# Publishing XBRL as Linked Open Data

Roberto García
Universitat de Lleida
Jaume II, 69
25001 Lleida, Spain
+34 973 702 742
rgarcia@diei.udl.cat

Rosa Gil
Universitat de Lleida
Jaume II, 69
25001 Lleida, Spain
+34 973 702 742
rgil@diei.udl.cat

## ABSTRACT

The XML Business Reporting Language (XBRL) is a standard for business and financial information reporting. It is based on XML so instance documents based on XBRL, e.g. a quarterly report, are highly constrained by the XML document-oriented nature. This makes more difficult to perform queries that mix information from filings from different dates, companies, or accounting principles than with a formalism based on a graph model instead of a tree model. Semantic Web technologies provide a graph model that facilitates mashing-up different XBRL sources. We have put into practice this approach mapping the XBRL filings available from the SEC's EDGAR program to Resource Description Framework (RDF) and the XML Schema taxonomies these filings are based on to Web Ontology Language (OWL). The resulting semantic metadata, though highly tied to the XML structure it is mapped from, benefits from Semantic Web technologies and tools in order to facilitate integration and cross-querying, even together with other parts of the Web of Linked Data.

## Keywords

Business, reporting, Semantic Web, Linked Data, Web 3.0, accounting, finance, interoperability.

## 1. INTRODUCTION

XBRL (eXtensible Business Reporting Language) is an XML language intended for modeling, exchanging and automatically processing business and financial information. XBRL is starting to be deployed in many different scenarios. For instance, there is the EDGAR [1] program promoted by the U.S Securities and Exchange Commission (SEC).

It performs automated collection, validation, indexing, acceptance and forwarding of submissions by companies and others who are required by law to file forms with the SEC. Filers may choose to voluntarily submit documents in XBRL format to accompany certain official filings. Three dozen companies, representing more than $1 trillion of market value, have joined the SEC's XBRL test group.

However, we have observed limited support for cross analysis of financial information in XBRL tools and applications, as it is detailed in the Related Work Section. This is not just among data based on different accounting principles, which are represented in XBRL using taxonomies. It even happens when comparing filings for different companies based on the same taxonomies or filings for the same company based on different versions of the taxonomies.

We argue that this limitation is inherited from the technologies underlying XBRL, especially XML. XML takes a document-oriented approach, where each document presents a tree structure. This makes it difficult for XML-based tools to provide functionalities that blur this separation into documents and that overcome the limitations of a tree structure when mashing-up data from different sources. Moreover, XBRL does not provide formal semantics that might help to integrate different taxonomies by using logic reasoners.

In any case, the integration of data contained in XBRL into comparable information is a strong requirement for the analysis of business and financial information at the global level. This might increase the efficiency and effectiveness of the decision making processes relying on this kind of information. For instance, bankruptcy prediction and other tasks related to the assessment of the solvency of a firm, a business sector or set of interrelated companies.

Many have already pointed to this issue and proposed Semantic Web technologies as a natural choice for XBRL data integration, cf. the Related Work Section. However, we think that this is not enough. Semantic Web provides the technologies for data integration but some principles are required that facilitate Web-wide deployment of highly interlinked XBRL data. Linked Data [2] provides these principles to publish data in the World Wide Web in a way that helps making it easily discoverable through the links that connect it to other pieces of data.

Despite these benefits, currently, financial and business data is being produced using XBRL and it seems that more and more XBRL data is going to be available in the future. It is been promoted by regulators and government agencies like the SEC and other entities like the European Union or the Spanish securities commission [3].

Consequently, we think that the best approach in order to get financial and business data to the Semantic Web is not to propose an alternative language based on Semantic Web technologies, but to apply methods to map existing XBRL to semantic metadata. This approach, its results and its validations are presented in the following sections, after XBRL is introduced.

## 2. XBRL

XBRL is based on two kinds of documents, instance documents and taxonomies. Instance documents report business facts and point to a set of taxonomies, which define the meaning of these


*LDOW2009*, April 20, 2009, Madrid, Spain.

facts, e.g. under what accounting principles they hold, what other facts they related to or what kind of things do they refer to.

## 2.1 Instances

More concretely, a XBRL instance document contains business Facts. An example of a Fact could be "sales in the last quarter". If the Fact is simple valued, like "the long term debt is 350,000" whose value is just a number, it is called Item. If the Fact has a more complex value, like "for the *preferred stock*, the *preferred stock par value per share* is 0 and the *preferred stock shares authorized* is 2000", it is called Tuple.

Items are represented in XBRL as a single XML element with the value as its content while Tuples are represented by XML elements containing nested Items or Tuples, i.e. subelements.

However, facts are not isolated entities and it is not enough to provide its value, it is also necessary to contextualize them. Consequently, more entities are introduced in the XBRL model:

- **Context**: it defines the entity (e.g. company or individual) to which the fact applies, the period of time the fact is relevant and an optional scenario. Contexts are referenced from Facts using the "contextRef" attribute, which specifies that the given Fact is valid for the Context entity, period and scenario.
- **Unit**: it defines a unit of measure, such as "USD" or "shares". They are referenced from Facts using the "unitRef" attribute.
- **Reference**: The kinds of facts under consideration are defined by taxonomies, which specify their meaning. These kinds of facts are then used in instance documents and linked to their definition in the taxonomies, typically through schema references.

The a.) row of Table 1 shows part of an instance document from the EDGAR program that contains a Context element which defines a company, a time period and the scenario "unaudited". Then, there is a Fact that holds in that context. The Fact references the Context and its value unit, while their content is the actual numeric value.

## 2.2 Taxonomies

Taxonomies are the other kind of XBRL document. A taxonomy defines a hierarchy of concepts, basically kinds of Facts, and captures part of their intended meaning. In XBRL there is a set of base taxonomies that define the core concepts and other ones that extend them in order to particularize these concepts for concrete accounting principles, application domains, etc. Additionally, it is possible to extend existing taxonomies and accommodate them to particular needs.

Taxonomies are based on XML Schemas, which provides the taxonomy building primitives and the extension mechanisms. Moreover, there are also the linkbases, which allow establishing links beyond the tree structure of a taxonomy by virtue of their use of XLink.

# 3. RELATED WORK

The U.S Securities and Exchange Commission (SEC) offers some online tools that allow interacting with the data available in XBRL form.

There is a tool called Interactive Financial Reports that allows viewing and charting companies financial information. It also provides some functionality that allows comparing different filings and different companies, thought it is hard to use and prone to even the slightest differences between the compared filing facts, even when there is just a name change for facts from filings of the same company.

**Table 1. a.) XBRL XML instance data example, b.) OpenLink XBRL sponge mapping and c.) XML2RDF XBRL mapping for the XBRL example**

**a.)**

```
<context id="AsOf20061201_Consolidated_Unaudited">
  <entity>
    <identifier scheme="http://www.sec.gov/CIK">796343</identifier>
    <segment><adbe:Consolidated /></segment>
  </entity>
  <period>
    <instant>2006-12-01</instant>
  </period>
  <scenario><adbe:Unaudited /></scenario>
</context>
...
<usfr-pte:CashCashEquivalents decimals="-3"
contextRef="AsOf20061201_Consolidated_Unaudited"
unitRef="USD">772500000</usfr-pte:CashCashEquivalents>
```

**b.)**

```
<sioc:Container rdf:about="AsOf20061201_Consolidated_Unaudited">
  <olsw:identifier>796343</olsw:identifier>
  <olsw:scheme rdf:resource="http://www.sec.gov/CIK"/>
  <olsw:instant>2006-12-01</olsw:instant>
  <olsw:CashCashEquivalents>772500000</olsw:CashCashEquivalents>
</sioc:Container>
```

**c.)**

```
<xbrli:contextType
rdf:about="AsOf20061201_Consolidated_Unaudited">
  <xbrli:entity>
    <xbrli:contextEntityType rdf:about="&semxbrl;CIK/796343">
      <xbrli:segment>
        <xbrli:segmentType>
          <adbe20080530:Consolidated rdf:parseType="Resource">
          </adbe20080530:Consolidated>
        </xbrli:segmentType>
      </xbrli:segment>
    </xbrli:contextEntityType>
  </xbrli:entity>
  <xbrli:period>
    <xbrli:contextPeriodType>
      <xbrli:instant>2006-12-01</xbrli:instant>
    </xbrli:contextPeriodType>
  </xbrli:period>
  <xbrli:scenario>
    <xbrli:contextScenarioType>
      <adbe20080530:Unaudited rdf:parseType="Resource">
       </adbe20080530:Unaudited>
    </xbrli:contextScenarioType>
  </xbrli:scenario>
</xbrli:contextType>
...
<usfr-pte:CashCashEquivalents>
  <xbrli:monetaryItemType>
    <xbrli:unitRef rdf:resource="http://dbpedia.org/resource/USD"/>
    <xbrli:decimals>-3</xbrli:decimals>
    <xbrli:contextRef
rdf:resource="#AsOf20061201_Consolidated_Unaudited"/>
    <rdf:value>772500000</rdf:value>
  </xbrli:monetaryItemType>
</usfr-pte:CashCashEquivalents>
```

There is also the Financial Explorer, which presents company financial data through very informative diagrams. In this case, it is just possible to show data from one company at a time. Finally, there is the Executive Compensation tool, which allows comparing just two facts, Public Market Capitalization and Revenue, across all filed companies.

Apart from the SEC tools, there are some other XBRL tools, most of them proprietary and with quite high licensing cost. Among them, the Fujitsu XBRL Tools[1] should be highlighted because this is one of the most popular ones and it is available for XBRL Consortium members and academic users. The tools comprise taxonomy and instance editors, viewers and validators.

The most powerful tool in this set, though still in beta and with many usability problems, is the Instance Dashboard. This application can consume multiple instance documents and, by specifying base taxonomy, users can perform some comparison analysis.

As it has been noted, the main limitation of XBRL tools is their limited support for cross analysis of financial information, not just among data based on different taxonomies, even when comparing filings for different companies based on the same taxonomies.

This limitation is inherited from the technologies underlying XBRL, especially XML. XML takes a document oriented approach, where each document presents a tree structure. This makes it difficult for XML-based tools to provide functionalities that blur this separation into documents and that overcome the limitations of a tree structure when mashing-up data from different sources.

Consequently, Semantic Web tools are being considered by people like Charles Hoffman, the father of XBRL: "*This field [W3C semantic standards] is rich with possibilities and stands as the next logical step in the natural progression of information technology to seek a higher value proposition*" [4].

This interest is materializing, and the combination of XBRL and the Semantic Web has been receiving some attention in different blogs, mailing lists and web groups [5,6,2]. However, it is difficult to find concrete results that put into practice Semantic Web technologies in the XBRL field.

Moreover, most of these results are specific for some parts of XBRL. For instance, there is an ontology about financial information based on XBRL that is specific for investment funds [7] or a tool that maps quarterly and semester accounting information submitted to the Spanish securities commission (CNMV) to Semantic Web technologies [3].

Moreover, both approaches are based on procedural code specially developed in order to extract specific patterns from the XBRL data. Consequently, they are difficult to scale to the whole XBRL specification and affected by even slight changes in it. We propose an approach that, instead of directly processing XBRL data, takes profit from the fact that it is expressed using XML and specified using XML Schemas.

[1] Fujitsu XBRL Tools, http://www.fujitsu.com/global/services/software/interstage/xbrltools/

[2] XBRL Ontology Specification Group http://groups.google.com/group/xbrl-ontology-specification-group

Finally, there is the work by OpenLink Software[3] based on "sponges" that extract semantic metadata from different kinds of content, among them XBRL. This is a quite recent and relevant work towards making XBRL available as linked open data. Therefore, it has been used in the validation section in order to compare the results obtained by our approach.

# 4. APPROACH

In order to move existing XBRL instances and taxonomies to the Semantic Web, and due to the fact that XBRL is based on XML and XML Schema, we have applied the XML Semantics Reuse methodology [8]. This methodology is implemented as two mappings by the ReDeFer project[4], the first one from XML Schema to OWL and the second one from XML to RDF.

This approach has already shown its usefulness with other quite big XML Schemas, especially in the multimedia metadata domain [9], where it has produced the more complete MPEG-7 ontology to date [10].

## 4.1 XSD2OWL Mapping

The XML Schema to OWL mapping is responsible for capturing the schema semantics. This semantics are determined by the combination of XML Schema constructs. The mapping is based on translating these constructs to the OWL ones that best capture their intended meaning. These translations are detailed in Table 2.

**Table 2. XSD2OWL translations for the XML Schema constructs and shared semantics with OWL constructs**

| XML Schema | OWL | Mapping motivation |
|---|---|---|
| element \| attribute | rdf:Property<br>owl:DatatypeProperty<br>owl:ObjectProperty | Named relation between nodes or nodes and values |
| element@substitutionGroup | rdfs:subPropertyOf | Relation can appear in place of a more general one |
| element@type | rdfs:range | The relation range kind |
| complexType\|group \| attributeGroup | owl:Class | Relations and contextual restrictions package |
| complexType//element | owl:Restriction | Contextualised restriction of a relation |
| extension@base \| restriction@base | rdfs:subClassOf | Package concretises the base package |
| @maxOccurs<br>@minOccurs | owl:maxCardinality<br>owl:minCardinality | Restrict the number of occurrences of a relation |

## 4.2 XML2RDF Mapping

Once all the metadata XML Schemas are available as mapped OWL ontologies, it is time to map the XML metadata that instantiates them. The mapping is based on modeling the XML structure, i.e. a tree, using RDF.

The fundamental translation is between relations, from *xsd:elements* and *xsd:attributes* to *rdf:Properties*. Concretely, *owl:ObjectProperties* for node to node relations and *owl:DatatypeProperties* for node to value ones.

Values are kept during the translation as simple types and RDF blank nodes are introduced in the RDF model in order to serve as the source and destination for properties. They will remain blank for the moment until they are enriched with semantic information.

[3] OpenLink Software, http://www.openlinksw.com

[4] ReDeFer project, http://rhizomik.net/redefer

The resulting RDF graph model contains all that we can obtain from the XML tree. It is already semantically enriched thanks to the *rdf:type* relation that connects each RDF property to the *owl:ObjectProperty* or *owl:DatatypeProperty* it instantiates. It can be enriched further if the blank nodes are related to the *owl:Class* that defines the package of properties and associated restrictions they contain, i.e. the corresponding *xsd:complexType*. This semantic decoration of the graph is formalised using *rdf:type* relations from blank nodes to the corresponding OWL classes.

## 5. RESULTS

First of all, we have generated an ontological infrastructure for the XBRL core, currently XBRL 2.1. It is composed by the ontologies resulting from mapping the XBRL core XML Schemas using the XSD2OWL mapping: XBRL Instance, XBRL Linkbase, XBRL XL and XBRL XLink. Apart from the previous schemas, the EDGAR Standard Taxonomies schemas have been also mapped in order to be able to map the XBRL data submitted to the XBRL voluntary program EDGAR.

Each filing for the companies participating in the EDGAR program contains an XBRL XML file representing the actual financial data and also a specific XML Schema extending the XBRL core. This schema provides specific guides for the corresponding financial data. Both files are mapped using XML2RDF and XSD2OWL respectively.

For instance, for Adobe Systems Inc filing on 2008-07-03, there are the adbe-20080616.xml file containing the instance data and the adbe-20080530.xsd schema for data structures specific for this filing. They are mapped, respectively, to the RDF file for instance data adbe-20080616.rdf and the OWL ontology adbe-20080530.owl for the schema.

All the previous ontologies are available from the BizOntos Business Ontologies web page[5] and the semantic data for all the processed filings can be queried and browsed from the Semantic XBRL site[6]. Currently, 489 filings have been processed from EDGAR. The combination of all these filings once mapped to RDF amounts slightly more than 1 million triples, concretely 1,023,929 triples.

Part "c.)" of Table 1 shows the RDF metadata resulting from applying the XML2RDF mapping to the XBRL context and facts shown in part "a.)" of the same table. The RDF metadata references classes and properties from the OWL ontology resulting from mapping the XML Schemas used in the XML instance. This includes the XBRL schemas and also those specific for the concrete filing being processes.

Finally, the generated data is published as Linked Open Data in the World Wide Web. The approach is based on generating XHTML plus RDFa [11]. In order to do that, we have used the Rhizomer platform that, apart from encapsulating the metadata store, also provides an RDF to XHTML+RDFa transformation and a RDF to HTML Form transformation that makes it possible for users to interactively edit the published data. The whole architecture is shown in Figure 1, which apart from the semantic data generation and publishing functionalities also features a linking one described in the next section.

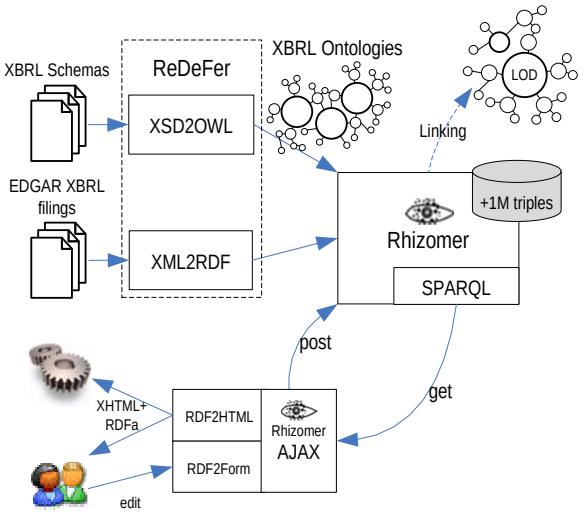


**Figure 1. Architecture of the proposed solution for semantic XBRL generation, linking and publishing**

### 5.1 Linking to Linked Open Data

The data on the EDGAR program is public. Anyone can access and download this information for free. Consequently, once mapped to RDF, it can be integrated into the Web of Linked Open Data.

In order to connect the EDGAR dataset with other ones in the Web of Linked Data, the entities in the XBRL model have been analyzed in order to detect those also described in other datasets. The more prominent and interesting ones are companies, a kind of EntityType present in most EDGAR filings.

XBRL data provides an identifier for these entities, the Central Index Key (CIK) number. It is a number given to an individual or company by the U.S. SEC and used to identify the filings of a company, person, or entity in several online databases, including EDGAR.

However, there are some EDGAR filings that do not use this identifier and use the "CompanyName" one instead. For most of them it is possible to get the corresponding CIK using EDGAR's CIK Lookup service[7]. Unfortunately, as the filings are directly submitted by the participant companies, there are some discrepancies between the names in the filings and those in the lookup service. For these, we are still trying to find some algorithm that allows us to relax the query when no results are returned or choose the appropriate one when more than one company is returned. In any case, we can get the appropriate CIK for most of the EDGAR filings.

Even when a CIK identifier is available, it might be impossible to directly connect it to company descriptions available in DBPedia because just some of them have the "secCik" property that links them to the company CIK. Due to these inconveniences, we have been able to directly link just 20 of the companies in the EDGAR dataset to DBPedia. Our work concentrates now on using the CIK Lookup service in order to obtain these identifiers for the DBPedia counterparts.

[5] BizOntos, http://rhizomik.net/ontologies/bizontos

[6] SemanticXBRL, http://rhizomik.net/semanticxbrl

[7] Search EDGAR: CIK Lookup, http://sec.gov/edgar/searchedgar/cik.htm

## 6. EVALUATION

The XSD2OWL and XML2RDF mappings have been validated in different ways. First, we have used OWL validators in order to check the consistency of the resulting ontologies. Once all the ontologies were validated, which also includes checking that all the dependencies among them are met, we proceeded to put them into practice, together with the semantic metadata generated by the XML2RDF mapping.

In parallel with our efforts, the ontologies we have generated for XBRL using the XSD2OWL mapping have being also used by OpenLink Software in their XBRL sponge that translates XBRL to RDF. Apart from an independent evaluation of the ontologies, their reuse in the XBRL sponge also facilitates comparing the RDF data it generates with that resulting from the XML2RDF mapping we propose.

First of all, there is a significant difference in the number of triples generated by the OpenLink XBRL sponge and XML2RDF. For instance, for the same EDGAR XBRL filing[8], the XBRL sponge produces 900 triples while XML2RDF produces 4739 triples. One possible reason for this difference is that we have followed quite different approaches relative to how the original XML tree structure is captured in the RDF graph.

For instance, Table 1 shows in the first row a portion of XBRL XML instance data from the previous filing. The second row contains the RDF generated by the OpenLink sponge. As it can be seen, not all the information in the XBRL is captured and the whole structure is flattened.

On the other hand, the "c.)" row in Table 1 shows the mapping for the same XBRL XML as generated by the XML2RDF mapping. As it can be seen, the result is much move verbose, even more than the original XBRL. However, it does capture all the original information and keeps the original structure. Even more, the original XBRL does not explicitly refer to the XML Schema *complexTypes* defined in the schemas and used in the instance data. This information is available in the XML2RDF semantic data and can be used, together with the hierarchical relations among complex types, when resolving semantic queries against this data.

## 7. CONCLUSIONS AND FUTURE WORK

As it has been shown, it is possible to map the XML data for XBRL filings in order to generate RDF semantic data that keeps all the original information and structure. This mapping also includes the involved XML Schemas that structure the XML data, which are mapped to Web ontologies.

This approach has been put into practice in the context of the SEC's EDGAR program that promotes XBRL filings for USA companies. It has been possible to apply the previous XML to RDF and XML Schema to Web ontology mappings to all the EDGAR filings and more than 1 million triples have been obtained.

Our approach has been partially adopted by OpenLink Software, a company that is currently using our XBRL ontologies in its own XBRL to RDF mapping product. However, OpenLink does not follow the same XML to RDF mapping approach. Their approach has been compared to ours showing that our proposal retains much more of the original XBRL information and structure.

We have also have made all this semantic information generated from the EDGAR program available online, so it can be queried and browsed using a Web user interface. The proposed semantic queries illustrate the benefits of the semantic integration available once XBRL data is translated to semantic data.

Our work concentrates now on linking the resulting semantic data to the rest of the Web of Linked Open Data, completing the links to companies in DBPedia. Moreover, we are considering restructuring the semantic model resulting from mapping the XBRL XML because it is not intuitive and usable enough from a Semantic Web point of view. For instance, in the current model resulting from directly mapping from XML to RDF, Facts are modeled as properties while it would be more intuitive and easier to query if modeled as resources.

[8] Adobe Systems Inc. EDGAR filing 2008-07-03, XBRL file: http://www.sec.gov/Archives/edgar/data/796343/000079634308000005/adbe-20080616.xml

## REFERENCES


[1] Electronic Data Gathering, Analysis, and Retrieval system, http://www.sec.gov/edgar.shtml

[2] Bizer, C., Heath, T., Idehen, K., Berners-Lee, T. 2008. Linked data on the web (LDOW2008). In Proceeding of the 17th International WWW Conference, ACM, 1265-1266.

[3] Núñez, S., de Andrés, J., Gayo, J. E., and Ordoñez, P. 2008. A Semantic Based Collaborative System for the Interoperability of XBRL Accounting Information. In Emerging Technologies and Information Systems for the Knowledge Society. LNCS Vol. 5288, Springer, 593-599.

[4] Hoffman, C. 2006. Financial Reporting Using XBRL: IFRS and US GAAP Edition. Lulu.com.

[5] DuCharme, B. 2008. Changing my mind about XBRL again. In Bob DuCharme's weblog, bobdc.blog. http://www.snee.com/bobdc.blog/2008/08/changing_my_mind_about_xbrl_ag.html

[6] Raggett, D. 2008. XBRL and RDF. Dave Raggett's Blog. http://people.w3.org/~dsr/blog/?p=8

[7] Lara, R., Cantador, I., and Castells, P. 2008. Semantic Web Technologies For The Financial Domain. In J. Cardoso and M. Lytras (Eds.), The Semantic Web: Real-World Applications from Industry. Springer, 41-74.

[8] García, R. 2006. XML Semantics Reuse. Chapter 7 in A Semantic Web Approach to Digital Rights Management, PhD Thesis, Universitat Pompeu Fabra, Barcelona, Spain. http://rhizomik.net/~roberto/thesis

[9] García, R., Perdrix, F., Gil, R., and Oliva, M. 2008. The Semantic Web as a Newspaper Media Convergence Facilitator. Journal of Web Semantics 6, 2, 151-161.

[10] García, R., Tsinaraki, C., Celma, O., and Christodoulakis, S. 2008. Multimedia Content Description using Semantic Web Languages. In Semantic Multimedia and Ontologies: Theory and Applications, Y. Kompatsiaris and P. Hobson Eds. Springer, 17-54.

[11] Halb, W., Raimond, Y., Hausenblas, M. 2008. Building Linked Data For Both Humans and Machines. In proceedings of the Linked Data on the Web Workshop (LDOW'08), Beijing, China.