Semantic, Terminological and Linguistic Interpretation of XBRL

Tobias Wunner*°, Paul Buitelaar*, Sean O'Riain°

*Unit for Natural Language Processing & °eBusiness Unit
Digital Enterprise Research Institute,
National University of Ireland, Galway
firstname.lastname@deri.org

Abstract. Standardization efforts in financial reporting have led to large numbers of machine-interpretable vocabularies that attempt to model complex accounting practices in XBRL (eXtended Business Reporting Language). Because reporting agencies do not require fine-grained semantic and terminological representations, these vocabularies cannot be easily reused. Ontology-based Information Extraction, in particular, requires much greater semantic and terminological structure, and the introduction of a linguistic structure currently absent from XBRL. In order to facilitate such reuse, we propose a three-faceted methodology that analyzes and enriches the XBRL vocabulary: (1) transform semantic structure by analyzing the semantic relationships between terms (e.g. taxonomic, meronymic); (2) enhance terminological structure by using several domain-specific (XBRL), domain-related (SAPTerm, etc.) and domain-independent (GoogleDefine, Wikipedia, etc.) terminologies; and (3) add linguistic structure at term level (e.g. part-of-speech, morphology, syntactic arguments). This paper outlines a first experiment towards implementing this methodology on the International Financial Reporting Standard XBRL vocabulary.

Key words:ontology-based information extraction, accounting language, XBRL semantics, XBRL terminology, linguistic analysis

1 Introduction

Accounting has its own domain-specific language, developed over time in response to national and international legal requirements. This language facilitates a common understanding across accounting community. Therefore, these agreements, known as *Generally Accepted Accounting Principles* (GAAP), are very complex in their semantics and have a highly stylized terminology. While the term "lease"[1], for instance, once

[1] Merriam-Webster Dictionary: *contract by which one conveys real estate, equipment, or facilities for a specified term and for a specified rent; also the act of such conveyance or the term for which it is made*

referred to a single concept, to describe the debt state between a lessee and a lessor, the current US-GAAP contains 516 different semantic specifications, which come in complex terminology such as *"minimum operating lease payments, recognized finance lease as assets"*. In order to promote standardization and automation, specifically for the use-case of financial reporting these GAAPs have been implemented in formalized machine-interpretable semantic vocabularies in XBRL[2] (eXtensible Business Reporting Language). Currently there exist worldwide over 50 different XBRL vocabularies which provide a vast amount of terminological and semantic information for the accounting domain. However, even with some semantics and terminology contained in XBRL vocabularies, because reporting agencies (e.g. SEC[3]) only require specific semantic characteristics (i.e. calculation formulas, presentation layout and hierarchy of concepts) these vocabularies do not come with full-featured semantics and cannot easily be reused in other use-cases. Therefore we developed a three-faceted methodology to enrich the (1) semantic (e.g. taxonomic, meronymic, synonymic, etc.), (2) terminological (term inclusion, sub-term structure, etc.) and linguistic (e.g. syntax, morphology, etc.) structure of the XBRL vocabulary. This enriched vocabulary can then be used for Ontology-based Information Extraction (OBIE), which requires a more complete and fine-grained semantic, terminological and linguistic analysis of the vocabulary. This methodology has been developed in the context of the Monnet[4] (Multilingual Ontologies for Networked Knowledge) project, which has financial and business domain use cases around XBRL, in particular OBIE from financial reports.
A typical OBIE task involves extracting information from a financial report. A financial report is a document with structured and unstructured information comprising of three closely related parts:

1. Financial Statement: highly structured with mostly tables
2. Notes: less structured part with short factual text and tables linked to concepts in the financial statement
3. Disclosure: mostly unstructured part with additional company information in long prosaic text

In order to extract meaningful facts from the unstructured i.e., text parts of such documents, it is necessary to generate all possible term variations of an XBRL concept (see also [1]). If we, for example, consider the term

[2] http://www.xbrl.org
[3] Securities and Exchange Commission (SEC), the US reporting agency
[4] http://www.monnet-project.eu

"income tax rate benefit", the following variations of the term might occur in a text:

Semantic (synonymic) *"income tax rate bonus"*
Terminological (acronymic) *"inc tax rate benefit"*
Linguistic (Syntax) *"benefit on income tax rate"*
Linguistic (Morphology) *"income tax rate benefits"*

In order to not confuse the reader with the frequently used expressions to describe aspects of XBRL vocabularies we define them as follows:

Definition 1 (word). *A word is a lexical unit which consists of a token (e.g. "tax", "taxes", "net", "cash") and refers to the linguistic level.*

Definition 2 (term). *A term is a word or a sequence of words which can again contain other terms (e.g. "net", "net-cash", "finance", "finance lease", etc.) , also called sub-terms or term inclusion, and describes entities at a terminological level.*

Definition 3 (terminology). *A terminology contains a set of terms and often describes a domain: e.g.*
FinanceDomain= {*finance*,*finance lease*}

Definition 4 (concepts and properties). *Concepts and properties are formal semantic descriptions of entities and their relationships independent of language. They can be described by a set of formal operators and relations of the formal language (e.g. negation, transitivity, taxonomic or meronymic relationships, etc.). Consider for example the concepts* c_{lessee} *and* c_{lessor} *which can be related by use of the property* $p_{leasepayments}$*. Then* $p_{leasepayments}$ *is a property which takes two arguments. Note that the terms associated with concepts and properties are often encoded by the variable name of the concept (e.g. "lease payments" for* $p_{leasepayments}$*).*

Definition 5 (ontology). *An ontology is a structure that consists of concepts and properties. An ontology with only one property namely the sub-class relationship is also called taxonomy* [5]*.*

[5] Note that the XBRL community uses the expression taxonomy to refer to a XBRL vocabulary which has specific semantic and terminological properties i.e., financial properties (referring to rules for classification), properties to define the label and the definition for terms and properties to encode the textual representation of these terms.

Definition 6 (vocabulary). *A vocabulary is a structure which describes semantic, terminological and linguistic aspects of a domain. When the vocabulary only describes semantic aspects it is the same as an ontology. When the vocabulary only describes terminological aspects it is the same as a terminology. When the vocabulary only describes linguistic aspects it is referred to as a dictionary or lexicon.*

The paper is structured in the following parts. First, in Section 2, we discuss related work. In Section 3 we propose a methodology for enriching the structure of XBRL vocabularies in a three-faceted (semantic, terminological and linguistic) way. In Section 4 we apply our methodology in an exploratory experiment on an international used XBRL vocabulary. In Section 5 we discuss the preliminary results mainly focusing on the terminological aspects In the last section we outline future work.

2 Related Work

In respect to Semantic Web applications of XBRL there are several studies in making its semantics explicit through a transformation into OWL/RDFS ([2], [3]), however these approaches are limited to the XML semantics rather than addressing the semantics of the underlying financial and business concepts. This may be the reason why the XBRL data set has not been used in the Semantic Web and Linked Data community so far.
However some semantic aspects of XBRL have been studied on the level of XML schemas, e.g., the analysis of semantic interoperability across different XBRL vocabularies ([4]) as well as the compatibility between different XBRL vocabularies from a business point of view, with an emphasis on the harmonization of the UK and US GAAP (Abdullah et al. [5]). The main conclusion of this paper was that harmonization requires further terminological and semantic analysis of these vocabularies to be undertaken.
Semantic, terminological and linguistic analysis of terms in the biomedical domain has been explored by Verspoor [6] by merging several domain-specific lexical resources The results of this study have shown that domain-specific lexical resources are of importance to domain-specific NLP tasks. An analysis of ontology labels in the biomedical domain using a semantic, terminological and linguistic lexicon model has been carried out by Afzal et al. [7], also finding that ontologies have rich implicit terminological and linguistic structure that needs to be made explicit for these use-cases.
The contribution of our paper builds on these approaches in establishing a deeper understanding of XBRL semantics through a methodology

for a layered analysis of its implicit semantics, terminology and linguistic aspects.

3 A Three-faceted Interpretation of XBRL

To enrich the XBRL domain vocabularies and enhance their semantic, terminological and linguistic structure we propose the following methodology:

1. Transform the vocabulary to RDF semantics (RDFS) and enhance its semantics by analyzing implicit business semantic relationships.
2. Interpret XBRL from the terminology point of view by decomposing XBRL terms in an incremental fashion. Decomposition starts with the XBRL vocabulary itself, i.e. decomposition of XBRL terms into XBRL sub-terms, followed by decomposition with domain related terminology resources and finally with general language resources.
3. Enrich the vocabulary with linguistic information and structure such as part-of-speech, morphology, syntactic arguments, etc. by analyzing them with lexical resources and linguistically annotated text corpora (i.e. a data set of financial reports).

An overview of this process is given in Figure 1, highlighting the technical presentation of the analysis results of each step. On the left (1) the original XBRL vocabulary in XML Linkbase is shown. This is then analyzed in a three-faceted way and results in: (2) the transformed semantics (OWL/RDF) and (3+4) terminological and linguistic structure encoded in a lexicon.

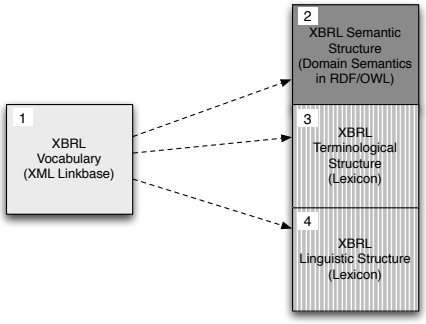

Fig. 1. Lexicalized XBRL as a three-step enrichment process.

3.1 Semantic Analysis and RDFS Transformation

In a semantic analysis step we attempt to enhance XBRL with explicit business semantics by making implicit business semantics explicit. In current work we are exploring the following first steps towards this, which are concerned with two aspects. First, on a more technical level, the XBRL vocabulary, which is encoded in XML Linkbase format, is transformed to RDF semantics (RDFS). Secondly, we map XBRL properties, by applying heuristics, to their most meaningful RDFS counterparts. This is done by a combination of manually defined heuristic rules and a subsequent manual selection of matched semantic relationships. We identified two XBRL properties (xbrl:parent-child, xbrl:summation-item), which we mapped to RDF-based semantics (rdfs:subClass and skos:relatedPartOf) as shown in Figure 2. The figure shows the semantic context of the example term *"minimum finance lease payments receivable, at present value, end of period not later than one year."*, its semantic transformation process and heuristic transformation rules. This example is taken from the International Finance Reporting Standard (IFRS[6]) XBRL vocabulary. Besides the obvious advantage of using RDFS for facilitating a deeper semantic analysis, the use of RDFS in general is also of importance for Semantic Web compliance.

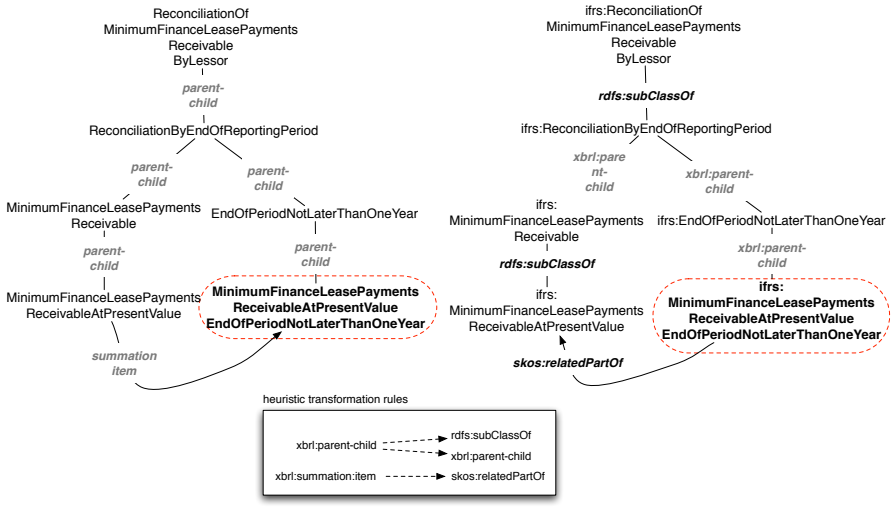

Fig. 2. Semantic analysis of an example term

[6] http://www.iasb.org/IFRSs

3.2 Terminological Analysis

Terms in XBRL vocabularies often consist of complex compositions of either (i) other XBRL vocabulary terms, (ii) more basic financial concepts or (iii) any other domain-independent terms. Therefore we suggest an incremental decomposition approach, which analyzes terms a domain-specific to a domain-independent level, using the following terminological resource types:

1. XBRL domain terminology itself
2. Domain-related terminologies such as financial vocabularies like SAPTerm[7] or MicrosoftTerm[8].
3. Domain-independent terminology resources such as Wikipedia, GoogleDefine or WordNet

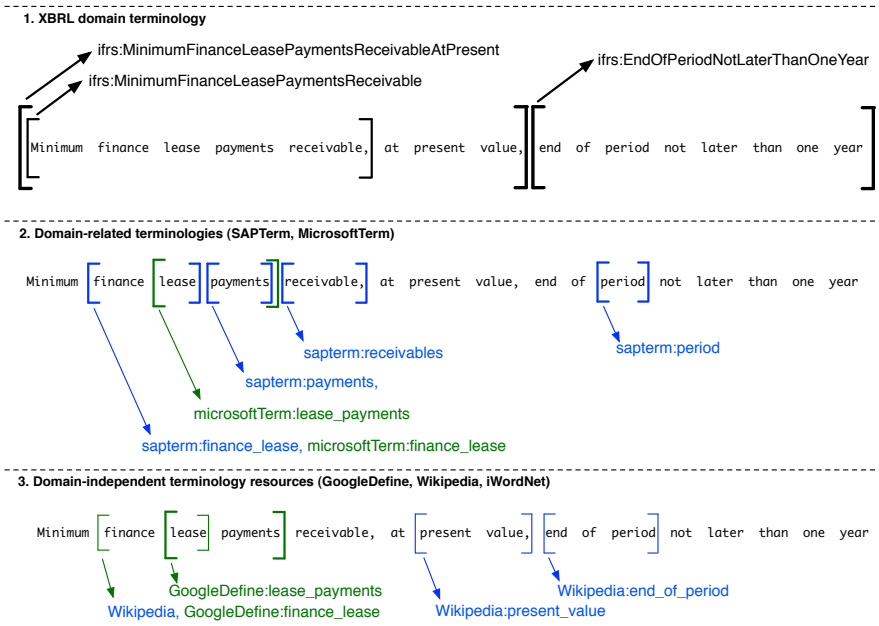

Fig. 3. Terminological analysis of an example term

Figure 3 outlines the terminological analysis of the example term. The analysis first decomposes the term at the most domain-specific level by

[7] http://help.sap.com/saphelp_glossary/en/
[8] http://www.microsoft.com/language/en/us/default.mspx

using XBRL (IFRS) terminology, which results in three sub-terms, i.e. *"minimum finance lease payments receivable"*, *"minimum finance lease payments receivable, at present value"*, *"end of period not later than one year"*. In the next step we used SAPTerm, GoogleDefine and MicrosoftTerm as domain-related terminology resources, which resulted into a further decomposition into five sub-terms, i.e. *"finance lease payments"*, *"lease payments"*, *"payments"*, *"receivable"*, *"period"*. In the last step, we further decompose the term by using domain-independent resources. We identified four more sub-terms. *"finance lease"*, *"lease payments"*, *"present value"* and *"end of period"* via GoogleDefine and Wikipedia.

3.3 Linguistic Information and Structure

In addition to the semantic analysis and terminological decomposition, the XBRL vocabulary will be further enriched with linguistic information and structure.
As briefly explained in the introduction this is needed for the generation and or identification of term variants, e.g. *"minimum finance lease payments receivable"* can also occur in financial reports as *"received minimum finance lease payments"*. Figure 4 illustrates the linguistic enrichment[9] of the example term.

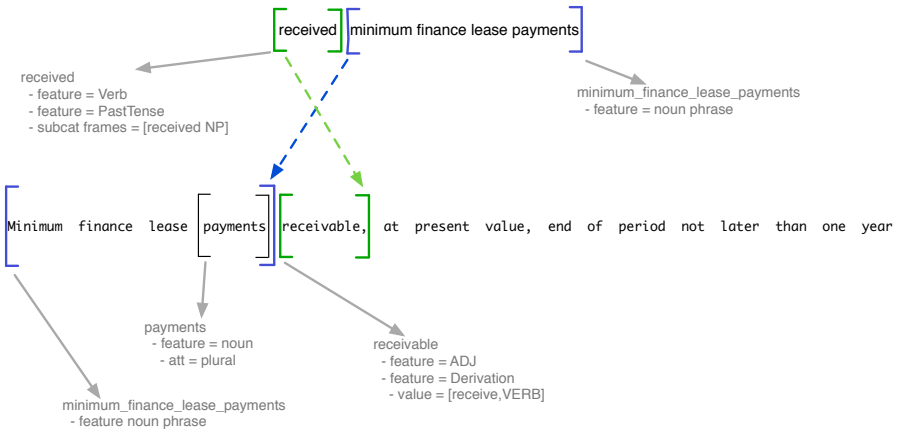

Fig. 4. Linguistic analysis of an example term

[9] For this we currently used Stanford Parser: http://nlp.stanford.edu/software/lex-parser.shtml

4 Experiment

In order to evaluate our approach, we defined an experiment in the analysis of an IFRS based data set. The task of the experiment is to incrementally enrich this data set on a semantic, terminological and linguistic level by use of a systematic decomposition algorithm. As data set we chose the IFRS vocabulary from 2009[10] with English labels. This data set consists of 2766 concepts that are semantically interlinked by a small set of properties[11] of which we list the six most important ones in Table 1. The extent of semantic interlinkage is shown by the number of occurrences of these properties. For the terminological analysis we used the IFRS vocabulary

Property name	URI	Num.
parent-child	http://www.xbrl.org/2003/arcrole/parent-child	13713
summation-item	http://www.xbrl.org/2003/arcrole/summation-item	7092
all	http://xbrl.org/int/dim/arcrole/all	9881
dimension-default	http://xbrl.org/int/dim/arcrole/dimension-default	13
dimension-domain	http://xbrl.org/int/dim/arcrole/dimension-domain	16
domain-member	http://xbrl.org/int/dim/arcrole/domain-member	9845
hypercube-dimension	http://xbrl.org/int/dim/arcrole/hypercube-dimension	16

Table 1. Distribution of semantic properties of XBRL definition, presentation and calculation layer of IFRS XBRL vocabulary.

(2766 terms), SAPTerm (32.875 terms) and WordNet (206.941 terms). A distribution of the number of words per term is given in Table 2.

Interestingly, as shown in Table 3, we can give a terminological structure for 70% of the IFRS vocabulary itself as terminology resource. Using SAPTerm we can provide a terminological structure for 26% of all IFRS terms. And finally the use of WordNet gives us an analysis of sub-terms for 25% of all IFRS terms. If we look closely at those terms not covered by the terminology resources we can observe for instance that many of these include sub-terms with plural forms which indicates a need for deeper, i.e. linguistic analysis.

[10] IFRS is released and published by the IASB (International Accounting Standards Board): http://www.iasb.org

[11] In XBRL Linkbase semantics these are referred to as link roleshttp://www.xbrl.org/lrr/lrr.xml

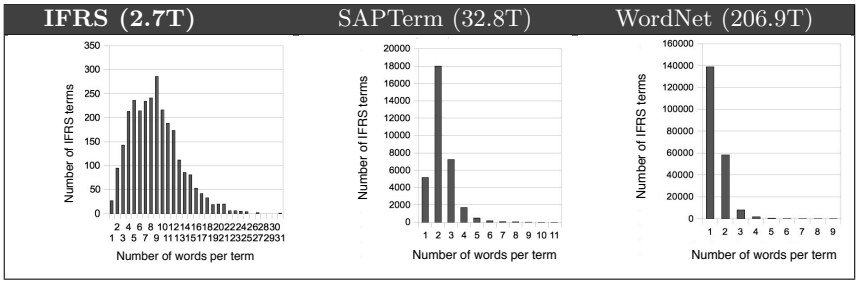

Table 2. Number of words distribution per term in different terminologies.

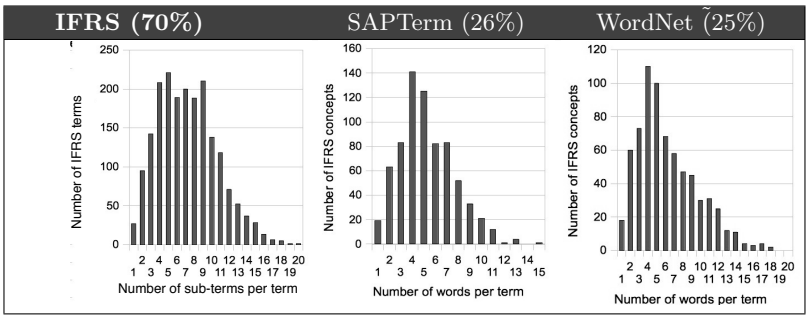

Table 3. IFRS term distribution according to the number of sub-terms explained by IFRS, SAPTerm, WordNet.

5 Conclusions and Future Work

We established a methodology for the systematic analysis of vocabularies with rich terminology and semantics such as XBRL, the use case for which we exemplified by ontology-based information extraction from financial reports. The proposed methodology consists of a three-faceted incremental analysis of vocabularies on the semantic, terminological and linguistic level. In order to verify this approach, experiments are needed with large data sets and with multiple algorithmic configurations, e.g., including or excluding certain resource types (domain-specific, domain-related, domain-independent), individual resources (SAPTerm, GoogleDefine, etc.), analysis methods (use of reasoners, parsers, transformation heuristics etc.). In this paper we made a first attempt at this by analyzing the IFRS data set in a systematic way using SAPTerm as domain-related resource and WordNet as domain-independent resource.

In future work we intend to further analyze the IFRS data set with additional resources, measuring their individual contribution to the semantic,

terminological and linguistic interpretation. The outcome of this process will enable us also to define a benchmark for the ontology-based information extraction task by comparing performance relative to the level of analysis of the vocabulary. A baseline will be formed by the original IFRS vocabulary with increased performance on each analysis level and its configuration. As the IFRS vocabulary is available in various languages and can therefore be used for cross-lingual ontology-based information extraction, we will extend our methodology to also capture this multilingual aspect. Finally, to ensure compatibility with existing Semantic Web standards we encode the semantic analysis results in an RDF ontology using the business semantics as described above. The results of the terminological and linguistic analysis are stored in the Semantic Web compliant lexicon model LexInfo[12] [8], which is linked to the RDF ontology.

Acknowledgements

This work is supported in part by the European Union under Grant No. 248458 for the Monnet project as well as by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

References

1. Federmann, C., Declerck, T.: Extraction, merging, and monitoring of company data from heterogeneous sources. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, European Language Resources Association (ELRA) (may 2010)
2. Declerck, T., Krieger, H.U.: Translating xbrl into description logic: an approach using protege, sesame and owl. In: In Proceedings of Business Information Systems (BISi). (2006)
3. Gil, R.G.R.: Linked data on the web (ldow2009). (April 2009)
4. Zhu, M.e.a.: Semantic integration approach to efficient business data supply chain: Integration approach to interoperable xbrl. Working papers, Massachusetts Institute of Technology (MIT), Sloan School of Management (2008)
5. Abdullah, A., Khadaroo, I., Shaikh, J.: Institutionalisation of xbrl in the usa and uk. International Journal of Managerial and Financial Accounting **1**(3) (2009) 292–304
6. Verspoor, K.: Towards a semantic lexicon for biological language processing: Conference papers. Comp. Funct. Genomics **6**(1-2) (2005) 61–66
7. H. Afzal, P. Buitelaar, P.C.J.M.T.W.: Tke conference 2010: Presenting terminology and knowledge engineering resources online: models and challenges. (August 2010)
8. Buitelaar, P., Cimiano, P., Haase, P., Sintek, M.: Towards linguistically grounded ontologies. In: 6th Annual European Semantic Web Conference (ESWC2009). (June 2009) 111–125

[12] LexInfo (http://lexinfo.net/). is a RDF model for describing linguistic information of ontologies