Journal of the Association for Information Systems (2020) **21**(4), 971-1000
doi: 10.17705/1jais.00626

RESEARCH ARTICLE

ISSN 1536-9323

X-IM Framework to Overcome Semantic Heterogeneity Across XBRL Filings

Dapeng Liu[1], Ugochukwu Etudo[2], Victoria Yoon[3]

[1] University of New South Wales, Australia, dapeng.liu@unsw.edu.au
[2] University of Connecticut, USA, ugochukwu.etudo@uconn.edu
[3] Virginia Commonwealth University, USA, vyyoon@vcu.edu

Abstract

Semantic heterogeneity in XBRL precludes the full automation of the business reporting pipeline, a key motivation for the SEC's XBRL mandate. To mitigate this problem, several approaches leveraging Semantic Web technologies have emerged. While some approaches are promising, their mapping accuracy in resolving semantic heterogeneity must be improved to realize the promised benefits of XBRL. Considering this limitation and following the design science research methodology (DSRM), we develop a novel framework, XBRL indexing-based mapping (X-IM), which takes advantage of the representational model of representation theory to map heterogeneous XBRL elements across diverse XBRL filings. The application of representation theory to the design process informs the use of XBRL label linkbases as a repository of regularities constitutive of the relationships between financial item names and the concepts they describe along a set of equivalent financial terms of interest to investors. The instantiated design artifact is thoroughly evaluated using standard information retrieval metrics. Our experiments show that X-IM significantly outperforms existing methods.

Keywords: XBRL Element, Ontology Mapping, Representation Theory, Theory of Ontological Clarity

Roger Chiang was the accepting senior editor. This research article was submitted on April 25, 2018 and underwent three revisions.

1 Introduction

In 2009, following a voluntary filing period, the SEC finalized its rule on interactive data to improve financial reporting, mandating the use of eXtensible business reporting language (XBRL) by all public companies in the United States for annual, quarterly and other reports (SEC, 2009). In preparation for and as part of the evaluation of this mandate, the SEC, XBRL US (the body contracted by the SEC to implement XBRL in the US jurisdiction) and the Financial Accounting Standards Board (FASB) developed the list of XBRL tags (i.e., XBRL elements) that would be used to classify and define the semantics of financial information (primarily financial statement line items) in accordance with SEC regulations and US generally accepted accounting principles (US GAAP) (SEC, 2009). This annually updated "list" of XBRL tags is known as the US GAAP Financial Reporting Taxonomy (UGT). Companies are expected to draw from the UGT when creating financial reports in XBRL. However, the SEC notes:

> *Occasionally, because filers have considerable flexibility in how financial information is reported under US reporting standards, it is possible that a company may wish to use a non-standard financial statement line item that is not included in*

> *the standard list of tags. In this situation, a company will create a company-specific element, called an extension.* (SEC, 2009)

A recent study examined 121 XBRL format financial statements developed based on the 2009 version of the UGT (Zhu & Wu, 2011). Using these statements, the study computes an interoperability metric for each of 7,260 pairs of XBRL tags extracted from the documents. The authors note that "a set of data instances is [interoperable] if the instances use the same set of data elements defined in a data standard" such that interoperability between two XBRL filings measures the degree of overlap in their use of UGT tags. The study reports that the average interoperability between two XBRL filings for the period investigated is 29.52% and falls to 17.35% when three XBRL filings are compared. The interoperability problem between XBRL filings suggests the presence of semantic heterogeneity. Semantic heterogeneity exists in the presence of "differences in the meaning and use of data that make it difficult to identify the various relationships between similar or related objects in different components" (Hammer & McLeod, 1993). The "components" in this context are individual XBRL filings. The difficulty in identifying relationships of equivalence between XBRL tags used in different filings can thus be characterized as semantic heterogeneity. The issue of semantic heterogeneity across XBRL filings in the US jurisdiction is well documented (Chowdhuri et al., 2014; Etudo & Yoon, 2015; Etudo, Yoon, & Liu, 2017; Zhu & Wu, 2014). Its effects are palpable. For instance, in 2013, CFO magazine published an article reporting on a comment letter sent by Rep. Darrell Issa to the SEC chair remarking that the SEC does not make use of the XBRL filings it collects. It instead reviews filings manually and purchases licenses for commercial databases such as Yahoo! Finance and Compustat (Hoffelder, 2013). More recently, in 2017, as the SEC proposed a new rule with respect to Inline XBRL, companies have commented that the existing standard is still far too problematic to justify additional rule making. Although 74% of XBRL financial statements contain custom tags (extensions), tagging remains a very error-prone process and downstream consumers of financial statements do not rely on the standard to collect financial data (Ernst & Young LLP, 2017). We note that the presence of semantic heterogeneity across XBRL filings precludes its automated consumption, especially given the need to compare companies' performance data along a set of financial concepts and measures.

The automated resolution of semantic heterogeneity across XBRL filings in the US jurisdiction is thus the focus of this research. Several studies have emerged proposing a diverse range of solutions. Unsurprisingly, many efforts rely on ontology mapping, a common approach to resolving semantic heterogeneity. Several researchers have proposed methods that "ontologize" XBRL by representing the semantics of financial reporting concepts unambiguously in an ontology language (Bao et al., 2010; Declerck & Krieger, 2006; Raggett, 2009; Recio-García, Quijano, & Díaz-Agudo, 2013; Spies, 2010). Such approaches often fall short because they focus on translating individual filings into description logics and formal semantics. The resulting representations still retain the heterogeneous tags and no mapping strategy is proposed to resolve this heterogeneity across filings. Some approaches do, however, provide mapping algorithms to link such ontologies (Chowdhuri et al., 2014; Etudo & Yoon, 2015) with so-called upper-level ontologies that define financial reporting concepts independent of any given financial report. While these approaches perform relatively well, there is much room for improvement where mapping accuracy is concerned. In addition to the suboptimal accuracy of these algorithms, their designs often lack explicit theoretical insight and do not contribute to generalizable knowledge in terms of the semantic interoperability of data standards.

Since XBRL cannot realize its intended benefits in the face of semantic heterogeneity, the research issues highlighted in the previous paragraph motivate the following research question: *How may a fully automatic algorithm be designed to accurately map XBRL tags to financial concepts defined in an upper-level ontology?* We answer this question by providing an indexing-based classifier that relies on a theoretically informed feature space for its classification task. The proposed approach ontologizes XBRL filings and abstracts financial concepts into an upper-level ontology. The upper-level ontology stores a collection of equivalence relationships between the abstracted financial concepts and XBRL tags for financial line items. We show how the theory of ontological clarity (Wand & Weber, 1995), also known as the representation model of representation theory, at least partially explains why US XBRL financial statements do not interoperate even in the face of a unifying taxonomy or grammar (i.e., the US GAAP taxonomy). Our work contributes to representation theory by showing how correcting ontological deficiencies in a grammar lead to more interoperable scripts generated from that grammar, which also extends the theory into the space of semantic interoperability.

We follow a design science research methodology (Peffers et al., 2007) that structures the resolution of the above question through articulating a process that moves through problem identification, specification of objectives, exposition of a design strategy, demonstration and evaluation, and, finally, discussion of implications (communication). XBRL in the US reporting jurisdiction cannot realize its intended

benefits in its current form (indeed it is disliked by filers and disregarded by downstream consumers of financial information). The information systems literature has recognized this failure and has offered solutions enabling the automated downstream consumption of financial reports published in XBRL format for real-world financial decision-making. Our objective is to build on these solutions by designing a precise and automated technique for resolving semantic heterogeneity in these filings. Our design represents a novel classification scheme that defines relationships of equivalence between terminologically disparate but semantically equivalent XBRL tags. We contribute to the theoretical understanding of semantic interoperability in data standards by linking interoperability with ontological clarity and show how a design that directly addresses the ontological clarity of XBRL will also improve its interoperability. We evaluate our efforts using standard classification evaluation metrics and, by adopting an "experimentation, observation and performance testing" philosophy (Nunamaker et al., 1990, p. 89), demonstrate a statistically significant and meaningful, improvement over previous attempts. At the end of this paper, we present the implications of our work to theory, research, and domain practice.

2 The XBRL Framework

Corporate regulators around the world, including the United States SEC, have adopted the XML-based XBRL framework for tagged financial data. The framework is composed of XBRL taxonomies and XBRL instance documents. Providing a collection of "tags" for financial concepts in a financial statement, XBRL taxonomies consist of an XML schema (or taxonomy schema) and a set of associated linkbases. A taxonomy schema describes and classifies the XBRL elements (tags) such that each XBRL element is uniquely defined by an XML element's syntax declaration. For example, *us-gaap_Assets* and *us-gaap_AccountsPayableCurrent* are XBRL elements designed to tag the financial concepts of *total assets* and *accounts payable*, respectively. Extended links in an XBRL taxonomy are organized into linkbases and provide multidirectional links between two or more XML snippets. Notice that the taxonomy document provided for an individual XBRL filing includes a subset of the US GAAP taxonomy as well as extension elements created by the filer. There are five types of extended links in XBRL taxonomies: calculation, definition, presentation, reference, and label links. A calculation linkbase defines a set of calculation relationships between XBRL elements, and a definition linkbase asserts relationships such as *general-special* or *requires-element* between pairs of XBRL elements. A presentation linkbase defines how XBRL elements are rendered for human viewing with respect to other XBRL elements. A reference linkbase describes relationships between XBRL elements and references to authoritative statements in the published document that give meaning to the elements. A label linkbase amalgamates human-readable text (label terms) with XBRL elements using special identifiers (i.e., @xlink:label) (Luna-Reyes et al., 2005). While XBRL taxonomies provide metadata regarding XBRL elements, XBRL instance documents assert *facts* (quantities) about those elements (e.g., net income = $55,000,000). In Table 1, we define some important XBRL related terms. Further details can be found in Chowdhuri et al. (2014) and Engel et al. (2013).

3 Literature Review

Our review of extant work consists of two parts. The first part reviews semantic integration in the literature, and the second part reviews prior work on XBRL interoperability. We highlight the novelty of our design artifact within the semantic interoperability space, in general, and the XBRL interoperability space, in particular, by exploiting gaps in the literature.

Table 1. Terms, Synonyms and Definitions

Term	Synonym(s)	Definition
XBRL element	XBRL Tag; US GAAP taxonomy element; US GAAP taxonomy tag	An XML element defined in a standard XBRL taxonomy to be used in the annotation (tagging) of XBRL-based financial reports/statements by any firms; this element is defined in the UGT.
XBRL extension element	Extension element	A custom XML element defined in a certain firm's XBRL taxonomy that is used in the annotation (tagging) of their XBRL-based financial reports/statements; at the time of their use in a filing, this element is not defined in the UGT.
Financial concept	Investor term	Widely recognized financial measure, relevant to statutorily mandated financial disclosures, and instantiated with a usually numeric value.

3.1 Semantic Interoperability

Semantic interoperability is primarily concerned with discovering ways to assert *relationships of equivalence* between data points from disparate sources (Heiler, 1995). Semantic interoperability is critical to applications and use cases that "need to query across [multiple] autonomous and heterogeneous data sources" (Halevy, Ordille, & Rajaraman, 2006, p. 9). The problem of XBRL interoperability is a special case of the problem of semantic interoperability. Multi-stakeholder efforts to provide a unified standard through which information from heterogeneous sources can be disseminated naturally require mappings from these disparate information sources onto a unified taxonomy or shared upper-level ontology. Since the introduction of the Semantic Web (Berners-Lee, Hendler, & Lassila, 2001), there has been a growing need for the design of systems that provide semantic interoperability. In response to this need, a wide range of approaches have been proposed to discover mappings between various applications. Consistent themes across the approaches are the presence of a meta-database or ontology that captures discovered mappings and a matching algorithm that exploits the available information in order to discover those mappings.

The semantic interoperability literature can be categorized into a bipartite framework: (1) data model development, and (2) semi/fully automatic semantic data integration. The approaches to data model development have primarily focused on detailing data standards in varied domains, such as e-health (Ure et al., 2009), emergency response management (Chen et al., 2008), internet of things (IoT) (Alaya et al., 2015), manufacturing systems ontology (Lin, Harding, & Shahbaz, 2004), pharmaceutical drug discovery (Williams et al., 2012), web services interoperability (Nagarajan, Verma, Sheth, Miller, & Lathem, 2006), and too many others to list here. The data model development alone does not provide interoperability between heterogeneous data sources. For instance, the US GAAP taxonomy is a data model for financial statement interoperability. However, financial statements created using the UGT do not automatically interoperate. Given a data model intended to promote interoperability between a set of heterogeneous systems, such as the IoT-O ontology proposed in Alaya et al. (2015), previously unseen IoT devices plugged into a network will each expose a different set of attributes and methods that must be mapped to IoT-O constructs. This mapping is not addressed in the data model development literature.

On the other hand, the semi/fully automatic semantic data integration literature directly addresses the mapping problem. In a recent survey of this literature, Thiéblin et al. (2018) use a framework that defines a bi-axial characterization of extant approaches—*outputs* and *process*.[1] Outputs are further subdivided to account for the nature of the output mappings that an approach provides. A semantic integration approach can output its mappings as *logical relations* (the approach maps two constructs by asserting a logical correspondence [mapping] between them—e.g., finding necessary and sufficient conditions for equivalence at the schema level), *transformation functions* (applicable only in certain domains where semantic integration involves identifying a necessary calculation), or *blocks* (an instance-level mapping output that asserts relationships of equivalence between groups of instances in the to-be-merged data sources). With respect to the process (i.e., the *how*), Thiéblin et al. (2018) identify the five categories used in the literature to generate equivalences across data sources: (1) Atomic pattern-based approaches work best with expressive data sources (such as OWL 2 DL), as they define exact rules based on the semantics of the to-be-merged data sources. (2) Composite pattern-based approaches find relations of equivalence by iteratively constructing compound matching rules. For instance, Parundekar, Knoblock, and Ambite (2012) match attribute pairs (an attribute pair is a relation with two arguments) in one ontology with attribute pairs in another ontology by iteratively compiling a union of acceptable values for the arguments using instance-level data. (3) Path-based approaches begin with simple mappings between to-be-merged data sources that are enriched into more complex mappings by exhaustively searching along the paths generated by the simple mappings. For instance, some studies obtain simple mappings by mining query logs for the to-be-integrated data sources at the schema level before discovering complex mappings at the instance level (Dou, Qin, & Lependu, 2010; Qin, Dou, & LePendu, 2007). (4) Tree-based approaches (e.g., Etudo et al., 2017) focus on the structural similarity between two to-be-merged data sources. This is distinct from the use of tree-based algorithms for classification. Tree-based approaches are the least common in the literature. Finally, (5) no-structure-based approaches do not depend on any of the above structures to discover correspondences. For example, the work by Hu et al., (2012) uses inductive logic programming to identify complex alignments.

An important distinction between our work and previous studies in the semantic integration literature concerned with automated data integration is that prior

[1] Also considered in Thiéblin et al., 2018 are visualization approaches, but these are not relevant to our work.

solutions do not address the unique problems posed by the XBRL context. First, these approaches tend to be general, that is, they are not standard specific (they provide good foundations for approaches such as ours) but may require significant refinement and modification. Second, these solutions attempt data integration in contexts exclusive of data standards that are used to create scripts or instances that do not interoperate. The existence of a standard rescopes the semantic interoperability problem and brings to the fore a different set of signifiers/semantics/features/models required to correct deficiencies that impair interoperability in both the underlying standard (taxonomy) and the instances generated from it. Third, these approaches do not contribute to general theories. We present a novel linkage between the representation model of representation theory and the semantic interoperability that offers four broad propositions (two of which we explicitly test). Each proposition argues that one of the four possible ontological deficiencies of a grammar will lead to scripts generated from that grammar that do not interoperate. We believe that this theoretical formulation is sufficiently general to apply to contexts besides XBRL.

3.2 XBRL Interoperability

We organize extant design science publications germane to the interoperability problem in XBRL using the information systems design theory (ISDT) framework (Gregor & Jones, 2007). We focus on the five components of the framework: purpose and scope, constructs, principles of form and function, justificatory knowledge, and principles of implementation. Our analysis maps each of these components to a dimension useful for characterizing approaches to XBRL interoperability. Purpose and scope capture the completeness of a study's approach toward an interoperability solution. Constructs enumerate the kernel-theoretical components deployed in the solution-specific IT artifacts or subartifacts. Principles of form and function describe how these constructs are mobilized toward the purpose and scope. Justificatory knowledge identifies the discipline-specific knowledge area used and principles of implementation relate to the instantiation and evaluation of the various artifacts. We review related work along these five components and present them in Table 2.

With respect to principles of implementation, we are most concerned with automaticity. It stands to reason that the ideal outcome in the implementation of any IT design is a fully automatic artifact, such that no substantive human intervention is required in its operation. Of the available approaches to interoperability in the literature, implementation tends to either be absent (not discussed), manual or semiautomatic. There are some exceptions where fully automatic approaches have been successfully evaluated (Etudo & Yoon, 2015; Etudo et al., 2017), both of which leverage related methods. An approach proposed by Yaghoobirafi and Nazemi (2019) is fully automatic and well evaluated; however, it is incapable of mapping more than two XBRL instance documents simultaneously. In addition, it is based on the IFRS taxonomy and not on the UGT. The literature thus lacks a healthy variety of fully automatic approaches to XBRL interoperability across multiple instance documents. As we show in the evaluation of our artifact, there is significant room for improvement over the state of the art in this space.

Semantic Web technologies are the dominant justificatory knowledge source in XBRL interoperability research. This paper relies on similar justificatory knowledge. Ontology modeling lends itself naturally to this problem space, as ontologies have traditionally been used to define explicit relationships of equivalence between disparately represented but semantically identical concepts. There is a robust literature on ontology integration (Wache et al., 2001) that has provided initial motivation for researchers seeking to define solutions to XBRL interoperability. Of the papers in this review, the largest share of justificatory knowledge concerns ontology modeling (Bao et al., 2010; Declerck & Krieger, 2006; Livieri, Zappatore, & Bochicchio, 2014; Luna-Reyes et al., 2005; O'Riain, Curry, & Harth, 2012; Radzimski et al., 2014; Spies, 2010). Scholars have also attempted to use justificatory knowledge from natural language processing and information retrieval to define XBRL interoperability artifacts and have combined these approaches in practice with ontology modeling to create fully automatic implementations: e.g., (Etudo et al., 2017).

Table 2 shows that the state of the art lies with fully automatic methods employing a mix of heuristics and machine learning to decipher relationships of equivalence between terminologically heterogeneous but semantically equivalent XBRL elements contained in the calculation linkbases across multiple XBRL instance documents (Etudo & Yoon, 2015; Etudo et al., 2017). However, these methods do not consider the natural language aspects of XBRL filings intended for human presentation and consumption, leaving much unleveraged information. Further, the precision and recall of these methods leave significant room for improvement. To fill this gap, we propose, instantiate, and evaluate a design artifact, X-IM that leverages human-readable label terms and structural (designative) features of US 10-K XBRL filings to map UGT and extension elements to an investor's ontology, a taxonomy of financial concepts commonly used in investment decision-making.

Table 2. Summary of Related Work

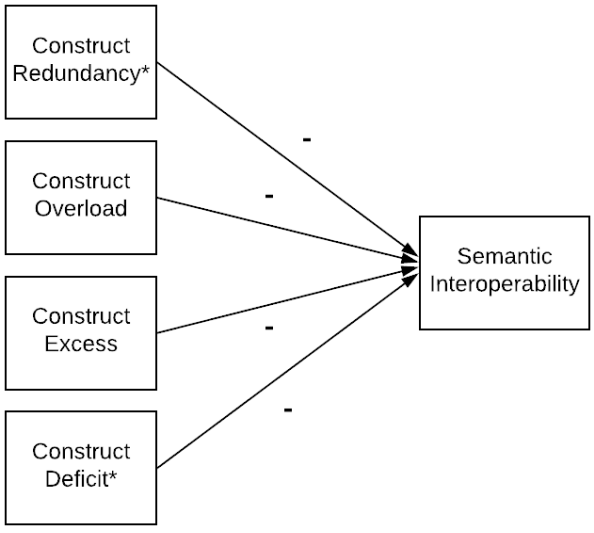
Paper	Purpose and scope	Constructs	Principles of form and function	Justificatory knowledge	Principles of implementation
Bao et al. (2010)	One-to-one mapping of XBRL elements	OWL 2 DL, XBRL	Creating shared ontology for XBRL specification	Ontology modeling	Unclear
Radzimski et al. (2014)	Mapping of XBRL elements to well-defined concepts and linked open data	SPARQL, RDF, Sesame, LOD, XBRL, Silk	Semantic representation of XBRL, links to LOD	Ontology modeling	Unclear
Wunner, Buitelaar, & O'Riain (2010)	Directly addresses XBRL interoperability	Part-of-speech tagging (POS), NLP, RDFS	Heuristics and machine learning	IR and NLP	Semiautomatic
Chowdhuri et al. (2014)	Directly addresses XBRL interoperability	RDF, XBRL, ReDeFer, SWRL, SPARQL	Heuristics and machine learning	IR and lexical processing	Semiautomatic
Zhu & Madnick (2007)	Directly addresses XBRL interoperability	Context interchange framework (COIN), XBRL	Heuristics and machine learning	IR and NLP	Semiautomatic
Declerck & Krieger (2006)	One-to-one mapping of XBRL elements	XBRL, PDF, text mining, OWL, XML, description logic (DL), RDF/RDFS	Creating shared ontology for XBRL specification	Ontology modeling	Manual
García & Gil (2009)	Mapping of XBRL elements to well-defined concepts and linked open data	RDF, XML semantics reuse methodology, OWL, ontology, Semantic Web, WoD	Creating shared ontology for XBRL specification	Ontology modeling	Semiautomatic
Livieri et al. (2014)	One-to-one mapping of XBRL elements	(KPIs), ontology, XML, basic competency questions (BCQs), complex competency questions (CCQs), XBRL, OWL, W3C time ontology	Creating shared ontology for XBRL specification	Ontology modeling	Unclear, likely manual
O'Riain et al. (2012)	Mapping of XBRL elements to well-defined concepts and linked open data		Semantic representation of XBRL	Ontology modeling	Unclear, likely manual
Debreceny et al. (2011)	Directly addresses XBRL interoperability		Heuristic-based approach	Practitioner-in-use	Manual
Spies (2010)	One-to-one mapping of XBRL elements	OWL, XBRL, UML, common warehouse metamodel (CWM), ontology definition metamodel (ODM)	Creating shared ontology for XBRL specification	Ontology modeling	Unclear
Etudo & Yoon (2015)	Directly addresses XBRL interoperability	RDF, XBRL, SPARQL	Heuristics and machine learning	IR and lexical processing	Automatic
Etudo et al. (2017)	Directly addresses XBRL interoperability	RDF, XBRL, SPARQL	Heuristics and machine learning	Channel theory	Automatic
Yaghoobirafi and Nazemi (2019)	Directly addresses XBRL interoperability	Bipartite graph	Ant colony optimization	Collective optimization	Automatic

4 Theoretical Background

We concur with the literature's characterization of the lack of interoperability across XBRL filings in the US jurisdiction as a *semantic interoperability* problem. Any large scale, distributed information system must be able to seamlessly exchange data between its components. This exchange must be based on agreed-upon protocols, grammars, taxonomies, and so forth (Heiler, 1995). As we mentioned in the literature review, semantic interoperability in such distributed sociotechnical systems requires that all parties have a shared understanding of the *meaning* of the data that flow between the parties (Heiler, 1995).

Unfortunately, the semantic integration literature does not provide a generalizable framework for understanding the basis of meaning-making in distributed systems. Here, we argue that the theory of ontological clarity, also known as the representational model (RM) of representation theory (RT) (Burton-Jones et al., 2017), can proffer a structured understanding of the meaning-making that undergirds distributed sociotechnical systems. The representation model of RT offers useful insights that allow us to link its constructs with semantic interoperability and, in turn, provides theoretical support for our design artifact.

Representation theory accepts that information systems constitute representations of real-world phenomena (Burton-Jones et al., 2017). The primary focus of RT is "the extent to which the deep structure of an information system provides and remains a faithful representation of the focal real-world phenomena" (Wand & Weber, 1995, p. 206). In examining their notion of faithful representations of focal real-world phenomena, Wand and Weber (1995) proposed three distinct but related models under the RT umbrella: the representation model (also known as the theory of ontological clarity), the state-tracking model, and the good decomposition model. Each model provides conditions necessary (but not sufficient) to ensure that an information system is and remains a faithful representation of real-world phenomena in spite of changes within its own components and changes in its environment. Our work extends their ideas into the domain of semantic interoperability. The state tracking model expands upon the proposition that an information system providing a good representation of its focal real-world phenomena must faithfully track changes in its focal real-world phenomena over time (Wand & Weber, 1995). The good decomposition model proposes a set of necessary conditions related to the decompositions of the focal real-world phenomena embodied by the information system (Wand & Weber, 1989). When met, the necessary conditions of the good decomposition model indicate that the information system is better capable of conveying the meaning of the focal real-world phenomena (Burton-Jones et al., 2017). In our assessment, neither state tracking nor good decomposition models are particularly relevant to the interoperability of XBRL-based financial statements. We focus instead on the representation model. The main thrust of our argument is that, with respect to data standards, in general, and the US implementation of XBRL for financial reporting, in particular, ontologically clear information systems produce semantically interoperable scripts.



*Relevant to UGT

Figure 1. Ontological Clarity and Semantic Interoperability Model

The theory of ontological clarity is concerned with the symbols that make up the scripts generated by a model (a data standard in our case). These symbols are drawn from a grammar that must be able to generate construct instances to represent real-world objects completely and clearly (Burton-Jones et al., 2017). The theory defines four ways in which a grammar can fall short of ontological clarity: (1) construct deficit, (2) construct excess, (3) construct redundancy, and (4) construct overload, as shown in Figure 1. *Construct deficit* indicates that the standard is missing constructs necessary to represent a real-world construct. *Construct excess* may arise if the grammar/standard contains constructs that do not map to any real-world construct. *Construct redundancy* is caused by two or more constructs that map onto the same real-world construct. This is also called *construct identity fallacy* (Larsen & Bong, 2016). *Construct overload* indicates that the representation contains constructs that map to multiple real-world constructs. We argue here that the presence of any of these four defects in a data standard will cause scripts generated by that standard to be non-interoperable. As a corollary, we also argue that remedying any of these defects will improve the interoperability of scripts generated by the faulty standard. In the following paragraphs, we provide our assessments of the UGT with respect to the four defects. To do this, we draw upon the accounts of practitioners, the existing literature, and our own experience with XBRL in the US financial reporting jurisdiction.

The generation of XBRL instance documents using the US GAAP taxonomy can be thought of as an ordered set of four tasks: (1) mapping, (2) extensions, (3) tagging, and (4) creating and validating (Bartley, Al-Chen, & Taylor, 2010). Our interpretation of the representation model of representation theory as well as the conclusions drawn in Zhu & Wu (2014) strongly suggest that the source of the XBRL interoperability problems lies in the mapping phase of XBRL preparation. The *mapping* process is increasingly performed by specialists within the firm (recommended by Bartley et al., 2010) or outsourced to specialized firms. The mapping function identifies and matches each financial concept in a firm's financial statement to a corresponding XBRL element in the US GAAP taxonomy. *Extensions* are another major source of errors. As we've discussed previously, the XBRL standard permits the creation of nonstandard XBRL elements to accommodate what preparers of financial statements believe to be idiosyncratic reporting situations. XBRL is a complex standard that implements the extensibility of XML technologies to produce rich metadata-enhanced representations. While the UGT elements already include XML markup and code to represent the relevant metadata, the extension process must specify the metadata from scratch, causing *tagging* and *creation/validation* processes to inadvertently introduce errors.

Given the empirical reality of XBRL implementation in the US financial reporting jurisdiction and the centrality of the US GAAP taxonomy to the functioning of the standard, we argue that poor interoperability in XBRL is, at least in part, explained by representation theory, in general, and its theory of ontological clarity, in particular. In the seminal formulations of the theory of ontological clarity (Wand & Weber, 1993), information systems are decomposed into scripts and grammars for the generation of those scripts. This breakdown can be applied to XBRL in an obvious way—the grammar is the UGT and the scripts generated from the UGT are the XBRL instance documents (individual filings expressed in XBRL and drawing from the grammar). The IS literature has examined a number of implications regarding the misspecification of a grammar; however, to our knowledge, semantic interoperability has never been explained as the result of ontologically unclear grammars.

Few studies examine the relationship between a grammar's compliance across the four requirements and the semantic interoperability of the scripts generated by that grammar. While this study does not seek to directly fill that gap, we show that, at least in the case XBRL, the interoperability does appear to be a function of the ontological clarity of the UGT and that an IT artifact developed to address the ontological clarity of the grammar using the scripts generated by same also enables the interoperation of those scripts. In the next sections, we show that a lack of ontological clarity in the UGT, specifically construct deficit and construct redundancy, leads to the generation of scripts (XBRL instances) that do not interoperate. We assess that construct excess and construct overload do not exist in the UGT. We subsequently formulate hypotheses that formalize our assertion that addressing ontological clarity with respect to XBRL and the UGT would improve the interoperability of XBRL instance documents.

4.1 Construct Deficit

In XBRL, construct deficit is intentionally built into the UGT standard in order to support the creation of new custom constructs (XBRL extension elements) that are specific to each filer's unique needs: "The higher the proportion of custom tags in a set of financial statements, the lower the comparability with other financial statements" (Henry et al., 2018). Recent studies continue to report high usage of these extensions among filers. Whereas a sample of 2010 filings showed that 12% of XBRL tags were custom extensions (Debreceny et al., 2011), a sample of 2015 filings showed that, on average, 7.3% of a company's XBRL tags are custom (Henry et al., 2018). The use of

extensions is so important that the SEC regularly releases figures on their use in firm disclosures.

Our approach is to explicitly define an "investor's ontology" that establishes financial concepts that are important to downstream consumers of financial information. Our investor's ontology thus explicitly defines the ontology onto which XBRL elements map. The incorporation of an investor's ontology directly addresses the built-in construct deficit of the UGT.

4.2 Construct Redundancy

In our assessment, it is improbable that construct redundancy objectively exists in the UGT. However, in its interpretation by filers (firms generating scrips using the grammar), the UGT subjectively displays signs of construct redundancy. For a given accounting concept (ontological construct), two filers interpreting the grammar (the UGT) may come to different conclusions about the element in the UGT that faithfully denotes the accounting concept. For example, for a given real-world financial concept *C* (e.g., net income), different XBRL terms t_1, t_2, …, t_n from the standard may be used to denote the concept *C* by various filers f_1, f_2, …, f_n. The multiple choices to interpret the same concept *C* leads to *construct redundancy*. That is, construct redundancy in this setting manifests in nonobvious mappings from an ontological construct to an XBRL element. We looked to practitioner accounts of their experiences with XBRL filing and found that the sheer scale of the UGT means that in determining the appropriate XBRL element for a financial concept, filers are often faced with several choices for the same financial concept. [2] Further evidence of this can be found by efforts undertaken by the SEC to manage the complexity of the UGT.[3] Figure 2 shows that the two different XBRL elements (*us-gaap_ProfitLoss* and *us-gaap_NetIncomeLoss*) may be used to quantify the financial concept of *net income*. Preparers of XBRL financial statements look to the grammar's metadata to determine the appropriateness of an XBRL element to the financial concept they wish to report and tag. In particular, preparers leverage the label information (human-readable text) used to describe an XBRL element. Different firms may use different label terms, resulting in a list of label terms for an XBRL element, as shown in Figure 2, but we intuit that these terms will be lexically close. We argue here that label information is therefore a useful signifier of the meaning of an XBRL element and that even disparate XBRL elements used to communicate the same financial concept will have similar labels. Indeed, lexical closeness has been shown to be a powerful conveyer of shared meaning (Gefen & Larsen, 2017).

Representation theory has been criticized as being built on an ontology that was never intended to model the objects of human perception (Allen & March, 2006). Rather, Bunge's ontology, upon which RT is built, is a conceptualization of the material world that is independent of human interpretation. This fact, Allen and March argue, precludes the application of Bunge's ontology to the conceptual realm of conceptual modeling. This exclusion, we note, encompasses the application of RT herein. While it is indisputable that Bunge's ontology explicitly excludes the conceptual world, it does so as a means of simplifying the task at hand. In their descriptions of how Bunge-Wand-Weber ontology leads to inappropriate proscriptions regarding conceptual world models, a clear theme emerges: the conceptual world is more complex to model than the physical world (i.e., the world existing independently of human interpretation) with the consequence that RT, based on Bunge's ontology, cannot possibly capture the richness of the conceptual world.

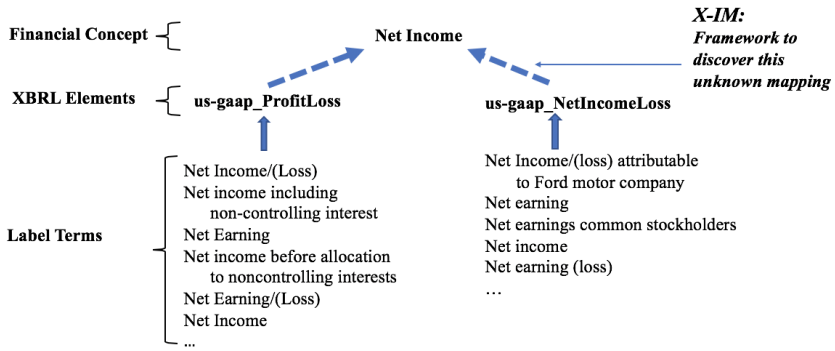


Figure 2: XBRL Elements Aligned by X-IM based on the Label Terms from Various Firms

[2] https://sfmagazine.com/wp-content/uploads/sfarchive/2013/07/XBRL-An-IMA-Member-Shares-His-XBRL-Filing-Experience.pdf

[3] https://haslam.utk.edu/sites/default/files/files/SECs_Increasingly_Sophisticated_Use_of_XBRL_Tagged_Data.pdf

However, this critique falls short in three key areas. First, a conceptual model is *always* a simplification of reality. If an ontology of the physical world can be projected into a useful approximation of the conceptual world, its foundation is supported. Second, if this useful approximation passes the muster of multiple attempts at empirical refutation (the work by Allen & March, 2006, is not an empirical refutation) and there exist multiple streams of evidence consistent with its propositions, RT remains useful and relevant. Finally, as noted in Burton-Jones et al, Allen and March (2006) offer no meaningful alternative formulation beyond RT for the evaluation of conceptual grammars. In this paper, we sought out representation theory because its constructs and propositions heighten our understanding of our focal phenomenon. Our work is a single instance of evidence that is consistent with the propositions of the representation model of representation theory. The constructs we examine are both enlightening and valid with respect to our focal phenomenon.

In applying representation theory to the XBRL case, we generate two sets of design principles that are generalizable beyond the current application. First, as regards construct redundancy, even in situations where the grammar is not redundant, it may generate scripts that suffer from construct redundancy or complexity in the grammar. Metadata from a grammar can be leveraged by automated agents to detect terminologically distinct but semantically equivalent uses of grammar constructs in scripts. Second, regarding construct deficit, grammars may be intentionally sparse and extensible. Extensions to such grammars generate scripts that do not interoperate. Such scripts may be made interoperable by the automated generation of ontologies/taxonomies that progressively augment the deficient grammar.

4.3 Hypotheses

Informed by the representation model of representation theory, our proposed design artifact, X-IM, addresses construct deficit and construct redundancy using the investor's ontology and XBRL label terms, respectively. The investor's ontology encapsulates a set of widely used equivalent financial terms of interest to investors and the designative information relevant to their respective financial concepts: e.g., short-term marketable securities, short marketable securities, and short-term investments. For the index ontology, the web crawler designed for X-IM automatically extracts *XBRL elements* (e.g., us-gaap_NetIncomeLoss), *their label terms* (e.g., net income), and *their corresponding designative information* (e.g., balance type and period type) from the SEC's website. The extracted information is represented in the index ontology. X-IM aligns each financial concept represented in the investor's ontology with its corresponding XBRL elements using label terms that are encapsulated in the index ontology. X-IM can be viewed as the framework for an ontology alignment between the investor's ontology and the index ontology to achieve XBRL interoperability. Figure 2 shows the high-level view of our proposed framework, X-IM, in terms of a financial concept, XBRL elements, and label terms used by various firms. The next section has a detailed description of X-IM.

Utilizing the investor's ontology as the means to address construct deficit, we formulate the following hypotheses:

H1a: X-IM with an investor's ontology (incorporating investor's standard terms and designative information) will outperform X-IM without an investor's ontology in terms of overall precision.

H1b: X-IM with an investor's ontology will outperform X-IM without an investor's ontology in terms of overall recall.

H1c: X-IM with an investor's ontology will outperform X-IM without an investor's ontology in terms of overall *F*-measure.

Additionally, we propose the use of label terms in the resolution of construct redundancy and thus formulate the following hypothesis:

H2a: X-IM employing label information will outperform an approach employing no label information for ontology mapping in terms of overall precision.

H2b: X-IM employing label information will outperform an approach employing no label information for ontology mapping in terms of overall recall.

H2c: X-IM employing label information will outperform an approach employing no label information for ontology mapping in terms of overall *F*-measure.

5 Framework Design: X-IM System for XBRL Ontologies Mapping

5.1 System Architecture

The system architecture for our proposed artifact, X-IM framework, is shown in Figure 3. X-IM consists of three components: the EDGAR web crawler, the IOnto generator, and the IBC learner. The EDGAR web crawler accesses the SEC's electronic data-gathering, analysis, and retrieval (EDGAR) website and automatically extracts XBRL elements (e.g., *us-gaap_AccountsPayableCurrent*), their label terms (e.g., "Accounts payable"), as well as their corresponding designative information (e.g., balance type and period type) from EDGAR's interactive

financial statements. The balance type classifies a financial concept as *duration* or *instant*, whereas the period type categorizes it as *debit* or *credit*. The IOnto generator generates the indexing ontology (IOnto) by integrating XBRL elements, their corresponding label terms, as well as their associated designative information. The IBC learner leverages IOnto and the investor's ontology to align heterogeneous XBRL elements (e.g., mapping between *us-gaap_AccountsPayableCurrent* and *us-gaap_AccountsPayableTradeCurrent*), capturing the mapping results in the derived ontology, X-Onto. Following are detailed descriptions of each component.

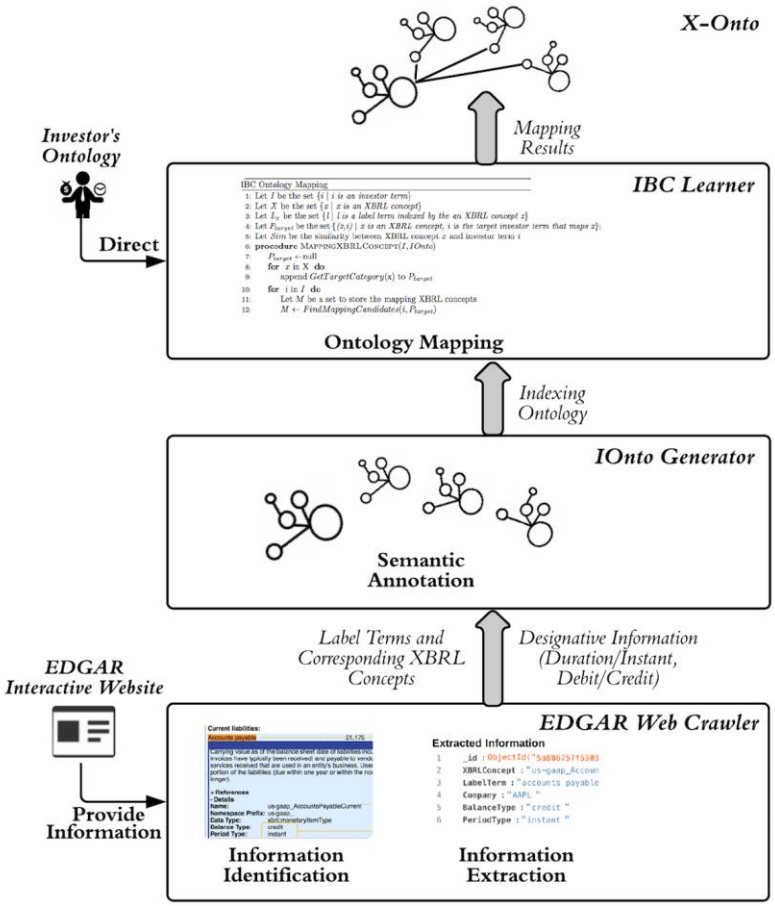


Figure 3. System Architecture of XBRL Indexing-Based Mapping (X-IM) Framework

5.2 The EDGAR Web Crawler

EDGAR collects, validates, and indexes individual XBRL filings. It provides public access to corporate financial information parsed from these XBRL filings, and, in particular, SEC forms 10-K and 10-Q. Our EDGAR web crawler (EWC) automatically collects, parses, and integrates the label terms, their corresponding XBRL elements, and the designative information from the 10-K interactive financial statements of the listed companies.

Figure 4 presents EWC's automatic information retrieval method. First, EWC, powered by a browser automation tool, locates the EDGAR interactive financial statements of a company by entering its ticker symbol (AAPL for Apple) in the EDGAR search portal,[4] as shown in Figure 5. Please note that the screenshots are manually obtained and shown here to clearly illustrate the multiple steps that our EWC automatically goes through in retrieving the information. Second, it retrieves the 10-K interactive filings of the company (see Figure 6). We use annual financial reports (10-K) as a test case in our approach to XBRL interoperability; annual reports contain a

[4] https://www.sec.gov/edgar/searchedgar/companysearch.html

richer collection of XBRL elements than quarterly reports (10-Q). Third, EWC identifies a specific financial statement, such as consolidated balance sheets or statements of cash flows. The left panel in Figure 7 shows a collection of interactive financial statements available on EDGAR, whereas the right panel presents the line items of the consolidated balance sheets for the period of September 24, 2011 to September 29, 2012 for Apple Inc. Figure 7 shows many line items for "Current assets" as well as for "Shareholders' equity." Note that we point out only the first three line items of "Current assets." In an interactive financial statement, each line item is the label term used to quantify the financial concept and is automatically read from Apple Inc.'s XBRL filing.

Web Crawler Collecting Indexing and Designative Information

```
1: EDGAR-CRAW(CompanyList*)
2:     for i in CompanyList do
3:         Locate the interactive financial filings of a specific company i
4:         Retrieve the 10-K filings of the company i
5:         Read specific financial statements from the set of 10-K filings
6:         for each label term in the financial statements do
7:             Collect the corresponding XBRL concept and designative infor-
8:             mation
9:     Return Collected information
```

Note: * *CompanyList* = a list of companies in the training dataset

Figure 4. Web Crawler Method

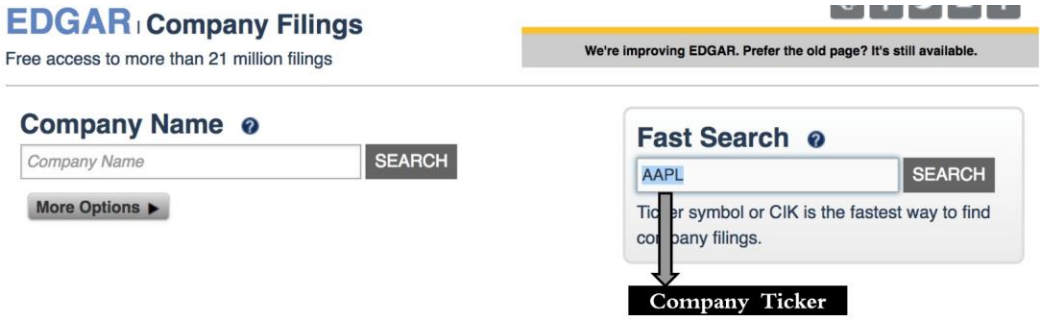


Figure 5. Screenshot of EDGAR's Company Search Portal

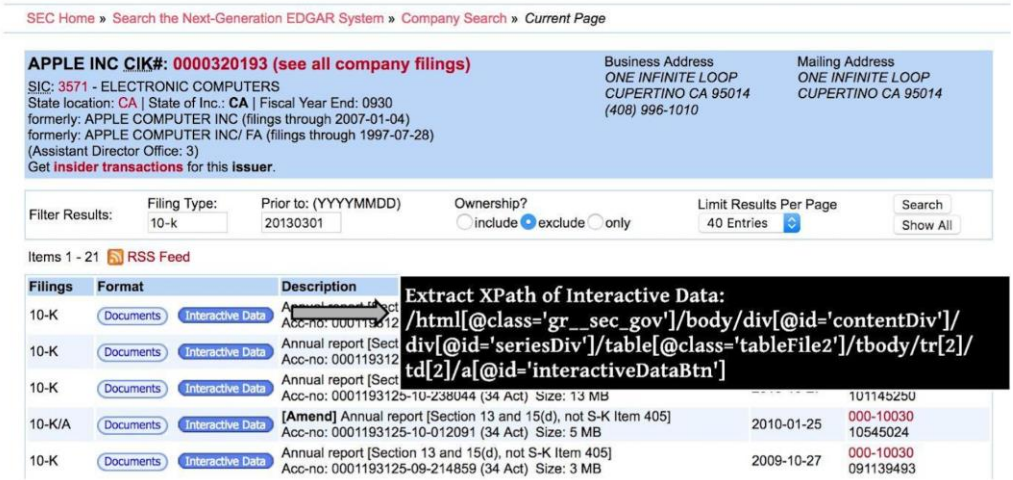
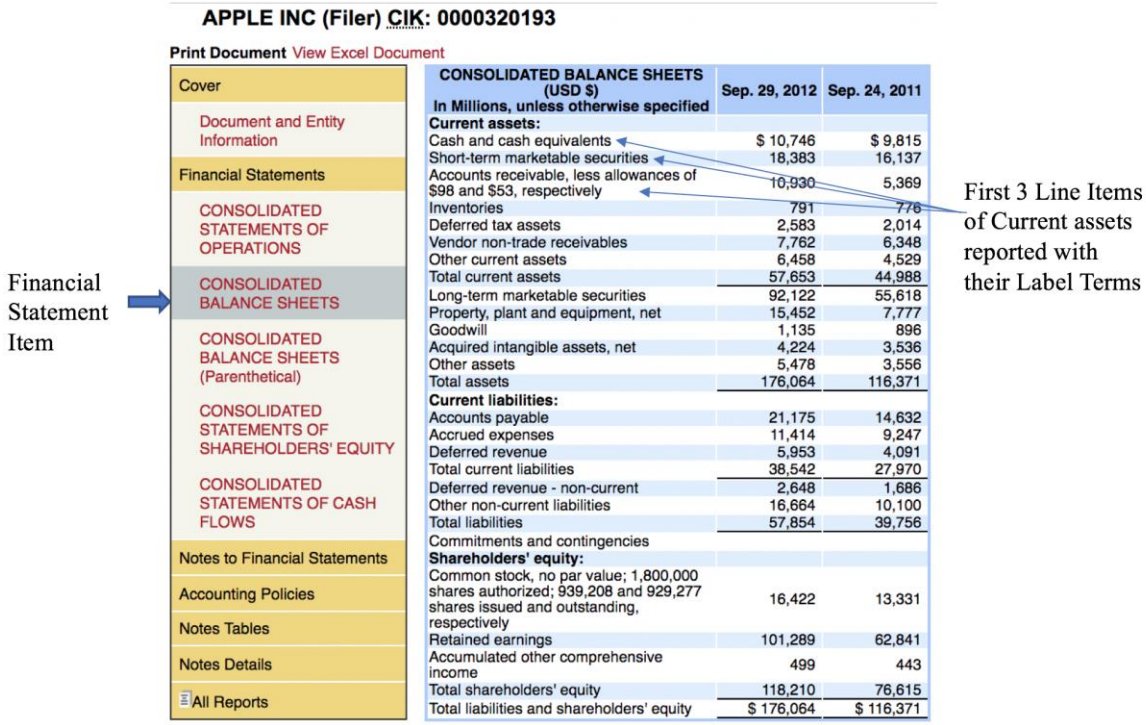


Figure 6. A List of Apple's 10-K Interactive Financial Documents



CONSOLIDATED BALANCE SHEETS (USD $) In Millions, unless otherwise specified	Sep. 29, 2012	Sep. 24, 2011
Current assets:		
Cash and cash equivalents	$ 10,746	$ 9,815
Short-term marketable securities	18,383	16,137
Accounts receivable, less allowances of $98 and $53, respectively	10,930	5,369
Inventories	791	776
Deferred tax assets	2,583	2,014
Vendor non-trade receivables	7,762	6,348
Other current assets	6,458	4,529
Total current assets	57,653	44,988
Long-term marketable securities	92,122	55,618
Property, plant and equipment, net	15,452	7,777
Goodwill	1,135	896
Acquired intangible assets, net	4,224	3,536
Other assets	5,478	3,556
Total assets	176,064	116,371
Current liabilities:		
Accounts payable	21,175	14,632
Accrued expenses	11,414	9,247
Deferred revenue	5,953	4,091
Total current liabilities	38,542	27,970
Deferred revenue - non-current	2,648	1,686
Other non-current liabilities	16,664	10,100
Total liabilities	57,854	39,756
Commitments and contingencies		
Shareholders' equity:		
Common stock, no par value; 1,800,000 shares authorized; 939,208 and 929,277 shares issued and outstanding, respectively	16,422	13,331
Retained earnings	101,289	62,841
Accumulated other comprehensive income	499	443
Total shareholders' equity	118,210	76,615
Total liabilities and shareholders' equity	$ 176,064	$ 116,371

Figure 7. A Collection of Interactive Financial Statements available on EDGAR (right) and Contents of Consolidated Balance Sheets (left)

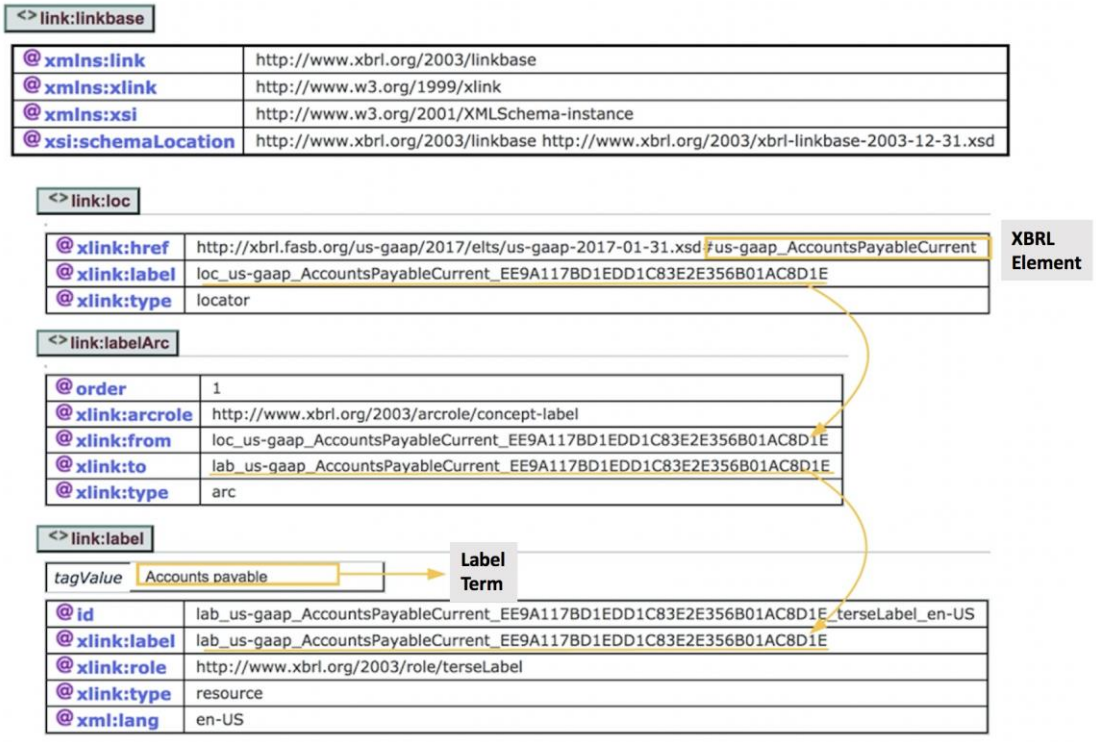


Figure 8. XBRL Label Links

Among the five different linkbases introduced earlier, this study leverages the label linkbases that link human-readable text (label terms) with XBRL elements using specific tags (i.e., @xlink:label) (Luna-Reyes et al., 2005). An exemplary label link in Figure 8 illustrates that Apple Inc. uses a readable label term "Accounts payable" for an XBRL element, "us-gaap_AccountsPayableCurrent." The labels given in a company's label linkbase are parsed by EDGAR to provide human-readable label terms in EDGAR interactive financial statements (SEC, 2010).

Additionally, each label term in an EDGAR interactive financial statement is rendered in hypertext so that a link is maintained with its corresponding XBRL element and designative information, such as balance type and period type, as shown in Figure 9. This designative information, which creates discrete categories for XBRL elements, assists X-IM in annotating and interpreting XBRL elements. Lastly, EWC iteratively retrieves each label term of interest to the investors, its corresponding XBRL element, as well as designative information one at a time. The top portion of Figure 9 presents the XBRL element, balance type and period type for the label term, "Accounts payable," whereas the bottom portion shows the information extracted by our EDGAR web crawler.

5.3 The IOnto Generator

The IOnto generator amalgamates XBRL elements, label terms, and designative information to construct the indexing ontology (IOnto). IOnto is an ontological representation of the indexing correlation between XBRL elements and label terms in which XBRL elements are depicted as indices, and label terms as references and interpretations of the associated XBRL elements. There is a precedent for an indexing ontology approach to the resolution of semantic heterogeneity (Doan et al., 2002; Kaza & Chen, 2008). In our approach, an ontology provides an effective means to represent indexed relationships, especially when relationships are sparsely distributed—not all XBRL elements have the same number of label terms. Further, an ontological representation enables us to annotate the designative information that structures our XBRL elements, facilitating ontology mapping.

We illustrate our IOnto generation algorithm for use on information extracted by EWC in Figure 10. Each XBRL element is a class in IOnto. If a company com_i uses a specific XBRL element C_{XBRLj} in its financial statements, the *IOntoGeneration* method adds a new individual $indi_k$ to the ontology class C_{XBRLj}. For the individual $indi_k$, we specify its data attributes (hasBalanceType and hasPeriodType) and object attributes (hasLabelTerm and comUseConcept). In this manner, we are indexing label terms for a corpus on their corresponding XBRL elements

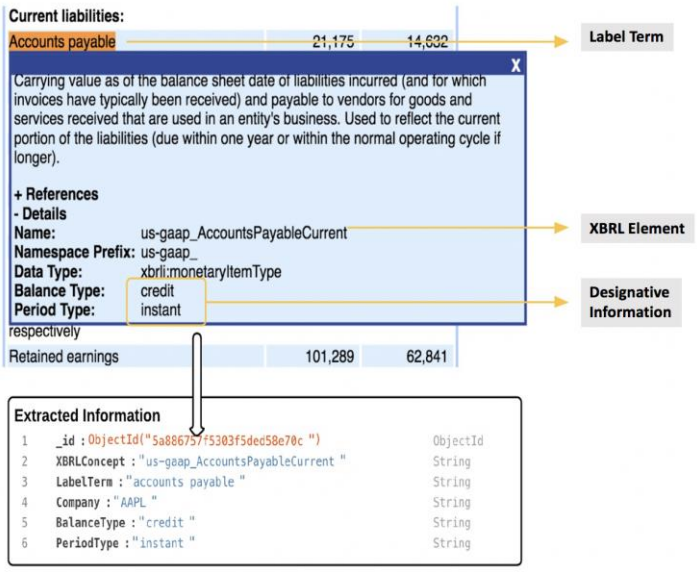


Figure 9. XBRL Element and Designative Information for Label Term, "Accounts payable"

IOnto Generation

Let C_{XBRL} be a set of XBRL elements (i.e., GrossProfit, SalesRevenueNet, AccountsReceivableNet, etc.) in the training set
Let *com* be a set of companies (i.e., APA, AAPL, AMZN, etc.) in the training set
IONTOGENERATION(*CrawledResults*)
    Create an ontology class, *Company*
    for each com_i in *com* **do**
        create a new individual, com_i
        add com_i to the class *Company* as its individual
        for each C_{XBRLj} in C_{XBRL} **do**
            create C_{XBRLj} as an ontology class
            if C_{XBRLj} is included in the financial statements of com_i **then**
                create a new individual $indi_k$
                add $indi_k$ to the class C_{XBRLj} as its individual
                specify data attribute $hasBalanceType$ for $indi_k$
                specify data attribute $hasPeriodType$ for $indi_k$
                specify object attribute $hasLabelTerm$ for $indi_k$
                specify object attribute $comUseConcept$ for $indi_k$

Figure 10. IOnto Generation Method

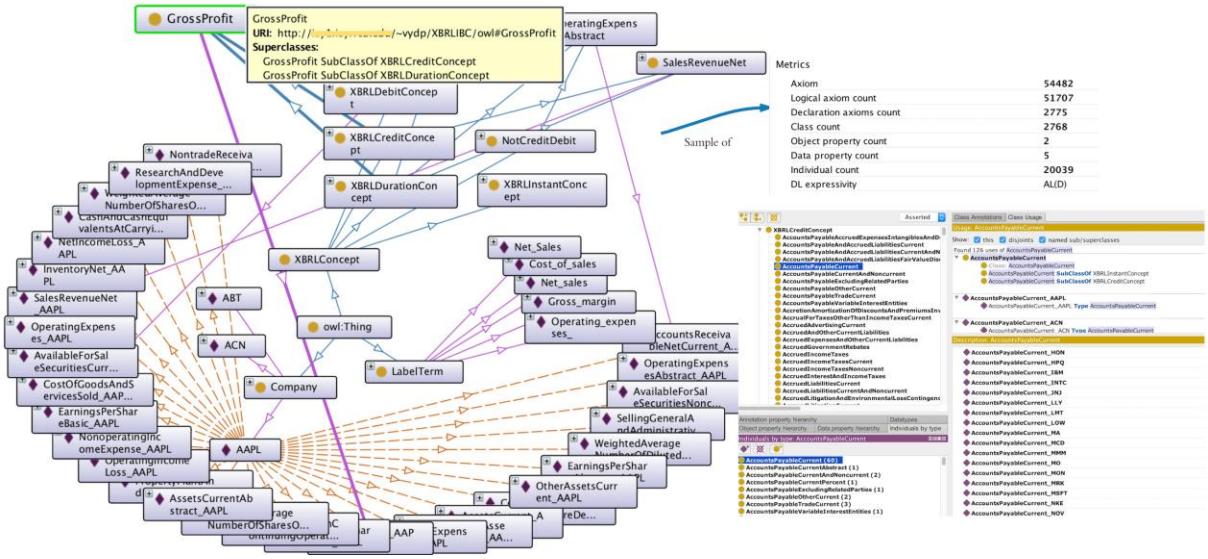


Figure 11. Graphical Representation of an IOnto Segment

Figure 11 graphically depicts a segment of IOnto leveraging label linkbases and designative information. The IOnto becomes the input to our IBC Learner that conducts ontology mapping. One benefit of generating IOnto is to avoid frequently accessing complex XBRL filing ontologies (recording label information and other numeric information, such as the annual or quarterly value of a statement item) to retrieve label information when the IBC learner is invoked. In this way, IOnto improves system efficiency. Another benefit that IOnto brings to our design is portability. It enables IOnto and the IBC learner to be transplanted in other XBRL ontology mapping environments (e.g., FinCEM in Etudo et al., 2017) without rebuilding the whole set of XBRL ontologies. Further, IOnto, which leverages label links in the XBRL label linkbases of various firms and builds up the indexing relationship between each XBRL element and its correspondent label terms, provides two major utilities to the IBC learner: (1) organizing the correspondence between an XBRL element and the label terms as its features in a vector form, and (2) enabling the IBC learner (discussed below) to classify XBRL elements into its target investor's term through calculating the similarity of an investor's term and its feature vector.

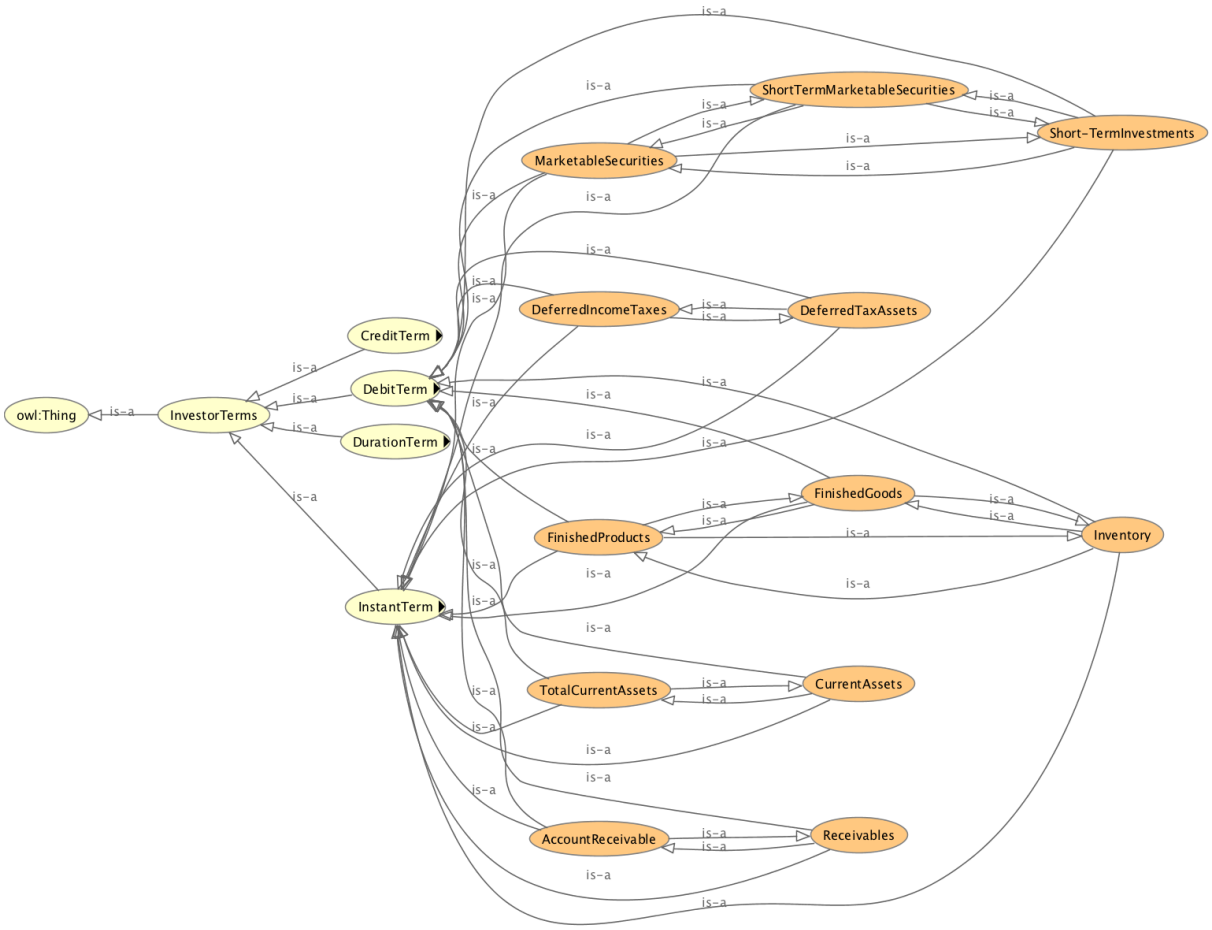


Figure 12. A Segment of Investor's Ontology

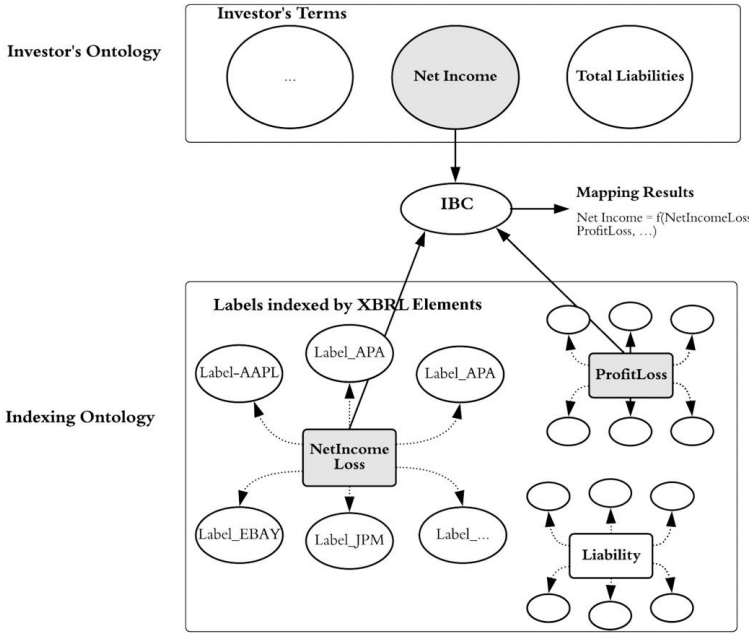


Figure 13. Conceptual Representation of Indexing-Based Classification

5.4 The IBC Learner

The IBC learner, with the assistance of IOnto, aligns all heterogeneous XBRL elements for each investor's term in the investor's ontology. As defined earlier, the investor's ontology encapsulates a set of widely used financial terms of interest to investors and their designative information. As different investors may use different terms according to their preferences and individual histories, our investor's ontology represents a set of the equivalent investor's terms that are extensively used in accounting and finance for each financial concept (Kieso, Weygandt, & Warfield, 2013). For example, some investors may use short-term marketable securities, but others may choose short marketable securities or short-term investments. Our investor's ontology denotes the equivalence among these three terms, as shown in Figure 12. See Appendix Table A1 for investor's terms encapsulated in the ontology.

Leveraging XBRL elements and their label terms in IOnto, the IBC learner conducts indexing-based classifications to map all heterogeneous XBRL elements that different companies use for each investor's term in the investor's ontology. The core of the IBC learner is our novel indexing-based classifier that functions across a set of XBRL elements, each coupled with its feature vector of correspondent label terms from various companies. We compute the semantic similarity between each investor's term and a vector of label terms indexed by an XBRL element. Prior studies have used lexicon-based and/or structure-based XBRL ontology mappings at the pairwise level. Although they have achieved relatively high accuracy and made notable contributions to the XBRL interoperability, our careful analysis reveals that a pair-wise mapping method has limitations. For example, it fails to map the XBRL element "ProfitLoss" to the investor's term "Net Income" because of a low lexicon similarity between them. In order to overcome this limitation, informed by IOnto integrating XBRL elements with their corresponding label terms along their designative information, we perform a vector-wise similarity calculation between a set of label terms and an investor's term rather than a pair-wise similarity measure. Incorporating the indexing relationships between XBRL elements and their corresponding label terms in IOnto, the task of mapping heterogeneous XBRL elements to a target investor's term can be conceptualized as a classification problem, as follows:

Let $Y = \{y1, y2, y3 \ ... \ ym\}$ be the set of targets. Let $X = \{x1, x2, x3 \ ... \ xn\}$ be the set of items to be classified. Let *f* be the classification function, which maps an item *x* with its target *y*. If a target element *y* has multiple mapping items {Xj, *X*k ... *X*r}, we can say Xj, *X*k ... *X*r have the same classification target *y*. Likewise, in the context of XBRL, given $Investor\_terms =$ {Investorterm1, Investorterm2 ... Investortermm } as the target set and $XBRL\_elements =$ {XBRLelement1, XBRLelement2 ... XBRLelementn} as the item set, we can find the mapping function between Investor_terms and XBRLelements, as shown in Figure 13: $\text{Investortermi} = f(\text{XBRLelementj})$. A given Investortermi may find several corresponding XBRL elements {XBRLelementj, XBRLelementk … XBRLelementr} through f , where XBRLelementj, XBRLelementk … XBRLelementr are equivalent to each other. For example, given an investor's term "Net Income" as the target, we can find multiple XBRL elements mapping to it through f, such as NetIncomeLoss, NetIncomeLossAvailableToCommonStockholders Basic, and ProfitLoss.

Figure 14 presents the logic of our indexing-based classification (IBC) method in detail. IBC Learner starts the classification process by invoking the MappingXBRLElement method, incorporating two crucial inputs: a set I of investor's terms of interest and IOnto. For each XBRL element (x) in IOnto, the IBC Learner uses the GetTargetCategory method to find its target investor's term.To find the target investor's term for a specific XBRL element *x*, our method traverses the set *I* and calculates the similarity between a XBRL element *x* and each investor's term *i*. The similarity between *x* and *i* can be determined based on Formula (1), which aggregates the similarity between *i* and each label term *l* used for *x*. The similarity between *i* and a label term *l* can be achieved through Formula (3), which calculates the Jaccard similarity of the two terms, and then (2) amplifies the Jaccard similarity signal by converting the range of [0, 1] to the scale of [0, ∞].

$$Similarity\ (x, i) = \sum_{n=1}^{length\ of\ label\ term\ vector\ for\ XBRL\ element\ x} AdjustedSimilarity\ (l_n, i) \quad (1)$$

$$AdjustedSimilarity\ (l, i) = \frac{JaccardSimilarity(l, i)}{1 - JaccardSimilarity(l, i)} \quad (2)$$

$$JaccardSimilarity\ (l, i) = \left(\frac{|l \cap i|}{|l \cup i|}\right) \quad (3)$$

Figure 15 shows the function of amplifying the Jaccard similarity signal. Before concluding that an XBRL element *x* is the candidate concept mapping to the investor's term *i*, the similarity score must exceed a certain threshold. As recognized by Etudo et al. (2017), which explored the effect of the varying thresholds, a proper threshold to filter out the noise may, in the operation phase, impact the precision and recall ratios. We conduct experiments to carefully examine the threshold effect on the precision and recall ratios.

After getting the classification target for each XBRL element, the IBC learner uses the *FindMappingCandidates* method to group the XBRL elements along with their target investor's terms. The XBRL elements and their target, *i*, are fetched into the same group as the mapping candidates for *I:* $\{x_m, i\}$, $\{x_n, i\}$, ... $\{x_k, i\}$. To filter out any invalid candidates, the IBC learner uses the *FiterFalseCandidates* method, which compares the designative information of the candidate with that of the investor's term. Those XBRL elements with disparate designative information are excluded from the final mapping list.

The above indexing-based classification function is trained with the annual financial reports of a small number of firms listed in the S&P 100 and rigorously tested with the financial reports over multiple years from a larger number of firms. The evaluation section presents the results of our performance analysis in detail.

IBC for Ontology Mapping

Let I be the set $\{i \mid i$ *is an investor term*$\}$
Let X be the set $\{x \mid x$ *is an XBRL element*$\}$
Let L_x be the set $\{l \mid l$ *is a label term indexed by an XBRL element* $x\}$
Let P_{target} be the set $\{(x,i) \mid x$ *is an XBRL element,* i *is the target investor term that maps* $x\}$;
Let Sim be the similarity between *XBRL element* x and investor term i
MAPPINGXBRLELEMENT(I, *IOnto*)
    $P_{target} \leftarrow$null
    for x in X **do**
        append $GetTargetCategory(x)$ to P_{target}
    for i in I **do**
        Let M be a set to store the mapping XBRL concepts
        $M \leftarrow FindMappingCandidates(i, P_{target})$
        $M \leftarrow FilteringFalseCandidates(i, M)$
GETTARGETCATEGORY(x)
    for i in I **do**
        $accumulatedSim \leftarrow 0$
        for l in L_x **do**
            $similarity(l,i) \leftarrow JaccardSimilarity\ (l,i)$
            $adjustedSim(l,i) \leftarrow similarity\ (l,i)/\ (1 - similarity(l,i))$
            $accumulatedSim \leftarrow accumulatedSim + adjustedSim(l,i)$
        $Sim(x,i) \leftarrow accumulatedSim/len(L_x)$
    if $Sim(x,i) >$ any $Sim(x,i')$ and $Sim(x,i) > threshold$ **then**
        return (x, i)
    else
        return null
FINDMAPPINGCANDIDATES(i, P_{target})
    traverse each pair in P_{target}
    find all pairs $\{x_m, i\}$, $\{x_n, i\}$, ..., $\{x_k, i\}$, where the target is i
    Conclude x_m, x_n, ... and x_k are the mapping candidates to i
    return $\{x_m, x_n, \text{ ... and } x_k\}$
FILTERFALSECANDIDATES(i, M)
    for x in M **do**
        Let d_x be the designative information of x
        retrieve d_x
        Let d_i be the designative information of i
        retrieve d_i
        if $d_x == d_i$ **then**
            Conclude x is valid
        else
            Conclude x is invalid
            exclude x from M
    return M

Figure 14. Logics of Indexing-Based Classification (IBC) Method

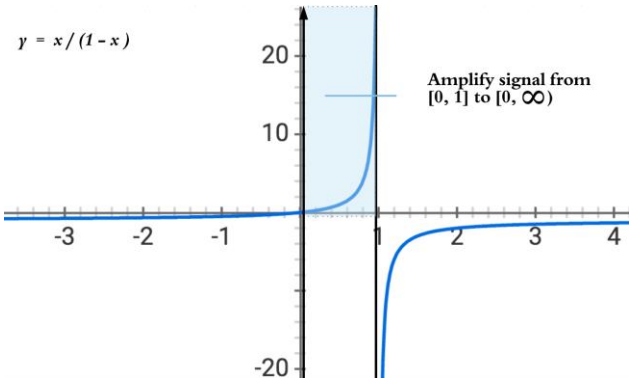


Figure 15. Function for Signal Amplification

5.5 Implementation Details

We employ several technologies to instantiate our artifact, X-IM. To bypass the anticrawler mechanism of the EDGAR interactive website, we use Selenium Python binding, which provides the interface to access all utilities of Selenium WebDriver. Selenium WebDriver enables our crawler to mimic the human behavior of visiting a website and collecting web pages.

XPath and regular expression are employed to locate and extract the exact information within the financial statements on web pages. In developing an instance of IOnto, we use a JAVA and JENA framework that provides an API for reading, writing, and processing ontologies. The IBC learner is implemented with Python 2.7, conducting similarity calculation and comparison. Finally, we use Protégé 5.17 in examining the RDF ontologies.

6 Evaluation and Discussion

6.1 Evaluation Method and Frame of Reference

We evaluate our artifact using the formal ontology evaluation method proposed in (Yu, Thom, & Tam, 2009). This method has been shown to be useful for evaluating ontology-driven applications (Etudo et al., 2017; Narock, Yoon, & March, 2014). Another reason for using the formal evaluation method is that it is the method employed to evaluate prior artifacts designed to achieve XBRL interoperability (Etudo et al., 2017). The formal ontology evaluation method is grounded in the experimental approach. The derived ontology (*DO*) to be evaluated represents a set of concepts DO_C, a set of instances DO_I, and a set of relationships DO_R between those concepts and instances: $DO = \{DO_C, DO_I, DO_R\}$. The target ontology (*TO*) encapsulates the set of concepts TO_C, the set of instances TO_I and the set of relationships TO_R between those concepts: $TO = \{TO_C, TO_I, TO_R\}$. In keeping with established methodology (Yu et al., 2009), this study presents the precision and recall metrics, evaluating the performance of our derived ontology, *DO*, with respect to the target ontology, *TO*.

The precision measure describes the extent to which all retrieved are relevant, whereas the recall describes the proportion of relevant that have been retrieved over all relevant. In the context of X-IM evaluation, the precision ratio measures the extent to which retrieved XBRL elements are correct, and the recall ratio depicts the extent to which XBRL elements claimed to be equivalent to a financial concept are retrieved through our methods. The *F*-measure, a weighted harmonic mean of precision and recall, is defined as the following (Powers, 2011):

$$F - measure = 2 * \left(\frac{Precision * Recall}{Precision + Recall}\right)$$

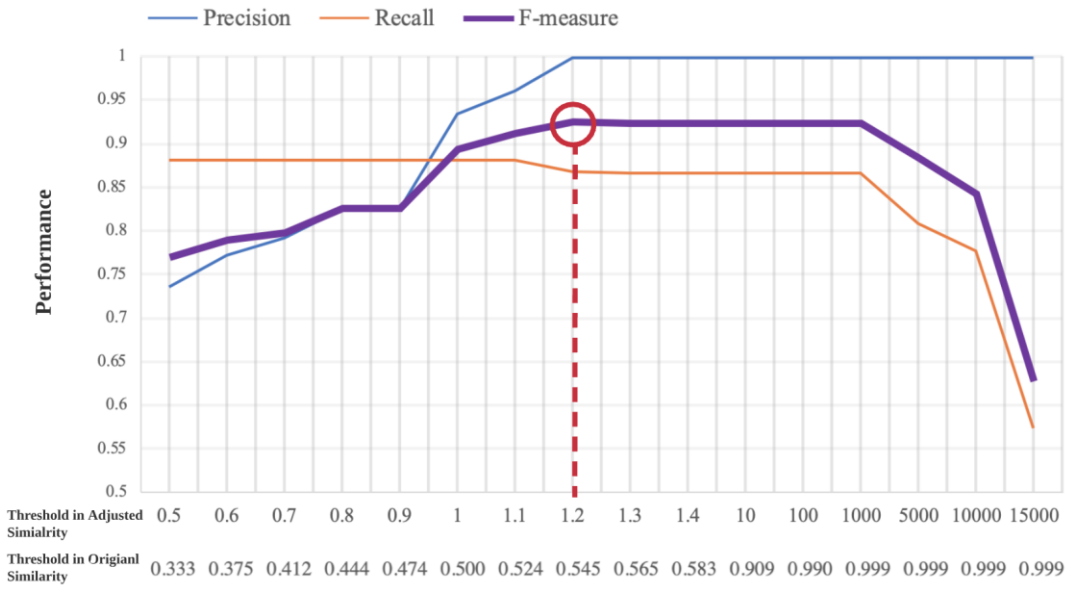
Suppose the correct XBRL elements of *Net Income* for five companies are *NetIncomeLoss, ProfitLoss, NetIncomeLoss, ProfitLoss,* and *NetIncomeLoss,* and the actual XBRL elements retrieved by a design artifact for those five firms are *NetIncomeLoss, Null, NetIncomeLoss, Null,* and *NetIncomeLoss,* respectively. A precision ratio is 100%, since three out of three retrieved XBRL elements are correct, whereas a recall ratio is 60%, because three out of five relevant XBRL elements are retrieved. Using the three measures, we assess the automatic mapping capability of X-IM. Performance along these metrics indicates the extent to which X-IM resolves the heterogeneity in XBRL elements to achieve XBRL interoperability.

This study used the annual financial reports (10-K) of the 92 firms listed in the S&P 100 for the two fiscal years ending in 2011 and 2012; eight were ruled out because they did not have XBRL filings in either 2011 and 2012. S&P 100 companies cover a variety of industries, introducing variation to the collection of tags in their XBRL filings.

Table 3: Specific Financial Concepts Included in Evaluation

Financial concepts in frame of reference	
FinCEM	**X-IM**
Cash from operations*	*FinCEM* +
Common stock	Stakeholders' equity*
Long-term debt	Accounts payable
Net income*	Accounts receivable*
Total assets	Deferred tax assets*
Total current assets*	Short-term marketable securities*
Total current liabilities	Cost of goods sold*
Total liabilities	Interest payments*
Total revenues*	Operating income*
	Inventory*

Note: *denote that the item in investor's ontology has multiple terms



Notes: The conversation between adjusted similarity and original similarity:

$$AdjustedSimilarity\ (l,\ i) = \frac{JaccardSimilarity(l,\ i)}{1 - JaccardSimilarity(l,\ i)}$$

Figure 16. Threshold Effects on *F*-Measures

Prior work draws upon three financial indicators (profitability, financial leverage/liquidity, and operating efficiency) as a realistic basis for evaluating the mapping capability of their proposed systems. These three indicators rely on the nine financial concepts resolved by FinCEM (Etudo et al., 2017), which is, to our best knowledge, the state of the art in XBRL interoperability. Our research follows this convention but provides a more comprehensive list of 18 extensively used financial concepts (Table 3). We conducted a series of experiments to test our hypotheses along these 18 financial concepts.

6.2 Experiment 1: Sensitivity Analysis of Threshold

The IBC learner will not be able to retrieve the value for a particular financial concept from a firm's XBRL instance document when the XBRL element used in the film's filing does not match any of the equivalent terms in X-Onto. To conclude that an XBRL element used by a company is the concept mapping to the investor's term, the similarity score between the XBRL element and the investor's term must exceed a certain threshold. A proper threshold assists the IBC learner in

filtering out noises, thereby impacting the precision and recall ratios. If a threshold is very small (e.g., toward 0), the learner may map an XBRL element with an investor's term even when there is little similarity between them, resulting in a low precision ratio. Meanwhile, if the threshold is too large (e.g., toward ∞), the algorithm may not identify the mapping relationship even though the similarity score between an XBRL element and an investor's term is high, yielding a low recall ratio. Therefore, prior to experimentally testing our hypotheses, we conducted a sensitivity analysis of the threshold to examine the trade-off between precision and recall in achieving the best overall performance (*F*-measure).

We started with the threshold of 0.1 (adjusted similarity) with 0.1 increments and calculated the precision, recall, and *F*-measure scores for each threshold. Figure 16 shows that different thresholds result in significantly different precision and recall ratios. Taking into account that the *F*-measures assess the overall performance, the threshold of 1.2 resulted in the highest performance score. Thus, we chose 1.2 (in adjusted similarity) as the threshold for X-IM implementation for our experiments.

6.3 Experiment 2: Effect of Investor's Ontology

To test hypotheses H1a, H1b and H1c, we conducted a Wilcoxon signed-ranks test, which is a non-parametric hypothesis test comparing two matched samples to assess whether there is a mean difference (Gibbons & Chakraborti, 2011). It can be used as an alternative to the paired sample *t*-test, especially when the population cannot be assumed to be normally distributed (e.g., when the sample size of an experiment is relatively small). In our experiment, the Wilcoxon signed-ranks test statistically compared the means of two related variables (e.g., X-IM precision with investor's ontology and X-IM precision without investor's ontology) along 18 financial concepts to determine whether the experimental intervention resulted in a significant difference in their means. The experimental intervention was the presence or absence of the investor's ontology. In this experiment, the two instances of X-IM were trained using XBRL format 10-K annual reports for 10 randomly selected firms for the fiscal year 2011 (FY 2011)—one instance with the investor's ontology and the other without it. We used XBRL format 10-K annual reports for all 92 firms in FY 2012 for comparing the performance of X-IM with and without the investor's ontology.

In Table 4 and Table 5, the mean precision of X-IM for 18 financial concepts increased from 0.962 to 0.998 when employing the investor's ontology. However, the increase is not statistically significant ($Z = 1.604$, $p = 0.151$); thus, we cannot reject the null hypothesis in H1a. A possible explanation for this is that X-IM is already quite precise (0.962) even without the investor's ontology. The *F*-measure and recall of X-IM with the investor's ontology are both significantly larger than those of X-IM without the investor's ontology at a significance level of 0.01 (recall: $Z = 2.934$, $p = 0.003$; *F*: $Z = 2.934$, $p = 0.003$), supporting H1b and H1c. The results bolster our arguments that by incorporating designative information and investor's standard terms, the investor's ontology enables X-IM to recall more relevant terms and to raise its overall performance (*F*-measure).

6.4 Experiment 3: Comparative Performance Analyses

To test hypotheses H2a, H2b and H2c, we evaluated the performance of X-IM in comparison with the state of the art in XBRL interoperability, FinCEM (Etudo et al., 2017). We used a randomly selected sample of S&P 100 firms. In order to test the generalizability of our proposed solution over multiple years, we chose a training data set limited to 2011 XBRL filings. The resultant training set consisted of the 2011 XBRL filings of 10 randomly selected firms. X-IM was trained using XBRL-based 10-K annual reports for 10 firms for the fiscal year 2011 (FY 2011). To demonstrate how drawing on rich natural language information in label linkbases can improve the performance of an XBRL element mapping system, we use XBRL format 10-K annual reports for 82 firms in FY2011 and for all 92 firms in FY 2012 to conduct a comparative performance analysis of X-IM vs. FinCEM.

Table 6 shows the results of our comparative performance analysis of X-IM vs. FinCEM. X-IM outperformed FinCEM for both fiscal year 2011 and 2012 in terms of the overall precision, recall, and *F*-measure. Particularly noteworthy is that X-IM achieved outstanding performance in terms of its precision; the overall precision ratios of all 18 XBRL elements are 99% for both 2011 and 2012. The results support our proposition that both label terms and designative information (period type, balance type, etc.) enable X-IM to interpret and map the heterogeneous XBRL elements over and above the state of the art. High precision is especially desirable in the context of financial information retrieval for business decision-making.

To experimentally test the hypotheses H2a, H2b, and H2c, we conducted another Wilcoxon signed-ranks test, examining the significance of the mean difference between two sets of observations (X-IM vs. FinCEM). We tested 36 observations for all 18 financial concepts in the combined data set of year 2011 and 2012. Table 7 presents the descriptive statistics of our test. As shown in the table, the experiment results reveal that X-IM provides better overall performance than FinCEM in terms of precision (0.998 compared with

0.878), recall (0.896 compared with 0.752), and *F*-measure (0.937 compared with 0.796). As shown in Table 8, the mean difference of precision is statistically significant (mean = 0.119, $Z = 2.519$, $p = 0.012$) at the significance level of 0.05, thus corroborating H2a. Despite the noticeable increase in the overall recall ratio by X-IM, the mean difference of recall (mean = 0.143, $Z = 1.606$, $p = 0.108$) is not statistically significant at the significance level of 0.05, leaving H2b unsupported. However, the Wilcoxon test results show that the mean difference of the *F*-measure is statistically significant at the significance level of 0.05 (mean = 0.141, $Z = 2.013$, $p = 0.044$), supporting H2c. The test results clearly demonstrate that our method informed by the representation model of representation theory considerably improves the overall XBRL mapping performance by significantly increasing precision and the *F*-measure. A plausible explanation is that the natural language information in label linkbases (employed in X-IM) is more discriminative with respect to the designation of a financial concept

Table 4. Descriptive Statistics

Measurement	**Investor's ontology**	***N***	**Mean**	***SD***
Precision	with	18	0.998	0.009
	without	18	0.962	0.106
Recall	with	18	0.869	0.111
	without	18	0.699	0.252
F	with	18	0.923	0.070
	without	18	0.779	0.202

Table 5. Wilcoxon Signed-Ranks Test for Equality of Means

Measurement (without vs. with)	**Differences**		***Z***	***p*-value**
	Mean	***SD***		
Precision	0.036	0.100	1.604	0.109
Recall	0.171	0.227	2.934	0.003**
F	0.143	0.183	2.934	0.003**

Table 6. Evaluation Results of X-IM and FinCEM (FY 2011 and 2012)

Fiscal year	**2011**						**2012**					
Design	**X-IM**			**FinCEM**			**X-IM**			**FinCEM**		
Measurement	P	R	F	P	R	F	P	R	F	P	R	F
Stakeholders' equity	1.00	1.00	1.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.54	0.70
Accounts payable	1.00	0.85	0.92	0.92	0.84	0.87	1.00	0.81	0.90	0.92	0.94	0.93
Account receivable	1.00	0.90	0.95	1.00	0.90	0.95	1.00	0.80	0.89	1.00	0.96	0.98
Current assets	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Total liabilities	1.00	1.00	1.00	0.89	1.00	0.94	1.00	0.97	0.98	0.89	1.00	0.94
Total current liabilities	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00
Deferred tax assets	1.00	0.96	0.98	0.00	0.00	0.00	1.00	0.86	0.92	0.00	0.00	0.00
Common stock	1.00	0.96	0.98	1.00	0.98	0.99	1.00	0.86	0.93	1.00	0.99	0.99
Short-term marketable securities	1.00	1.00	1.00	1.00	0.47	0.63	1.00	0.83	0.91	1.00	0.38	0.55
Inventory	1.00	1.00	1.00	1.00	0.92	0.96	1.00	0.97	0.99	1.00	0.91	0.95
Total assets	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Cost of goods sold	1.00	0.79	0.88	1.00	0.72	0.84	1.00	0.84	0.91	1.00	0.86	0.92
Interest payment	1.00	1.00	1.00	0.00	0.00	0.00	1.00	0.87	0.93	1.00	0.86	0.92
Operating income	1.00	0.42	0.59	1.00	0.42	0.59	1.00	0.59	0.74	1.00	0.68	0.81
Net income	1.00	0.97	0.99	1.00	0.57	0.72	1.00	0.88	0.94	1.00	0.58	0.73
Cash generated by operating activities	1.00	0.93	0.96	1.00	1.00	1.00	1.00	0.81	0.89	1.00	1.00	1.00
Long-term debt	0.96	0.99	0.97	1.00	1.00	1.00	0.96	0.69	0.80	1.00	1.00	1.00
Total revenues	1.00	0.82	0.90	1.00	0.78	0.88	1.00	0.88	0.89	1.00	0.78	0.88
Overall	0.998	0.922	0.951	0.823	0.700	0.743	0.998	0.869	0.923	0.934	0.804	0.850

Notes: P = precision, R = recall, F = *F*-measure

Table 7. Comparisons on Descriptive Statistics of X-IM and FinCEM

Measurement	Model	Mean	*N*	*SD*
Precision	X-IM	0.998	36	0.092
	FinCEM	0.878	36	0.316
Recall	X-IM	0.896	36	0.129
	FinCEM	0.752	36	0.327
F	X-IM	0. 937	36	0.085
	FinCEM	0.796	36	0.311
Note: Observations from all 18 terms in fiscal year 2011 and 2012				

Table 8. Wilcoxon Signed-Ranks Test for Equality of Means

Measurement (X-IM vs. FinCEM)	Differences		*Z*	*p*-value
	Mean	*SD*		
Precision	0.119	0.317	2.519	0.012*
Recall	0.143	0.342	1.606	0.108
F	0.141	0.323	2.013	0.044*

Table 9. *F*-Measure with Training Sizes of 10, 20, and 40 Companies

Concept	*F*-Measure		
	40	20	10
Stakeholders' equity	1.00	1.00	1.00
Accounts payable	0.91	0.90	0.90
Account receivable	0.92	0.92	0.89
Current assets	1.00	1.00	1.00
Total liabilities	0.97	0.97	0.98
Total current liabilities	0.99	0.99	0.99
Deferred tax assets	0.91	0.91	0.92
Common stock	0.98	0.93	0.93
Short-term marketable securities	0.85	0.85	0.91
Inventory	0.99	0.99	0.99
Total assets	1.00	1.00	1.00
Cost of goods sold	1.00	1.00	0.91
Interest payment	0.94	0.94	0.93
Operating income	0.74	0.74	0.74
Net income	0.94	0.94	0.94
Cash generated by operating activities	0.89	0.89	0.89
Long-term debt	0.78	0.78	0.80
Total revenues	0.89	0.89	0.89
Average *F*-measure	0.928	0.924	0.923

6.5 Experiment 4: Effect of Training Data Sizes

To compare our artifact with the state of the art, FinCEM, this study trains X-IM by using label linkbases from 10 randomly selected companies in the S&P100. However, incorporating more label terms from more companies may assist X-IM in improving the performance of interpreting XBRL elements and mapping them. Therefore, this study takes a further step to examine and analyze the effect of training data sizes on the X-IM performance. In this experiment, we generated three training corpora, each of varying size. Using 2011 label linkbases from randomly selected S&P100 companies, we generated a 10-company sample, a 20-company sample, and a 40-company sample for training. For this experiment, we used 2012 label linkbases for testing. The values of the *F*-measure regarding each training set are listed in Table 9.

The results show that the overall *F*-measure does not improve when the training set expands from 10 to 20 companies. We only observed a slight increase of 0.542% from 0.923 to 0.928 when the training set was expanded to 40 companies. However, as the training set expanded, the computational time cost increased dramatically (up to 51.74%), from 117.9s to 140.7s and then to178.9s. We conclude that a training corpus of 10 firms is efficient and effective for our purposes.

7 Conclusion and Discussion

This study makes significant contributions toward true XBRL interoperability. As implemented in the US reporting jurisdiction, XBRL filings suffer from terminological ambiguity across firms' filings. The interoperability problem critically precludes full automaticity in the business reporting pipeline. Downstream consumers of financial reports still do not have open source options for automated cross-firm comparisons of financial data. The information symmetry promise of XBRL, therefore, has not been fulfilled. We believe that the work presented in this paper offers a viable, practical solution to this problem.

We contribute to theory by presenting a tangible information technology artifact that instantiates the representation model of representation theory and demonstrate that by reducing construct deficit and construct redundancy in the taxonomy of a data standard, instances generated by the standard become more interoperable. Our work also constitutes one of several kernel theories for reducing construct deficit and construct redundancy toward more interoperable representations. Specifically, we argue and decisively illustrate that there is rich semantic information encoded within label linkbases. This information is capable of providing mappings between disparately termed but semantically identical XBRL elements using an upper-level (investor's) ontology.

The SEC's XBRL mandate has not fared well in practice. While upstream entities in the financial reporting pipeline participate in XBRL report production by mandate, downstream consumers of financial reports have choices. We indicated that the SEC itself is not a downstream consumer of XBRL data and that data quality issues underscored by the semantic heterogeneity problem are clearly at fault. Solutions such as those proposed here are critical for the continued survival of the mandate. To be sure, upstream participants incur significant costs as a result of the mandate, whereas there is little or no evidence that downstream consumers actually use the standard. Yet the potential to democratize the availability of structured is undeniable (currently users have to pay for expensive databases, which thus favors institutional investors). As such, systems such as X-IM may prove critical in yielding downstream value from the mandate, thus justifying its tenability going forward. As the source for X-IM is open, we hope to contribute to openly available software that leverages openly available data (XBRL filings) to provide structured financial data to noninstitutional and institutional consumers.

One potential limitation of our research is the fact that some financial concepts are not represented in financial reports. For those concepts, it may be possible to access them by making certain calculations. Therefore, a future research direction would be the augmentation of the artifact with further inference ability to calculate the absent financial concepts in a specific report. While the X-IM system employs label linkbases, our method could be further augmented with topic analysis of applicable definitions of XBRL elements from various accounting education resources, which would be another promising area for future development. Another future research avenue would be the construction of a multilevel learner that leverages the information in both label linkbases and calculation linkbases.

Finally, we acknowledge the potential of our design and approach to other semantic integration research. The applicability of our method in the artifact goes beyond the XBRL interoperability problem and proposes an alternative solution to existing ontology-based semantic integration approaches. The efficacy and efficiency of our proposed approach rely on one core condition—that there exists sufficient parallel nominal information with respect to one concept. When this core condition is met, our proposed methods offer domains efficacy and efficiency.

References

Alaya, M. Ben, Medjiah, S., Monteil, T., & Drira, K. (2015). Toward semantic interoperability in oneM2M architecture. *IEEE Communications Magazine*, *53*(12), 35-41.

Bao, J., Rong, G., Li, X., & Ding, L. (2010). Representing financial reports on the semantic web. In M. Dean, J. Hall, A. Rotolo, & S. Tabet (Eds.), *Semantic web rules* (pp. 144-152). Springer.

Bartley, J., Al-Chen, Y. S., & Taylor, E. (2010). Avoiding common errors of XBRL implementation. *Journal of Accountancy*, *209*(2), 46-51.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*(5), 34-43.

Burton-Jones, A., Recker, J., Indulska, M., Green, P., & Weber, R. (2017). Assessing representation theory with a framework for pursuing success and failure. *MIS Quarterly*, *41*(4), 1307-1333.

Chen, R., Sharman, R., Chakravarti, N., Rao, H. R., & Upadhyaya, S. J. (2008). Emergency response information system interoperability: Development of chemical incident response data model. *Journal of the Association for Information Systems*, *9*(3), 200-230.

Chowdhuri, R., Yoon, V. Y., Redmond, R. T., & Etudo, U. O. (2014). Ontology based integration of XBRL filings for financial decision making. *Decision Support Systems*, *68*, 64-76.

Chowdhuri, R., Yoon, V. Y., Redmond, R. T., & Etudo, U. O. (2014). Ontology based integration of XBRL filings for financial decision making. *Decision Support Systems*, *68*, 64-76.

Debreceny, R. S., Farewell, S. M., Piechocki, M., Felden, C., Gräning, A., & d'Eri, A. (2011). Flex or break? Extensions in XBRL disclosures to the SEC. *Accounting Horizons*, *25*(4), 631-657.

Declerck, T., & Krieger, H.-U. (2006). Translating XBRL into description logic. An approach using Protege, Sesame & OWL. *Proceedings of the International Conference on Business Information Systems*.

Doan, A., Madhavan, J., Domingos, P., & Halevy, A. (2002). Learning to map between ontologies on the semantic web. *Proceedings of the Eleventh International World Wide Web Conference.*

Dou, D., Qin, H., & Lependu, P. (2010). OntoGrate: Towards automatic integration for relational databases and the semantic web through an ontology-based framework. *International Journal of Semantic Computing*, *4*(1), 123-151.

Engel, P., Hamscher, W., Shuetrim, G., Kannon, D., & Wallis, H. (2013). *Extensible business reporting language (XBRL) 2.1*. http://www.xbrl.org/Specification/XBRL-2.1/REC-2003-12-31/XBRL-2.1-REC-2003-12-31+corrected-errata-2013-02-20.html

Ernst & Young LLP. (2017). *Re: Inline XBRL filing of tagged data (Release Nos. 33-10323, 34-80133; file no . S7-03-17)*. https://www.sec.gov/comments/s7-03-17/s70317-1755258-152006.pdf

Etudo, U., & Yoon, V. (2015). Leveraging XBRL calculation linkbases to overcome semantic heterogeneity across XBRL fillings: the multi-ontology multi-concept matrix (M3). *Proceedings of the International Conference on Information Systems: Exploring the Information Frontier*.

Etudo, U., Yoon, V., & Liu, D. (2017a). Financial concept element mapper (FinCEM) for XBRL interoperability: Utilizing the M3 Plus method. *Decision Support Systems*, *98*, 36-48.

Etudo, U., Yoon, V., & Liu, D. (2017b). Financial concept element mapper (FinCEM) for XBRL interoperability: Utilizing the M3 Plus method. *Decision Support Systems*, *98*, 36-48.

García, R., & Gil, R. (2009). Publishing XBRL as linked open data. *CEUR Workshop Proceedings*.

Gefen, D., & Larsen, K. (2017). Controlling for lexical closeness in survey research: A demonstration on the technology acceptance model. *Journal of the Association for Information Systems*, *18*(10), 727-757.

Gibbons, J., & Chakraborti, S. (2011). Nonparametric Statistical Inference. In *International Encyclopedia of Statistical Science* (pp. 977-979). Springer.

Gregor, S., & Jones, D. (2007). The anatomy of a design theory. *Journal of the Association for Information Systems*, *8*(5), 312-335.

Halevy, A., Ordille, J., & Rajaraman, A. (2006). Data integration : The teenage years. *Artificial Intelligence*, *41*(1), 9-16.

Hammer, J., & McLeod, D. (1993). An approach to resolving semantic heterogenity in a federation of autonomous, heterogeneous database

systems. *International Journal of Cooperative Information Systems*, *2*, 51-83.

Heiler, S. (1995). Semantic Interoperability. *ACM Computing Surveys*, *27*(2), 271-273.

Henry, E., Liu, F.-C., Yang, S. Y., & Zhu, X. (2018). *Structural comparability of financial statements* (Stevens Institute of Technology School of Business Research paper). https://doi.org/10.2139/ssrn.3133324

Hoffelder, K. (2013). House to SEC : Use XBRL. *CFO Magazine*, *November*, 16-17.

Hu, W., Chen, J., Zhang, H., & Qu, Y. (2012). Learning complex mappings between ontologies. In J. Z. Pan et al. (Eds.), *The semantic web* (pp. 350-357). Springer.

Kaza, S., & Chen, H. (2008). Evaluating ontology mapping techniques: An experiment in public safety information sharing. *Decision Support Systems*, *45*(4), 714-728.

Kieso, D., Weygandt, J. J., & Warfield, T. (2013). *Intermediate Accounting* (15th ed.). Wiley.

Lin, H. K., Harding, J. A., & Shahbaz, M. (2004). Manufacturing system engineering ontology for semantic interoperability across extended project teams. *International Journal of Production Research*, *42*(24), 5099-5118.

Livieri, B., Zappatore, M., & Bochicchio, M. (2014). Towards an XBRL ontology extension for management accounting. *Proccedings of the 33rd International Conference on Conceptual Modeling*.

Luna-Reyes, L. F., Zhang, J., Ramón Gil-García, J., & Cresswell, A. M. (2005). Information systems development as emergent socio-technical change: A practice approach. *European Journal of Information Systems*, *14*(1), 93-105.

Nagarajan, M., Verma, K., Sheth, A. P., Miller, J., & Lathem, J. (2006). Semantic interoperability of Web services: Challenges and experiences. *Proceedings of the International Conference on Web Services*.

Narock, T., Yoon, V., & March, S. (2014). A provenance-based approach to semantic web service description and discovery. *Decision Support Systems*, *64*, 90-99.

Nunamaker, J. F., Chen, M., & Purdin, T. D. (1990). Systems development in information systems research. *Journal of Management Information Systems*, *7*(3), 89-106.

O'Riain, S., Curry, E., & Harth, A. (2012). XBRL and open data for global financial ecosystems: A linked data approach. *International Journal of Accounting Information Systems*, *13*(2), 141-162.

Parundekar, R., Knoblock, C. A., & Ambite, J. L. (2012). Discovering concept coverings in ontologies of linked data sources. *Proceedings of the International Semantic Web Conference*.

Peffers, K., Tuunanen, T., Rothenberger, M. a., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45-77.

Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*(1), 37-63.

Qin, H., Dou, D., & LePendu, P. (2007). Discovering executable semantic mappings between ontologies. *Proceedings of the OTM Confederated International Conferences: On the Move to Meaningful Internet Systems*.

Radzimski, M., Sanchez-Cervantes, J. L., Garcia-Crespo, A., & Temiño-Aguirre, I. (2014). Intelligent architecture for comparative analysis of public companies using semantics and XBRL data. *International Journal of Software Engineering and Knowledge Engineering*, *24*(05), 801-823.

Raggett, D. (2009). How can we exploit XBRL and Semantic Web technologies to realize the opportunities? *Proceedings of the 19th XBRL International Conference*.

Recio-García, J. a., Quijano, L., & Díaz-Agudo, B. (2013). Including social factors in an argumentative model for group decision support systems. *Decision Support Systems*, *56*(1), 48-55.

SEC. *Interactive data to improve financial reporting* (2009). https://www.sec.gov/rules/final/2009/33-9002.pdf

SEC. (2010). *Important information about EDGAR*. https://www.sec.gov/edgar/aboutedgar.htm

Spies, M. (2010). An ontology modelling perspective on business reporting. *Information Systems*, *35*(4), 404-416.

Thiéblin, É., Haemmerlé, O., Hernandez, N., & Trojahn, C. (2018). Survey on complex ontology matching. *Semantic Web*, http://www.semantic-web-journal.net/content/survey-complex-ontology-matching

Ure, J., Hartswood, M., Wardlaw, J., Procter, R., Anderson, S., Lin, Y., … Ho. (2009). The development of data infrastructures for

eHealth: A socio-technical perspective. *Journal of the Association for Information Systems*, *10*(5), 415-429.

Wache, H., Voegele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., & Hübner, S. (2001). Ontology-based integration of information: A survey of existing approaches. *Proceedings of IJCAI-01 Workshop: Ontologies and Information Sharing*.

Wand, Y., & Weber, R. (1989). A model of systems decomposition. *Proceedings of the International Conference on Information Systems*.

Wand, Y., & Weber, R. (1993). On the ontological expressiveness of information systems analysis and design grammars. *Information Systems Journal*, *3*(4), 217-237.

Wand, Y., & Weber, R. (1995). On the deep structure of information systems. *Information Systems Journal*, *5*(3), 203-223.

Williams, A. J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E. L., … Mons, B. (2012). Open PHACTS: Semantic interoperability for drug discovery. *Drug Discovery Today*, *17*(21-22), 1188-1198.

Wunner, T., Buitelaar, P., & O'Riain, S. (2010). Semantic, terminological and linguistic interpretation of XBRL. *Proceedings of the Reuse and Adaptation of Ontologies and Terminologies Workshop at the 17th International Conference on Knowledge Engineering and Knowledge Management*.

Yaghoobirafi, K., & Nazemi, E. (2019). An approach to XBRL interoperability based on ant colony optimization algorithm. *Knowledge-Based Systems*, *163*, 342-357.

Yu, J., Thom, J. A., & Tam, A. (2009). Requirements-oriented methodology for evaluating ontologies. *Information Systems*, *34*(8), 766-791.

Zhu, H. H., & Madnick, S. (2007). *Semantic integration approach to efficient business data supply chain: Integration approach to inter-operable XBRL* (Working Paper CISL No. 2007-10, MIT Sloan School of Management).

Zhu, H., & Wu, H. (2011). Quality of data standards: Framework and illustration using XBRL taxonomy and instances. *Electronic Markets*, *21*(2), 129-139.

Zhu, H., & Wu, H. (2014). Assessing the quality of large-scale data standards: A case of XBRL GAAP Taxonomy. *Decision Support Systems*, *59*, 351-360.

Appendix A

Table A1. References of Equivalent Investor's Terms Appearing in Investor's Ontology Based on Kieso et al., 2013

Financial concept	Balance|period	Equivalent investor's terms
Stakeholders' equity	Credit|instant	Stockholders' equity (p.89)
		Total stockholders' equity (p.106)
		Total shareholders' equity (p.250)
Accounts payable	Credit|instant	Accounts payable (p.96, p.250)
Accounts receivable	Debit|instant	Accounts receivable (p.351)
		Receivables (p.15)
Current assets	Debit|instant	Total current assets (p.116, p.416)
Total liabilities	Credit|instant	Total liabilities (p.110)
Total current liabilities	Credit|instant	Total current liabilities (p.116)
Deferred tax assets	Debit|instant	Deferred tax assets (p.283, p.1125)
		Deferred income taxes (p.416)
Common stock	Credit|instant	Common stock (p.89, p.109)
Short-term marketable securities	Debit|instant	Short-term marketable securities (p.571)
		Marketable securities (p.272)
		Short-term investments (p.219)
Inventory	Debit|instant	Inventory (p.416)
		Finished products (p.200)
		Finished goods (p.250)
Total assets	Debit|instant	Total assets (p.110)
Cost of goods sold	Debit|duration	Cost of goods sold (p.115)
		Cost of sales (p.210)
		Cost of products sold (p.249)
Interest payment	Debit|duration	Interest payment (p.320)
		Interest expense (p.106, p.139)
Operating income	Credit|duration	Operating income (p.170)
		Operating profit (p.152)
		Operating earnings (p.166)
		Income from operations (p.178)
Net income	Credit|duration	Net income (loss) (p.276)
		Net income from continuing operations (p.151)
		Net earnings (p.249)
		Net income (p.178, p.110)
Cash generated by operating activities	NA|duration	Net cash provided by operating activities (p.246, p.234)
		Net cash flow from operating activities (p.585, p.1351)
Long-term debt	Credit|instant	Long-term debt (p.214, p.216)
Total revenues	Credit|duration	Revenues (p.109)
		Net sales (p.167)

Appendix B

Table B1. Example of Similarity Calculations: Financial Concept, XBRL Elements, and a Vector of Label Terms Used by Various Firms

Financial concept	**Net income**	
XBRL element	**us-gaap_profitloss**	**us-gaap_netincomeloss**
	0.0^{++}	**0.667**$^{++}$
Label terms used by various firms	Net income/(loss), 0.333* Net income including noncontrolling interest, 0.4* Net earnings, 0.333* Net income before allocation to noncontrolling interests, 0.286* Net income including noncontrolling interest, 0.4* Net income (loss), 0.667* Net earnings (loss), 0.25* Net income, 1.0* Consolidated net income, 0.667* Net earnings including noncontrolling interests, 0.167* Profit of consolidated and affiliated companies, 0.0* Net income from consolidated operations, 0.4* Net income including noncontrolling interests, 0.4*	Net income/(loss) attributable to Ford Motor Company, 0.125* Net earnings, 0.333* Net earnings common stockholders, 0.2* Net income, 1.0* Net earnings (loss), 0.25* Wells Fargo net income, 0.5* Net income (loss), 0.667* Net income attributable to common shareowners, 0.333*
	0.408**	**0.379****

Notes:
++ A Jaccard similarity between the financial concept and the XBRL element
* A Jaccard similarity between the financial concept and the label term
** An average of Jaccard similarities between the financial concept and a vector of label terms

About the Authors

Dapeng Liu is a lecturer in the School of Information Systems and Technology Management at the University of New South Wales, Sydney. He received his PhD in business with an information systems concentration from Virginia Commonwealth University in 2019. His research interests include business intelligence, knowledge modeling and management, information security and privacy, and e-government adoption. His work has been supported by the Microsoft Azure Research Award and VCU PeRQ fund. His research appears in *Decision Support Systems*.

Ugochukwu Etudo is an assistant professor of operations and information management at the University of Connecticut School of Business. He received his PhD in business with an information systems concentration from Virginia Commonwealth University in 2016. His research investigates pro- and antisocial uses of online social platforms. Ugo's work has investigated the detection of terroristic framing on the web, information warfare and online social movements. Ugo's methodological interests center on the application of deep and machine learning to natural language processing tasks and Semantic Web technologies. His work has been published in *Decision Support Systems*.

Victoria Yoon is a professor in the Department of Information Systems at the Virginia Commonwealth University. She received her MS from the University of Pittsburgh and her PhD 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*, *Journal of Management Information Systems*, *Decision Support Systems, and Communications of the ACM*.