Chapter 3

SEMANTIC WEB TECHNOLOGIES FOR THE FINANCIAL DOMAIN

Rubén Lara[1], Iván Cantador[2] and Pablo Castells[2]
[1]*Tecnología, Información y Finanzas (TIF), 28010 Madrid, Spain – rlara@afi.es*
[2]*Escuela Politécnica Superior, Universidad Autónoma de Madrid, 28049 Madrid, Spain – ivan.cantador@uam.es, pablo.castells@uam.es*

Abstract: Data, information and knowledge management are key activities in modern economies, and considerable efforts and resources are devoted to them by organisations world-wide. An optimal handling of information assets is especially critical in the financial field, a conceptually rich domain where information is complex, huge in volume, and a highly valuable business product by itself, and where the exchange and integration of information for its posterior analysis is a key task for financial analysts. The volume, complexity and value of economic and financial information make finance a strategic area for research and innovation on information modelling, exchange and integration and, consequently, there is an increasing interest in evaluating what semantic technologies can contribute to this domain. In this chapter, we present two applications of semantic technologies to the financial domain, namely: a) the management of economic and financial information, and b) the building of explicit information models for the exchange of information in the investment funds market, comparing the use of XBRL and the use of semantic languages such as OWL. With the description and analysis of these applications, we shall attempt to illustrate and analyse the possibilities for exploiting semantic technologies in the financial domain, the achieved and expected benefits therein, and the problems and obstacles to be overcome in the future.

Key words: knowledge management, information integration, information exchange, semantic web, investment fund, taxonomy, ontology, XBRL, OWL

1. INTRODUCTION

Data, information and knowledge management are key activities in modern economies, consuming a considerable amount of the efforts and resources in organisations and businesses (Alexiev 2005). Information management involves, in most cases, the integration of data from disparate and heterogeneous channels, including information from third parties. An optimal handling of information assets is especially critical in the financial field, a conceptually rich domain where information is complex, huge in volume, and a highly valuable business product by itself, and where the exchange and integration of information for its posterior analysis is a key task for financial analysts. The volume, complexity and value of economic and financial information make finance a strategic area for research and innovation on information modelling, exchange and integration. For this reason, there is interest in evaluating what the emerging semantic-based knowledge technologies can achieve in this context, and what is needed for them to be adopted by businesses in this area.

Along these lines, we have worked on two applications of Semantic Web technologies to the financial domain, namely: a) the management of economic and financial information (Castells et al. 2004), and b) the building of explicit information models for the exchange of information in the investment funds market, comparing the use of XBRL and the use of semantic languages such as OWL (Lara et al. 2006). These two applications, and our experiences and lessons learnt, are presented in this chapter, along with a general analysis of the possible uptake of Semantic Web technologies in finance, the potential benefits of such uptake, and the requisites for it to happen. The goal of the chapter is thus to illustrate and analyse the possibilities for exploiting Semantic Web technologies in the financial domain, the achieved and expected benefits therein, the problems and obstacles to be overcome, and an analysis and reflection on the potential of Semantic Web technologies for finance in the future.

2. SEMANTIC TECHNOLOGIES FOR ECONOMIC AND FINANCIAL INFORMATION MANAGEMENT

2.1 Description of the domain

A huge amount of valuable economic and financial information is produced world-wide every day, but its interpretation and processing is a hard and time-consuming task, as a major part of the information generated is mainly textual and, therefore, the possibilities for automated management are quite limited, whereby a considerable amount of human involvement in the loop is needed. Moreover, the manual management of information resources is error-prone and time consuming (Ding and Fensel 2001), searches are often imprecise or yield an excessive number of candidate matches through which users need to clear their way in order to find the sought information. In order to overcome these problems, efficient filtering, search, and browsing mechanisms are needed by information consumers to access the relevant contents for their needs or business, and find their way through in an effective way. On the information provision side, efficient production, management and delivery technologies are needed as well.

Semantic Web technologies are foreseen as a possible solution to such problems by providing an explicit and formal representation of the semantics underlying information sources and by exploiting such formal semantics (Berners-Lee et al. 2001). Ontologies have been proposed as a backbone technology for the Semantic Web and, more generally, for the management of formalised knowledge in the context of distributed systems (Gruber 1993). They enrich content with machine-processable semantics, which can be communicated and processed by software programs. The main principle in this vision is to make information understandable for computers, thus enabling new, more powerful information processing and management capabilities, and reducing the involved costs.

As an experience towards this vision, we have developed an ontology-based platform for managing economic and financial data. TIF, a Spanish provider of information technology solutions for the financial domain, was involved in this research. One of the core business activities of the corporation TIF is part of[1] is the creation, management and delivery of added-value economic and financial contents, such as market research, market analysis, or investment recommendations. Thus, the purpose of the research presented herein is to evaluate the improvements the application of

[1] www.grupoanalistas.com

semantic technologies can bring to current information management and delivery practices in such areas.

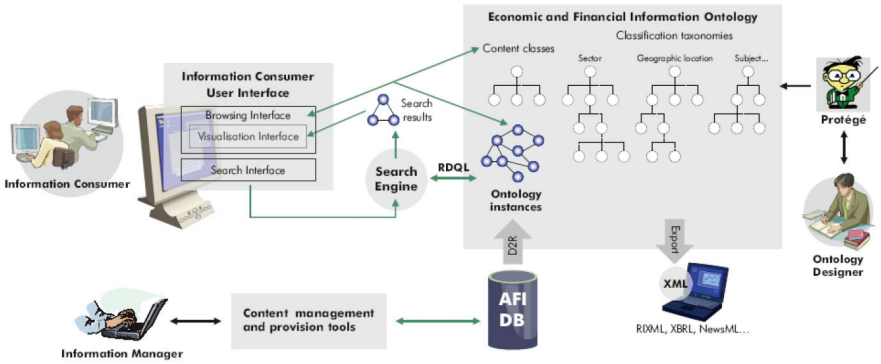

Figure 3-1. Platform architecture

The main components of the developed platform are shown in Figure 3-1:

- A domain ontology that formally models the economic and financial information produced, managed, and delivered to customers by TIF.
- Import (from corporate databases) and export (to different formats) facilities.
- Content management and provision tools.
- A visualisation component for information consumers and managers.
- A search engine.

Each of these components is described in the sections that follow.

2.2 An ontology for content management in the economic and financial domain

The information generated by providers and consumed by requesters always refers to some particular context. However, the common situation is that this information does not refer to any explicit model of such domain, or it only refers to an ad-hoc model. Furthermore, if a model is available, it is in most cases purely syntactic, which limits the automatic processing of information that would conform to such model, in particular, reusing data in another context or transforming it according to another model.

Before the project we describe in this section was accomplished, TIF made use of a partially explicit, syntactic domain model to create, manage and deliver economic and financial information and data. In particular, a custom content management system was used to, based on this model,

create, manage and deliver, through different channels (e.g. Web sites, XML syndication), contents to customers with different profiles such as banks and financial institutions, SMEs which use the information for decision making and foreign trade activity, and distributors who publish the information in printed and digital specialised media.

In this context, the first task we faced in the project was the design of an explicit and formal model for economy and finance, i.e., the creation of an ontology for the economic and financial domain covering the business needs of the company.

2.2.1 Ontology definition

The creation of the ontology started with a study existing domain models which could be reused. However, no suitable ontologies for our target domain were found, and the use of taxonomies such as the Global Industry Classification Standard (GICS)[2] for the classification of industry sectors was discarded as it did not meet the requirements of TIF. Thus the domain ontology was essentially defined from scratch. The design procedure was incremental, interacting with domain experts in order to produce refined versions of the ontology. Two main steps were followed in the creation process:

1. A first version of the ontology was created based on the existent, semi-explicit domain model and on the corporate databases scheme.
2. We interacted with domain experts in order to refine the ontology, adding missing concepts, relations and properties, and to evaluate to what extent the target domain was covered by the defined ontology.

The interaction with domain experts was the most crucial step for a successful design of the ontology, as they contributed with numerous and valuable improvements to the first version of the ontology. The first step was motivated by the need to make the current model fully explicit, so that a starting point for the interaction with domain experts was available.

2.2.2 Root ontology classes

The interaction with our financial and technical staff led us to consider four distinct kinds of concepts (classes) in the ontology:

1. *Contents*. They reflect different types of documents and contents created by domain experts, such as reports, analysis, studies, recommendations, etc. All economic and financial contents generated by TIF are modelled as an instance of a content class.

[2] www.mscibarra.com/products/gics/

2. *Classification categories*. These categories serve to classify the contents generated according to topics, sectors, etc.
3. *Entities*. They represent contents which are not economic and financial contents themselves, but which are referred to in the economic and financial contents generated by financial experts. Examples of these concepts are companies, banks, organisations, people, information sources, events, etc.
4. *Enumerated types*. They provide sets of values (controlled vocabularies) for certain properties. These concepts have a fixed set of instances.

From our experience in ontology engineering for information systems, the consideration of these four kinds of concepts is an interesting and recurrent distinction that arises in many, if not most, information management systems in diverse domains. In fact, a similar approach can be found in information exchange standards like RIXML[3] and other standards in the controlled vocabulary community (Fast et al. 2002). However, this distinction is sometimes not a sharp line.

The resulting ontology provides explicit connections between contents, categories, and other concepts, which were only semi-explicit in the current content management system. These relations are now well characterised, and can be further described in as much detail as needed and allowed by the Semantic Web technology employed. As it will be later described, the defined ontology is exploited in our platform to support more expressive and precise search capabilities, and for the automation of the generation of user interfaces (query input forms, content presentation views, and content provision forms).

2.2.3 Ontology language

For the description of the ontology, RDF(S) was used (Brickley and Guha 2004). The reasons for this choice were:

1. RDF(S) was a World Wide Web Consortium (W3C)[4] recommendation, which was a guarantee of its maturity and stability. OWL (Bechhofer et al. 2004) was also considered, but at the time the ontology was defined this language was still in the process of becoming a W3C recommendation.
2. RDF(S) had the widest tool support at the time the ontology was created. This reduced the risks in the development and the time required to implement our platform.
3. The expressivity of RDF/RDFS was considered enough for a first evaluation of the benefits of an ontology-based platform for content

[3] Research Information Exchange Language, http://www.rixml.org.
[4] http://www.w3.org

management. The transitive closure of *subPropertyOf* and *subClassOf* relations in RDF(S), domain and range entailments, and the implications of *subPropertyOf* and *subClassOf* have been the inference mechanisms exploited by the developed platform.

The OWL ontology language and its foreseen extensions can be considered in the future for the evolution of the platform. In fact, OWL has been used in other developments we have later undertaken, as will be described in the section 3.

For the definition of the ontology, Protégé[5] was used, as it offers a complete and well-tested set of capabilities for ontology modelling.

2.3 Semantic content description and exploitation

2.3.1 Integration of legacy content

Once an ontology is available that captures and formalises the domain in which the company produces and manages information, new contents can be created in the form of instances of concepts of the ontology from the outset. This way, newly created contents conform to a well-defined, formal, and agreed upon model, greatly facilitating the automated processing of contents by semantic-aware software programs, as we will describe later in this section. However this solves only a part of the problem, since huge volumes of information are already stored in corporate databases, based on conventional (relational) information models. In order for these legacy assets to benefit from the ontology-enabled semantic-based search, visualisation, and generation capabilities which will be described in this section, new ontological descriptions of the old contents need to be added to the management system.

We used the open source tool D2R[6] for this purpose. D2R can extract information from relational databases supporting JDBC or ODBC and, using an XML mapping file, generate RDF instances. An XML mapping file has to be created for each concept in the ontology. This mapping defines how the results of an SQL query over corporate databases are mapped to attributes of an instance of a particular concept in the ontology. We have defined such mappings, based on which the available contents stored in corporate databases have been translated into domain ontology instances. These instances have been in turn stored in corporate databases for persistency. In

[5] http://protege.stanford.edu
[6] http://sites.wiwiss.fu-berlin.de/suhl/bizer/d2rmap/D2Rmap.htm

particular, Jena[7] has been used to retrieve RDF instances of the ontology from the files generated using D2R and to store them in a database back-end.

Our platform aims at improving current content management systems, but without interfering with existing production systems. This means that all the information (both old and new) should remain available in the old database model for its use by production systems. Therefore, two versions of the produced contents are maintained, one conforming to the old scheme, and one described in terms of the domain ontology. In order to reuse the already defined mappings for the annotation of existing contents, newly generated contents are first stored in the old model; afterwards, the mappings defined are used by D2R to annotate these contents, and the resulting RDF instances are stored in the database back-end.

2.3.2 Ontology-based search

Our platform provides a search module which can be used by customers, content providers, content managers and administrators to query for contents in terms of the defined ontology. In this way, users can go beyond keyword-based search, full-text search, and structured search based on a semi-explicit, syntactic model. In particular, our search module supports full structured search in terms of any dimension of the ontology, and allows setting different levels of detail for expressing the search query (different partial views of the ontology) depending on the user profile.

Furthermore, the inference capabilities enabled by RDF(S) are used to obtain search results. In particular, the transitive closure of *subPropertyOf* and *subClassOf* relations in RDF(S), domain and range entailments, and the implications of *subPropertyOf* and *subClassOf* are exploited to obtain contents that match a user query.

In our system, the user interacts with an HTML search form where he can select concepts in the ontology (content classes), and provide search values for properties of the selected concept. Thus, the user can formulate his needs in terms of concepts, properties, and relations among concepts as defined by the ontology.

Search forms are automatically generated from the definition of ontology concepts. In particular, we have defined a search form generation mechanism which provides a default procedure to generate forms adapted to the structure and the types of properties of concepts in the ontology. In the default procedure, the properties of content classes have a Boolean *searchable* metaproperty, used by ontology designers to control whether the generated search forms should include an input control where search values

[7] http://jena.sourceforge.net/

for this property can be provided. If a property of a class is searchable, the search form generation procedure selects different HTML/JavaScript controls depending on the type of the searchable property.

Furthermore, it is usually necessary to create a custom search form following particular design and brand image considerations. This is achieved in our system by creating form templates for each content class, where all aspects of the design can be defined in as much detail as desired. Our template definition language is based on JSP, where custom tags have been defined to reference properties of concepts in the ontology and to include other ontology graph traversal expressions. The language also includes primitives to easily specify HTML or JavaScript input components, and facilities to define global layout constructs. Wherever details are not explicitly indicated in the template, the system tries to provide appropriate default solutions. An example of the search form generated for a particular concept in the ontology can be seen in Figure 3-2.

It is generally more adequate to provide customers with fairly simple and easy to use search interfaces (Green et al. 1990), whereas experts and content managers, who have better knowledge of the domain, can benefit from more complex and powerful search facilities. This is supported in our platform by creating different templates for different user profiles and usage modes, thus enabling the creation of as large and varied an array of power levels and modalities as needed in a highly modular and extensible way (see (Hearst 1999) for an overview of user interface approaches for searching).

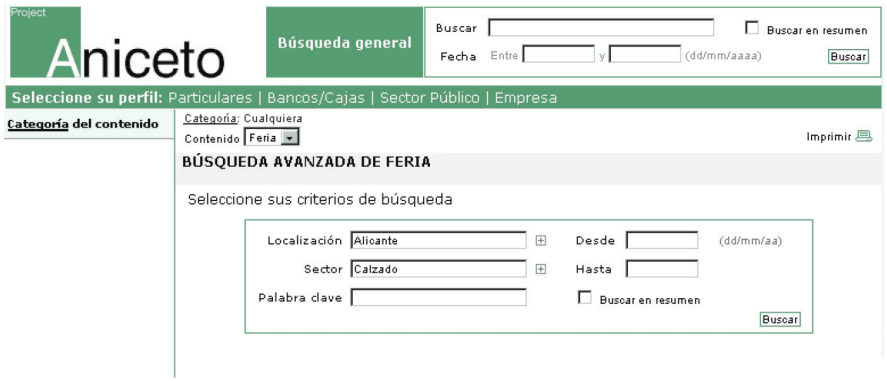

Figure 3-2. Search form generated for a "trade fair" content

The possibilities to use the model defined by the ontology to formulate search queries go beyond specifying property values for content concepts. The search module allows the user to combine direct search, using content

classes and properties, with navigation through the classification categories defined by the ontology. This approach follows the classic combination of searching and browsing in systems like Yahoo! and others (Hearst et al. 2002). In particular, users can restrict their search queries to selected classification categories. Furthermore, search results indicate the categories returned contents belong to, which allows the user to narrow or widen his search query to particular categories, or to browse the contents in the same category as returned contents.

The search module converts the information query conveyed by the user into an RDQL (Seaborne 2004) query, which is executed against the ontology and the knowledge base, yielding the set of RDF instances which match the query. These instances are presented to the user, who can view their detailed description. The visualisation of ontology instances is controlled by a visualisation module which is described next.

2.3.3 Ontology-based information visualisation

Our platform includes a specialised module for the presentation of ontology instances (contents), i.e., for the visualisation of information units and for the navigation across units. This module is based on our early work on the Pegasus tool (Castells and Macías 2001).

The visualisation module dynamically generates Web pages from the description of ontology instances. How contents are visualised depends on the definition of the concepts they are instance of, which includes the definition of their properties and relations. Instead of hard-wiring this treatment in visualisation code, our platform allows defining the presentation of ontology concepts declaratively, using one or several visualisation models per concept.

The presentation model defined for each concept establishes the parts of an instance that have to be shown, in what order, and under what appearance. This model is defined with a fairly simple language which permits referencing and traversing the parts of the ontology the instance refers to. The presentation engine dynamically selects the appropriate view of the content depending on the concept the content is an instance of. The visualisation module also takes care of presenting on the same page other instances related to the content currently presented, or of generating hyperlinks to them in order to navigate across ontology relations.

The presentation language is based on JSP, with a library of custom tags which includes a) ontology access expressions, b) HTML / JavaScript primitives to display ontology constructs, and c) layout constructs. The presentation language also offers the possibility to express conditions on user profiles, the access device, the state of the application, or the

characteristics of the information to be presented. These conditions can determine the choice of one or other presentation model for an instance, or at a more detailed level, establish the aspect of small parts of the presentation, the inclusion or not of certain information fragments, the generation of hyperlinks, or the selection of user interface components (lists, tables, trees, etc.).

Three presentation models have been defined: extended view, to show instances with maximum level of detail on a single page; summary view, to show lists of instances e.g. search results; and minimum view, to be used for example as the text of the link to an instance. Figure 3-3 shows an example of an extended view of an instance of the Fundamental Analysis concept.

User profiles have been defined, referring to the domain ontology created, to express preferences on specific categories or content classes. The user profiles defined include: a) professional profiles, and b) subscription profiles. The subscription profile defines access permissions to different parts of the ontology; when instances are visualised, only parts to which the user has been granted access will be shown. The professional profile defines a scale of interests for different categories and types of contents, which determines the order (priority) and amount of information that is presented to the user, depending on the typology and relevant subject areas for his profile.

Figure 3-3. Extended view of an instance of the "Fundamental Analysis" concept

2.3.4 Content managers

Precision in locating the right contents, and ease of navigation through them, are essential for authors who create, classify, maintain, or link contents. To this end, the content management system used at TIF has been adapted to incorporate the creation and management of contents in terms of the domain ontology defined, making use of the search and visualisation modules developed.

For the creation and edition of different types of contents, appropriate Web forms are automatically generated based on the definition of the content class in the domain ontology. These forms are dynamically generated in the same way search forms were: there is a default generation mechanism which takes into account the definition of the concept and the type of its properties to generate appropriate input controls, and custom forms can be defined for each content class. The main difference between content creation forms and search forms is that search forms will usually correspond to a partial view of the content class as defined by the searchable meta-property, i.e., not all class attributes are interesting for searching contents but, usually, all properties of a content have to be provided when the content is created.

Similarly, the set of instance properties presented to end users is typically a superset of the properties which appear in a search form for this instance class, and a subset of all fields required by a content creation form.

2.4 Experience and results

The development of the platform presented in this section and its application to the management of economic and financial contents at TIF gave raise to an initial knowledge base of 180,831 instances and 2,705,827 statements. With this knowledge base, we have been able to evaluate the benefits achieved by adapting our content management system to make use of an explicit and formal domain model, exploiting some of the features of semantic technologies.

In particular, the following improvements are achieved:

1. Definition of a completely explicit information model: the building of a completely explicit model has necessarily driven to a review of the existing model. Furthermore, both old and new systems and applications have now a clear and shared information model to follow, which can considerably ease information integration and sharing among internal applications. In general, the existence of an ontology serves as reference for communication (both among persons and computers) and helps to improve data quality based on a shared conceptualisation of the domain.
2. Improved search capabilities: by describing economic and financial contents in terms of a well-defined and formal domain model, we can:
   a) Automatically generate user interfaces for search from the concept definitions provided by the ontology.
   b) Apply standard inference mechanisms to obtain richer search results.
   c) Easily interleave structured search and browsing.
   d) Declaratively and based on explicit models adapt search results to different user profiles and, in general, to the context of the search.
3. Improved visualisation: contents can be dynamically and following general procedures visualised according to the definition of concepts in the domain ontology. Furthermore, different visualisation modes can be dynamically selected based on declarative descriptions of user profiles.
4. Improved management of information: by exploiting the search and visualisation capabilities resulting from the use of a domain ontology, more efficient management of information can be achieved. Furthermore, the existence of a clear information model, evaluated by domain experts, helps manage contents better; contents are created, linked and maintained following an agreed and explicit model.

However, some problems have been encountered for the development of the platform, as well as some limitations:

1. Maintenance of the model: the maintenance and evolution of the model is a critical task and can become a bottleneck if it is not properly monitored.
2. Scope of the model: while the domain model has been defined to cover the contents created by TIF and by the group of companies it is part of, it is insufficient for the general exchange of information with other parties, as other parties might use a different (most likely syntactic and semi-explicit) model. Therefore, mediation mechanisms or the joint definition of an ontology by major actors in the market is required for easing the exchange of information between parties. Otherwise, the model defined can be exploited internally but its benefits are reduced when used for the exchange of information.
3. Creation of ontologies from scratch: the creation of an appropriate and *shared* domain model is a fundamental task for achieving the benefits semantic technologies can offer. However, we have detected a lack of existing models or, at least, of existing semantic models in the financial domain. This makes the definition of new models necessary, and reaching an agreement with other parties for the use of a shared model more difficult.

The ontology defined has served to generate and manage contents created by TIF. However, it does not cover the reception of information from other parties, its processing, and its possible delivery. In the next section, we discuss the definition of a model for the reception, integration, processing and delivery of information in the Spanish investment funds market. We focus on building a model to be used for the exchange of information across organisational boundaries and on the comparison of the XBRL language (Engel et al. 2005), which is being widely adopted and promoted in the financial domain, to a semantic language such as OWL.

3. SEMANTIC TECHNOLOGIES FOR THE EXCHANGE OF INFORMATION IN THE INVESTMENT FUNDS MARKET

3.1 Description of the domain

The analysis of the investment funds market requires the availability of harmonised information on the considered funds, including both last-minute and historical data, which is usually generated and provisioned by different parties and in heterogeneous formats. Furthermore, added-value information

such as risk-profitability ratios of commercialised funds is demanded by different customer profiles. For example, final investors demand this kind of information for supporting their investment decisions, so funds managers do in order to compare the evolution of their funds with respect to the general market behaviour.

Current mechanisms for the exchange of information among the different actors in the investment funds market (investment firms, management firms, stock markets, analysts, investors, market supervisors) are not based on uniform and explicit information models, hampering an agile exchange and requiring important efforts to process and integrate such information. Furthermore, a situation where the exchange and processing of information is time-consuming and error-prone leads to a reduction of market transparency.

In this context, the gathering and integration of information from disparate, heterogeneous sources becomes a key task that can be considerably eased by the availability of explicit and shared information models. Moreover, the analysis process leads to the generation of analytic, added-value information, the consumption of which by other parties can also benefit from the existence of agreed information models.

TIF, in conjunction with AFINet Global[8], is the leading provider of analytical information of investment funds in the Spanish market. For providing this service, TIF continuously receives and aggregates information from the national stock markets, from firms managing investment funds, and from the national market supervisor (the CNMV)[9], covering all the investment funds currently commercialised in Spain and counting with a 10-years historical base (over 6000 investment funds at the time of writing). The information received includes all the descriptive aspects of a fund when it starts to be commercialised (entity commercialising the fund, investment policy, commissions, etc.), changes on any of these aspects, and the Net Asset Value (NAV) of the fund at different points in time.

The different parties from which TIF receives and aggregates information currently use heterogeneous information models and formats. This makes the reception, validation, and aggregation of the information a difficult task, and requires ad-hoc validation procedures and a rather costly maintenance, as providers sometimes introduce changes on their information models and formats. In this setting, when heterogeneous information about a certain fund or group of funds is received, it has to be validated first (sometimes it has not been properly validated in origin) and transformed so that it follows a uniform information model. After that, the analytical indicators associated to

[8] http://www.grupoanalistas.com
[9] http://www.cnmv.es

these funds are (re)calculated and published via different channels, currently including XML syndication and direct access via a number of information portals[10].

The part of the investment funds information life-cycle relevant for TIF is depicted in Figure 3-4. Descriptive information about investment funds commercialised in the Spanish market is provided by the CNMV, and periodical information such as the NAV of a fund is provided by the national stock markets (Madrid, Barcelona, Bilbao and Valencia) and by the firms managing the funds. This information is validated, converted and aggregated, leading to the creation of an aggregated and consistent information base that is ready for analysis. The analysis process leads to analytical, added-value information which is consumed by agents such as management firms, sellers, or directly by investors.

A gain in efficiency in the life cycle of Figure 3-4 can be achieved if the validation and conversion process, instead of dealing with heterogeneous information, would receive information according to a shared model so that ad-hoc processing can be avoided and maintenance needs are reduced. Furthermore, if the analytical, added-value information produced also follows an agreed model, the consumption of such information by different agents can be considerably eased. The ontology for economic and financial information presented in the previous section covered all the contents generated by TIF, but not contents like investment funds information, aggregated from information received from other parties. Therefore, we have accomplished the definition of a domain model for this type of information.

[10] See http://www.invertia.com/fondos/default.asp for an example

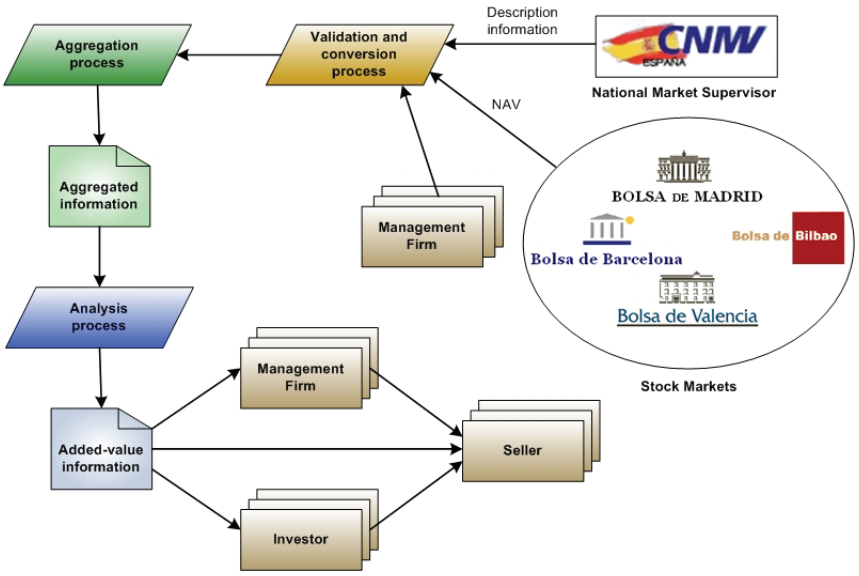

Figure 3-4. Investment funds information life cycle

The Spanish market supervisor (the CNMV) is considering the definition of XBRL (Engel et al. 2005) taxonomies for the descriptive and regulatory information on investment funds, which would have to be naturally adopted by all agents in the Spanish market. However, these models would not initially include analytical information. Furthermore, a semantic language such as OWL has not been considered so far as an alternative for defining shared information models for investment funds. In this setting, we have worked on: a) an XBRL taxonomy that includes not only descriptive but also analytical information of funds and that can serve as a basis for possible future developments led by the CNMV or for their extension, and b) on the evaluation of OWL as an alternative to XBRL. The reason for building an XBRL taxonomy first is that XBRL is being promoted as the language of choice for the modelling of financial information. However, it lacks one of the key features the Semantic Web has: information has formal semantics. Therefore, we evaluate in which way OWL ontologies and XBRL taxonomies are related, how XBRL taxonomies can be translated into OWL ontologies, and what benefits and limitations have OWL ontologies with respect to XBRL taxonomies.

3.2 Creation of a domain model based on XBRL

3.2.1 XBRL in a nutshell

XBRL is a language that builds on top of XML, XML Schema and XLink to provide users with a standard format in which information can be exchanged, enabling the automatic extraction of information by software applications (Engel et al. 2005). For that purpose, XBRL defines **taxonomies**, which provide the elements that will be used to describe information, and **instances**, which provide the real content of the elements defined. These are introduced below.

3.2.1.1 XBRL taxonomies

An XBRL taxonomy is constituted by an **XML Schema** and the **XLink linkbases** contained in or directly referenced by that schema. In XBRL terminology, the XML schema is known as the taxonomy schema.

Concepts describing reporting facts are exposed as XML Schema element definitions. A concept is given a *name* and a *type*. The type establishes the kind of data allowed for those facts described according to the concept definition. For example, the NAV concept of an investment fund would typically have a monetary type, declaring that when a NAV is reported, its value will be monetary. On the other hand, the legal name of the fund would usually have a string type so that, when it is reported in an XBRL instance, its value is interpreted as a string of characters. Besides these two attributes, additional constraints on how concepts can be used (e.g. instant/duration *period*, debit/credit *balance*) are documented by other XBRL attributes on the XML Schema element definitions.

Linkbases in a taxonomy provide further information about the meaning of the concepts by expressing relationships between concepts (inter-concept relationships) and by associating concepts to their documentation. Taxonomies make use of five different types of XLink linkbases, namely: definition linkbases, calculation linkbases, presentation linkbases, label linkbases and reference linkbases. The first three types contain different kinds of relations between elements, whereas the last two types contain documentation of elements.

Definition links describe relations among concepts in a taxonomy, such as generalisation and specialisation relations, that provide information on what an element actually is e.g. the specialisation of some other concept. Calculation linkbases provide information on how some elements are calculated in terms of some other elements, which can be exploited for data

validation. Presentation linkbases contain relations such as parent-child that are exclusively used for presentation purposes e.g. a given element will be shown as the child of some other.

The last two types of links do not define relations among elements but document elements in a taxonomy. Label links provide labels in natural language with the purpose of facilitating the understanding of data by a human user. XBRL is equipped with multilingual support and enables the user to associate labels in different languages to the same element. Reference links point to legal or other type of documentation that explains the meaning of a given taxonomy element.

Usually, it is necessary to consider multiple related taxonomies together when interpreting an XBRL instance. The set of related taxonomy schemas and linkbases is called a Discoverable Taxonomy Set (DTS). The bounds of a DTS are determined by starting from some set of documents (instance, taxonomy schema, or linkbase) and following DTS discovery rules. Taxonomies are interconnected, extending and modifying each other in various ways.

3.2.1.2 XBRL instances

A taxonomy defines reporting concepts but does not contain the actual values of facts based on the defined concepts. These fact values are included in XBRL instances. The way XBRL organises the reporting information within a certain instance is based on two main elements: **XBRL items** and **XBRL tuples**.

- **XBRL items**. Defined as extensions of primitive data types (String, Integer, Boolean, etc.), XBRL items represent atomic information elements of an XBRL instance. Items can also reference XML Complex Types in the XBRL Instance Schema[11], or extensions of these types defined in existing taxonomies. In XBRL taxonomies, complex types are typically used to provide the set of possible values a data type can hold.
- **XBRL tuples**. In XBRL, a data model is built through tuples or blocks of information. While most business facts can be independently understood, some facts are dependent on each other and they must be grouped for a proper and/or complete understanding. For instance, in reporting the information of an investment fund, each deposit entity name has to be properly associated with a correct deposit entity identifier. Such sets of facts (deposit entity name, deposit entity

[11] http://www.xbrl.org/2003/xbrl-instance-2003-12-31.xsd

identifier) are called tuples. Tuples have complex content and may contain both items and other tuples.

In addition to the actual values of a fact, such as "NAV is 50", XBRL instances provide contextual information necessary for interpreting such values e.g. "NAV is 50 today". Furthermore, for numeric facts, XBRL instances can also document measurement units e.g. "NAV is $50".

- **XBRL context elements**. Context elements include information about the entity being described, the reporting period and the reporting scenario (additional metadata taxonomy designers might want to associate to items), all of which are necessary for understanding a business fact captured as an XBRL item. The *period* element contains the instant or interval of time for reference by an item element. The sub-elements of the period element are used to construct one of the allowed choices for representing date intervals. For an item element with *periodType* attribute value equal to “instant”, the period must contain an instant element that indicates the particular point in time in which the fact is valid. For an item element with the *periodType* attribute value set to “duration”, the period must contain “forever” or a valid sequence of *startDate* and *endDate* elements, indicating the start and end points of the interval of time in which the value is valid.
- **XBRL unit elements**. Unit elements specify the units in which a numeric item has been measured. The content of a unit element must be either a simple unit of measure expressed with a single *measure* element, a ratio, or a product of units of measure. Some examples of simple units of measure are EUR (Euros), meters and kilograms. Some examples of complex units of measures are Earnings per Share or Square Feet.

3.2.2 Description of the XBRL taxonomy of investment funds

The lack of explicit and shared models for exchanging information in the investment funds market and the promotion and increasing adoption of XBRL by Spanish regulators and supervisors, e.g. Bank of Spain and CNMV, led us to consider XBRL as an interesting language for creating an explicit information model for the Spanish funds market and to create a taxonomy of investment funds.

For building this taxonomy we started by evaluating and reviewing the information model used by TIF and AFINet Global in order to define a revised model that could meet the needs of different agents in the market. For that purpose, we counted on the cooperation of Analistas Financieros

Internacionales[12], a leading company in the analysis of the Spanish financial market, and Gestifonsa[13], a firm that manages a number of funds commercialised in Spain.

The resulting model, agreed and approved by all parties, has been described using XBRL. For that purpose, the possible reuse of existing XBRL taxonomies was evaluated. In particular, the IPP[14] taxonomy from CNMV, the DGI[15] taxonomy, and the ES-BE-FS[16] taxonomy from Bank of Spain were evaluated. The result of the evaluation has been that parts of the DGI taxonomy can be reused for the description of certain elements of the investment funds information model, especially those elements describing the entities that commercialise or manage a given fund. Figure 3-5 shows the DTS of the taxonomy built, where *dgi-lc-es-2005-03-10.xsd* contains the information elements of the imported DGI taxonomy in Spanish and its respective linkbases, and *dgi-lc-int-2005-03-10.xsd* contains the international elements of the DGI taxonomy.

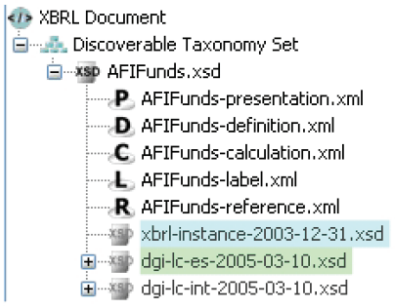

Figure 3-5. DTS of the investment funds taxonomy

The information elements of the created taxonomy have been divided into the following groups:

- **Descriptive information**: models all the descriptive aspects of a fund, such as the name of the fund, the entity managing the fund, the data relative to the registration by the CNMV, etc.

[12] http://www.afi.es
[13] http://www.cajacaminos.es/
[14] http://www.xbrl.org.es/informacion/ipp.html
[15] http://www.xbrl.org.es/informacion/dgi.html
[16] http://www.xbrl.org.es/informacion/es_be_fs.html

- **Relevant facts information**: models relevant facts about a given fund, such as changes in its investment policy. They allow for keeping historical track of the relevant changes of the fund.
- **Periodic descriptive values**: model descriptive information that is periodically updated, such as the NAV of the fund or the number of investors that own some shares of the fund.
- **Analytic information**: models the analytic values associated to a fund, such as performance measures, the rating of the fund in its category, its ranking, or different types of ratios (volatility, risk-profitability relation, etc).

The reason for identifying these four distinct groups of information (being the root of each group an XBRL tuple) is that the information they contain has a different nature, the sources providing the information are different, and the periodicity with which each group of information is produced is diverse. Besides the information elements created, the following linkbases have been defined:

- **Presentation linkbase** (*AFIFunds-presentation.xml* in Figure 3-5) defines how the information elements are presented. An extended link has been created for each of the information groups, and a parent-child hierarchy has been defined for the presentation of the elements of each of the groups.
- **Label linkbase** (*AFIFunds-label.xml* in Figure 3-5) defines labels for each information element. Only labels in Spanish have been defined so far, as the taxonomy is intended to be used in the Spanish market.
- **Calculation linkbase** (*AFIFunds-calculation.xml* in Figure 3-5) only links to validate that the percentage of the different types of assets sums up a 100% have been created. As it will be explained in the next subsection, other links could not been defined as the current version of XBRL does not provide enough expressivity for it.
- **Reference linkbase** (*AFIFunds-reference.xml* in Figure 3-5) associates references to information elements, providing a detailed explanation of their meaning.

Definition links have not been used because: a) the use of links of type *requires-child* is not recommended in (Hamscher 2005), b) there are no equivalent elements in the taxonomy, so links of type *essence-alias* have not been used, c) no use was found for links of type *general-special*, and d) there are no similar tuples for which a link of type *similar-tuples* makes sense.

The last version of the taxonomy can be found at http://www.tifbrewery.com/tifBrewery/resources/XBRLTaxonomies.zip.

3.2.3 Limitations of calculation links

XBRL provides calculation links that allow for the description of the mathematical relation between different (numerical) information items. However, the current version of the XBRL specification has some important limitations in what can be expressed by such links.

First, the investment funds taxonomy should include validations that involve the evaluation of information items in different contexts. For example, we want to validate that a given NAV is not more than a 15% higher or lower than the previous NAV known for that fund. That requires expressing some mathematical relation between the same information element e.g. NAV at different points in time given by XBRL contexts. However, the current XBRL specification does not allow for this kind of validation, and calculation links are defined between information items independently of their context.

Second, XBRL calculation links only allow for the summation of items. However, there are some analytical values whose calculation from descriptive values is much more complex, involving the use of a wider range of mathematical operators. This is the case of, for example, the calculation of most of the performance measures used.

Future versions of XBRL are expected to overcome these limitations, and the requirements for future formula linkbases that extend the current calculation linkbases are already an XBRL candidate recommendation (Hamscher 2005).

3.3 Translating XBRL taxonomies into OWL ontologies

OWL is a potential alternative to the use of XBRL which presents some features that are of practical interest in the investment funds market. For this reason, we have developed a generic translation process from XBRL taxonomies to OWL ontologies so that existing and future taxonomies can be easily converted into OWL ontologies with the purpose of exploring the advantages of models with formal semantics with respect to XBRL taxonomies. In this section, we present the translation process designed and a discussion on the advantages and disadvantages of using OWL.

3.3.1 Description of the translation process

XBRL taxonomies provide explicit and shared information models and, thus, they are very similar to ontologies except that they do not have a formal semantics for all the aspects of the model. Similarly, XBRL instances

can be seen as ontology instances and expressed as such. Therefore, we have designed a translation process of XBRL taxonomies into OWL ontologies, and of XBRL instances into OWL instances. In the following, we will restrict ourselves to the translation of taxonomies into ontologies.

An automatic translator has been implemented based on the process that will be presented. It has been tested by translated not only our funds taxonomy but also other XBRL taxonomies available at the International[17] and Spanish[18] XBRL official Web pages. Specifically, DGI, IFRS-GP[19], ES-BE-FS and IPP taxonomies have been translated. The last version of the obtained ontologies can be found at http://www.tifbrewery.com/tifBrewery/resources/OWLOntologiesv2.zip.

In Figure 3-6 we show the architecture of the translator. As XBRL is an XML[20] based technology, the first step in the translation process is to parse the XML elements. Using JDOM[21], the XML parsing module obtains the XML elements in the XBRL taxonomies, instances, and links to be translated. The translation steps that will be described below are then applied to the obtained elements, resulting in a Jena[22] model that provides us with a programmatic environment to OWL. The model, corresponding to the OWL ontologies and instances derived from the XBRL taxonomy and instances, is finally saved to text files.

The different types of XBRL elements, the hierarchy and relationships between elements within a common taxonomy, and the relationships among several taxonomies will establish their order of translation in our proposal. In the following we describe the steps involved in the automatic translations, which are summarised in Table 1-1. For the sake of simplicity, we will reference the DGI taxonomy in the explanations. The transformation process for other taxonomies follows the same structure.

[17] http://www.xbrl.org
[18] http://www.xbrl.org.es
[19] http://xbrl.iasb.org/int/fr/ifrs/gp/2005-05-15
[20] http://www.w3c.org/XML
[21] http://www.jdom.org
[22] http://jena.sourceforge.net

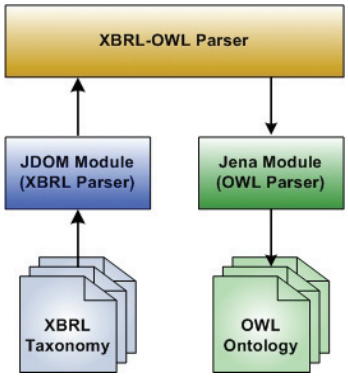

Figure 3-6. Syntactic translator architecture

Table 3-1. Summary of parsed taxonomy element translations

Parsed taxonomy element	Root OWL class	Direct OWL subclasses
XML complex type	DGI_ComplexType	A subclass for each complex type
XBRL tuples	DGI_Element	DGI_Tuple
XBRL items		DGI_Item
XLink links	DGI_Link	DGI_LabelLink DGI_PresentationLink DGI_CalculationLink
XBRL contexts	Context (the ranges of its properties are subclasses of ContextElement)	Subclasses of ContextElement: ContextEntity ContextEntityElement (Identifier) ContextPeriod ContextScenario
XBRL units	Unit (the ranges of its properties are subclasses of UnitElement)	Subclasses of UnitElement: UnitMeasure

1. **Declaration of a root OWL class Element** from which complex (tuples) and simple (items) information parts of the taxonomy will inherit, named DGI_Element for the DGI taxonomy. This class has associated a property *xbrl_id*, corresponding to the XBRL attribute *id* common to all XBRL elements.
2. **Declaration of DGI_Tuple and DGI_Item, subclasses of DGI Element**. XBRL tuples and items correspond to OWL subclasses of DGI_Tuple and DGI_Item, respectively. The attributes of XBRL Item are translated into the OWL properties: *xbrl_balance*, with possible values “credit” and “debit”; *xbrl_periodType*, with possible values “instant” and “duration”; *xbrl_contextRef*, whose range is the OWL class

Context (step 11); and *xbrl_unitRef*, whose range is the OWL class Unit (step 12).

3. **Declaration of a root OWL class DGI ComplexType**. XML complex types are translated into subclasses of DGI_ComplexType, having OWL properties: *xml_name* to store the name of the complex type, *xbrl_periodType*, with possible values "instant" and "duration", and *xbrl_contextRef*, whose range is the Context class.
4. **Syntactic translation of XML complex types into OWL subclasses of DGI_ComplexType**. The names of the obtained subclasses are those stored in the XML attribute *name* of the complex type elements. Each subclass of DGI_ComplexType has a property whose name is the concatenation of the complex type name and the word "value", and whose type is the primitive data type associated to the complex type (xsd:string, xsd:integer, xsd:boolean, etc.). Additionally, they contain those properties defined in the primitive XBRL data types (xbri:stringItemType, xbrli:integerItemType, xbrli:booleanItemType, etc.). For example, in the DGI taxonomy, the class AddressFormatCodeItemType has the property *length* with a fixed value of 2, indicating that the possible values of the data type can only have 2 characters.
5. **Syntactic translation of XBRL Items into OWL subclasses of DGI_Item**. The names of the obtained subclasses are those stored in the XML attribute *name* of the item elements. Each subclass of DGI_Item has a property for storing the value of the item, and whose range is the type of the XBRL Item.
6. **Record XBRL Tuples as OWL subclasses of DGI_Tuple**. Initially, the classes are created empty, and their properties are added in step 7. The reason is that tuple properties will reference other tuples, which might be not yet created and which will have to exist in the OWL model that is being built.
7. **Syntactic translation of the XBRL tuple attributes into OWL object properties**. The attributes of the tuples are added to the subclasses of DGI_Tuple as OWL object properties. These properties will have as range a class associated to a complex type of step 4, a class created in step 5 or a class recorded in step 6.
8. **Declaration of a root OWL class DGI_Link**. Its instances, which correspond to the XLink links of the XBRL taxonomies, contain the properties: *xlink_from*, created for the translation of the XLink attribute *from*, stores the origin element of the link; *xlink_to*, created for the translation of the XLink attribute *to*, indicates the destination element of the link; *xlink_role*, created for the translation of the XLink attribute

role, indicates the role assigned to the link: “label”, “calculation”, “presentation”, etc.

9. **Declaration of OWL subclasses of DGI_Link**. Subclasses of DGI_Link are built for each type of link: DGI_LabelLink, DGI_PresentationLink, DGI_CalculationLink, DGI_ReferenceLink, and DGI_DefinitionLinks.
10. **Syntactic translation of XBRL linkbases into instances of the corresponding subclasses of DGI_Link**. Links in XBRL linkbases are translated into OWL instances of the different subclasses of DGI_Link (for reasons of space, only the translation of label, presentation and calculation linkbases is presented):
    - Label links are translated into OWL instances of DGI_LabelLink. In addition to the common link properties (*from*, *to*, *role*), label links have properties: *xbrl_label*, obtained from the translation of the XBRL attribute *label* and used to store the text of the label, and *xml_lang*, obtained from the translation of the XML attribute *lang* and used to indicate the language of the label.
    - Presentation links are translated into instances of DGI_PresentationLink. Besides common link properties, presentation links have properties: *xbrl_order*, from the translation of the attribute *order* and used to store the relative position of the destination element within the presentation of the origin element, and *xbrl_preferedLabel*, obtained from the translation of *preferedLabel*.
    - Calculation links are translated into OWL instances of DGI_CalculationLink. Additionally to common link properties, calculation links have properties: *xbrl_order*, obtained from the translation of the XBRL attribute order and used to store the relative position of the destination element value within the calculation of the origin element value, and *xbrl_weight*, obtained from the translation of the XBRL attribute weight and used to store the weight of the destination value within the calculation of the origin element value.
11. **Syntactic translation of XBRL contextRef elements**. In order to translate XBRL contexts, a new ontology has been created, which will be imported by the ontologies resulting from the translation of XBRL taxonomies. This ontology contains a main class Context. The Context class has the following properties: a) *xbrl_id*, of type xsd:ID, for the translation of the XBRL attribute *id* to identify each context, b) *xbrl_entity*, of type ContextEntity, defined for the translation of *entity*, c) *xbrl_period*, of type ContextPeriod, defined for the translation of *period*,

and d) *xbrl_scenario*, of type OWL Thing, and defined for the translation of *scenario*. Other classes such as ContextEntityElement, ContextPeriod (with subclasses ContextForeverPeriod, ContextInstantPeriod, and ContextStartEndPeriod), and ContextScenario are defined corresponding to the types of values that define an XBRL context.

12. **Syntactic translation of XBRL unitRef elements**. For the translation of units defined in an XBRL taxonomy, an independent OWL ontology has been created. This ontology will be imported by ontologies resulting from the translation process. Its main class is Unit, which has a property *xbrl_unitMeasure* of type UnitMeasure and whose content is the definition of the associated unit. The UnitMeasure class, used to define the units added in a given context, does not have properties. Its subclasses distinguish the different types of units:
    - **Divide** for units defined by means of a ratio (with properties *xbrl_unitNumerator* and *xbrl_unitDenominator*).
    - **Measure** for simple units (with property *xbrl_measure*).

As mentioned before, besides the order of steps presented above, the hierarchy and relationships between elements within a taxonomy, and the relationships among different taxonomies, will define their translation order.

3.4 COMPARISON

The translation process presented in the previous sections helps to identify similarities and differences between XBRL taxonomies and OWL ontologies, which are described below.

- *XBRL items and tuples*. There is a natural correspondence between XBRL items, and tuples and OWL classes. While XBRL items correspond to classes that only have one value (besides information such as the period, context, etc.), XBRL tuples correspond to classes with object properties that store the constituent parts of the tuple. In this sense, XBRL items and tuples can be naturally represented by OWL classes.
- *XBRL contexts and units*. An important feature of XBRL is the possibility of associating contexts and units to XBRL elements. This can also be done in OWL by creating ontologies for contexts and units, as presented in the previous subsection, and by including appropriate object properties in OWL classes representing XBRL items and tuples. Therefore, we conclude that this type of information can be easily ontologically represented.

- *Reference and label links*. Reference and label links can be represented in OWL by creating appropriate classes and instances, as it has be done by our translation process. Notice that these links are intended for documentation purposes, and no formal semantics is associated to them. Furthermore, no application of a possible formal semantics for this type of links envisioned.
- *Definition links*. Definition links can be represented by creating instances of the classes introduced in the previous subsection. Special attention deserves the representation of *general-special* definition links which, even though they are currently translated into instances of definition link classes, naturally correspond to subclass relations in ontologies. However, existing taxonomies e.g. IPP, DGI, or IFRS-GP hardly make use of general-special definition links. A reason for this is that this type of links is not exploited by current XBRL tools to infer additional information, as this kind of relation does not currently have a formal semantics. We believe that the formalisation of subclass relations can be of interest in practical applications, and that general-special definition links could be given formal semantics by using OWL.
- *Calculation links*. Calculation links can be represented in the way outlined in the previous section. However, these links have a formal, mathematical semantics in XBRL, while in OWL this semantics is not supported. Therefore, we believe that for OWL ontologies to be adopted in the financial domain in general and in the investment funds market in particular, where mathematical relations are highly relevant for data validation, linking OWL to some form of mathematical support would be required.
- *Presentation links*. Presentation links can be represented as described by our translation process. However, OWL tools should exploit this presentation information for data visualisation. Therefore, visualisation tools should be adapted to take into account presentation information, not currently available in OWL.
- *Open-World Assumption (OWA) vs. Closed-World Assumption (CWA)*. The semantics of OWL is based on classical First-Order Logic, FOL (Fitting 1996), and the OWA is made, i.e., information is not assumed to be false if it cannot be proven to be true. However, in an industrial setting the CWA is widely made e.g. in relational databases. In fact, XBRL users are expected to intuitively make the CWA when, for example, querying for particular information of an investment fund. Due to his background, an average user would most likely see natural a “no” answer to the question “Is the investment fund *myFund* classified in category *myCategory*?” if, according to the available information, the

investment fund is not classified under this category. Locally closing the world using an epistemic operator for OWL could be a solution to this problem (Donini et al. 1998; Heflin et al. 2002). In addition, OWL does not define constraints but restrictions, as explained in (de Brujin et al. 2005b). However, for validation purposes we believe that the use of constraints, and not of restrictions, is required.

Summarising, the major advantage we see from the use of OWL is its formal semantics, which can be exploited for the automatic classification of funds if general-special relations are used and represented as OWL subclass (or subsumption) relations. As implicit subsumption relations can be automatically inferred using Description Logics reasoners (Nardi et al. 2003), customers or analysts can e.g. formally define the characteristics of funds they are interested in and appropriate funds will automatically and precisely be found. In particular, we are investigating the application of formal semantics to personalisation in the reception of information in the investment funds market and to the automated classification of funds. For this purpose, we can analyse subsumption relations present in current taxonomies but not explicitly declared. However, the Open-World semantics of OWL and the use of restrictions instead of constraints can hamper the use of OWL for querying investment funds information and for validating information reported.

Extensions of OWL to incorporate and automatically validate mathematical relations in the style of XBRL should be built, and current OWL tools should incorporate presentation information in ontologies so that they can be visualised according to different presentation specifications.

Other alternative languages for the formal description of investment funds can be considered, like the WSML family of languages (de Brujin et al. 2005a), which provides a basic interoperability layer and extensions in the direction of Description Logics and in the direction of Logic Programming.

4. CONCLUSIONS

Semantic Web technologies promise an improvement in how information is currently described, managed, integrated, searched and exchanged based on the definition of explicit, shared and formal models of a given domain. The financial domain, in which information is complex and valuable, and where big volumes of information are daily exchanged, can naturally benefit from the use of explicit domain models, shared by all actors in different financial markets, and for which standard inference mechanisms can be

applied for e.g. improve search results or better adapt information to investor profiles.

We have presented in this chapter the results of two investigations we have conducted: the development of an ontology-based platform for enhancing current practices in the management of economic and financial information, and the definition of a domain model for the investment funds market. These investigations have demonstrated some of the benefits of using semantic technologies, which are not exclusive for the financial domain but of which businesses in different fields can take advantage of.

While semantic technologies offer improvements in different aspects of information integration, management and exchange, these improvements are not possible without the definition of a shared and explicit model. However, the task of building such a model is challenging, especially if the model is meant to be shared beyond organisational boundaries to improve information exchange and (data) interoperability with external systems.

In fact, the biggest benefit of a semantic approach to information management is the construction of a model shared and agreed by all actors in the market. In this sense, models not necessarily semantic such as XBRL taxonomies being developed in finance are a valuable outcome; bringing commercial banks, financial institutions, central banks and other actors together to define shared models, as the XBRL community is doing, is good news for the achievement of an improvement in current information management and exchange practices. In fact, XBRL is being promoted by public institutions such as the Committee of European Banking Supervisors (CEBS)[23], which includes high level representatives from the banking supervisors and central banks of the European Union. CEBS has promoted the creation of working groups that have the mission of defining XBRL taxonomies to be later adapted and used for the financial reporting that banks and other institutions have to submit periodically to the banking supervisors.

While the models created by the XBRL community lack formal semantics, they possibly reflect the most difficult think to achieve when using semantic technologies: building a model most parties in a business domain can agree upon. Translation processes from non-semantic models to formal ontologies, such as the one we have presented in this chapter, become crucial in this context, as we can *ontologise* agreed models (Hepp et al 2006) and, thus, apply semantic search, visualisation, etc. to these models.

We can see, by following the activities of the XBRL community, how the awareness of the need of explicit and agreed domain models is dramatically increasing in the financial domain. This can ease the uptake, in the near

[23] www.c-ebs.org

future, of semantic technologies in finance. However, there are some barriers for such uptake, mainly:

1. the formal semantics of current languages such as OWL is not straightforward neither for business users in finance nor for IT staff; especially, the Open World Assumption made by the OWL language, while reasonable in the context of an open and distributed source of information like the Web, it is a bit unnatural in a more closed context like financial markets,
2. tool support and expertise in semantic technologies by IT developers is still not sufficient,
3. the Semantic Web community has not paid so far enough attention to the achievements of other communities, especially of the XBRL community, in building shared models; while the languages used by these communities are different, they share the goal of building explicit models enabling a better processing of information,
4. semantic languages such as OWL have been designed as general purpose languages, i.e., as languages to cover the description of any possible domain. However, domain-specific extensions are required in the financial domain, such as the support of complex calculation relations.

In a nutshell, companies and other institutions in the financial domain, as well as customers, can benefit from the advantages of using semantic technologies, namely: a better processing, management, search, visualisation and exchange of information. However, for such benefits to be achieved, and for semantic technologies to be widely adopted in the financial field, the problems discussed above have to be overcome.

5. QUESTIONS FOR DISCUSSION

Beginner:

1. Why are current content management systems not semantic?
2. How would you currently find e.g. an investment fund or a mortgage meeting your needs? How could this search be improved if firms commercialising these products would describe them using an ontology?
3. What kind of information does your company (or school/university/institution) manage? Is there an explicit model of this information which can be (or is) communicated to users? What kind of language is used to represent it?

Intermediate:

1. What are the differences and similarities between an XML Schema, an XBRL taxonomy, and an OWL ontology?
2. How do tagging, taxonomies, ontologies and Semantic Web relate? What role can these different concepts/technologies play in improving e.g. the search of a loan meeting certain requirements? And in improving the exchange of information?
3. What initiatives exist for improving regulatory reporting to central banks by using shared, explicit models? How can these initiatives be extended beyond regulatory reporting and semi-formal models? TIP: visit www.xbrl.org.

Advanced:

1. What type of applications/domains can benefit from the Open World Assumption (OWA) made by languages such as OWL? And from the Closed World Assumption (CWA)? Is the OWA or the CWA made by XBRL?
2. Do you think banks and investment firms will be willing to semantically annotate their products and make these descriptions publicly available? What reasons would they have for and against this initiative? How would you convince these institutions to follow this initiative? Are intermediate solutions possible?
3. How would you extend OWL to incorporate XBRL calculation links?

Practical Exercises:

1. Imagine you are a financial analyst who wants to launch a new Web-based service to guide users on where to invest their money depending on their profile (especially on how much risk they are willing to assume and for how long they are willing to put their money in some investment instrument):
    a. Find sites and companies who supply information about different investment products (investment funds, pension plans, stock markets, deposits, etc.)
    b. Analyse how you can integrate information from all these sources in order to have a complete knowledge base of investment products you can use for your investment recommendations to users.
    c. Describe how your integration approach would react to changes in the structure of the information provided by one of your sources.
    d. Describe how would you model and manage user profiles, and how you would match them against product descriptions.

e. Analyse how the definition of a model of investment products shared by all your information sources would improve your new service and its profitability.
f. Define a simple ontology of investment products.
g. Think of applications of the formal semantics of your ontology to improve your investment recommendation service.

6. SUGGESTED ADDITIONAL READING

- Alexiev, V., Breu, M., de Bruijn, J., Fensel, D., Lara, R. and Lausen, H. (2005). Information Integration with Ontologies: Experiences from an Industrial Showcase, Wiley, 2005. This book describes the application of semantic technologies in the automotive industry, including the annotation of information, the modelling of the domain, and the benefits achieved by the use of semantic technologies.
- Singh, M. P. (2004). The Practical Handbook of Internet Computing, Chapman & Hall/CRC, 2004. The third part of this book is devoted to different information management techniques, giving a good overview of different approaches, including the use of formal semantics.

7. REFERENCES

Alexiev, V., Breu, M., de Bruijn, J., Fensel, D., Lara, R. and Lausen, H. (2005). Information Integration with Ontologies: Experiences from an Industrial Showcase. Wiley, 2005.

Bechhofer, S., Harmelen, F. V., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F. and Stein, L. A. (2004). OWL Web Ontology Language Reference. Technical report, W3C recommendation. http://www.w3.org/TR/2004/REC-owl-ref-20040210.

Berners-Lee, T., Handler, J., Lassila, O. (2001). The Semantic Web, Scientific American, **64**(5):34-43.

Brickley, D. and Guha, R. V. (2004). RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation, 10 February 2004.

de Brujin, J., Lausen, H., Krummenacher, R., Pollers, A., Predoiu, L., Kifer, M. and Fensel, D. (2005a). The Web Service Modelling Language WSML. Technical report, WSML, 2005.

de Bruijn, J., Polleres, A., Lara, R. and Fensel, D. (2005b). OWL DL vs. OWL Flight: Conceptual Modelling and Reasoning for the semantic Web. In Proceedings of the 14th World Wide Web Conference (WWW 2005), Tokyo, Japan, May 2005.

Castells, P. and Macías J. A. (2001). An Adaptive Hypermedia Presentation Modelling System for Custom Knowledge Representations. World Conference on the WWW and Internet (Web-Net'2001). Orlando, 2001.

Castells, P., Foncillas, B. and Lara, R. (2004). Semantic Web Technologies for Economic and Financial Information Management. In Proceedings of the 1st European Semantic Web Symposium (ESWS 2004), Heraklion, Greece, May 2004.

Ding, Y. and Fensel, D.: Ontology Library Systems (2001). The key to successful Ontology Re-Use. In Proceedings of the 1st Semantic Web Working Symposium (SWWS 2001). California, USA, July 2001.

Donini, F. M., Lenzerini, M., Nardi, D., Nutt, W. and Schaerf, A. (1998). An Epistemic Operator for Description Logics. Artificial Intelligence, **100**(1-2):225-274.

Engel, P., Hamscher, W., Shuetrim, G., Kannon, D. V. and Wallis, H. (2005). XBRL eXtensible Business Reporting Language. Technical report, XBRL International recommendation. http://www.xbrl.org/Specification/XBRL-RECOMMENDATION-2003-12-31+Corrected-Errata-2005-11-07.htm.

Fast, K., Leise, F. and Steckel, M. (2002). What Is A Controlled Vocabulary? Boxes and Arrows, December 2002.

Fitting, M. (1996). First Order Logic and Automated Theorem Proving. Springer Verlag, 2nd edition, 1996.

Green, S. L., Delvin, S. J., Cannata, P. E. and Gómez, L. M. (1990). No Ifs, ANDs or Ors: A study of database querying. International Journal of Man-Machine Studies, 32 (3), pp. 303-326, 1990.

Gruber, T. (1993). A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, **5**(2):199-220.

Hamscher, W., Shuetrim, G. and Kannon, D. V. (2005). XBRL Formula Requirements. Technical report, XBRL International candidate recommendation. http://www.xbrl.org/technical/requirements/Formula-Req-CR-2005-06-21.htm.

Hearst, M. (1999). User Interfaces and Visualization. In Modern Information Retrieval. Addison-Wesley, pp.257-323, 1999.

Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K. and Yee K. (2002). Finding the Flow in Web Site Search, Communications of the ACM, 45 (9), September 2002.

Heflin, J. and Muñoz-Ávila, H. (2002). LCW-based Agent Planning for the Semantic Web. In Proceedings of the AAAI Workshop on Ontologies and the Semantic Web, Palo Alto, CA, USA, July 2002.

Hepp, M., Lytras, M. D., and Benjamins, V. R. (2006). Preface for OIS 2006. In: John F. Roddick et al.: Advances in Conceptual Modelling - Theory and Practice, ER 2006 Workshops. Springer Verlag LNCS 4231, 2006, pp. 269-270.

Lara, R., Cantador, I. and Castells, P. (2006). XBRL Taxonomies and OWL Ontologies for Investment Funds. In the 1st International Workshop on Ontologizing Industrial Standards at the 25th International Conference on Conceptual Modelling (ER2006), Tucson, AZ, USA, November 2006.

Nardi, D., Baader, F., Calvanese, D., McGuinness, D. L. and Patel-Schneider, P. F. (eds.). (2003). The Description Logic Handbook. Cambridge, 2003.

Seaborne, A. (2004). RDQL - A Query Language for RDF. W3C Member Submission, 9 January 2004.