Ontology based Integration of XBRL Filings for Financial Decision Making

Romilla Chowdhuri, Victoria Y. Yoon, Richard T. Redmond, Ugochukwu O. Etudo

PII: S0167-9236(14)00232-2
DOI: doi: 10.1016/j.dss.2014.09.004
Reference: DECSUP 12531

To appear in: Decision Support Systems

Received date: 18 September 2013
Revised date: 15 September 2014
Accepted date: 15 September 2014


This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
Ontology based Integration of XBRL Filings for Financial Decision Making

Romilla Chowdhuri, Victoria Y. Yoon*, Richard T. Redmond, Ugochukwu O. Etudo
Department of Information Systems, Virginia Commonwealth University
301 W. Main St. P. O. Box 844000, Richmond, VA 23284-4000

Abstract

In 2008, the U.S Securities and Exchange Commission (SEC) mandated all tier-1 public companies to report their financial statements in eXtensible Business Reporting Language (XBRL). XBRL has the potential to improve the efficiency and accuracy of financial disclosures, and thereby reduce costs. However, semantic heterogeneity across XBRL filings poses a challenge for individual (non-institutional) investors seeking to make inter-firm comparisons using XBRL data. In order to extend the benefits of XBRL to individual investors, this paper presents a design artifact, Ontology-based Framework for XBRL-mapping and Decision-making (OFXD), which provides interoperability between different XBRL filings. Specifically, OFXD resolves the semantic heterogeneity of the element and context definitions in various XBRL filings and generates an XBRL ontology - XBRLOnt.

The results of this study show that the proposed design artifact is capable of addressing semantic heterogeneities between different XBRL filings. Using the concepts of the information value chain, this paper discusses the implications of XBRL interoperability on financial decision-making.

Keywords: eXtensible Business Reporting Language, XBRL, Semantic Web, Ontology, Financial decision-making

* Corresponding Author
1. Introduction

With the passage of time, more individual or non-institutional investors have been drawn to the financial markets; however, very few consult a financial advisor [12]. Instead, such individual investors rely on publicly available data sources to make investment decisions; however, the accuracy of traditional data sources and online financial data is questionable [11, 18]. In 2008, the United States Securities and Exchange Commission (SEC) mandated all tier-1 public companies to report their financial statements in eXtensible Business Reporting Language (XBRL) format [33]. XBRL adoption is expected to improve accessibility, transparency, and efficiency with respect to the dissemination of financial data [2,22]. Previous studies have asserted that the benefits of XBRL extend to consumers including economic stakeholders inside and outside of the organization [2, 5, 14, 21, 23, 27]. However, with respect to financial decision-making, XBRL filings pose cognitive and technical challenges to investors. [7]. In particular, the semantic heterogeneity between XBRL filings becomes problematic when the goal is to compare financial performance between companies. Scholars consider this as a major impediment for investors seeking to make inter-firm comparisons using XBRL filings [see 36]. This calls for more research into understanding the source of heterogeneity and subsequent efforts to systematically resolve towards better comparability of XBRL data across companies.

Informed by the concepts of the information value chain [25,31], the most formidable challenge to leveraging the value of XBRL has been the processing hurdle. From the financial decision-making perspective, three levels of information processing are required to realize the value of XBRL. (1) Information Extraction: Investors need to identify and extract relevant financial elements from XBRL documents. Investors will require software support to view and extract XBRL data elements, since XBRL as a data structure is technically challenging. (2) Information Interpretation: The mere extraction of financial elements (individual data points) is insufficient to make well informed investment decisions; investors need to contextualize XBRL data in order to extract information, for example, by calculating financial indicators relevant to a firm’s ability to settle current liabilities without incurring additional debt or unduly constraining cash flows. Hence, software tools are
required to process XBRL data to inform varying application contexts. (3) Comparative Evaluation: Investors routinely compare relevant financial indicators of performance across multiple companies. However, it is cognitively strenuous to decipher the relationships between data elements found in annual reports [35]. In assuaging this cognitive overload, investors may choose to leverage software-enabled capabilities to compare financial data across companies. To that effect, we developed a design artifact, specifically, an Ontology-based Framework for XBRL mapping and Decision-making (OFXD) which addresses the issue of semantic heterogeneity thereby improving the usability of XBRL data for investors.

OFXD maps semantically heterogeneous XBRL elements and provides a set of rules to calculate ratios based on three financial conditions: profitability, financial leverage/liquidity, and operating efficiency [24]. In designing the proposed artifact, we selected 68 companies from the Standard and Poor’s 100 (S&P 100) index. Filings were accessed using the SECs online Electronic Data-Gathering, Analysis, and Retrieval (EDGAR) system. The design artifact was trained using XBRL format 10-K annual reports for 25 companies within the US jurisdiction for the fiscal year 2011 (FY 2011). We then evaluated the artifact against XBRL format 10-K annual reports for 43 companies in FY2011 and for all 68 companies in FY 2012. In order to illustrate the benefits of the artifact compared to professional data sources, we also evaluated it against data retrieved from Thomson Reuters Data Stream (TRDS). This study constitutes design science research [15] because it conceptualizes, designs, and evaluates the artifact that is informed by the concepts of information value chain. We extended Information Systems research by opening up a new field of inquiry that is important not only to realize the government’s vision for financial reporting, but also to extend its benefits to individual investors.

The remainder of the paper is organized as follows. First, a brief background of XBRL and related work is given followed by a description of the proposed design artifact, OFXD. Next, we describe instance development and the evaluation approach. Following this, we present an analysis of the results and discuss their implications. Finally, we conclude the discussion and identify areas for future research.

2. XBRL
2.1 XBRL Pragmatics

eXtensible Business Reporting Language (XBRL) is a standard for the electronic communication of business and financial data. As a derivative of eXtensible Markup Language (XML), XBRL inherits the key benefits of XML, including platform independence and flexibility. XBRL has been adopted by many organizations in global jurisdictions. XBRL International\(^1\), a non-profit consortium comprised of over 600 organizations across the globe, oversees the evolution of XBRL specifications and coordinates the efforts of local consortia representing international jurisdictions [39, 40]. The XBRL U.S consortium conforms to the American jurisdiction and oversees the development of the taxonomies per United States Generally Accepted Accounting Principles (US GAAP). Likewise, many countries have adopted the standard and have designed taxonomies applicable to their respective jurisdictions. The XBRL specification provides a technical framework for machine-readable financial reports composed of XBRL taxonomies and XBRL instance documents.

**XBRL Taxonomy**

XBRL taxonomies can be viewed as dictionaries of ‘tags’ where a tag defines a particular reporting item in a financial statement. More specifically, the taxonomy document consists of an XML schema (a.k.a. taxonomy schema) and a collection of associated linkbases. The taxonomy schema is a mandatory component of the taxonomy document and must be a valid XML schema document. A taxonomy schema is a logical model that describes and delineates financial reporting concepts. Per XBRL specifications [41], an XBRL concept is the definition of a reporting item such as Net Income or Assets, and each business reporting concept is uniquely defined by an XML element’s syntax declaration.

Linkbases define the extended links in XBRL taxonomies and provide multidirectional links between two or more XML snippets. Extended links within a taxonomy express relationships between concepts, or associate concepts with additional documentation. Five types of extended links can be used in taxonomies: calculation, definition, presentation, label, and reference links. While the first three links define inter-concept relationships, the last two links associate concepts with additional

\(^1\) [http://www.xbrl.org/AboutXBRL](http://www.xbrl.org/AboutXBRL)
documentation. Taxonomy linkbases may be included in the same document as the taxonomy schema or referenced as external resources.

The US SEC defines the taxonomy based on the US GAAP. The GAAP taxonomy provides an exhaustive set of financial concepts sufficing all disclosure requirements associated with US GAAP and the SEC itself. The taxonomy structure is designed to be quite flexible, enabling every company to extend and create its own version of the base taxonomy. This feature allows companies to add concepts specific to their respective businesses, as federal regulators cannot practically generate a complete list of financial reporting items. Per the guidance of the XBRL US consortium, every company filing XBRL documents must extend the base taxonomy before tagging data in the corresponding instance documents. The consortium notes, “… even if every line item on the face of the financial statements of a company had a corresponding element in the XBRL US GAAP Taxonomy, the preparer would still need to create an extension. This is because an extension taxonomy has more in it than just new elements to represent company-specific line items. It also specifies the form of statements (direct vs. indirect cash flow statement, for example), the ordering of line items, the disclosures that are relevant, the names of segments, and other reporting details.”

**XBRL Instances**

The XBRL instance document puts to use the reporting framework defined by a combination of base and extended taxonomies. The purpose of the XBRL taxonomy is to define reporting concepts. The purpose of the XBRL instance document is to quantify facts. A fact is a single reportable piece of information such as Revenues or Net Income. Multiple XBRL elements are required to specify the value of a fact. At minimum, a *container element*, one *item element*, and a *context element* containing sub-elements are required. A container element, the XBRL element, serves as a root element. An item element represents a single business fact or measure. An item element includes attributes such as *name*, *data type*, *reporting period*, etc. A context element contextualizes each instance of the item element by defining the reporting item’s relevant reporting period, company identifier, period,

---

industry segment, and reporting scenario. Together, an item element and its corresponding context element define a fact. Each reporting item includes a contextRef attribute that links the reporting item with the appropriate reporting context.

2.2 XBRL Data Interoperability

The XBRL standard is designed to transmit financial data electronically. When compared to other mediums of information exchange associated with financial reporting, XBRL offers several advantages. As alluded to earlier, XBRL is platform independent and extensible. The extensibility feature of XBRL allows companies to add new elements quite easily without altering the accounting standards. Further, XBRL not only displays information on the web just like HTML, but also adds context to it. Unlike today’s web engines that return irrelevant search results very quickly, XBRL searches the content within context, which not only makes the search faster but also increases the probability of returning relevant results.

Although XBRL has increased accessibility and availability of financial data, there are several impediments to the realization of its full potential [2, 37, 42, 43]. One such impediment is that XBRL is very technical and cannot be used directly by investors who do not have the relevant technical expertise. A second hurdle is that although the extensibility feature allows companies to customize the taxonomies to suit their unique reporting requirements, differences between multiple XBRL taxonomies and instance documents have resulted in serious challenges to the interoperability of XBRL data. Companies extend their taxonomies without following any common convention. Consequently, varying definitions for elements and contexts make it difficult to cross-compare filings.

The integration of proprietary and incompatible formats required by different organizations is an expensive effort. Third, due to the non-uniformity of accounting requirements across industries, standard taxonomies vary from company to company. Vasarhelyi et al. [36] argue that changing accounting policies is not only a hindrance to data comparison but can also lead to incorrect decisions. In addition, taxonomies and instance documents developed using the same XBRL standard differ in terms of element names, structure, contexts, and units.
Of the many sources of heterogeneity in XBRL data across firms, element definition and contextual conflicts are the focus of this research. An Element definition conflict occurs when different taxonomies follow disparate element naming conventions. For example, to report cash generated by operating activities, Apple, Inc. uses the XBRL element label \textit{NetCashProvidedByUsedInOperatingActivities} whereas IBM uses the element label \textit{NetCashProvidedByUsedInOperatingActivitiesContinuingOperations}. Contextual conflict occurs when companies use different conventions to identify context elements. Every XBRL fact is associated with a context element via a contextRef attribute, which provides information pertaining to reporting period, unit of measurement, use of decimal points, precision, and data value. For example in reporting \textit{Assets}, Apple uses the context reference definition to define a reporting period as \texttt{<us-gaap:Assets contextRef="eol_PE2035----1110-K0007_STD_0_20110924_0" ...>116371000000</us-gaap:Assets>}, whereas Dell uses \texttt{<us-gaap:Assets contextRef="I2011Q4" ...>38599000000</us-gaap:Assets>}. Although both Apple and Dell follow the 2011 US GAAP taxonomy and conform to SEC regulations, the two references to contextual data vary markedly in convention.

3. Related Work

The US SEC offers a number of online tools to view and edit XBRL data. The XBRL US consortium also provides tools to help its member organizations file reports with the SEC. Besides open source tools, some commercial XBRL tools such as Fujitsu’s Interstage XWand and Oracle’s Hyperion Disclosure Management are available for the XBRL consortium members and academic users. These tools offer features to edit, view, and validate taxonomies and instance documents; however, they provide limited support to analyze financial information across different XBRL filings even within the same jurisdiction. This limitation stems from the technology suite underlying XBRL, i.e. XML. Inherent in XML are disadvantages in terms of semantic interoperability. Previous research demonstrates the advantages of using Resource Description Framework (RDF) over XML where the actual knowledge is defined as an RDF ontology [8].

A number of approaches to converting XBRL taxonomies and instance documents into their RDF or OWL equivalents have been proposed [3, 9, 13, 30]. One early instance is Declerck and Krieger [9]
who manually created an XBRL OWL base taxonomy using the Protégé editor and defined a scheme for translating relevant accounting principles. Recognizing XBRL’s lack of formal semantics, Lara et al. [17] proposed a process to translate XBRL taxonomies into OWL ontologies; however, they failed to provide a method for mapping XBRL instance documents originating from disparate sources. In a related study, Rabbi et al. [28] also demonstrated the utility of ontologies in terms of their potential to represent financial information in a semantically rich format. To enrich the domain vocabulary, and enhance the semantics, terminology, and structure of XBRL data, Wünner et al. [38] proposed a three-step methodology. XBRL is first converted to RDFS. Then the semantic vocabulary is decomposed from the domain terminology perspective. Lastly, the vocabulary is enriched by analyzing and annotating the data with lexical resources.

Bao et al., [3] leveraged OWL in their efforts to translate the XBRL data model into a semantic web representation. By transforming concepts into classes and arcroles into properties, the study produced three ontologies: XBRL ontology, taxonomy ontology, and instance ontology. However, the derived ontologies were not evaluated for comparing data across different filings. Zhu and Madnick [42] proposed a conceptual model designed to resolve schematic and contextual differences between different taxonomies and instance documents related to different jurisdictions. The authors proposed to first generate an ontology based on the taxonomies used by an instance document and then identify the equivalent concepts using (unelaborated) mapping algorithms. The proposed framework is a skeletal and conceptual architecture, which has not been instantiated. Carretié et al. [6] proposed another approach to transfer XBRL taxonomies and instance documents to the semantic web. Their methodology first converts XML schema into OWL representation i.e. XSD to OWL, and then converts XML to RDF representation. This approach is evaluated for only a single company; as such, its scalability remains in question. Likewise, Garcia and Gil [13] demonstrated an approach to move XBRL data to the semantic web by exploiting XML semantics. However, as the authors assert, their approach is not intuitive and usable enough from a semantic web perspective as the translation fails to preserve the semantic relationships among classes [see 3].
Previous research recognizes the lack of semantics in XBRL data and attempts to translate XBRL data into ontological structure. Representing XBRL data in a semantically rich format (ontology) is an important first step, [see 34, 37]; however, the semantic heterogeneity issue between different XBRL filings remains unresolved. Specifically, there is a need to compare XBRL instance data across companies. This apparent gap in extant research frames the scope of this study. We first set out to identify the sources of heterogeneity and subsequently to design an artifact that improves the comparability of XBRL data for the benefit of individual investors.

4. Ontology-based Framework for XBRL-mapping and Decision-making (OFXD)

4.1 Finance Ratios and Data Set

Financial statements are perceived to be an important source of information about a company’s financial health and are used by investors for financial decision-making [10, 29]. Studies show that relevant historical financial information can be used to distinguish between those firms that are strong performers and those that are under-performers [24]. Borrowing certain concepts from Piotroski [24], our artifact aims to analyze XBRL statements for the calculation of financial ratios. Six financial ratios are used to measure three of the financial conditions: profitability, financial leverage/liquidity, and operating efficiency. Profitability indicates the ability of a company to generate funds internally and is used in the measurement of Return on Assets, Net Cash Provided by Operating Activities, and Accrual. Financial leverage provides a snapshot of a company’s capital structure and its ability to meet future debt service obligations. The two indicators of financial leverage are Current Ratio and Common Stock. Operating efficiency is indicated by Asset Turnover. Table 1 lists the ratios and the corresponding formula. The formulas are sourced from Thomson Reuter’s financial database.

<table>
<thead>
<tr>
<th>Financial Condition</th>
<th>Ratio</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profitability</td>
<td>ROA</td>
<td>Return On Assets (ROA) = Net Income/Total Assets*100</td>
</tr>
<tr>
<td></td>
<td>CFO</td>
<td>Net Cash Provided by Operating Activities</td>
</tr>
<tr>
<td></td>
<td>ACCRUAL</td>
<td>Accrual = (Net income - Cash generated by operating activities) / Total Assets</td>
</tr>
<tr>
<td>Financial Leverage</td>
<td>LIQUID</td>
<td>Current ratio = current assets/current liabilities</td>
</tr>
<tr>
<td>or Liquidity</td>
<td>EQ_OFFER</td>
<td>Common Stock</td>
</tr>
<tr>
<td>Operating Efficiency</td>
<td>TURN</td>
<td>Asset turnover = Net Sales or Revenues/Total Assets *100</td>
</tr>
</tbody>
</table>

Table 1. Variables and Formulas per Thomson Reuters
In developing the artifact, OFXD, we used the S&P 100 as a starting point for the selection of a representative set of companies. After eliminating a number of “volatile” industries (finance and banking, for example) and those companies which had, at the time of data collection, not filed their statements in XBRL format for FY 2011 and FY 2012, a final set of 68 companies representing more than 20 industries was derived. The companies were further divided into two sets: a training set consisting of 25 companies only for FY 2011 and a test set consisting of 43 companies for FY 2011 and all 68 companies for FY 2012. Since all 68 companies were based on US specifications, we ensured a representative mix of industries while dividing the companies between two sets. Based on our analysis of the 25 companies for FY 2011 in the training set, we developed the proposed design artifact, OFXD.

4.2 High-Level Conceptual Model of OFXD

The high-level conceptual model of OFXD is illustrated in Fig.1. To start, OFXD collects XBRL instances (.xml) along with their corresponding taxonomies (.xsd). Using the standard US GAAP taxonomy as well as any relevant custom taxonomies, OFXD retrieves financial values from each XBRL instance document and integrates them into the XBRL Ontology – XBRLOnt. The XBRLOnt generated by OFXD reconciles semantic heterogeneity, enabling end users to retrieve financial values regardless of the elements or contextual definitions used in XBRL instances built from multiple taxonomies.

Fig. 1. Conceptual Model of OXFD
4.3 Architecture of OFXD

Fig. 2 shows the architecture of OFXD which consists of an XBRL repository, RDF Converter, a Company Instance Ontology, a Name Mapper, a Context Mapper, a Value Retriever, a Ratio Rule Engine, and an XBRLOnt for public companies.

**Fig. 2. Architecture of OFXD**

**XBRL Repository**

The SEC has adopted the US GAAP Taxonomy as the standard base taxonomy for financial reporting. In addition to the elements defined in the base taxonomy, companies are encouraged to extend it by defining additional elements. This provides an opportunity for a company to not only conform to the mandatory disclosure requirements, but also represent its unique financial situation more accurately. A company’s XBRL instance document thus includes elements from both the standard base taxonomy and its custom taxonomy. As a result, each XBRL instance contains elements linked to multiple taxonomies. As previously stated, we retrieved XBRL taxonomies and instances from EDGAR, the electronic filing platform provided by the SEC. EDGAR also makes financial statements publicly available. The XBRL Repository stores XBRL taxonomies as well as XBRL instances for the set of 68 public companies for FY 2011 and FY 2012.

**RDF Converter and Instance Ontology**
The original XBRL instance document containing a company’s financial disclosure information conforms to XML format. However, not all companies use the same element names specified in the standard taxonomy. For example, to report profits (or losses), a company may use NetIncomeLoss whereas others may choose ProfitLoss. The ability to retrieve the value of semantically similar (equivalent) elements is essential in comparing financial data from different companies. An ontology is capable of representing the equivalence among classes. To address semantic heterogeneity by encapsulating equivalence, the RDF Converter transforms an XBRL instance document in XML form into an XBRL instance ontology in RDF format. Fig. 3A shows a snapshot of a Net Income element (named NetIncomeLoss) in the raw XML-based XBRL document. Fig. 3B illustrates its corresponding conversion in the RDF-based XBRL document. Essentially the structure and element names are changed. This conversion is necessary because semantic web technologies, which use RDF-based documents, are required to identify and capture the equivalent finance classes: i.e. SalesRevenueNet and OilAndGasRevenue. An RDF ontology consists of classes, individuals, and properties; an XML element becomes a class in an ontology; an instance of an XML element becomes an individual; their relationship is defined by a property. See [1, 19] for the details of an ontology. The term ‘class’ will be used hereinafter.

![Fig. 3A. NetIncomeLoss in the Original XML-based XBRL Document](image)

![Fig. 3B. NetIncomeLoss in the RDF-based XBRL Document](image)

**Name Mapper and XBRLOnt I**

The Name Mapper is designed to automatically construct the XBRL Ontology for public companies, XBRLOnt-I, which hierarchically represents a variety of classes used in reporting the values of the seven financial items: *Net Income, Total Assets, Cash Generated by Operating Activities, Net Sales, Total Current Assets, Total Current Liabilities, and Common Stock*. In
developing Name Mapper, we chose 25 companies and used their FY 2011 RDF-based XBRL instance documents as the training data set. We then manually retrieved the values of seven financial items from the HTML-based financial statements of those companies and created a two-dimensional Value Table; rows list the companies and columns list the seven financial variables; each cell records the value of financial variables for a particular company. The Name Mapper utilizes the Value Table along with the RDF instance ontology to create XBRLOnt I. The algorithm is detailed in Table 2.

<table>
<thead>
<tr>
<th>for</th>
<th>i = 0 to n do, where n is the number of financial items.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Let $F_i$ be the ith financial item</td>
</tr>
<tr>
<td></td>
<td>Create a subclass of FinanceItem class for $F_i$ in XBRLOnt I</td>
</tr>
<tr>
<td></td>
<td>Generate a vector of words $VF_i$ for $F_i$</td>
</tr>
<tr>
<td>for</td>
<td>j = 1 to m do, where m is the number of companies</td>
</tr>
<tr>
<td></td>
<td>Retrieve the value $V_{ij}$ of $F_i$ for the jth company from the Value Table</td>
</tr>
<tr>
<td></td>
<td>Retrieve all classes $E$, whose value is equal to $V_{ij}$, from the jth company’s RDF instance ontology</td>
</tr>
<tr>
<td>for</td>
<td>k = 1 to s do, where s is the number of classes in $E$</td>
</tr>
<tr>
<td></td>
<td>Generate a vector of words $VE_k$ for $E_k$</td>
</tr>
<tr>
<td></td>
<td>Calculate Jaccard Similarity Coefficient ($VF_i$, $VE_k$)</td>
</tr>
<tr>
<td></td>
<td>end for</td>
</tr>
<tr>
<td></td>
<td>Choose $E_k$ with the largest Jaccard Similarity Coefficient</td>
</tr>
<tr>
<td></td>
<td>Store $E_k$ if it is not already in Equivalent Class Array (ECA)</td>
</tr>
<tr>
<td></td>
<td>end for</td>
</tr>
<tr>
<td></td>
<td>Create each class in ECA as a subclass of $F_i$</td>
</tr>
<tr>
<td></td>
<td>Make all subclasses of $F_i$ as equivalent classes</td>
</tr>
<tr>
<td>end for</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2** Class Identification Algorithm

For each of the seven financial items $F_i$, Name Mapper reads its value of the jth company $V_{ij}$ from Value Table and then retrieves all the classes that report $V_{ij}$ from the company’s RDF-based XBRL instance ontology using SPARQL query language. SPARQL Protocol and RDF Query Language (SPARQL) are the most widely used query languages when searching an ontology. The syntax of SPARQL is somewhat similar to that of SQL. However, SPARQL is used for querying an ontology, whereas SQL is used for querying a relational database. (See [26] for the details on SPARQL) The following SPARQL query is generated to search for all classes that report a particular value $V_{ij}$:
SELECT ?y WHERE { ?x rdf:type ?y . ?x rdf:value \( V_{ij} \)};

Following the basic RDF triple format which consists of a subject, a predicate, and an object, the above SPARQL statement is composed of the two triples: “?x rdf:type ?y” and “?x rdf:value \( V_{ij} \)” A prefix “?” refers to a variable in SPARQL and a period between the two triples defines a compound condition. The “rdf” is the namespace for “type,” which represents the instances of a class. The query result is the bindings of ?y, which are the classes whose instance (?x) has the value \( V_{ij} \). For American Electronic Power in which the value of the \textit{Common Stock} in the HTML-based financial statements is $3,274,000,000, the above SPARQL query returns the following four classes:

- \textit{PrepaidExpenseAndOtherAssetsCurrent}
- \textit{StockholdersEquityIncludingPortionAttributableToNoncontrollingInterest}
- \textit{IncreaseDecreaseInAccountsPayable}
- \textit{CommonStockValue}

Since XBRL elements are, per convention, defined by concatenating multiple words, the Name Mapper separates each word in an element class, deletes stop words, such as “And” and “To”, and generates a vector of meaningful words for each retrieved class, \( E_k \). It then assesses the semantic similarity between the vector of each class retrieved \( V_{Ek} \) and that of the financial item \( VF_i \). We chose Jaccard Similarity Coefficient (JSC) \cite{32} to measure semantic similarity between finite sample sets. JSC is defined as the size of the intersection divided by the size of the union of the sample sets, as denoted below:

\[
J \left( VF_i, VE_k \right) = \frac{\left( VF_i \cap VE_k \right)}{\left( VF_i \cup VE_k \right)}, \quad 0 \leq J \left( VF_i, VE_k \right) \leq 1
\]

Where \( VF_i \) is the vector of words for the \( i \)th finance item, and \( VE_k \) is the vector of words for the \( k \)th retrieved class that reports the given value.

For American Electronic Power, the Name Mapper applies the JSC method to measure the similarity between the financial element \textit{Common Stock} and each of the above four classes retrieved, as shown in Table 3.

\[
F_i \leftarrow \{\text{CommonStock}\};
E \leftarrow \{\text{PrepaidExpenseAndOtherAssetsCurrent, IncreaseDecreaseAccountsPayable ... }\};
J : \text{set of similarity coefficients} \leftarrow \{ \};
\text{for each } e_j \in E \text{ do }
\text{ } \quad V_{E_j} \leftarrow e_j;
\]

14
\[ V_F_i \leftarrow F_i; \]
\[ \text{intersect}_{ij} \leftarrow V_F_i \cap V_E_j; \]
\[ \text{union}_{ij} \leftarrow V_F_i \cup V_E_j; \]
\[ \text{sim}_{ij} \leftarrow \text{intersect}_{ij}/\text{union}_{ij}; \]
\[ J \leftarrow J \cup \{\text{sim}_{ij}\}; \]
\[ \text{end} \]
\[ \text{return } \max J; \]
\[ \text{end} \]

**Table 3** Semantic Similarity Measure

Having the largest number of common words with *Common Stock*, the fourth class, *CommonStockValue*, has the highest JCA among the four classes, as shown above. The Name Mapper thus chooses the *CommonStockValue* class for American Electronic Power and stores it in the Equivalent Class Array (ECA). If *CommonStockValue* is recognized as the most similar class name with the *Common Stock* of another company, Name Mapper does not store it in ECA since it is already in ECA. Meanwhile, the value of Amgen’s *Common Stock* in its financial statement is $27,777,000,000, and the search returns the two classes: *StockholdersEquity* and *CommonStockIncludingAdditionalPaidInCapital*. Between these two classes, *CommonStockIncludingAdditionalPaidInCapital* is more similar to *Common Stock* than *StockholdersEquity* as measured by JCA. Name Mapper stores *CommonStockIncludingAdditionalPaidInCapital* in ECA since it is not already stacked in ECA. Name Mapper iterates these steps for the seven financial items and identifies a total of sixteen classes. For each of these sixteen classes, Name Mapper assigns a respective parent class and marks all the subclasses under its parent class as equivalent classes.

The *FinanceItem* class encloses seven subclasses corresponding to seven financial items chosen for this study. Each of the sixteen classes identified by Name Mapper are represented as subclasses under their respective parent classes. Furthermore, XBRLOnt-I represents equivalence among sixteen finance classes. For example, one set of equivalent classes includes the elements *CommonStockValue*, *CommonStockOutstanding*, and *CommonStockIncludingAdditionalPaidInCapital*. This example is illustrated in Fig.4; a snippet of XBRLOnt where the equivalence of the classes is indicated by two-way arrows marked as equivalent classes. The appendix shows the entire set of equivalent classes for
all seven financial classes.

Fig. 4. A Set of Equivalent Classes

**Context Mapper**

The objective of the context mapper is to leverage the context references in XBRL instance documents to retrieve relevant financial classes. An XBRL 10-K instance document reports quarterly as well as annual data over multiple years. On average, a company reports three years of data in a single XBRL instance document. This necessitates the identification of data by fiscal year. The attribute ‘contextRef’ does just that. It defines the reporting period for financial items. On average there are three different context references corresponding to one financial item. These must be reconciled.

As expected and previously noted, values of the contextRef attribute vary significantly from company to company; there is no common pattern for its definition. Table 4 lists a few contextRef definitions identified in instance documents. Additionally, some financial data are reported on a cumulative basis whereas others are reported for a moment in time (e.g. balance sheet data are for a moment in time).

<table>
<thead>
<tr>
<th>Company</th>
<th>Context Reference Syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple Inc.</td>
<td>eol PE2035----1110-K0007_STD_364_20110924_0</td>
</tr>
<tr>
<td></td>
<td>eol PE2035----1110-K0007_STD_0_20110924_0</td>
</tr>
<tr>
<td>Google</td>
<td>eol PE633170--1110-K0018_STD_365_20111231_0</td>
</tr>
<tr>
<td></td>
<td>eol PE633170--1110-K0018_STD_0_20111231_0</td>
</tr>
<tr>
<td>Hewlett Packard/IBM</td>
<td>D2011</td>
</tr>
<tr>
<td></td>
<td>I2011</td>
</tr>
<tr>
<td>Boing</td>
<td>Duration_1_1_2011_To_12_31_2011</td>
</tr>
<tr>
<td></td>
<td>As_Of_12_31_2011</td>
</tr>
<tr>
<td>American Electric Power/Intel</td>
<td>FROM_Jan01_2011_TO_Dec31_2011</td>
</tr>
<tr>
<td></td>
<td>AS_OF_Dec31_2011</td>
</tr>
<tr>
<td>Baker Hughes</td>
<td>TwelveMonthsEnded_31Dec2011</td>
</tr>
<tr>
<td></td>
<td>BalanceAsOf_31Dec2011</td>
</tr>
</tbody>
</table>
As such, we sought approaches to transcend these difficulties. We developed two solutions to these problems: one solution for cumulative items (income statement and statement of cash flows members) and another solution for those items that capture a moment in time (balance sheet items).

**Cumulative Items:** The DocumentType class is a mandatory element that all companies must use to report both the type of financial document (e.g. 10-K, 10-Q) and the latest cumulative period reported in the financial document. As time periods in XBRL are encapsulated by contextRef attributes, the DocumentType element includes a contextRef attribute. This contextRef attribute corresponds to those of the cumulative financial classes of interest: Net Income, Cash Generated by Operating Activities, and Net Sales, as shown in Fig. 5A and 5B.

![Fig. 5A. ContextRef of DocumentType Class](image)

![Fig. 5B. ContextRef of NetIncomeLoss Class](image)

**Moment in Time:** It was also found that the contextRef of the AccountsPayableCurrent class matches those of other financial items that are reported for a moment in time, such as Total Assets, Total Current Assets, Total Current Liabilities, and Common Stock. While a majority of the twenty-five companies in the training set use the AccountsPayableCurrent class, a few companies employ a similar class name, such as AccountsPayableTradeCurrent. Accordingly, we constructed a set of
equivalent classes with respect to AccountsPayableCurrent. Since many firms report multiple years of financial data in a single instance document, we use the value of the reporting year of interest (e.g. 2011) to identify the relevant contextRef associated with AccountsPayableCurrent class or its equivalents. Fig. 6A and 6B show that the contextRef of AccountsPayableCurrent class on September 24, 2011 is the same as that of CommonStockValue on the same day. The Context Mapper thus retrieves the contextRef where the relevant class name belongs to the set of classes equivalent to AccountsPayableCurrent.

![Fig. 6A. ContextRef of AccountsPayableCurrent Class](image)

![Fig. 6B. ContextRef of CommonStockValue Class](image)

**Value Retriever**

The Value Retriever is designed to retrieve the values of relevant financial classes. Table 5 presents the value retrieval algorithm.

---

Read XBRLOnt I  
Create the Company class in XBRLOnt I  
for \( i = 0 \) to \( c \) do, where \( c \) is the number of companies  
Create an instance of the Company class for the \( i \)th company, \( \text{Company}_i \)  
Read the \( i \)th company’s RDF-based instance ontology, \( \text{XBRL}_i \)  
Retrieve the ContextRef of a cumulative financial item, \( \text{CxtRef}_{cum} \)  
Retrieve the contextRef of a daily financial item, \( \text{CxtRef}_{day} \)  
for \( j = 0 \) to \( p \) do, where \( p \) is the number of financial items.
Let $F_j$ be the $j$th financial item.

Retrieve all subclasses $S$ of $F_j$ from XBRLOnt I.

for $k = 1$ to $q$ do, \(q\) is the number of subclasses

Let $S_k$ be the $k$th subclass.

if $F_j$ is a cumulative financial item

Retrieve the value $V$ of $S_k$ with $\text{CxtRef}_{\text{cum}}$ from XBRL;

else

Retrieve the value $V$ of $S_k$ with $\text{CxtRef}_{\text{day}}$ from XBRL;

Attach the value $V$ to Company.

end for

doi
end for

---

**Table 5 Value Retrieval Algorithm**

First, the Value Retriever reads XBRLOnt I and creates the Company class as a sibling class of the FinanceItem class. For each company, the Value Retriever creates an instance (a.k.a. individual) of the Company class and then receives the company's RDF-based XBRL instance ontology along with all taxonomy documents as input. Next, the Value Retriever retrieves the contextRef for cumulative financial items as well as those corresponding to a moment in time from its RDF-based XBRL instance ontology using the following SPARQL queries:

```sparql
SELECT ?CxtRef_cum
WHERE {?x rdf:type :DocumentType . ?x xbrli:contextRef ?CxtRef_cum};

SELECT DISTINCT ?CxtRef_day
WHERE {?className rdf:type owl:Class .
FILTER ( regex(xsd:string(?className), "AccountsPayable") &&
regex (xsd:string(?className), "Current").
?x rdf:type ?className .
?x xbrli:contextRef ?CxtRef_day} ;
```

$\text{CxtRef}_{\text{cum}}$ and $\text{CxtRef}_{\text{day}}$ bind the contextRef of DocumentType class and that of the class whose name contains “AccountsPayable” and “Current”, respectively. Second, the Value Retriever retrieves all the subclasses of each of the financial items from XBRLOnt I. The SPARQL query to search for all subclasses is shown below:

```sparql
SELECT ?x
WHERE { ?x rdfs:subClassOf financialItem} ;
```

Third, the Value Retriever reads the value of each of the retrieved subclasses with a contextRef, either $\text{CxtRef}_{\text{cum}}$ or $\text{CxtRef}_{\text{day}}$, depending on the type of the financial item. Once the non-zero value of a subclass is retrieved, the value is attached to the instance of the Company class, Company. The Value Retriever loops through this logic for the seven finance items whose values are to be retrieved for the
25 companies for FY 2011 during the training phase, and for the 43 companies for FY 2011 and all 68 companies for FY 2012 during the testing phase.

**Ratio Rule Engine**

OFXD calculates the six financial ratios once Value Retriever completes the search for the values of all equivalent classes for the seven finance items. Depending on the class used to report the values of the concepts in its XBRL instance document, the rules to calculate the ratios could significantly vary. For example, companies may report Revenues, SalesRevenuesNet, or both. As per Thomson Reuters, to calculate a company’s Asset Turnover, the value of Revenues gets precedence over SalesRevenuesNet. We used Semantic Web Rule Language (SWRL) to encapsulate the ratio calculation rules of Thomson Reuters. Like rules in general, SWRL rules are can be interpreted as: whenever the conditions specified in the antecedent hold true, then the actions specified in the consequent are executed. See [16] for the details on SWRL.

To illustrate the formal description of SWRL rules, we present a SWRL rule for calculating a company’s Asset Turnover and Accrual ratios. For the Asset Turnover ratio, if the value for Revenues is not zero, then it should be used to calculate the Asset Turnover ratio for the company. In those cases where a company does not report its Revenues but reports its SalesRevenuesNet, then SalesRevenuesNet should be used to calculate the turnover. We defined two rules for calculating Asset Turnover. RuleRev specifies that if a company’s Revenues are not zero, then \( \text{turnover} = \frac{\text{Revenues}}{\text{Assets}} \times 100 \). RuleSalesRev states that if a company’s Revenues are zero but its SalesRevenuesNet is not zero, then \( \text{turnover} = \frac{\text{salesrevenuenet}}{\text{assets}} \times 100 \). The SWRL rules are formally described as:

\[
\text{RuleRev: } \text{company (c)} \wedge \text{hasRevenues (c, ?revenues)} \wedge \text{swrlb:notEqual(?revenues, 0)} \wedge \text{hasAssets (c, assets)} \wedge \text{swrlb:divide (?temp, ?revenues, ?assets)} \wedge \text{swrlb:multiply (?turn, ?temp, 100)} \Rightarrow \text{hasTurn (c, ?turn)}
\]

\[
\text{RuleSalesRev: } \text{company (c)} \wedge \text{hasRevenues (c, ?revenues)} \wedge \text{swrlb:equal(?revenues, 0)} \wedge \text{hasSalesRevenueNet (c, ?salesrevenuenet)} \wedge \text{swrlb:notEqual(?salesrevenuenet, 0)} \wedge \text{hasAssets (c, assets)} \wedge \text{swrlb:divide (?temp, ?revenues, ?assets)} \wedge \text{swrlb:multiply (?turn, ?temp, 100)} \Rightarrow \text{hasTurn (c, ?turn)}
\]
For Accrual, RuleAccural specifies the rule to calculate it:

\[
\text{RuleAccural}: \text{company}(c) \wedge \text{hasNetIncomeLoss}(c, \text{income}) \wedge \text{hasNetCash}(c, \text{cash}) \wedge \text{hasAssets}(c, \text{assets}) \wedge \text{swrlb:subtract}(\text{temp}, \text{income}, \text{cash}) \wedge \\
\text{swrlb:divide}(\text{accural}, \text{temp}, \text{assets}) \\
\rightarrow \text{hasAccural}(c, \text{accural})
\]

**XBRLont-II**

The XBRL Ontology for public companies, XBRLont-II, has evolved from XBRLont-I to which the values of the financial classes and financial ratios for each company were added. At the top level, XBRLont-II contains three classes: FinanceItem, Company, and Industry. The FinanceItem class encloses seven subclasses that correspond to seven financial items. Each subclass contains subordinate classes in a hierarchical fashion. The Company class has sixty-eight instances corresponding to sixty-eight companies. Each instance records the values of seven finance classes and six ratios for FYs 2011 and 2012. The left-hand-side of the screen capture in Fig.7 shows the hierarchical relationship of all the financial classes and the equivalent relationships among the subclasses of each financial class, whereas the right-hand-side of the window displays sixty-eight instances of the Company class (rectangles in circle). The box in the middle of the right-hand-side window shows the values of financial classes and financial ratios relevant to Apple Inc. The Industry class represents the industries associated with sixty-eight companies.

By representing the values of XBRL instance documents comprised of multiple taxonomies, XBRLont-II resolves semantic heterogeneity and in so doing permits the interoperability of XBRL data. As a result, investors may compare the financial disclosures of different companies regardless of class names and context definitions used in XBRL instance documents.
4.4. Instance Development

In developing an instance of OFXD, we used several semantic web technologies. Specifically, the RDF converter uses an RDF-aware utility called ReDeFer\(^3\) to convert XML documents into RDF ontologies. To extract the classes and their contextRef in XBRL instance ontologies for Name and Context Mapper, we used the JENA framework. JENA is a Java-based framework for building a Semantic Web application. It provides an API for reading, processing, and writing RDF data. It also provides a SPARQL query engine, and a rule-based inference engine for reasoning with RDF and OWL data. However, JENA does not support the SWRL rule engine yet. The SWRL rules for the calculation of finance ratios are implemented in JENA rule format; this conversion requires only a minor change in rule syntax. We also used Protégé 4.1 in examining the RDF ontologies [20]; Figures 3-7 are captured within Protégé.

5. Evaluation

In evaluating an instantiated artifact, an examination of its utility with respect to its intended application is the best approach [4]. The design artifact, OFXD, produces XBRLOnt – an ontology to support financial decision making. As such, evaluation of the artifact is conducted in the domain of financial decision-making. As previously noted, we sourced a total of 68 companies for the training

\(^3\) http://rhizomik.net/html/redefer/
and evaluation of OFXD. The artifact was trained on the annual report data of 25 companies for FY 2011. The utility of OFXD is evaluated using 43 (the total data set less the training set) of the total 68 companies for FY 2011 and all 68 companies for FY 2012. We adopted a two-phase approach in our evaluation of the instantiated OFXD artifact. In the first phase we evaluate the values of seven financial classes retrieved from OFXD with respect to the values of financial classes manually extracted from EDGAR 10-K annual filings. We then used the manually extracted values to calculate the six financial ratios, which were used to evaluate the values of six financial ratios as calculated by OFXD. In the second phase, we repeated the evaluations of the values of financial classes and financial ratios with respect to the data extracted from the Thomson Reuters Data Stream (TRDS). While TRDS provides the relevant pre-calculated ratios, it makes only the most recent year’s data available. Thus, just like in phase 1, the ratios used in this phase were manually calculated using TRDS data. Together, the two phases evaluated the utility of the artifact by comparing the values of seven financial elements and six ratios for two fiscal years against two trusted data sources, i.e. SEC’s EDGAR and Thomson Reuters Data Stream.

Following the evaluation guidelines by Hevner et al. [15], we adopted two metrics to measure the utility of our proposed design artifact: Accuracy, and Root Mean Square Error (RMSE). Specifically we evaluated the correctness of finance classes and the performance of ratio calculation rules. We considered OFXD output for a given company vis-à-vis a financial class as accurate if the OFXD value matches, with exact precision, the corresponding EDGAR and TRDS values. As such, accuracy is the percentage of companies for which the correct value for a financial class is retrieved. Likewise, we consider the OFXD calculated ratio as accurate if it matches, with exact precision, the ratios calculated using EDGAR and TRDS values. The RMSE measure contributes to a determination of the utility of the artifact by going beyond the binary classification of a simple percentage accuracy measure. RMSE corresponds to the average spread of the calculated values or ratios around the benchmarked values or ratios. Individual investors may have various levels of tolerance for the difference of OFXD calculated values from the actual values. The results of financial classes and ratios with respect to each data source are presented and discussed in the proceeding sections.
Financial Classes

EDGAR: Table 6 lists the percentage accuracy of the retrieved values for the financial classes and shows that the training set returned each value corresponding to each company and financial class with 100% accuracy, as expected. In comparison, the test sets for both FY 2011 and 2012 attained high accuracy for almost all the financial classes. For FY 2011, while NetIncomeLoss and SalesRevenueNet attained the maximum accuracy of 100%, the remaining five financial classes had accuracy rates between 86% and 97%. For FY 2012, the accuracy rates of all seven financial classes fall between 94% and 99% nearing perfect accuracy.

TRDS: The percentage accuracy of the values of seven financial classes as returned by OFXD ranges between 83% and 95% for FY 2011 test data and between 79% and 91% for FY 2012. We note a marked decrease in accuracy when OFXD is evaluated against TRDS data. For example, with respect to Assets, accuracy drops to 83.72% when evaluated against TRDS data compared to 90.70% accuracy when evaluated against EDGAR data for FY 2011. Likewise, OFXD accuracy vis-à-vis Assets drops to 88% when evaluated against TRDS data compared to 95.59% accuracy when evaluated against EDGAR for FY 2012 data. Other financial classes such as AssetsCurrent and Revenues also report a lower accuracy. As discussed in the next section, this change is due to inaccuracies stemming from TRDS data.

<table>
<thead>
<tr>
<th>Financial Classes</th>
<th>Source: EDGAR</th>
<th>Source: TRDS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training Set</td>
<td>Test Set</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2011</td>
</tr>
<tr>
<td></td>
<td># of Firms</td>
<td>% Accuracy</td>
</tr>
<tr>
<td>Net_Income</td>
<td>25 100</td>
<td>43 100</td>
</tr>
<tr>
<td>Total_Assets</td>
<td>25 100</td>
<td>43 100</td>
</tr>
<tr>
<td>Cash_Generated_By_Operating_Activities</td>
<td>25 100</td>
<td>43 90.70</td>
</tr>
<tr>
<td>Net_Sales</td>
<td>25 100</td>
<td>43 100</td>
</tr>
<tr>
<td>Total_Current_Assets</td>
<td>25 100</td>
<td>43 100</td>
</tr>
<tr>
<td>Total_Current_Liabilities</td>
<td>25 100</td>
<td>43 100</td>
</tr>
<tr>
<td>Common_Stock_Value</td>
<td>25 100</td>
<td>43 86.05</td>
</tr>
</tbody>
</table>

Table 6 Accuracy of the Values retrieved for Finance Classes
We also analyzed the frequency of the equivalent classes found in our dataset of 68 instance documents for the seven financial classes. Table 7 shows the results of our analysis. Financial class names in the right column represent the equivalent classes of their base finance class in the left column. Several idiosyncrasies were observed. While most of the companies use either Revenues or SalesRevenueNet, a few companies such as eBay and Caterpillar report both. Furthermore, companies such as Apache and Devon use unique element labels, OilandGasRevenue, to report revenues. Companies such as American Electronic Power and Anadarko use both NetIncomeLoss and NetIncomeLossAvailableToCommonStockholdersBasic while others use one of the two labels. Additionally, as stated before, the contextRef definition for the financial classes varies significantly across companies (see Table 4). However, these serious conflicts in element naming and context reference definition can be resolved by Name Mapper and Context Mapper as shown in Table 7.

<table>
<thead>
<tr>
<th>Finance Class</th>
<th>Equivalent Classes and Percentage Frequency*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Net_Income</td>
<td>NetIncomeLoss (88.24%), NetIncomeLossAvailableToCommonStockholdersBasics (14.71%), ProfitLoss (1.47%);</td>
</tr>
<tr>
<td>Total_Assets</td>
<td>Assets (100%);</td>
</tr>
<tr>
<td>Cash_Generated_By_Operating</td>
<td>NetCashProvidedByUsedInOperatingActivities (86.76%), NetCashProvidedByUsedInOperatingActivitiesContinuingOperations (11.76%);</td>
</tr>
<tr>
<td>Activities</td>
<td></td>
</tr>
<tr>
<td>Net_Sales</td>
<td>Revenues (48.53%)</td>
</tr>
<tr>
<td></td>
<td>OilAndGasRevenue (2.94%)</td>
</tr>
<tr>
<td></td>
<td>SalesRevenueGoodsNet (4.41%)</td>
</tr>
<tr>
<td></td>
<td>SalesRevenueNet (50%);</td>
</tr>
<tr>
<td>Total_Current_Assets</td>
<td>AssetsCurrent (97.06%);</td>
</tr>
<tr>
<td>Total_Current_Liabilities</td>
<td>LiabilitiesCurrent (97.06%);</td>
</tr>
<tr>
<td>Common_Stock_Value</td>
<td>CommonStockValue (79.41%)</td>
</tr>
<tr>
<td></td>
<td>CommonStockValueOutstanding (5.88%);</td>
</tr>
<tr>
<td></td>
<td>CommonStockIncludingAdditionalPaidInCapital (11.76%);</td>
</tr>
</tbody>
</table>

*percentage is based on 68 companies for FY2011

Table 7 Frequency of XBRL Element Names

Ratios
**EDGAR:** The results of the evaluation of the OFXD artifact vis-à-vis financial ratios are summarized in Table 8. OFXD correctly calculated the selected ratios in the training set for all firms. Using FY 2011 testing data, TURN achieved higher accuracy of 91% and lower RMSE of 0.88. For FY 2012, TURN achieved a higher accuracy of 91% and lower RMSE of 0.43. Overall, for FY 2011, the accuracy of ratios ranged between 84% and 98% whereas for FY 2012, the accuracy of ratios ranged between 91% and 97%. Consequently, RMSE ranged between 0.01 and 0.88 for FY 2011 and between 0 and 0.43 for FY 2012. Finally, as we compared the overall improvement in accuracy and RMSE between FY2011 and FY2012, the accuracy of FY 2012 data was generally higher than that of FY2011 with the exception of CFO, which dropped from 98% to 97%. The RMSE measures were also much lower for FY 2012 than FY 2011.

**TRDS:** Evaluation performance for OFXD with respect to financial ratios benchmarked against TRDS was markedly lower than OFXD performance benchmarked against EDGAR. The inaccuracies in the values of financial class data observed in TRDS were transferred and compounded in the ratio results. Overall, for the FY 2011 testing data, the accuracy ranged between 79% and 90% and RMSE ranged between 0.26 and 0.95. For FY 2012, the accuracy ranged between 83% and 95% and RMSE ranged between 0.12 and 0.47. TRDS benchmarked performance was particularly lower for 2011 test data wherein ACCRUAL and TURNOVER values were calculated with 79% accuracy and RMSE of 0.5516 and 0.9514, respectively.

<table>
<thead>
<tr>
<th>Ratios</th>
<th>Source: EDGAR</th>
<th>Source: TRDS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training Set</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test Set 2011</td>
<td>Test Set 2012</td>
</tr>
<tr>
<td></td>
<td>% Accuracy</td>
<td>% Accuracy</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>RMSE</td>
</tr>
<tr>
<td>ROA</td>
<td>100 0.00</td>
<td>86 0.59</td>
</tr>
<tr>
<td>CFO</td>
<td>100 0.00</td>
<td>98 0.01</td>
</tr>
<tr>
<td>ACCRUAL</td>
<td>100 0.00</td>
<td>81 0.57</td>
</tr>
<tr>
<td>TURN</td>
<td>100 0.00</td>
<td>88 0.88</td>
</tr>
<tr>
<td>LIQUID</td>
<td>100 0.00</td>
<td>91 0.60</td>
</tr>
<tr>
<td>EQ_OFFER</td>
<td>100 0.00</td>
<td>91 0.26</td>
</tr>
</tbody>
</table>

**Table 8** Accuracy of Finance Ratios
While the OFXD process is quite accurate in terms of its ability to extract financial line items from XBRL instance documents in spite of the considerable issue of semantic heterogeneity, it is important to acknowledge that the artifact does not operate at 100% accuracy. To verify the reasons for incorrect ratio results, we analyzed the values of the financial elements used to calculate financial ratios. Specifically, for every incorrect ratio, we compared the value of a financial element retrieved by OFXD with respect to the value retrieved from EDGAR and TRDS. Based on our findings, we summarized the reasons for incorrect mapping and ratios into three categories:

1. **Erroneous Baseline**: In many instances the values retrieved by OFXD were the same as those found in the XBRL instance documents; however, TRDS may report different values. For example, for DOW Chemicals, EDGAR reports a value of ‘69224000000’ for Assets; whereas, TRDS reports ‘721240000000’. In another case, for Target Corp., it was found that the values returned by TRDS were only relevant to FY 2010. After further verifying the values with EDGAR, it was found that the values reported by TRDS were completely different than those of XBRL and EDGAR filings for this company. Previous research also reports such data quality issues in databases such as Compustat [11]. The discrepancies between TRDS and EDGAR data are suggestive of data quality issues in professional data sources; however, addressing such discrepancies is out of the scope of this study.

2. **Missing Aggregation**: In a few instances, the companies did not report an aggregated value for financial elements. For example, both TRDS and EDGAR statements reported Common Stock value for Accenture as 7400. However, in the XBRL instance document, this figure was split into three contexts with values as 16000, 1000, and 57000. Likewise, Ford Motor Company reported two values of Common Stock corresponding to different members: CommonStockMember value of ‘37000000’ and the CommonClassBMember value of ‘1000000’. However, both EDGAR and TRDS reported an aggregate number. Since our retrieval logic did not account for this condition, the model did not return it. In the future, the Value Retriever algorithm could be modified to first iteratively return values corresponding to the total number of contextRef definitions of a financial element and then aggregate all returned values.
3. **Inconsistent Convention**: Since the Value Retriever algorithm was based on the element and contextRef definitions found in the training set, it is obvious that any different definition found in the test set will not be identified. The algorithm could be continuously improved to map any new element or contextRef definition; however, the element labels can only be considered synonymous if they indeed refer to the same accounting term. Our analysis revealed that some companies use completely different convention. For example, a few companies use unique labels such as OilandGasRevenue or ProfitLoss. While the rule could be modified to map such labels, one needs to ensure if the terms could be used interchangeably and what would be the impact on the calculation of other financial measures.

In the future, OFXD will be extended to include Reinforced Learner, which aims to learn unseen class names for inclusion in XBRLOnt (in this sense, the Reinforced Learner augments XBRLOnt). When, for a given company, the value retriever does not find a value for one of the seven financial classes (or any set of financial classes OFXD is trained on), it will defer to the Reinforced Learner. For each of the seven financial classes OFXD is currently trained on, the Learner houses a separate but related set of financial classes useful in calculating the value of a financial class. The context-relevant values corresponding to these *helper classes* are retrieved for each unseen class using the OFXD process given an XBRL instance where one or more of the seven financial classes cannot be retrieved. Based on the retrieved values, the Reinforced Learner may calculate upper and lower bounds of the possible values of the relevant unseen financial class. The Learner will subsequently retrieve all values from the relevant instance document corresponding to the relevant context which fall between the upper and lower bounds. The Learner thus narrows the area of consideration to a set of likely classes. The Learner will then calculate the Jaccard-based semantic similarity between each subclass of the problematic financial class(es) and each reporting item in the set of likely classes. The most similar class name will be added to XBRLOnt as a subclass of the unseen financial class thus expanding the set of equivalent classes in XBRLOnt.

6. **Discussion**
An important consideration in mandating XBRL by the US SEC is its potential to enhance the efficiency and accuracy of the electronic communication of business and financial data [7]. In this regard, the XBRL standard holds great promise for the improvement of the financial disclosure process. However, in order to realize the full potential of the XBRL standard, the semantic heterogeneity present between different XBRL filings must be resolved. Such heterogeneity poses a challenge when comparing financial information sourced from multiple companies, becoming a major hurdle for investors wishing to use XBRL data.

A clear contribution of this research is extending the value of XBRL financial information to individual investors. The proposed design artifact satisfies the design science research requirements [13]. The results show that the design artifact is capable of addressing semantic heterogeneity by mapping the financial classes. Investors could use the ratios calculated based on the mapped financial classes in making their financial decisions.

This research is the first of its kind to automatically build an ontology based on values extracted from XBRL-format financial statements towards addressing the issue of semantic heterogeneity in XBRL filings. We contributed to developing a method that measures semantic similarity using the Jaccard Similarity Coefficient between the financial data names in EDGAR’s HTML-based financial statements and their equivalent names in XBRL instance documents. Furthermore, the method is capable of mapping the semantically similar class names in different XBRL filings as equivalent classes.

We also contributed to illustrating how the proposed artifact can be instantiated using several semantic web technologies, including JENA, SPARQL, and SWRL. We evaluated the utility of the proposed artifact with XBRL filings for the FY 2011 and FY 2012. Considering the fact that XBRL filings for the FY 2012 were not included in the training set, the accuracy rate of results clearly shows the generalizability of the proposed artifact for multiple years. Moreover, we analyzed the features of XBRL standard that could prove technically challenging and cognitively overloading for investors. The proposed framework provides insights about how the perceived potential of XBRL could be extended for the benefit of individual investors.
From a practical perspective, the research provides insights into why the benefits of XBRL are not reaped and what can be done to increase the possibility of leveraging its benefits. By comparing XBRL filings, we identified certain sources of heterogeneity. These findings provided awareness to the practitioners about the type of issues that could be faced in comparing XBRL filings. The design artifact is evaluated against EDGAR, a public financial data source, and TRDS, a professional financial data source. Individual investors who could not afford access to such professional data sources would benefit from this framework. Finally, this research finding could prove helpful to companies and XBRL consortiums. Companies can use these findings as a guide to report their statements more consistently. The issues found may be of use to the XBRL in improving the XBRL standard. Specifically, the consortium needs to improve its taxonomy rules in terms of standardizing certain common fields, defining conventions for elements and context definitions, and adding tags to provide additional descriptions for elements.

7. Conclusions

Addressing the issue of semantic heterogeneity is the first step towards realizing and extending the potential benefits of the XBRL standard to individual investors. This study addresses the fundamental issue of XBRL interoperability and presents a design artifact to improve XBRL interoperability for financial decision-making. Our results offer promising steps toward addressing the issue of semantic heterogeneity present in inter-firm comparisons of XBRL data elements. In addition, this study provides several rich opportunities for future research. One such area concerns extending the proposed artifact by integrating XBRL and non-XBRL data. Such a model will allow investors to compare potential companies not only based on financial metrics, but also on other characteristics such as governance scores, stakeholder assessments, market shares, etc. Furthermore, the model could be enhanced to incorporate extant methods that enable the prediction of future performance based on historical information. Additionally, a user interface could be provided for investors to interact with the framework, which in turn would enable us to perform user-involved evaluations. To that end, our framework for integrating XBRL data provides a useful starting point for future efforts aimed at leveraging the promise of XBRL to individual and institutional investors.
Acknowledgement

The authors would like to thank Dr. Ruth Epps in the Department of Accounting, School of Business, Virginia Commonwealth University, for sharing her domain knowledge and helping the authors to conceptualize accounting principles.

References


Appendix – Seven Parent Finance Classes and their Equivalent Subclasses in XBRL Ont
Biography

Romilla Chowdhuri is a doctoral student in the Department of Information Systems at the Virginia Commonwealth University. She received her M.S in Information Systems from Virginia Commonwealth University, in 2011. Her research interests are decision support systems, information security, and social networks.

Victoria Yoon is Professor in the Department of Information Systems at the Virginia Commonwealth University. She received her M.S. from the University of Pittsburgh and her Ph.D. from the University of Texas at Arlington. Her primary research area has been the application of Artificial Intelligence to business decision-making in organizations and technical and social issues surrounding such applications. She has published articles in such leading journals as MIS Quarterly, Decision Support Systems, Communications of the ACM, and Journal of Management Information Systems.

Richard Redmond is Associate Professor in the Department of Information Systems at the Virginia Commonwealth University. He received his DBA from Kent State University. He specializes in software engineering, database systems, expert systems and applications of AI to business. His research and publications have been in the areas of software productivity, expert systems, applications of AI to business, database/imagebase design theory and compression theory. He has published articles in Communications of the ACM, Expert Systems with Applications, and Journal of Information Technology Theory And Application.

Ugochukwu Etudo is a doctoral student in the Department of Information Systems at Virginia Commonwealth University. He received his M.S in Information Systems from Virginia Commonwealth University in 2013. His research interests include text mining, decision support systems and socio-technical systems.
Highlights:

- We address the semantic heterogeneity issue and thereby improve the usability of XBRL data for investors.

- We propose an Ontology-based Framework for XBRL-mapping and financial decision-making (OXFD).

- OXFD maps semantic heterogeneous XBRL elements and provides a set of rules to calculate financial ratios.

- We contribute to developing a method that measures the semantic similarity using Jaccard Similarity Coefficient between the financial data names in HTML-based financial statements and their equivalent names in XBRL instance documents.

- We evaluate the utility of the proposed artifact with XBRL filings for the FY 2011 and FY 2012. The accuracy rate of results clearly shows the generalizability of the proposed artifact for different XBRL filings for multiple years.

- Our framework for integrating XBRL data provides a useful starting point for future efforts aimed at leveraging the promise of XBRL to individual and institutional investors.