Leveraging XBRL Calculation Linkbases to Overcome Semantic Heterogeneity across XBRL Fillings: The Multi-Ontology Multi-Concept Matrix (M³)

Completed Research Paper

Ugochukwu Etudo
Virginia Commonwealth University
Department of Information Systems
301 W. Main Street Richmond, VA
etudouo@vcu.edu

Victoria Yoon
Virginia Commonwealth University
Department of Information Systems
301 W. Main Street Richmond, VA
vyyoon@vcu.edu

Abstract

In 2008, the US Securities and Exchange Commission (SEC) mandated that large accelerated filers issue financial statements in eXtensible Business Reporting Language (XBRL). One purported benefit of issuing financial statements in a machine readable language is the facilitation of automated inter firm comparisons. However, XBRL, an XML-based language, is extensible by each individual filer. This extensibility compromises the ability of an automated agent to make meaningful comparisons between firms on any given metric or financial concept by introducing semantic heterogeneity across filings. In this paper we make an account of the development of an automated method for the generation of ontologies for XBRL calculation linkbases. In addition, we demonstrate a novel method for resolving semantic heterogeneity which leverages the expressivity of XBRL calculation linkbases. Our major premise is that the meaning of a given financial concept is captured by its relative position in a calculation hierarchy. The representation of calculation linkbases using an ontology language thus becomes the basis of our efforts towards the resolution of semantic heterogeneity across XBRL filings in the US jurisdiction.

Keywords: XBRL, Semantic Heterogeneity, Ontologies, Design Science

Introduction

Individual and institutional investors alike rely on financial statements for investment decision making. Investment decision making (as opposed to speculation1) is a process that culminates in estimates regarding the expected valuation of a firm by the stock market. The valuation process is multifaceted and each investor must determine an idiosyncratic estimate of the value of a set of firms. To do so, information on the intrinsic properties of the firm is required. Such information may be found within the financial statements and are considered to be fundamental indicators. For instance, Fernández (2007) provides a compendium of 10 different methods of firm valuation using cash flow data found in financial statements. An analysis of financial statements is a necessary step in any investing activity.

1 Speculation is often referred to as technical trading. Speculators are not particularly interested in the financial statements and make decisions by studying stock price patterns.
If financial statement data is crucial to investment decision making, then it follows that there must exist a means through which the financial statements of one company may be compared to those of others. As such, the ability of an investor to make decisions based on data is contingent on the ability to compare firms on the same metrics. While institutional investors are able to access expensive and proprietary databases to that end, the individual or smaller institutional investor remains in a challenging situation. For the discerning individual investor seeking to make inter-firm comparisons, a number, \( i \), of financial concepts \( C_i \) must be evaluated for each of \( n \) firms, \( C_{i,n} \), in a prospective portfolio such that conceptually, \( C_{i,1} = C_{i,2} = C_{i,n} \). For small \( n \) and small \( i \), the investor may leverage investment expertise and visually inspect the corresponding financial statements to retrieve values for all relevant instances of \( C_{i,n} \). However, as \( n \) and \( i \) become large, the prospect of visual inspection becomes increasingly tenuous. At this point, automation is necessary. This very reality is presented as one of the motivations behind eXtensible Business Reporting Language (XBRL). XBRL is a machine readable standard through which financial statements and other business information can be accurately and definitively stated. XBRL is automatically consumed by regulators, companies, accountants, data providers, analysts and investors. While XBRL is charged with standardizing the business reporting supply chain,\(^2\) problems do exist. More specifics regarding XBRL are presented later in this paper.

Inherent to the standard is the notion of extensibility. That is, where definitions of an idiosyncratic business reporting situation do not exist, the reporting entity may implicitly and explicitly declare them within their discoverable taxonomy set (DTS). By doing so, the nuances of the reporting situation are implicitly expressed in the structure of XBRL documents and explicitly expressed in concept declarations. However, this flexibility (a necessity for the ubiquitous acceptance of XBRL technology) results in heterogeneous reporting environments even within the same reporting jurisdiction. As recently as 2012, Zhu & Wu in a series of experiments demonstrated that 63% of XBRL concepts across US filings were not interoperable. The authors also state that only 27% of financial information across XBRL filings is comparable even when only financial element names from the standard generally accepted accounting principles (GAAP) taxonomy are considered. Prior works recognize the heterogeneous nature of XBRL data and attempt to translate XBRL data into ontological structure to resolve the semantic conflicts (Carretié et al. 2012; Spies 2010; Wunner et al. 2010). A recent study conducted by Chowdhuri et al. (2014) presents an ontology-based approach to integrate the XBRL filings of multiple companies over years, enabling investors to perform comparative performance analyses cross companies. Although this work provides a useful starting point for leveraging the promise of XBRL to investors, there exists a need for a more effective resolution of the semantic heterogeneity in XBRL filings. In response to this need, this paper makes an account of a method for the automatic generation of calculation linkbase ontologies towards this objective. We see the formal expression of calculation linkbase semantics as a necessary first step towards the automated resolution of business reporting heterogeneity where XBRL is concerned. This belief is predicated on the assumption that there exist within a calculation linkbase, calculation hierarchies both necessary and sufficient to express the meaning of any given financial concept. Specifically, the meaning, \( M_i \), of a financial concept \( C_i \) is encapsulated in its calculation hierarchy \( H \) such that \( H \Leftrightarrow C \). Formula 1 summarizes these statements.

**Formula 1:** \[ M_i \Leftrightarrow H_i \Leftrightarrow C_i \]

Although prior work developed schema matching for XML DTDs (Document Type Descriptors) based on XML hierarchies (Doan et al. 2003), to our knowledge, the academic literature has neither recognized nor leveraged the potential of calculation linkbase hierarchies encapsulated in Xlink to delineate the meaning of a financial concept expressed in XBRL. Therein lies the novel contribution of this work. Our research question is as follows: *is the hierarchical information contained within XBRL calculation linkbases able to discriminate between financial concepts in an XBRL filing to the extent that it can constitute an adequate solution to the problem of semantic heterogeneity?*

The remainder of this paper is structured as follows. We describe the modalities of the XBRL standard in brief and describe previous work on the resolution of semantic heterogeneity across XBRL filings of the same type. Next, we describe our method for the automatic generation of calculation linkbase ontologies and our method for leveraging these ontologies in the resolution of semantic heterogeneity across XBRL.

\(^2\) https://www.xbrl.org/the-standard/what/an-introduction-to-xbrl/
filings. We evaluate our approach so as to provide a definitive answer to our research question. We conclude by presenting the future direction of our research.

Related Work

Wunner et al. (2010) take on the challenges associated with the reuse of XBRL vocabularies as defined in XML linkbase format. Their approach employed a tripartite strategy of first using taxonomic, meronymic and other types of relationships to transform the semantic structure of XBRL taxonomy documents. Secondly, their approach employs domain specific dictionaries to articulate the terminological structure of XBRL terms. Finally, terms are structured by tagging them with parts-of-speech and other linguistic markers. This method involves much human intervention, is concerned with IFRS implementation of the XBRL standard and is of questionable generalizability. Zhu and Madnick (2007) propose a context and schematic matching approach to resolving semantic heterogeneity across XBRL taxonomies developed for different financial reporting jurisdictions. While their solution architecture appears promising, the authors have not, to our knowledge, proffered an instantiated solution to date. Declerck and Krieger (2006) manually create a base ontology of XBRL in OWL and produce guidelines for translating XBRL documents into this base ontology in the German financial reporting jurisdiction. This approach is a manual one in the sense that changes in financial reporting rules in the German reporting jurisdiction would require manual, human intervention with respect to the presented artifact. García and Gil (2009) instantiate an automatic approach for converting XBRL documents into RDF based ontologies. However, these authors had a central goal of expressing XBRL data as linked open data on the semantic web such that they did not take steps to resolve semantic heterogeneity. Very much preliminary and exploratory in nature, the work presented in Raggett (2009) navigates the opportunities available to XBRL on the semantic web. No artifact is created as a result of that work. Carretié et al. (2012) provide an automated method to generate RDF based ontologies from XBRL documents with the end goal of establishing equivalences across jurisdictions and firms. The mappings however are done manually, a major limitation for the mass consumption of XBRL data. Bao et al. (2010) provide a semantically rich framework for translating XBRL documents into OWL. Their method allows for the expression of actual XBRL semantics in OWL as opposed to the taxonomic representations on which the papers reviewed herein depend. However, it is unclear the extent to which their method may be automated. Further, no approach is proposed in their work regarding the reconciliation of semantic heterogeneity across XBRL filings. Finally, Chowdhuri et al. (2014) proposed a framework for overcoming semantic heterogeneity across XBRL filings. Their method relies solely on lexical similarity coefficients between financial concepts. Further, the accuracy of their method needs to be improved.

To conclude our review of related work, we remark upon two major deficiencies in the literature. First is the lack of automated methods for ontology generation and subsequent consumption from XBRL documents. Second, we observe that no adequate methods exist to resolve the problem of semantic heterogeneity for investors with the exception of Chowdhuri et al. (2014). While Chowdhuri et al address the issue of semantic heterogeneity we believe that the reliance on lexical similarity alone fails to produce the high accuracy required for investment decision making process. Our work begins to fill these gaps by instantiating a design artifact which introduces additional, previously unused, hierarchical information present in XBRL calculation linkbases.

The XBRL Standard

We briefly describe those aspects of the XBRL standard that are relevant to the present research. An XBRL filing of an annual report is made up of a number of linked XML documents. Broadly, these documents fall into two categories: Instance Documents and Taxonomies. Within an instance document are XBRL facts, that is, values associated with XBRL reporting concepts. For example, an XBRL fact may be: company X reported net income of Y for the period starting at t1 and ending at t2. Reporting concepts are defined in an XML schema document referred to as the taxonomy schema. Instance documents refer to taxonomy schemas for concept definitions. The taxonomy schema is further specified by a series of linkbases. Linkbases are repositories of XML extended links (XLink) that specify some relationship between the concepts defined in the taxonomy schemas. In this research, our focus is on the calculation linkbase, an XML document containing links to reporting concepts in the taxonomy schema and
calculation arcs which describe calculation relationships between concepts. A calculation linkbase can be thought of as a calculation hierarchy, expressing summation relationships between reporting concepts.

**Design Theory**

The core of the design described in this paper is expressed as a theory of design. Herbert Simon introduced a distinction between natural science and the science of design (Simon 1969). For Simon, natural science encompasses bodies of knowledge regarding classes of things which occur in the natural world. A natural science has as its goal the exposition of the properties, behaviors and interactions of a particular class of things which occur in the natural world (Simon 1969). A science of design is concerned with a class of artificial things, that is, man-made objects. A science of design is necessarily focused in the fulfillment of purpose (goals). Three interrelated items are involved in goal oriented adaptation (i.e. the focus of a science of design): “the purpose or goal, the character of the artifact, the environment in which the artifact performs” (Simon 1969). Natural science knowledge is relevant to design to the extent that the natural world imposes on the artifact an environment and determines the range of possibilities regarding the character of the artifact. To support the notion of design as a science, there exists a need for a basis on which to theorize about design. Such a basis “supports the cumulative building of knowledge, rather than the re-invention of design artifacts and methods under new labels” (Gregor and Jones 2007). The theoretical contributions of a design study such as this one can be expressed across eight dimensions: purpose and scope, constructs, principles of form and function, artifact mutability, testable propositions, justificatory knowledge, principles of implementation, and expository instantiation (Gregor and Jones 2007). These eight dimensions of design theory (more specifically, information systems design theory or ISDT) allow for the “prescription of guidelines for further artifacts of the same type” (Gregor and Jones 2007) such that an ISDT is a generalizable set of principles which inhere in a goal-seeking IS design artifact. Gregor and Jones add that an ISDT may refer to either the form of a design or the act of implementing a design in the real world. The ISDT presented in this paper regards the form of an artifact, where this form can be characterized along the eight dimensions of ISDT and can be generalized onto similar design domains.

Until this point, we have described the motivations of our artifact, related work, its application domain, and its major design premise. In the sections to follow, the overarching framework and solution approach to our problem domain are expounded upon. Following this, we evaluate the efficacy of our design with respect to its stated purpose. We then articulate the design as a contribution to ISDT by describing it along the eight dimensions in Gregor and Jones (2007). Finally, we conclude and discuss applications to practice and future research directions.

**Proposed Framework**

Figure 1 is a diagrammatic representation of the proposed framework. The framework consists of training XBRL documents, Calculation Linkbase Ontology Generator, Investor Vocabulary, M3, and FinOnts. The design artifact presented herein commences its operation with the input of XBRL documents, specifically, calculation linkbases. These calculation linkbases are fed into a Calculation Linkbase Ontology Generator, the inner workings of which are detailed in following section. For each calculation linkbase, a corresponding calculation linkbase ontology is generated. Upon the generation of these ontologies, the M3 algorithm is invoked. The algorithm takes as arguments a set of calculation linkbase ontologies. The initial set of calculation linkbase ontologies are used to train the algorithm with respect to enabling a learning mechanism that allows for the extraction of XBRL terms which correspond to investor terms. Of course, the investor terms (i.e. a user input regarding a financial concept for which values are to be extracted) are fed to the algorithm. The algorithm learns a set of XBRL terms, a single term for each investor term supplied by an end user. The result of the algorithm is FinOnt, an ontology which relates investor terms with XBRL concepts.
Figure 1 – Proposed Framework for the Resolution of Semantic Heterogeneity with Calculation Hierarchies

The framework presented in figure 1 is distinguished from existing approaches to the resolution of semantic heterogeneity across XBRL filings in three important ways. First, it is the only approach to our knowledge to automatically generate ontological representations of XBRL calculation linkbase taxonomies for the purpose of extracting the defining a subsumption hierarchy depicting the calculation relationships between financial terms. Second, this approach is the first attempt in this domain to take into account investor vocabulary.

**Ontologies**

Perhaps the most apt definition of ontology may be found in the work of Guarino (1998, p. 2) who wrote that: “in its most prevalent use in [artificial intelligence] AI, an ontology refers to an *engineering artifact*, constituted by a specific *vocabulary* used to describe a certain reality, plus a set of explicit assumptions regarding the *intended meaning* of the vocabulary words [...] In the simplest case, an ontology describes a hierarchy of concepts related by subsumption relationships; in more sophisticated cases, suitable axioms are added in order to express other relationships between concepts and constrain their intended interpretation.” Our use of ontologies in this paper is in lockstep with Guarino’s definition as we seek to capture a specific vocabulary of financial terms to be described as a hierarchy of concepts related by subsumption relationships. Formally conceived, an ontology becomes “a set of logical axioms designed to account for the intended meaning of a vocabulary” (Guarino 1998, p. 4). We use ontologies as a key informing kernel theory (Gregor and Jones 2007; Walls et al. 1992) due to the fact that we want to formally conceptualize the meaning of financial concepts especially as they relate to investor vocabulary. Ontologies and ontology modelling languages are explicitly designed for the representation of meaning or semantics. The question is often asked, “why ontologies?” Indeed, there exist other techniques to represent hierarchical or structured data, eXtensible Markup Language (XML) being the most popular and ubiquitous. There are several reasons why ontologies are an ideal technology to instantiate our high level framework.

Ontology languages such as Resource Description Format (RDF) and Web Ontology Language (OWL) are designed for data interchange and interoperability. So is XML. Both sets of languages can be used to represent hierarchical data. However, the similarities stop there. Understanding the choice between an ontology language and XML involves an understanding of the difference between XML’s purpose as a
representation language and the purpose of ontology languages as conceptualizations of meaning. Consider the example of a steak dinner. Steak dinners may be purchased and consumed at high end steakhouses, they may be made at home on a grill, and they may be purchased at middle of the range restaurants and so on. Represented in an ontology language, steak dinner, a concept, is the same concept regardless of where it is purchased. In XML, the manner in which steak dinner is represented (i.e. where it is bought or made) is what matters. Ontology languages represent the inherent meaning of a concept, regardless of its representation. New data instances, for example, can be inferred as members of an ontological concept when their attributes are found to be representative of that concept's underlying meaning.

In this paper we seek to capture the underlying semantics of a financial concept, that is, their meaning so as to resolve terminological differences to the same semantic. Accordingly, our problem domain is that of meaning representation and an ontology solution provides the basic mechanism for doing so.

**Calculation Linkbase Ontology Generator**

The first step in our method is the automatic generation of ontologies based on XBRL calculation linkbases. A single ontology is generated for each calculation linkbase per company. Our calculation linkbase ontologies are created by a mapping algorithm which examines `<calculationLink>` elements in order to build a taxonomy, the subsumption hierarchies of which correspond to calculation hierarchies. Of course, this taxonomy is expressed in Web Ontology Language (OWL). `<calculationLink>` elements have as sub-elements `<loc>` and `<calculationArc>` elements. `<loc>` elements are XLink locators which represent links to financial concepts defined in a taxonomy schema document. `<calculationArc>` elements describe relationships between the financial concepts declared in `<loc>` elements. A single `<calculationArc>` relates two financial concepts. `<calculationArc>` elements include information about the nature of the relationships that they encapsulate, such as whether or not the relationship is a summation and the weight assigned to the summation operation. Within the `<calculationArc>` element are two attributes of particular relevance to our purposes here. The attributes are `@xlink:to` and `@xlink:from`. The `@xlink:to` attribute specifies the label (a meaningful name) associated with a financial concept that is used in computing (subsumed by) the financial concept delineated by the label used in the `@xlink:from` attribute. In other words, in our taxonomy, the financial concept corresponding to the value of an `@xlink:to` attribute is always the child of the financial concept corresponding to the value of an `@xlink:from` attribute within the same `<calculationLink>` element. Figure 2 illustrates this concept. The generation of calculation linkbase ontologies is performed automatically without human intervention. All financial concepts are subclasses of the FinanceItem class. The subsumption hierarchy between financial concepts is also observable from Figure 2.
Multi-Ontology Multi-Concept Matrix (M³)

At the heart of our artifact’s novel contribution is its ability to reconcile investor terms with the more technical and precise concepts found in XBRL filings. Such that, for any investor term which corresponds to a financial concept, our method maps the term to the underlying concept, automatically. This problem is concisely represented in a logical expression (formula 2 below) where it is an investor term, xt, is an XBRL term and C is a financial concept.

**Formula 2:** \[ \exists it, xt; it \neq xt, it \equiv C, xt \equiv C \]

We begin by constructing a training set of 2011 XBRL calculation linkbases for 7 publicly traded companies randomly selected from the S&P 100: Apple, Abbott, American Electric Power, Boeing, Baker Hughes, Costco, and IBM. Each of the 7 XBRL calculation linkbases is automatically converted into a calculation linkbase ontology using the automatic ontology generation approach outlined in the previous section. We then train our artifact (to be described in the paragraphs below) against an initial list of 7
investor terms (i.e. it), namely, “Net Income,” “Total Assets,” “Cash from Operating Activities,” “Net Sales,” “Current Assets,” “Current Liabilities,” and “Common Stock.” We build an OWL ontology (FinOnt) with 7 root classes (i.e. subclasses of owl:Thing). Each of the 7 classes corresponds to an investor term. Subclasses of these investor-term-classes are conceptually equivalent XBRL terms. The similarity algorithm we employ is shown below.

Let \( CS \) be a weight assigned to children similarity
Let \( LS \) be a weight assigned to lexical similarity

for \( i = 0 \) to \( n \), where \( n \) is the number of investor defined terms
   Let \( IT_i \) be the \( i \)th investor defined term
   for \( j = 0 \) to \( m \), where \( m \) is the number of linkbase ontologies
      Let \( CO_j \) be the \( j \)th calculation linkbase ontology
      For \( k = 0 \) to \( p \), where \( p \) is the number of financial concepts in \( CO_j \)
         Let \( C_{jk} \) be the \( k \)th financial concept in the \( j \)th calculation ontology
         Calculate Jaccard Similarity \( C_{jk, \text{sim}} \) between \( C_{jk} \) and \( IT_i \)
         if \( C_{jk, \text{sim}} > 0 \) then // if there is at least one common word between \( C_{jk} \) and \( IT_i \)
            Let \( \text{Sim}(C_{jk}, C'_{jk}) := (\text{ChildrenSim} * CS + \text{LexicalSim} * LS) \) between \( C_{jk} \) and \( C'_{jk} \), where \( C_{jk} \neq C'_{jk} \)
            Let \( V_{jk} \) be a vector of values for \( \text{Sim}(C_{jk}, C'_{jk}) \)
         end if
   end for
Let \( M^3 \) be the set of all \( V_{jk} \) //the set of all \( V_{jk} \) constitutes the \( M^3 \) matrix
for each \( V_{jk} \in M^3 \) do
   Let \( \text{Sum}_{jk} \) be the sum of all \( \text{Sim}(C_{jk}, C'_{jk}) \in V_{jk} \)
end for
Select as the XBRL term which most likely corresponds to \( IT_i \), the \( C_{jk} \) that corresponds to \( \text{Max}(\text{Sum}_{jk}) \)

M3 Algorithm

Let \( CV_{jk} \) be a vector of \( C_{jk} \) children (subclasses)
Let \( CV'_{jk} \) be a vector of \( C'_{jk} \) children
intersect\( _{jk} \in C_{jk} \cap C'_{jk} \)
union\( _{jk} \in C_{jk} \cup C'_{jk} \)
ChildrenSim\( _{jk} \in \text{intersect}_{jk} / \text{union}_{jk} \)

Children (Subsumption Hierarchy) Similarity Algorithm of \( M^3 \)

To calculate the similarity between subsumption hierarchies, equal weights of 0.5 are applied for \( CS \) and \( LS \). We also used the Jaccard Similarity Coefficient (Van Rijsbergen 1979) to calculate the lexical similarity between \( C_{jk} \) and \( C'_{jk} \), where the co-occurrence of words in two finance concepts are counted to measure the similarity. The training process results in an XBRL term for each investor term, such that XBRL terms become subclasses of their respective investor terms in FinOnt. We hypothesize, and later show in our evaluation, that these terms are superior candidates for extracting financial concepts from XBRL documents (outside of the training set of documents) that correspond to the initial investor terms. We employ the same algorithm (i.e. Algorithm 1) with an important exception. Instead of beginning with the set of investor terms we begin with their subclasses, XBRL terms which correspond to investor terms.
Instantiation

We instantiated the proposed framework using several semantic web technologies. First, the XBRL calculation linkbases are converted into RDF statements using TopBraid Composer\(^3\). We then used the JENA framework to extract the classes in RDF statements for Calculation Linkbase Ontology Generator. JENA\(^4\) 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 to query on RDF statements. Additionally, we used Protégé\(^5\) in examining the ontologies. Figure 3 shows FinOnt – an ontology that integrating the XBRL filings of seventy companies. The Company class in Figure 3 has seventy instances, each encapsulating the financial facts of all retrieved terms for a firm. The financial performance information of Oracle is shown inside the rectangle in Figure 3.

![Figure 3 - The instantiation of M\(^3\), Results for Oracle Corporation](image)

Evaluation

The purported utility of the M\(^3\) approach lies in the resolution of the semantic heterogeneity that is known to exist across XBRL filings. Logically, and in keeping with the traditions of information systems design science research (Hevner et al. 2004), we evaluate the extent to which our design artifact resolves semantic heterogeneity in XBRL. To do so, we use our expertise and knowledge of financial accounting to manually extract from SEC documents values for 70 firms corresponding to the investor terms we identified above. Of the 70 firms, we use a randomly selected hold-out sampling method where our model training set consists of 10% or 7 of the 70 firms in the full sample (see tables A1 and A2 in the Appendix). An appropriate threshold for a hold-out sample is complicated choice and is indeed very domain specific. To validate our use of a 10% hold-out, we examined various thresholds. We found that there was no benefit to a training set greater than 10%.

The selection of these firms was based on Standard & Poor’s 100 (S&P 100) index. The S&P 100 is a list of 100 blue-chip stocks which must be listed on any of the following exchanges: NYSE, NYSE Arca, NYSE MKT, NASDAQ Global Select Market, NASDAQ Select Market and NASDAQ Capital Market. Further, the S&P 100 only includes firms which file 10-K annual reports with the SEC and are primarily domiciled within the United States financial reporting jurisdiction (our method uses 10-K annual reports only). S&P

\(^{3}\)http://www.topquadrant.com/
\(^{4}\)http://jena.apache.org/
\(^{5}\)http://protege.stanford.edu/
100 stocks are selected from the broader S&P 500 index stocks. The focus of S&P 100 is on firms with the highest market values where stocks are selected with an emphasis on industry balance (S&P Dow Jones Indices 2015). As such, using the S&P 100 guarantees that our artifact is evaluated on: (1) a sufficiently broad spectrum of industries, (2) stocks which are of much interest to investors and (3) a set of well known, well established companies. Each of the three justifications is important for application domain because (1) our artifact, in order for maximum generality, should be industry agnostic, (2) it should reflect those companies in which investors are interested, (3) and should be shown to perform well on a set of market leaders. We further refine the S&P 100 by excluding banking, insurance, and automotive stocks. We use XBRL filings for the 2011 reporting period. As such, some S&P 100 companies had not, at the time successfully adopted the standard. These companies were excluded. We use the SEC’s Electronic Data Gathering and Retrieval system (EDGAR) to extract these values. We use the SEC's EDGAR to extract these values. We only use values for annual reports filed in the year 2011. The values sourced from EDGAR serve as our benchmark.

The extent to which the figures we extract using M^3 supplied terms agree in value to those extracted manually provides evidence supporting the assertions in this paper. Specifically, that the proposed method accurately associates financial terms from disparate XBRL filings to the financial concepts embodied in investor terms. Further, we compare the ontologies generated (i.e. FinOnt) by a process of lexical similarity matching alone with the ontology generated by our procedure which introduces a measure of similarity between subsumption hierarchies. The comparison is conducted on the basis of correctness where we verify that the XBRL terms extracted from our calculation linkbase ontologies are conceptually the same as the investor defined terms to which they correspond.

Table 1 contains the result of the M^3 performance evaluation on the testing data set, which consists of 63 firms (90% of the entire data set). Performance across the various investor terms is accurate. With the exception of Net Income and Common Stock, the proposed approach performs with over 90% accuracy. We find this performance to be quite adequate as alternative commercial databases tend to be fraught with errors (Du and Zhou 2012). The results in Table 1 demonstrate that the approach presented here is well suited to its purpose. The primary utility of the approach as previously stated is resolution of semantic heterogeneity across XBRL filings. This resolution is shown to have been successful by the accurate retrieval of values associated with investor terms. Where the semantics expressed in investor terms and across XBRL filings are not resolved, such figures are impossible to generate.

<table>
<thead>
<tr>
<th>Investor Term</th>
<th>Count</th>
<th># Correct</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Net Income</td>
<td>63</td>
<td>56</td>
<td>88.89%</td>
</tr>
<tr>
<td>Total Assets</td>
<td>63</td>
<td>62</td>
<td>98.41%</td>
</tr>
<tr>
<td>Cash From Operating Activities</td>
<td>63</td>
<td>61</td>
<td>96.83%</td>
</tr>
<tr>
<td>Net Sales</td>
<td>63</td>
<td>57</td>
<td>90.48%</td>
</tr>
<tr>
<td>Current Assets</td>
<td>63</td>
<td>63</td>
<td>100.00%</td>
</tr>
<tr>
<td>Current Liabilities</td>
<td>62</td>
<td>61</td>
<td>98.39%</td>
</tr>
<tr>
<td>Common Stock</td>
<td>63</td>
<td>53</td>
<td>84.13%</td>
</tr>
</tbody>
</table>

Table 1 – M^3 Accuracy

1 One of the companies in the testing set did not report current assets

However, table 1 alone does not sufficiently test the propositions inherent in our approach. There exists a need to demonstrate the utility of the hierarchical approach with respect to the resolution of semantic heterogeneity. We perform a simple yet enlightening controlled experiment to test the proposition encapsulated in Formula 1. The essence of this proposition is that the calculation hierarchy associated with a given financial concept is sufficient to capture the meaning or semantics of that concept. In other words, a representation of the subsumption relationships regarding a concept in a hierarchy can be used to discriminate between concepts in different instantiations of such a hierarchy within the same domain. The domain of interest here is the calculation hierarchies used in XBRL filings of 10-K annual reports in the United States reporting jurisdiction. We can therefore test the proposition that our subsumption hierarchy similarity approach discriminates between financial concepts by examining its end product (this end product is the upper level ontology we termed FinOnt) when hierarchical similarity is used, and when...
it is not used. As such, the use of hierarchy similarity becomes the researcher controlled experimental manipulation. All other potentially confounding variables are held constant.

We conducted the test described above and compare the two versions of our upper level FinOnt ontology, as shown in Figure 4. Version 1 of the ontology is generated using lexical similarity and version 2 of the ontology is generated using subsumption hierarchy as well lexical similarity information. Both versions are quite similar with some very important exceptions. First, for the investor term Net Income, version 1 of FinOnt asserts that the XBRL concept OperatingIncomeLoss is equivalent. Second, for the investor term Total Liabilities, version 1 of FinOnt asserts that the XBRL term, DerivativeLiabilities, is equivalent. OperatingIncomeLoss and DerivativeLiabilities are incorrectly classified in version 1. Conceptually, OperatingIncomeLoss appears prior to NetIncome on an income statement, as the final figure NetIncome figure includes items which are not considered to be costs of doing business (operating costs). OperatingIncomeLoss, therefore, refers to a financial concept that differs substantially from the financial concept embodied in the investor term NetIncome. DerivativeLiabilities are but a subset of the TotalLiabilities figure and therefore refers to a different financial concept than the one implicit in the investor term Total Liabilities. While we observe these errors in version 1, we do not observe them in version 2. Due to the fact that the only difference between the two versions is the use of subsumption hierarchy similarity, we can attribute the greater conceptual correctness of version 2 to the use of hierarchical information following the logic of experimental design.

![Figure 4 – Two Versions of FinOnt](image)

**Theoretical Contribution**

Earlier in this paper, under the section titled “design theory” we staked our position with respect to the theoretical contribution of a design in the tradition of information systems (IS) design science. We drew upon a comprehensive, eight-part framework for articulating an information systems design theory (ISDT) (Gregor and Jones 2007). The various components are explained in brief in the paragraphs to follow and our ISDT is expressed in those terms.

The first component of the Gregor and Jones view of ISDT is the purpose and scope of the design. Design is at its essence, “goal adaption” (Simon 1969). Accordingly, any ISDT must be goal oriented, and must be clearly and appropriately scoped. The purpose of this work is to develop a method for the resolution of semantic heterogeneity across XBRL filings scoped in the US reporting jurisdiction.
Constructs are the basic units within a theory’s scope. Constructs encompass the entities of interest in a theory. As such, they must be clearly defined and applied consistently with no ambiguity as to what they represent. In this way, the design theory can meet its most basic purpose of being built upon by other researchers. The constructs used in this work include: Ontology; financial statement; semantic heterogeneity; semantic similarity; lexical similarity; hierarchical information; financial item; financial concept; investor; investor vocabulary; XBRL calculation linkbases and instances; subsumption hierarchy;

The principles of form and function that underlie a design may be seen as the proposed organization of its constructs towards the defined goal, and subject to the defined scope of the design. Jones and Gregor call this the “blueprint” of the design, principles which define its architecture and construction. In this paper we present a process whereby XBRL calculation linkbases are represented as an ontology hierarchy to facilitate the extraction of subsumption relationships. The approach herein is focused on capturing the relative hierarchical position of a given financial item in a financial statement. Hierarchical and lexical similarity measures are used to reconcile financial items to financial concepts as understood by an individual investor.

Artifact mutability brings to the fore the fact that IS design artifacts are always involved in change, where that change may be from within or without the artifact. It is here that the generalizability of a design onto different situations may be expressed. Our approach is sufficiently general so as to be applicable to many situations which can be reduced to problems of semantic heterogeneity and data integration. Prerequisites for the application of this ISDT to a problem situation are that the data to be merged must be amenable to hierarchical representation, that data sources are hierarchically represented under a similar framework (i.e. a subsumption relationship means the same thing across data sources) and that there exist users of these data sources with a vocabulary that differs from that used in data representation.

Testable propositions are essential in the development of an ISDT as they form the basis of evaluations with respect to the artifact’s utility against its stated purpose(s). An ISDT, or any theory for that matter (see Lee et al. 2014), should be falsifiable. Falsifiability should be with respect to the formative propositions of the theory. A universal, testable proposition that lies at the core of this ISDT is that the hierarchical arrangement of financial concepts in a calculation hierarchy is informative with respect to the semantics of any financial concept in the hierarchy. The evaluation section of this paper tests this proposition by comparing the mapping of XBRL terms to investor concepts produced by a lexical measure of similarity and mapping produced by both a lexical similarity measure and a subsumption hierarchy similarity measure.

Justificatory knowledge refers to the kernel theories employed in the development of a design. In order to ascertain the value of a concept, technology, material et cetera to a design, it is necessary to understand the behavior that concept, technology or material. Justificatory knowledge represents such understanding of key enabling components of the design taken from existing work. This process to overcome semantic heterogeneity draws from the general form of extant approaches to information retrieval. However, the primary kernel theory in use is made up of the ideas sketched out in the work done on conceptualizing a semantic web (Berners-Lee et al. 2001). We apply these ideas by expressing XBRL calculation linkbases in an ontology language such that the information therein is “given well defined meaning, better enabling computers and people to work in cooperation” (Berners-Lee et al. 2001). We also do this with FinOnt. The use of ontologies to structure knowledge lies at the heart of the techniques presented in this paper.

Gregor and Jones also argue for the inclusion in ISDT of certain principles of implementation. Such principles link the design with its purpose by including details regarding the manner(s) in which the design’s target audience may employ the artifact. Given a set of XBRL filings and a corresponding DTS, a semantic representation can be made of the hierarchical information contained within the linkbase taxonomy. Using this information, a user of the proposed approach (an investor) can resolve their own, idiosyncratic, set of financial terms to universal financial concepts across a corpus of XBRL filings.

Finally, a theorized design must include an expository instantiation. The purpose of this, as Gregor and Jones note is theory exposition or representation, similar to the operationalization of a behavioral theory in the form of a statistical model to be fitted with data. Such instantiation makes possible the testing of design propositions. This ISDT has been instantiated on a corpus of 70 XBRL filings divided into a holdout sample of 7 training and 63 testing filings.
## Discussion and Conclusion

We return to the question that motivated this work: “is the hierarchical information contained within XBRL calculation linkbases able to discriminate between financial concepts in an XBRL filing to the extent that it can constitute an adequate solution to the problem of semantic heterogeneity?” In our evaluation of our method, we demonstrate that the hierarchical information represented in XBRL calculation linkbases is indeed efficacious towards our ends. By translating the XBRL calculation linkbase into an OWL ontology, we are able to represent calculation hierarchies as ontological taxonomies. These taxonomies are then consumed by our M³ algorithm which leverages the hierarchical information therein. By basing our approach on ontologies, we are able to incrementally encode the relationships between investor vocabulary and XBRL concepts. Our design provides strong evidence consistent with our belief in the potential of calculation linkbases as mechanisms to overcome the issue of semantic heterogeneity across XBRL filings. Going forward, we shall exploit more of the semantics encoded in calculation linkbases to increase the accuracy of the M³ algorithm and examine the applicability of this approach across filing jurisdictions.

This research is highly applicable to practice. Amongst the stated target audience of the XBRL standard are investors who seek to make inter-firm comparisons with respect to performance. Of course, financial statement data is available from third party sources, but these sources are known to be inaccurate, fraught with errors and misinterpretations (Du and Zhou 2012). Directly consuming financial statement data from original company filings allows for the most accurate possible data on firm performance. The approach presented here provides a means through which this can be done. Incremental improvements to the process will also increase the accuracy of data retrieval. Further, our method allows for an investor specific vocabulary to be resolved against financial concepts, capability that is, to our knowledge, not available in third party data sets.

Ongoing work in this vein will be centered on the development of an intuitive user interface where end-users may generate data sets for any number of publicly listed firms in the US jurisdiction, using their own vocabulary to describe data points. Another stream of work that may stem from our approach is the development of an application programming interface (API) against which applications may make calls. Again, applications would be able to request data on particular financial items across US filings of annual reports using their own, idiosyncratic vocabulary. Finally, we plan to make our approach more flexible so as to accommodate specialized reporting situations such as those found in the annual reports of banks, utilities and other uniquely regulated industries.

### Appendices

<table>
<thead>
<tr>
<th>Company Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>3M Company</td>
</tr>
<tr>
<td>Accenture plc</td>
</tr>
<tr>
<td>Altria Group</td>
</tr>
<tr>
<td>Amazon.com</td>
</tr>
<tr>
<td>Amgen Inc.</td>
</tr>
<tr>
<td>Anadarko Petroleum Corporation</td>
</tr>
<tr>
<td>Apache Corp.</td>
</tr>
<tr>
<td>AT&amp;T Inc.</td>
</tr>
<tr>
<td>Baxter International Inc</td>
</tr>
<tr>
<td>Bristol-Myers Squibb</td>
</tr>
<tr>
<td>Texas Instruments</td>
</tr>
<tr>
<td>United Parcel Service Inc</td>
</tr>
<tr>
<td>United Technologies Corp</td>
</tr>
<tr>
<td>Verizon Communications Inc</td>
</tr>
<tr>
<td>Company</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>Home Depot</td>
</tr>
<tr>
<td>Intel Corporation</td>
</tr>
<tr>
<td>Johnson &amp; Johnson Inc</td>
</tr>
<tr>
<td>Lockheed-Martin</td>
</tr>
<tr>
<td>McDonald’s Corp</td>
</tr>
<tr>
<td>Raytheon Co (NEW)</td>
</tr>
<tr>
<td>Target Corp.</td>
</tr>
</tbody>
</table>

Table A1 – List of Companies in Testing Set

<table>
<thead>
<tr>
<th>Company Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple</td>
</tr>
<tr>
<td>Boeing</td>
</tr>
<tr>
<td>International Business Machines</td>
</tr>
<tr>
<td>Abbott</td>
</tr>
<tr>
<td>Baker Hughes</td>
</tr>
<tr>
<td>American Electric Power</td>
</tr>
<tr>
<td>Costco</td>
</tr>
</tbody>
</table>

Table A2 – List of Companies in Training Set

References


Wunner, T., Buitelaar, P., and O’Riain, S. 2010. “Semantic, terminological and linguistic interpretation of xbrl,” *Reuse and Adaptation of Ontologies and Terminologies Workshop at 17th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*.