (Thompson, 2009) is an early system advocating the use of multi-step query construction based on menus. Faceted browsing (Yee et al., 2003) while less expressive in terms of the complexity of queries, has become popular and is used on e-commerce sites such as Ebay.com. Polaris (Stolte et al., 2002) provides complex query and aggregation operations, but operates over relational data and thus requires a-priori knowledge about the schema used. The systems most closely related to ours in terms of features are GRQL (Athanasis, et al., 2004), Humboldt (Kobilarov and Dickinson, 2008) and Parallax (Huynh and Karger, 2008). GRQL relies on schema information rather than automatically

deriving the schema from the data itself, a feature required for web data which does not necessarily adhere to the vocabulary definitions. GRQL lacks keyword search, a useful feature when operating on arbitrary data since keywords are independent of any schema but was one of the earlier systems to support set-based navigation. Rather than permitting arbitrary facets, GRQL allows for restrictions based on the rdf:type predicate. Parallax (Huynh and Karger, 2008) is a recent system which exhibits browsing features similar to VisiNav. However, Parallax operates over the Freebase dataset which is manually curated; VisiNav operates over RDF data collected from the web. Parallax lacks ranking, a crucial feature when operating on web data. VisiNav prioritises facets, navigation axes and results based on global ranks. Although Parallax uses multiple result sets, the connections between the result sets are not propagated to the level of the user interface; our system maintains result paths in the results trees. Regarding methodology, VisiNav can be described in terms of the Semantic Hypermedia Design Method (de Moura and Schwabe, 2004). Tools and methods developed in the field of

semiotics (Goguen, 2003) can help to formalise user interaction models and signs, which supports consistency and thus understandability in the user interface. Work on display of information (Tufte, 2001) is related to the visual appearance of the user interface and influences the structure of the user interaction; similarly, design plays a role as well (Norman, 2002).

BEST PRACTICES FOR INTREGRATING WEB AND ENTERPRISE DATA

Our best practice recommendations are a combination of adoption standards, architectural abstraction, practical guidance on process steps from our case study implementation and insight from field research conducted on Linked Data adoption in the financial, pharmaceutical, and advertising sectors. All discussion is directed toward making data easier to publish, consume and make better sense of for business insight and intelligence purposes.

Making Data Available

Start by publishing your data in a format that can be easily shared, then progress to RDF. Begin to apply the Linked Data principles. Use URIs to name and access resources. First introduce the W3C recommended representational standards RDF, followed later by RDFS and OWL. The now mandatory XBRL or existing standards such as RIXML can be mapped to RDF for integrated oversight and federated querying. XBRL in particular represents an opportune area as financial institutions and corporates have yet to begin exploiting its full potential. Note, making the data available in Linked Data format is equally applicable for internal consumption and external publication.

Architectural Approach to Utilising Linked Data

Having to cater for data as opposed to documents will require changes to the data data abstraction selected and its implementation. Introduce ontology modelling as the common abstraction and access layer. As most enterprises should have sufficient database skills in house transition should be relatively easily accommodated. Use HTTP lookups for data requests and have them provide data in RDF. Scripts can be used to convert data dumps (e.g. CVS to RDF), wrappers for on-demand conversion and relational data can leverage tools such as D2R, which maps between a database schema and an RDFS vocabulary. Modify queries to queries to use SPARQL rather than SQL.

Link data to other data items. Link text to data, establish links between the text and structured descriptions of entities. Tools such as OpenCalais (http://www.opencalais.com/) detects entities and Wikipedia Miner Wikifier (http://wikipedia-miner. sourceforge.net/) recognises topics, linking them to Wiki articles. Reuse URIs to facilitate merging of data from multiple sources. OWL constructs such as 'sameas' supports entity consolidation and can assist integrating fragmented data sets. With identifiers in place, should a data conversion script change, dereferencing the URI will remain unaffected.

When querying data across multiple sources begin to incorporate a distributed query processing approach with your data warehousing approach. Base the UI on operations combinations that are capable of exploiting a semantically integrated data set such as keyword search, entity search, entity path traversal and facet selection. Incorporate and use RDF data store as your repositories where relevant. For a current list of repositories refer to http://esw.w3.org/SemanticWebTools.

Overall start small, beginning with legacy data and progressively rewrite data publication. Guide data model refinement with your existing legacy data schemas and incrementally refine as

you integrate. Emphasise that Linked Data can co-exist with existing legacy systems, and need not be another large all-encompassing integration project that it the dread of IT Departments. Ensure that all parties are on board as making data readily available and accessible will be a position shift for most people, the benefits of doing should be clear.

FINANCIAL DATA ECOSYSTEM

The recent W3C Workshop on Improving access to Financial Data on the Web (http://www.w3.org/2009/03/xbrl/cfp.html) focussed on challenges and use cases for realizing an ecosystem of financial related services (Curry et al., 2010). The workshop brought together experts from government, financial services and academia to discuss Web APIs for financial data and ecosystem support for value add players in the space. Financial reports were noted as having both financial figure and financial statement notes, with the figures being those most readily used. The filing statements along with news regarding economic, business and sectoral influences information provide valuable comment and insight that cannot be easily integrated as part of an analysis activity. As a softer type of information, automated extraction is difficult and human judgement is relied upon. If person or task generated content could be contributed to, or some relevant source included into an ecosystem, the business value and benefits would be immediate.

Integrating data from multiple Web sources requires a common data platform from which search, browsing, analysis, and interactive visualisation can take place. Consolidation in Semantic Web terms leads to an aggregated source view or a coherent graph amalgamated, 'mashed up' from potentially thousands of sources, where an entity centric approach can provide a powerful single view point allowing information filtering and cross analysis. The key challenge for any information system operating in this space is the

need to perform semantic integration of structured and unstructured data from the open Web and monolithic data sources such as enterprise XML database dumps and large static datasets. This can be achieved using a hybrid data integration solution which amalgamates the data warehousing and on-demand approaches to integration.

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

This loosely coupled mashed up data set, that can accommodate additional data sets, and previous analysis results to provide a single information access point for interactive exploration and complex querying, we refer to as a Financial Data Ecosystem.

Financial Information Consumers

A large number of information consumers have varying degrees of interest in financial data. The integration and augmenting of financial information is of significant benefit for financial and business analysis as the following use cases illustrate:

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

- **Fraud Detection:** The relatively new branch of forensic economics (HBR, 2009) is an area which would benefit from the availability of financial, government and regulatory linked data. Interested in spotting patterns or conditions that suggest fraud, the economists look at the benefits of criminal activity as guides to spotting data footprint that suggest the activity is taking place. Identifying suspect activity sooner would be possible if regulation required companies to make relevant data available in a format that could easily be linked and aggregated for investigation. Financial regulators and fraud investigations could leverage such linked data within their forensic tools to improve their capacity to monitor regulatory compliance or early fraud detection. Individual and institutional investors alike spend considerable time looking at company and investment fund returns for investment potential.

- **Figures Comparison:** Financial results are often purposely presented in a convoluted manner, making direct comparisons difficult even for the financial professionals. Integrating the financial data in a common format with semantic mark-up would make similar financial instrument mapping and comparison easier. Overall it would increases the level of transparent and provide better information that would assist with the "what's the best performing fund?" or "what are the better values shared to buy?" type of financial analysis.

The principle behind each is that XBRL information extracted from SEC filings receives a semantic metadata lift allowing the data to then be published in RDF format. The Rhizomik project (http://rhizomik.net/semanticxbrl), OpenLinks Sponger (http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/VirtSponger), and W3C recent efforts (http://sourceforge.net/

projects/xbrlimport) are Boxs of first offerings for mapping XBRL to RDF. As the Linked data Principles previously outlined RDF can be linked and augmented with additional information from other financial data extracted from sources such as those previously mentioned as a consolidated financial 'mash-up'.

Financial Information Providers

As mentioned a prominent provider of public domain financial information is the Securities and Exchange Commission (SEC) which through their EDGAR web site makes freely available a wide range of company filings. Data from the SEC ranges from information about executives reporting the sale of equity in their companies (Form 4) to detailed annual reports (Form 10-K). Filings are in the older SGML format, free-text (HTML and PDF), and more recently XBRL based on the US-GAAP taxonomy. A wide range of other governmental or intergovernmental organisations publish data in various formats. For Box, central banks use RSS-CB (http://www.cbwiki.net/wiki/index.php/RSS-CBMain) for publishing currency exchange rate data. The United Nations, World Bank, Eurostat, and the OECD are working towards a standard data format for publishing statistical information in SDMX, Statistical Data and Metadata Exchange format (http://sdmx.org/). Finally, there is considerable (financial) information, regarding companies and their executives available in Wikipedia which DBpedia publishes in RDF.

Importance of Web Data

The introduction first mentioned the current trend towards open government and eGovernment data initiatives for public sector information, statistics data and economic indicators as part of information transparency and community exploitation purposes. Currently there exists a large volume of financial and economic data sets available as

Table 5. Open data financial and economic data catalogue providers, validated 12/22/10

Data Catalogues	No. Open Data Sets	No. LD Sets
CKAN (General) • Financial/Accounting, Economic	1554 55/47, 111	216, 0
Infochimps.org (General) Financial/Accounting, Economic	496 117/130, 411	43, 0
Numbrary.com • Financial, Economic	943, 777	
Freebase.com Business data defined in terms of types and instances number. Data is integrated based on freebase schema	54 types, 592K instances	
Eurostat (General)	4981	
World Bank Data sets given in terms of indicators e.g. Global Development Finance	>2000	
Data.gov.uk (General) • Finance, Economic & finance, Economy	5616 357, 300, 312	
Data.gov (General) • Banking finance, insurance, business, federal finance, foreign commerce	3046 220	29

Open Data with limited amounts available as Linked Data. The Comprehensive Knowledge Archive Network (CKAN, http://www.ckan. net/), Freebase (http://www.freebase.com/), Infochimps.org and numbrary.com are Boxs of comprehensive public data catalogues that provide access to a large number of structured economic and financial related data sets. Table 5 provides a listing of non-typical (e.g. central banks, census. gov, SEC) but established open data catalogues which concentrate on the provision of financial and economic related data sets.

With search engine providers beginning to index RDF formatted Web pages e.g. yahoo supports eRDF and RDFa a generic RDF metadata format for its search monkey (http://developer. yahoo.com/searchmonkey/) application and Google supports RDFa, these data sets will searchable and application usable. There is also evidence of corporates within the Pharmaceutical area (http://www.pistoiaalliance.org/) engaging in the practice of pre-competitive open data sharing (Curry et al., 2010). This can occur where data deemed to be of no commercial value is shared with competitors in some neutral hosting space,

and the maintenance resourcing cost is also shared. Only when the data set is re-introduced to the corporate intranet is proprietary data merged to generate competitive advantage.

The Semantic Web adhering to the Linked Data principles of accessibility and usability provides a global means to publish and interlink structured data and with the Web HTTP infrastructure, accessibility is easily extended beyond the enterprises boundaries when dealing with partners, outsourcing or just getting access to relevant open data. Linked Data principles when applied correctly to either Open Data or legacy data, allow its treatment as an integrated information resource, in effect helping to provide a 'Wikipedia for Data'.

CONCLUSION

There is ever increasing pressure on leveraging financial source data to distil better actionable information to drive policy, compliance and support decisions within tighter time frames. Key to better insight begins with having data structured and more thoroughly and accurately interlinked.

With current automated integration approaches in the main based upon ETL, information analysts have little ability to selectively include additional data to supplement analysis. Business analysts will increasingly require this level of information agility and visibility over data sets from within their own corporate or from trustworthy online sources of financial data. Whether assisted by data curators or as part of an extended team the accompanying expectation is that their information systems will cater for this. One of the largest barriers to the application of sophisticated analysis methods and algorithms for use in financial analysis, fraud or regulatory activities is the lack of available comprehensively integrated financial data sources. Achieving any holistic data set is however centred around the successful integration of the sources that will provide a more complete picture. In this chapter we have highlighted the data integration challenges facing the provision of Web based financial information and where Semantic Web standards and their modelling primitives can be of direct benefit in addressing those challenges. An architectural approach based upon a case study using these standards outlines how an abstracted data access layer can be modelled to provide a loosely coupled data mash up that can i) provide a single information access point for interactive exploration and complex querying; ii) accommodate additional data sets as they become available or if required. The components necessary to support the implementation of such an ecosystem are also provided in detail.

Active research areas that would also benefit the Linked Data based approach to Financial Data Integration are i) the development of a data provenance model (Freitas et al., 2010), particularly for dealing with aggregated data ii) data interaction and exploration based on combination approaches such as KOPF iii) post consolidation analytics for activities such as event and risk monitoring and iv) development of business models that exploit Open Data.

REFERENCES

Aberer, K., Cudre-Mauroux, P., & Despotovic, Z. (2005). On the convergence of structured search, information retrieval and trust management in distributed systems. In *Multiagent System Technologies, Third German Conference,* MATES 2005, (pp. 1–14).

Athanasis, N., Christophides, V., & Kotzinos, D. (2004). Generating on the fly queries for the Semantic Web: The ICS-FORTH graphical RQL interface (GRQL). In *3rd International Semantic Web Conference,* (pp. 486-501).

Bartley, J., Chen, Y., & Zalkin-Taylor, E. (2010, February 18). *A comparison of XBRL filings to corporate 10-Ks - Evidence from the voluntary filing program*.

Bast, H., Chitea, A., Suchanek, F., & Weber, I. (2007). ESTER: Efficient search on text, entities, and relations. In *30th ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 671–678).

Berners-Lee, T. (2009). *TED talk*. Retrieved from http://www.ted.com/ talks/tim_ berners_ lee_on_ the_next_web.html

Berners-Lee, T., Chen, Y., Chilton, L., Connolly, D., Dhanaraj, R., & Hollenbach, J. … Sheets, D. (2006). Tabulator: Exploring and analyzing linked data on the semantic web. In *Proceedings of the 3rd International Semantic Web User Interaction Workshop* (SWUI06), (p. 6).

Bouquet, P., Stoermer, H., Mancioppi, M., & Giacomuzzi, D. (2006). OkkaM: Towards a solution to the identity crisis on the Semantic Web. In *Proceedings of SWAP 2006, the 3rd Italian Semantic Web Workshop*, volume 20-1 of CEUR Workshop.

Brewer, E. A. (1998). Combining systems and databases: A search engine retrospective. In Hellerstein, J. M., & Stonebreaker, M. (Eds.), *Readings in database systems* (4th ed.).

Brickley, D. (2002). *RDFWeb notebook: Aggregation strategies*. Retrieved from http://rdfweb.org/2001/01/design/smush.html

Cheng, G., Ge, W., & Qu, Y. (2008) Falcons: Searching and browsing entities on the semantic web. In *WWW '08: Proceeding of the 17th international Conference on World Wide Web*, (pp. 1101–1102). ACM.

Constantine, L. (2002). *The emperor has no clothes: Naked objects meet the interface.*

Curry, E., Freitas, A., & O'Riain, S. (2011). The role of community-driven data curation for enterprises. In Wood, D. (Ed.), *Linking enterprise data* (1st ed.). doi:10.1007/978-1-4419-7665-9_2

Curry, E., Harth, A., & O'Riain, S. (2010). *Challenges ahead for converging financial data.* Workshop on Improving Access to Financial Data on the WEB, Co-organised by W3C and XBRL International, hosted by FDIC, Arlington, Virginia

Cyganiak, R., Catasta, M., & Tummarello, G. (2009). *Towards ECSSE: Live web of data search and integration*. In Semantic Search 2009 Workshop, located at the 18th Int. World Wide Web Conference

de Moura, S. S., & Schwabe, D. (2004). Interface development for hypermedia applications in the semantic web. In *Joint Conference 10th Brazilian Symposium on Multimedia and the Web & 2nd Latin American Web Congress*, (pp. 106–113).

Decker, S., Erdmann, M., Fensel, D., & Studer, R. (1998). Ontobroker: Ontology based access to distributed and semi-structured information. In *DS-8: Proceedings of the 8th Working Conference on Database Semantics.*

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., & Jhingran, A. … Zien, J.Y. (2003). SemTag and seeker: Bootstrapping the Semantic Web via automated semantic annotation. In *Proceedings of the Twelfth International World Wide Web Conference*, (pp. 178–186).

Ding, L., Finin, T., Joshi, A., Pan, R., Scott Cost, R., & Peng, Y. … Sachs, J. (2004). Swoogle: A search and metadata engine for the Semantic Web. In *Proceedings of the 13th ACM Conference on Information and Knowledge Management.*

Economist. (2009, May 28). *Overhauling financial regulation: The regulatory rumble begins*. Retrieved from http://www.economist.com/businessfinance/ displayStory.cfm? story_id= 13743435

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *19*(1), 1–16. doi:10.1109/TKDE.2007.250581

European Commission. (2003). *European Information Society, public sector information - Raw data for new services and products*. Retrieved from http://ec.europa.eu/information_society/policy/psi/index_en.htm

European Commission. (2006). *European information society, MESPIR study*. Retrieved from http://ec.europa.eu/information_society/policy/psi/mepsir/index_en.htm

Freitas, A., Knap, T., O'Riain, S., & Curry, E. (2010). *W3P: Building an OPM based Provenance Model for the Web*. Future Generation Computer Systems.

Frohn, J., Himmeroder, R., Kandzia, P., Lausen, G., & Schlepphorst, C. (1997) FLORID: A prototype for F-logic. In W. A. Gray & Per-Ake Larson (Eds.), *Proceedings of the Thirteenth International Conference on Data Engineering*, April 7-11, Birmingham UK, (p. 583). IEEE Computer Society.

Goguen, J. A. (1993). Keynote: On notation. In *10th International Conference on Technology of Object-Oriented Languages and Systems*, (pp. 5–10).

Guardian. (2009). Investigate your MP's expenses. *Guardian*. Retrieved from http://mps-expenses.guardian.co.uk/

Harth, A. (2009). VisiNav: Visual Web data search and navigation. *Proceedings of the 20th International Conference on Database and Expert Systems Applications*, Linz, Austria, Springer-Verlag.

Harth, A., & Decker, S. (2005). Optimized index structures for querying RDF from the Web. In *Proceedings of the 3rd Latin American Web Congress*, (pp. 71–80). IEEE Press.

Harth, A., Kinsella, S., & Decker, S. (2009). Using naming authority to rank data and ontologies for web search. In *Proceedings of the 8th International Semantic Web Conference*.

Harth, A., Umbrich, J., & Decker, S. (2006). Multicrawler: A pipelined architecture for crawling and indexing semantic web data. In *Proceedings of the 5th International Semantic Web Conference*, (pp. 258-271).

Harth, H., Hogan, A., Delbru, R., Umbrich, J., O'Riain, S., & Decker, S. (2007). *SWSE: Answers before links!* In Semantic Web Challenge, 6th International Semantic Web Conference.

HBR. (2009). *The rise of forensic economics*. Retrieved from http://hbr.harvardbusiness.org/web/2009/hbr-list/rise-of-forensic-economics

Heflin, J., Hendler, J., & Luke, S. (1999). *SHOE: A knowledge representation language for internet applications*. Technical Report CS-TR-4078, Dept. of Computer Science, University of Maryland.

Henzinger, M. (2007). Search technologies for the internet. *Science*, *317*(5837), 468–471. doi:10.1126/science.1126557

Hildebrand, M., Ossenbruggen, J., & Hardman, L. (2006). Facet: A browser for heterogeneous semantic web repositories. In *5th International Semantic Web Conference*, (pp. 272–285).

Hogan, A., Harth, A., & Decker, S. (2007). Performing object consolidation on the semantic web data graph. In *Proceedings of 1st I3: Identity, Identifiers, Identification Workshop*.

Huynh, D. F., & Karger, D. (2009). *Parallax and companion: Setbased browsing for the data web*. Retrieved from http://davidhuynh.net/media/papers/2009/www2009-parallax.pdf.

Hyvnen, E., Mkel, E., Salminen, M., Valo, A., Viljanen, K., & Saarela, S. (2005). Museum Finland – Finnish museums on the semantic web. *Journal of Web Semantics*, *3*(2), 25.

IDG. (2009, August 12). Companies offer services to crunch gov't raw data. Retrieved from http://www.pcworld.com/article/170105/companies_offer_services_to_crunch_govt_raw_data.html

Kementsietsidis, A., Arenas, M., & Miller, R. J. (2003). Managing data mappings in the hyperion project. In *Proceedings of the 19th International Conference on Data Engineering* (ICDE), (pp. 732–734).

Kobilarov, G., & Dickinson, I. (2008). *Humboldt: Exploring linked data*. In Linked Data on the Web Workshop.

Lipkie, D. E., Evans, S. R., Newlin, J. K., & Weissman, R. L. (1982). Star graphics: An object-oriented implementation. In *SIGGRAPH '82: Proceedings of the 9th Annual Conference on Computer Graphics and Interactive Techniques*, (pp. 115–124). ACM.

Manola, F., & Miller, E. (2004). *RDF primer*. W3C Recommendation, February 2004. Retrieved from http://www.w3.org/TR/rdf-primer/

Mendes, P. N., McKnight, B., Sheth, A., & Kissinger, J. C. (2008). TcruziKB: Enabling complex queries for genomic data exploration. *International Conference on Semantic Computing*, (pp. 432-439).

Michalowski, M., Thakkar, S., & Knoblock, C. A. (2003). Exploiting secondary sources for automatic object consolidation. In *Proceeding of 2003 KDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation.*

Norman, D. A. (2002). *The design of everyday things*. Basic Books.

Oren, E., Delbru, R., & Decker, S. (2006). *Extending faceted navigation for RDF data*. In 5th International Semantic Web Conference.

Park, J., & Durusau, P. (2006). Towards subject-centric merging of ontologies. In *Proceedings of the 9th International Protege Conference.*

Schumaker, R., & Che, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems*, *27*(2). doi:10.1145/1462198.1462204

Shneiderman, B., Byrd, D., & Croft, W. C. (1998). Sorting out searching: A user-interface framework for text searches. *Communications of the ACM*, *41*(4), 95–98. doi:10.1145/273035.273069

Sinha, V., & Karger, D. R. (2005). Magnet: Supporting navigation in semi-structured data environments. In *ACM SIGMOD International Conference on Management of Data*, (pp. 97–106).

Stolte, C., Tang, D., & Hanrahan, P. (2002). Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, *8*, 52–65. doi:10.1109/2945.981851

Thompson, C. W., & Ross, K. M. Tennant, H. R., & Saenz, R. M. (1983). Building usable menu-based natural language interfaces to databases. In *9th International Conference on Very Large Data Bases*, (pp. 43–55).

Tufte, E. (2001). *The visual display of quantitative information*. Graphics Press.

Wired Magazine. (2009). *Road map for financial recovery: Radical transparency now!* Retrieved from http://www.wired.com/techbiz/it/magazine/17-03/wp_reboot

Yee, K., Swearingen, K., Li, K., & Hearst, M. (2003). Faceted metadata for image search and browsing. In *CHI '03: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* (pp. 401–408). New York, NY: ACM Press.

KEY TERMS AND DEFINITIONS

RDF: an XML based W3C standard for describing resources that exist on the Web. RDF builds on URI (Uniform Resource Identifier) technologies, using the URI to identify and make statements about resources.

RDFS: allows the creation of vocabularies that describe groups of related RDF resources and their relationships.

SPARQL, the Simple Protocol and RDF Query Language: defines a standard query language and data access protocol for use with the RDF data model. SPARQL provides a means to build complex mash-up sites or search engines that include data mapped from any RDF data source.

Semantic Web: W3C's vision of a Web of linked data. RDF, RDFS, and SPARQL are part of a standards palette offered by the W3C, http://www.w3.org/standards/semanticweb/ to achieve that vision

Linked Data: best practice methods and technologies for exposing, sharing and connecting data via URIs on the Web.

Open Data: a philosophy and practice that makes freely available certain data for consumption without restriction as part of government public transparency and information dissemination initiatives.

ENDNOTE

[1]   http://purl.org/linked-data/cube