The Knowledge Engineering Review, Vol. 22:4, 379–403.
doi:10.1017/S0269888907001130 Printed in the United Kingdom

Text-based domain ontology building using Tf-Idf and metric clusters techniques

YACINE REZGUI

Informatics Research Institute, University of Salford, Maxwell Building, M5 4WT, Salford, Greater Manchester; e-mail: Y.Rezgui@salford.ac.uk

Abstract

The paper describes the methodology used to develop a construction domain ontology, taking into account the wealth of existing semantic resources in the sector ranging from dictionaries to thesauri. Given the characteristics and settings of the construction industry, a modular, architecture-centric approach was adopted to structure and develop the ontology. The paper argues that taxonomies provide an ideal backbone for any ontology project. Therefore, a construction industry standard taxonomy was used to provide the seeds of the ontology, enriched and expanded with additional concepts extracted from large discipline-oriented document bases using information retrieval (IR) techniques.

1 Introduction

Ontologies have always formed an important area of traditional philosophy concerned with "the set of things whose existence is acknowledged by a particular theory or system of thought" (Lowe, 1995). They attracted interest in the early 1980s from the artificial intelligence community, and towards the end of the 1990s became a buzzword used in a variety of areas and applications, including enterprise modelling, e-commerce and knowledge management (KM) (Swartout, 1999; Welty and Guarino, 2001). In recent years, ontologies have been applied to a wide range of purposes, including semantic integration of heterogeneous databases (Partridge, 2002), content-based retrieval of yellow pages as well as product catalogues (Guarino and Welty, 2000), and management of corporate memory (CoMMA Consortium, 2002; Euzenat, 1996). In knowledge-retrieval systems—i.e. systems where the user can accurately represent content structures and formulate sophisticated queries—an ontology provides the primitives needed to formulate these queries as well as the necessary resource descriptions (Guarino *et al.*, 1999; Rezgui and Zarli, 2006; Rezgui, 2007a). More generally, in KM, ontologies provide a perspective to migrate from a document to a content-oriented view, where knowledge items are interlinked, combined and used (Staab *et al.*, 2001).

Various definitions of what forms an ontology have been formulated and have evolved over time. A good description of these can be found in Corcho *et al.* (2003). From the author's perspective, the best definition that capture's the essence of an ontology is the one given by Gruber (Gruber, 1994): "an ontology is a formal, explicit specification of a shared conceptualization". As elaborated in Studer *et al.* (1998): "*Conceptualization* refers to an abstract model of some phenomenon in the world which identifies the relevant concepts of that phenomenon. *Explicit* means that the types of concepts used and the constraints on their use are explicitly defined. *Formal* refers to the fact that the ontology should be machine processable". An ontology defines the basic terms and relations that form the vocabulary of a topic area (Neches *et al.*, 1991). An ontology may have very high-level terms or be dedicated to a specific domain or application (Swartout *et al.*, 1997).

The paper describes one of the first experiences in developing an ontology in the construction domain. Construction is a knowledge-intensive industry characterized by some unique work settings and virtual organization-like modus operandi that differentiate it from other industrial sectors (Rezgui, 2001, 2007b). These are summarized below:

- The industry is project-oriented with a tendency for actors to be involved concurrently in several projects with similar, but sometimes different, roles and responsibilities.
- Each construction project is a one-off and unique prototype with distinct characteristics, including the choice of construction system, materials, site topography and geology, and local environmental factors.
- The industry is fragmented and structured into a variety of disciplines (e.g., Architecture, Civil Engineering, Building Services) that have their own regulations and use their own technical jargon.
- A construction project may involve up to 100 non-collocated organizations, with various levels of IT maturity and capability.
- Values, norms and cultures tend to vary from one organization to another in the sector, which is reflected in their work practices and business processes. The industry is dominated by small and medium-sized enterprises.
- The business relationships between organizations tend to be temporary and often of a short-term nature.
- Information and Communication Technology (ICT) provision tends to be fragmented (i.e. application- and discipline-oriented, dedicated to engineering functions), with little full project lifecycle support. This has led to the development of a plethora of discipline-oriented semantic resources.
- Information is conveyed using mainly unstructured documents, which promote the creation and sharing of human-interpretable information. A typical medium-sized project generates thousands of documents.

This overall context has often resulted in information inconsistencies, business process inefficiencies and change control and regulatory compliance problems. Actors working in the sector have to deal with information and knowledge originating from different sources. These sources can be classified into three main categories, namely *Domain knowledge* (administrative information, standards, technical rules, product databases, etc.), *Corporate knowledge* (the intellectual capital of a company) and *Project knowledge* (the knowledge that each company has about the project as well as the knowledge created through interaction with other companies).

Several initiatives at a national and European level have tried to overcome the above limitations by developing dictionaries, thesauri and several linguistic resources focused on construction terms to facilitate communication and improve understanding between the various stakeholders operating on a project. However, these initiatives tend to be country-specific and not adapted to the multinational nature of the sector.

The research presented in this paper illustrates a methodology, which was used in practice, to develop a significant construction domain ontology that can help overcome the process and knowledge management limitations faced in the multinational and multidisciplinary setting of a construction project. Despite the large number of semantic resources developed in the sector, these tend to be:

- Discipline-oriented and geared towards servicing specific user community needs or applications (e.g. structural analysis and thermal simulation);
- incomplete in terms of the concepts and relationships between the concepts covered and conveyed (generally, only relationships of type ‘IS-A’ are described, as is the case in construction classification systems).

Therefore, none of these (a) enable effective knowledge management across projects and organizations, (b) are accepted across disciplines and (c) can play the pivotal role of an ontology.

The paper describes the methodology used to develop a domain ontology, taking into account the wealth of existing semantic resources in the sector. While the approach relies on established information retrieval (IR) methods, the contribution of the research lies in (a) the way in which these were used and combined to provide a suitable solution to ontology building, in a domain (construction) where all related efforts have exclusively relied on human experts, and (b) the proposed ontology architecture developed specifically to address the characteristics and requirements of the construction sector. The ontology server application developed to maintain the ontology is briefly described. A dedicated set of knowledge management services centred on the ontology was developed, supporting the knowledge management needs of practitioners in the construction sector (Rezgui, 2006).

2 Related work

2.1 Ontology research

As reported by Corcho *et al*. (2003) and Cristani and Cuel (2005), a variety of methodologies have been developed for dealing with ontologies. These include:

(a) methodologies for ontology building (Gruninger and Fox, 1995; Blazquez *et al*., 1998; Pinto and Martins, 2000; Holsapple and Joschi, 2002; Kayed and Colomb, 2002; Sugumaran and Storey, 2002; Haava, 2004);
(b) methodologies for ontology reengineering (Klein, 2001);
(c) methodologies for ontology learning (Kietz *et al*., 2000; Aussenac-Gilles *et al*., 2001; Cimiano and Volker, 2005);
(d) methodologies for ontology evaluation (Kalfoglou and Robertson, 1999; Kietz *et al*., 2000; Guarino and Welty, 2000; Gomez-Perez, 2001)
(e) methodologies for ontology evolution (Klein *et al*., 2002; Noy and Klein, 2002; Stojanovic and Motik, 2002); and
(f) methodologies for ontology merging (Klein, 2001; Stumme and Maedche, 2001; Ganter and Stumme, 2003; Euzenat, 2004).

The coexistence of numerous methodologies suggests that a consensual methodology is difficult to establish due either to the lack of maturity of the field (Lopez, 1999) or to the difficulty of developing a methodology adaptable to different applications, sectors and settings. For instance, most of these methods and methodologies do not consider the collaborative and distributed construction of ontologies. Few methods (Euzenat, 1996; Holsapple and Joschi, 2002) include a proposal for collaborative construction of consensual ontologies primarily using domain experts. They include a protocol for agreeing new pieces of knowledge with the rest of the knowledge architecture.

Existing methodologies for building ontologies can be distinguished by two main aspects:

(a) the degree of dependency of the ontology on its application field (i.e. application-dependent, semi-application-dependent or generic) and/or
(b) the mechanism for deriving the ontology: generalization or specialization.

Furthermore, in relation to the aspect (a), some methods use the application field as a starting point for building the ontology and are therefore application-dependent. This is the case of the knowledge about complex technical systems (KACTUS) project (Bernaras *et al*., 1996) and the On-To-Knowledge methodology (Staab *et al*., 2001). Other methodologies can be described as semi-application-dependent as they have the aspiration to be generic while using a given application as a reference. This is the case of the Gruninger and Fox methodology (Gruninger and Fox, 1995) and the method based on Sensus (Swartout *et al*., 1997). Finally, some methodologies are generic, hence application-independent, since the ontology development process is totally independent of the uses of the ontology. This is the case of Cyc (Lenat and Guha, 1990), Uschold and King methods (Uschold and King, 1995) and methontology (Fernandez-Lopez *et al*., 1999). With few exceptions, Methontology (Fernandez-Lopez *et al*., 1999) and Tove (Gruninger and Fox, 1995), it is worth noting that most reported methodologies proceed by building the ontology as a one-off exercise (Holsapple and Joschi, 2002).

In terms of the mechanism to derive the ontology (aspect b), if we compare the KACTUS and the Sensus methods, in the former the ontology is built by means of an abstraction process from an initial knowledge base, while, in the latter, an ontology skeleton (presented in the form of a simple taxonomic structure) is automatically generated from a huge ontology developed by merging and extracting the information from existing electronic resources, including the semantic database of WordNet and the English Dictionary.

A plethora of ontology management tools and environments have also been developed providing, essentially, graphical interfaces for managing the ontologies augmented with various functionalities (e.g. ontology integration and merging) or reasoning capabilities. The latter category includes KAON2 (`http://kaon2.semanticweb.org/`), SWOOP (Kalyanpur *et al.*, 2005), OntoEdit (Sure *et al.*, 2002), Oiled (Bechhofer *et al.*, 2001), Protégé (Kietz *et al.*, 2000; Noy *et al.*, 2000), Ontolingua (Farquhar *et al.*, 1996) and Ontosaurus (Lenat and Guha, 1990). Furthermore, as described in Guarino and Welty (2000), the former category (environments supporting ontology integration and merging) includes FCA-Merge (Stumme and Maedche, 2001), OntoMorph (Euzenat, 1996), Prompt (Noy and Musen, 2000) and Chimaera (McGuiness *et al.*, 2000). These ontology management environments usually provide support for one or several ontology representation and structuring models, including ontology interchange language (Horrocks and van Harmelen, 2001; Partridge, 2002), RDF (resource description framework) and OWL (ontology web language—based on RDF).

The initiatives identified in this section were reviewed with the objective of identifying suitable methodologies that can be employed to develop the construction ontology. In that respect, while similar efforts have been either generic (Lenat *et al.*, 1990), discipline/application-specific (Kalfoglou and Robertson, 1999) or applied to a different domain using different techniques [e.g. the medical sector using description logics (Hahn and Schultz, 2003)], the proposed approach makes use of established term frequency-inverse document frequency (tf-idf) (Baeza-Yates and Ribeiro-Neto, 1999; Salton and Buckley, 1988) and Metric Clusters (Baeza-Yates and Ribeiro-Neto, 1999) techniques to identify relevant ontological concepts and their relationships, as described later in the paper.

2.2 Semantic resources in the construction sector

A variety of semantic resources have been developed in the construction sector ranging from domain dictionaries to specialized taxonomies. The most notable efforts include the following:

- The BS6100 (Glossary of Building and Civil Engineering terms) (BS6100, 1992), produced by the British Standards Institution (the independent national body responsible for preparing British Standards). This is a rich and complete glossary that provides a comprehensive number of synonyms per term that can contribute towards any ontology development effort in the sector.
- The *bcXML* (eConstruct, 2001) is an extensible markup language XML vocabulary developed by the eConstruct IST project for the construction industry. Through bcXML, eConstruct has enabled the creation of 'requirements messages' that can be interpreted by computer applications and then find suitable products and services that meet those requirements.
- The international organization for standardization (ISO) 12006-2 (ISO 12006-2, 2001) is concerned with current classification needs, and builds on the experience of developing and using conventional classification systems.
- The IFC (IAI, 2006) (industrial foundation classes) model, developed by the IAI (international alliance for interoperability), has produced a specification of data structures supporting an electronic project model enabling data sharing across software applications.
- The omniclass (`http://www.occsnet.org/`) construction classification system, developed in Canada by the Construction Specifications Institute, addresses the construction industry's information management needs through a coordinated classification system.

Amongst the resources listed earlier, the IFC model is certainly the one that has the ability to provide the closest base to build a construction ontology. The architecture of the IFC model comprises four layers, namely the Resource Layer, the Core Layer, the Interoperability Layer and the Domain Layer. The Resource Layer includes low-level concepts or objects that are independent of both application and domain needs, but rely on other classes in the model for their existence. The core layer includes two levels of generalization, namely Kernel and Extensions. The Kernel provides the high-level concepts required within the scope of the current IFC release. It can be seen as a 'template' that defines the form in which all other schema within the model are developed. The core extensions provide a kind of specialization for the concepts defined in the Kernel. In other words, each Core Extension is a specialization of a class defined in the Kernel. The Interoperability Layer provides the schemata defining the concepts common to two or more domain models. Finally, the Domain Layer supports the creation of domain models (e.g. architecture, structure and quantity surveying).

While the semantic resources mentioned earlier represent the most notable and widely acknowledged efforts in the construction field, the table below gives a wider snapshot of the related semantic resources used today in Europe. These include: classification systems and their associated product libraries (used to define and classify construction product components), taxonomies (providing hierarchical conceptualization of construction disciplines) and various dictionaries/glossaries providing comprehensive definitions of construction terms and concepts.

3 The eCognos ontology development approach

As mentioned earlier, several ontology development methodologies have been reported in the literature. In fact, the choice of any methodology is very much dependent on the nature and characteristics of the targeted domain and its various applications, as well as the resources and development time available and the required depth of analysis of the ontology. A critical analysis of the semantic resources available in construction, including those listed in Table 1, combined with an understanding of the characteristics of the sector, have helped formulate a set of requirements that ought to be addressed in order to maximize the chances of a wide adoption of any ontology project in the construction sector. While these requirements are construction-specific in that they were derived from the characteristics of the sector summarized in Section 1, a subset of these may apply more generally to any ontology project. These requirements are listed below:

1. The ontology should not be developed from scratch, but should make as much use as possible of the established and recognized semantic resources in the domain.
2. The ontology should be built *collaboratively* in a *multi-user* environment: the construction sector involves several disciplines and communities of practice that use their own jargon and have specialized information needs.
3. There is a need to ensure *total lifecycle support*, as the information produced by one actor within one discipline should be able to be used by others working in related disciplines.
4. The ontology must be developed *incrementally* involving the end-users. This is important given the *multi-disciplinary* and *multi-project* nature of the industry, and the fact that each project is a *one-off prototype*.
5. The ontology should be *flexible* and *comprehensive* enough to accommodate different business scenarios used across projects and disciplines.
6. The ontology should be *user-friendly*, i.e. easy to use and providing a conceptualization of the discipline/domain being represented that *embeds the technical jargon* used in the sector.
7. The ontology should be a *living system* and should allow for future expansion.

Given the following factors:

(a) the fragmented and discipline-oriented nature of the construction sector;
(b) the various interpretations that exist of common concepts by different communities of practice (disciplines);

Table 1 Key semantic resources in Europe

	Uniclass	BS6100	Lexicon	BC building definitions	BARBi	SDC	Edibatec
Type	Classification system	Glossary	Vocabulary of terms	Taxonomy	Data library	Dictionary	Dictionary
Access	Free	Online access available to subscribers	Free	GNU public license	Free access after registration	Free	Free
Location	Thesaurus.foraec/data/uniclass/or RIBA Bookshop	http://Bsonline.techindex.co.uk	http://www.stabu-lexicon.com	http://www.bcxml.org	http://Edmserver.Epmtech.jotne.Com/barbi/index.jsp	http://www.sdc.biz	http://www.Edibatec.Org/Accueil/Default4.htm
Developed by	National building specification services	British standards institution	STABU foundation	IST e-construct project	Norwegian building research institute	Gencod EAN France	Edibatec
Published	1997	Various depending on parts	1997	2002	2003	2002	2000
Scope	Architecture and civil engineering	Building and construction	Building and construction	Building and construction	Building and construction	Building and construction	HVAC
Application	Product libraries	Terms definition	Product catalogues	e-procurement	Electronic commerce	e-commerce	e-catalogues

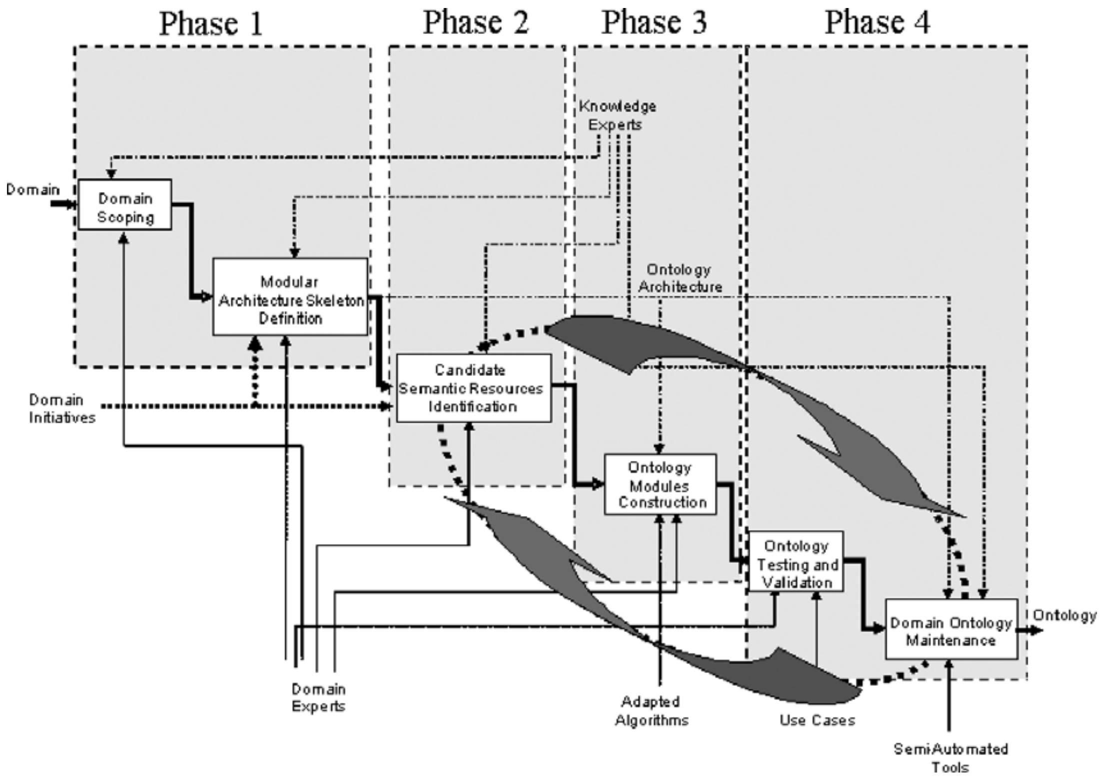

Figure 1 The various stages of the methodology

(c) the plethora of semantic resources that exist within each discipline (none of which have reached a consensual agreement);
(d) the lifecycle dimension of a construction project with the information being produced and updated at different stages of the design and build process with a strong information-sharing requirement across organizations and lifecycle stages,

a suitable methodology should accommodate the fact that the ontology should be specific enough to be accepted by practitioners within their own discipline, while providing a generic dimension that would promote communication and knowledge sharing amongst these communities.

The ontology should be developed incrementally in a collaborative way, involving representatives from the various disciplines, in order to promote ontological commitments and provide a mechanism whereby stakeholders share and exchange their perspectives and expertise (Hahn and Schultz, 2003). As mentioned earlier, textual documents have an important role in sharing and conveying knowledge and understanding. Therefore, a new methodology adapted to the construction sector was developed (referred to as the eCognos methodology), which draws on the strengths of several established methodologies described in Section 2.1. The proposed methodology, illustrated in Figure 1, comprises the following stages: domain scoping, modular architecture skeleton definition, candidate semantic resources identification, ontology modules construction, ontology testing and validation and ontology maintenance. For pragmatic reasons, these stages were grouped into four main phases and described as such in the following sections: Phase 1 (Domain scoping and architecture definition), Phase 2 (Candidate semantic resources identification), Phase 3 (Ontology modules construction) and Phase 4 (Ontology validation and maintenance).

Phase 1 addresses requirements 2, 3 and 7. It helps identify the boundary conditions of the construction domain, and defines a suitable ontology development architecture. Phase 2 addresses requirement 1. It takes as input the ontology architecture from Phase 1, as well as the domain

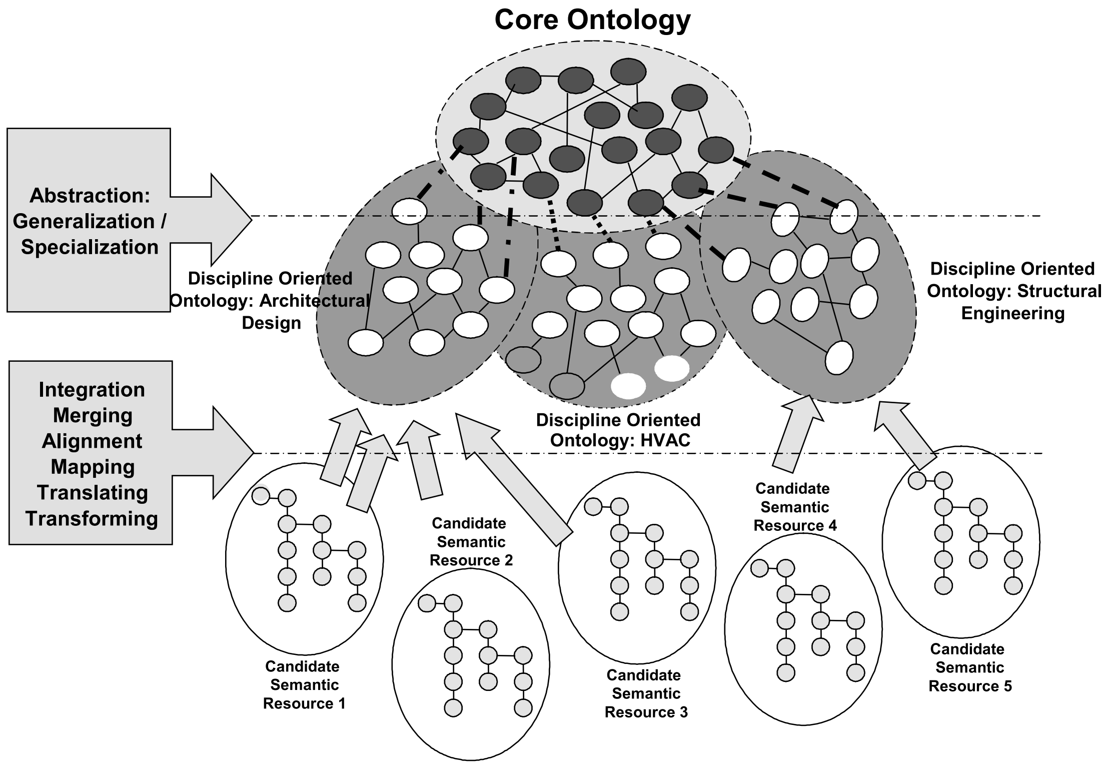

Figure 2 The eCognos ontology architecture

boundary conditions, and identifies relevant candidate semantic resources. Phase 3 addresses requirements 2, 3, 5 and 6. It takes as an input the identified candidate semantic resources from Phase 2, as well as the ontology architecture from Phase 1, and produces the ontology. Finally, Phase 4 addresses requirements 6 and 7. It tests, validates and maintains the ontology resulting from Phase 3. The decision was taken that the ontology would conform to an underlying knowledge model involving concepts, attributes and relations, defined using OWL.

4 Domain scoping and architecture definition

The scope of the ontology to be developed can be summarized by the following statement (eCognos Consortium, 2002): "In the context of a *Project*, a group of *Actors* uses a set of *Resources* to produce a set of *Products* following certain *Processes* within a work environment (*Related Disciplines*) and under well defined conditions (*Technical Topics*)". As such, the scope of the ontology can be conveyed through seven major themes: (1) Project, (2) Actor, (3) Resource, (4) Product, (5) Process, (6) Technical Topics (conditions) and (7) Related Disciplines (work environment).

The first five domains coincide with major themes in the IFC model. The last two domains include the related issues that are not directly covered by the IFC.

Given the requirements identified in Section 3, the eCognos ontology is structured into a set of discrete, core and discipline-oriented sub-ontologies referred to as modules (Figure 2). Each module features a high cohesion between its internal concepts, while ensuring a high degree of interoperability between them. These are organized into a layered architecture with, at a high level of abstraction, the core ontology that holds a common conceptualization of the whole construction domain enabled by a set of interrelated generic core concepts forming the seeds of the ontology. These generic concepts enable interoperability between specialized discipline-oriented modules defined at a lower level of abstraction. This middle layer of the architecture provides discipline-oriented conceptualizations of the construction domain. Concepts from these sub-ontologies are linked with the core concepts by generalization/specialization (commonly known

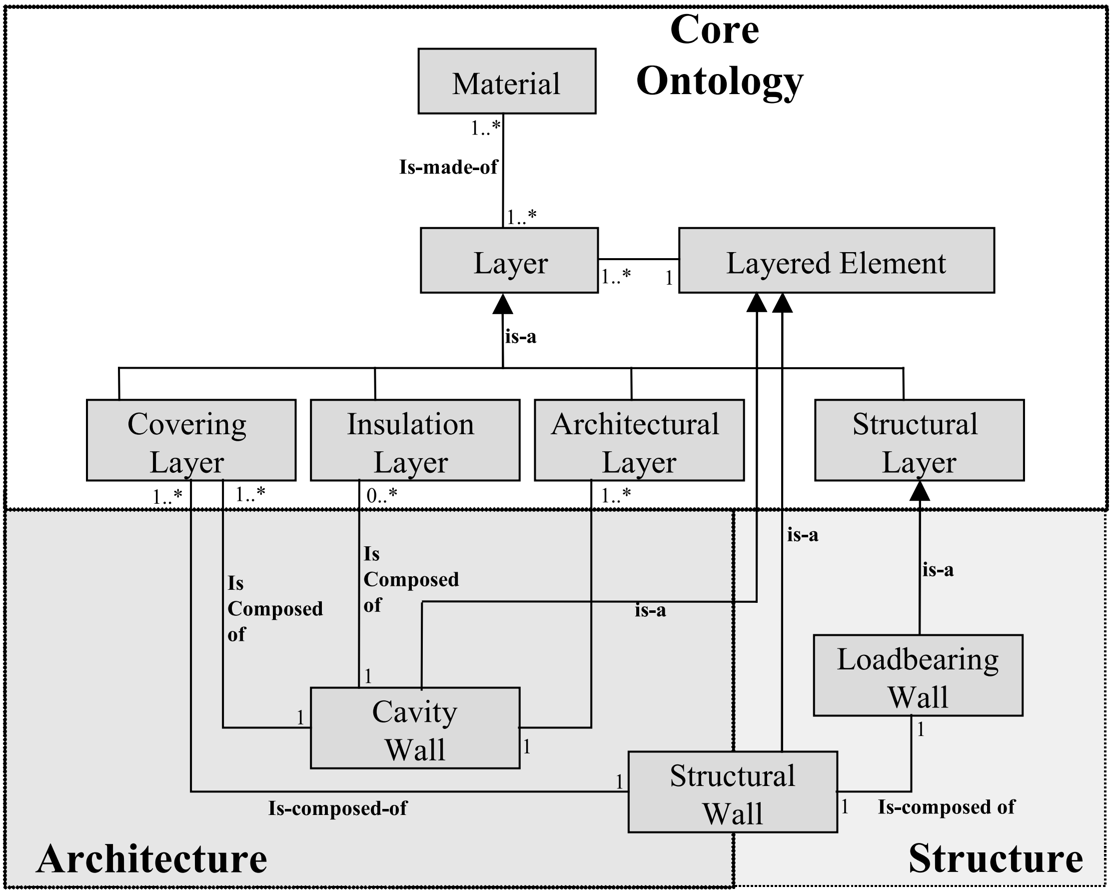

Figure 3 Illustration of interoperability around the concept of 'Wall' involving the 'Core', 'Architecture' and 'Structure' ontologies

as IS-A) relationships. The third and lowest level of the architecture (Figure 2) represents all the semantic resources currently available, which constitute potential candidates for inclusion into eCognos either at the core or discipline level.

Figure 3 illustrates the two upper levels of the ontology (Core ontology and discipline sub-ontologies) and the ways in which interoperability is supported. While architects (Sub-ontology Architecture) are responsible for the design of the building, they rely on structural engineers (Sub-ontology structure) to dimension all structural elements of the project. They, however, share key concepts such as the ones of *Loadbearing Wall* (which falls under the responsibility of the structural engineer, but forms an integral part of the architectural design) and *Cavity Wall* (which falls under the responsibility of the architect and does not require dimensioning by structural engineers, but is still important to them as may influence the dimensioning of the structure). Interoperability between concepts from both disciplines is ensured through the core ontology, whereby a *Loadbearing Wall* is a sub-type of *Structural Layer*, which in turn is part of a *Layered Element*. A layer is used to model the various constituents composing a finished wall, ranging from painting and finishes to cavity walls. The latter will directly be of interest to architects as they handle construction elements in their entirety to provide the most accurate dimensioning of their design.

5 Candidate semantic resources identification

There are a large variety of available semantic resources that can form the basis for building the eCognos Core Ontology. These range from classification systems to taxonomies. The latter do

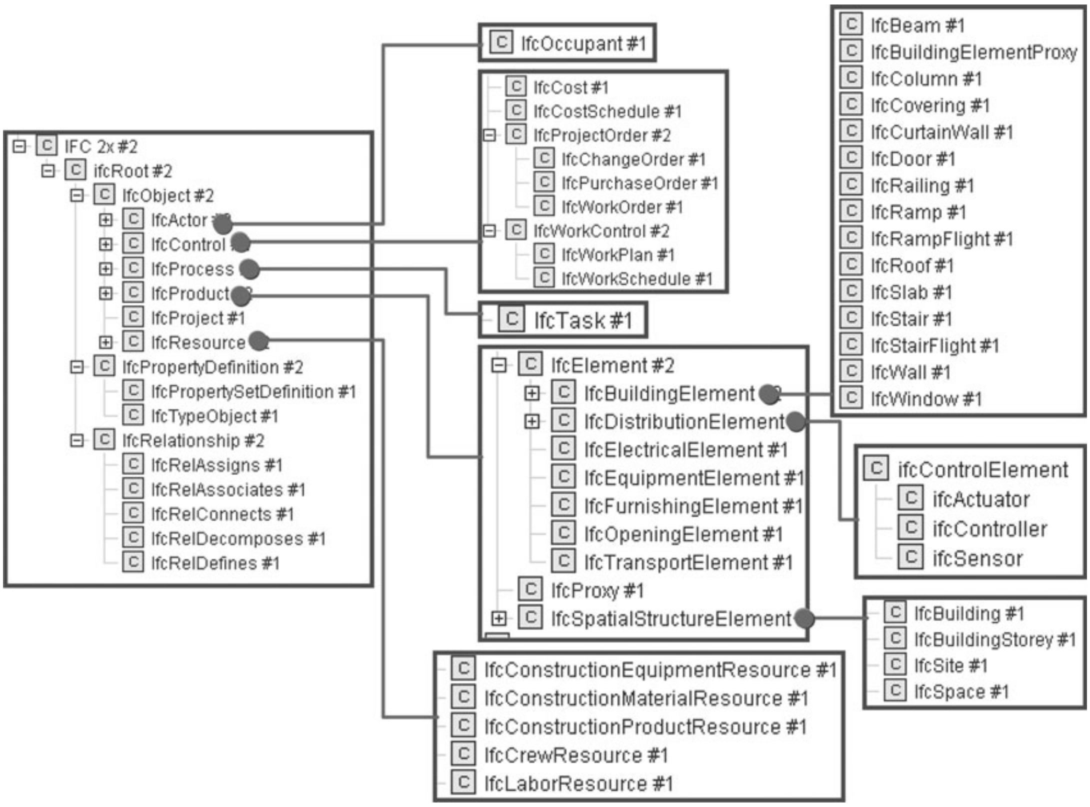

Figure 4 Illustration of the IFC kernel

deserve particular attention as argued in Guarino and Welty (2000) and Welty and Guarino (2001). Some of the principal roles of taxonomies are to facilitate human understanding, impart structure on an ontology and promote tenable integration. Furthermore, properly structured taxonomies:

- help bring substantial order to elements of a model;
- are particularly useful in presenting limited views of a model for human interpretation; and
- play a critical role in reuse and integration tasks. Improperly structured taxonomies have the opposite effect, making models confusing and difficult to reuse or reintegrate (Welty and Guarino, 2001).

IFC, being more recent and also the closest taxonomy currently in use in the sector, is therefore the preferred candidate semantic resource that can provide the skeleton on which such a core ontology can be built. The 85 major concepts composing the IFC Kernel (as described below) therefore formed the basis upon which to build the skeleton of the eCognos ontology. The wealth of information in related semantic resources, including BS6100, is used to enrich the eCognos ontology. As argued in the previous section, only generic concepts used across disciplines should be included in the core ontology. These concepts should hold minimal semantics necessary to establish a sound foundation for the ontology, supporting interoperability between the various discipline-oriented sub-ontologies.

The IFC model comprises several schemas organized according to the layer to which they belong. The schema IfcKernel defines the most abstract part within the IFC architecture, which contains the most abstract IFC entity—the IfcRoot. Each entity—defined in the core, interoperability or domain layer of the IFC model—inherits characteristics from the IfcRoot entity. There are three fundamental entity types in the IFC model derived from the IfcRoot, which form the first level of specialization within the IFC class hierarchy (Figure 4), namely: object (represents the generalization of any semantic entity within the IFC model), relation (represents the generalization of any relationship between IFC entities) and property (represents the generalization of any semantic item characteristic).

The entities of the *IfcKernel* schemata are organized in a taxonomy as illustrated in Figure 4. This was used as the skeleton of the eCognos Core Ontology.

The main advantages in using the IFC model include the following:

- The IFC model is an established standard that was strongly promoted in the construction sector.
- Each IFC entity has a defined set of mandatory attributes that define its minimal structure.
- A taxonomy, as argued in Guarino and Welty (2000) and Welty and Guarino (2001), provides an ideal backbone for any ontology project.

On the other hand, the IFC entities are regarded as 'very abstract concepts' that do not reflect the terms and language (jargon) used in end-users' daily work. In other words, they are not semantically rich enough to provide a true and user-friendly conceptualization of their discipline, and are therefore not sufficient to form a suitable ontology.

As far as the discipline-oriented ontologies are concerned, if a related conceptualization of the discipline exists, including a taxonomy, then this is selected as a candidate for the ontology. This process is detailed in the following section.

6 Ontology modules construction

This stage involves the technical evaluation of the previously identified candidate semantic resources by domain experts, with the objective of determining their suitability for reuse as discipline ontologies in accordance with the proposed modular architecture. In theory, once the semantic resources have been selected, several operations can be used in constructing an ontology. These include:

- *Integration*: the integration operation involves the integration of one ontology into another through their overlapping parts. While the integration operation provides a linkage and tight coupling between the two ontologies, their initial scope is preserved.
- *Merging*: the merging operation involves creating a new ontology from two or more existing ontologies, and bringing them as a result into mutual agreement, making them consistent and coherent. The initial ontologies cease existing in isolation and are replaced with the new ontology.
- *Translation*: this operation involves changing the representation formalism of an ontology while preserving its semantics.
- *Transformation*: this operation involves changing the semantics of an ontology to make it suitable for purposes other than the one for which it was originally conceived.

While these operations can be used to address the coexistence of several candidate semantic resources to construct and converge into a unique domain ontology (as illustrated in the lower layer of the architecture in Figure 2), the research argues that a taxonomy should be used wherever available as the preferred semantic resource for building a discipline-oriented ontology (Guarino and Welty, 2000; Welty and Guarino, 2001). In the absence of a taxonomy, the semantic resource that has clear support from industry and discipline practitioners should be retained. This can then be enriched and expanded using additional terms and relationships.

A particular approach was adopted for expanding the ontology. This involves selecting and making use of a large documentary corpus used in the discipline and ideally produced by the end-users. Expanding or building an ontology from the index terms extracted from commonly used documents in a given discipline requires a few operations organized into steps. These are described below and are applied to each document of the selected documentary corpus.

Step 1: document cleansing

This step aims at reducing the document to a textual description that contains nouns and associations between nouns that carry most of the document semantics. This involves the following steps:

- Lexical analysis of the text in order to deal with digits, hyphens, punctuation marks and the case of letters.
- Elimination of stopwords: this has the objective of filtering out non-content words, such as 'the' and 'of'.

Step 2: key word extraction

This step aims at providing a logical view of a document through summarization via a set of semantically relevant key words. These are referred to, in this stage, as index terms. The purpose is to gradually move from a full-text representation of the document to a higher-level representation. In order to reduce the complexity of the text, as well as the resulting computational costs, the index terms to be retained are all the nouns from the cleansed text. Several authors (Broglio *et al.*, 1995; Baeza-Yates and Ribeiro-Neto, 1999) argue that nouns, as opposed to verbs, adjectives and adverbs, carry out most of the semantics of a text document. This is assumed to be the case for the construction domain where technical documentation refers to noun-based terms, such as products, construction materials and resources. It is argued in this research that the type of semantics carried out by verbs in the construction sector describes and refers to relationships between ontology concepts. These relationships are identified in Step 4 of the proposed approach. However, concepts in the construction sector are sometime carried out by more than one term, such as 'Curtain Wall, and can therefore be conveyed by bigrams. More generally, the notion of an *n*-gram can be found in the literature when a sequence of *n* terms is regarded as a single object. While *n*-grams play a key role in capturing term collocations (Tomovic *et al.*, 2006), it is also worth noting that computation of *n*-gram probabilities does not accurately cater for rare term collocation events (Omelayenko, 2001). The approach presented here combines nouns co-occurring with a null syntactic distance (i.e. the number of words between the two nouns is null) into a single indexing component, regardless of their collocation frequency. These are referred to as noun groups (non-elementary index terms).

The result of this step is a set of elementary and non-elementary key words, which are representative of the discipline being conceptualized.

Step 3: integrating index terms into core and sub-ontology

Two types of concept integration are possible (Omelayenko, 2001): concept-level integration and syntactic-level integration. Concept-level integration requires inference over the domain ontology to make a decision about the integration of a particular pair of concepts. Syntactical integration defines the rules in terms of class and attribute names to be integrated. Such integration rules are conceptually blind, but are easy to implement and develop (Omelayenko, 2001). The approach used for concept-level integration makes use of a pivotal semantic resource, ideally a thesaurus or a taxonomy as illustrated in Figure 5.

As highlighted in Baeza-Yates and Ribeiro-Neto (1999), a glossary or thesaurus can provide a controlled vocabulary for the extension of the ontology based on the identified key words. A controlled vocabulary presents the advantage of normalized terms, the reduction of noise and the possibility of turning key words into concepts with clear semantic meaning. The construction BS6100 glossary, which is also structured as a taxonomy, was used.

For each identified key word, it is important to quantify the degree of importance (in terms of semantics) it has over not only the document but also the entire documentary corpus selected for

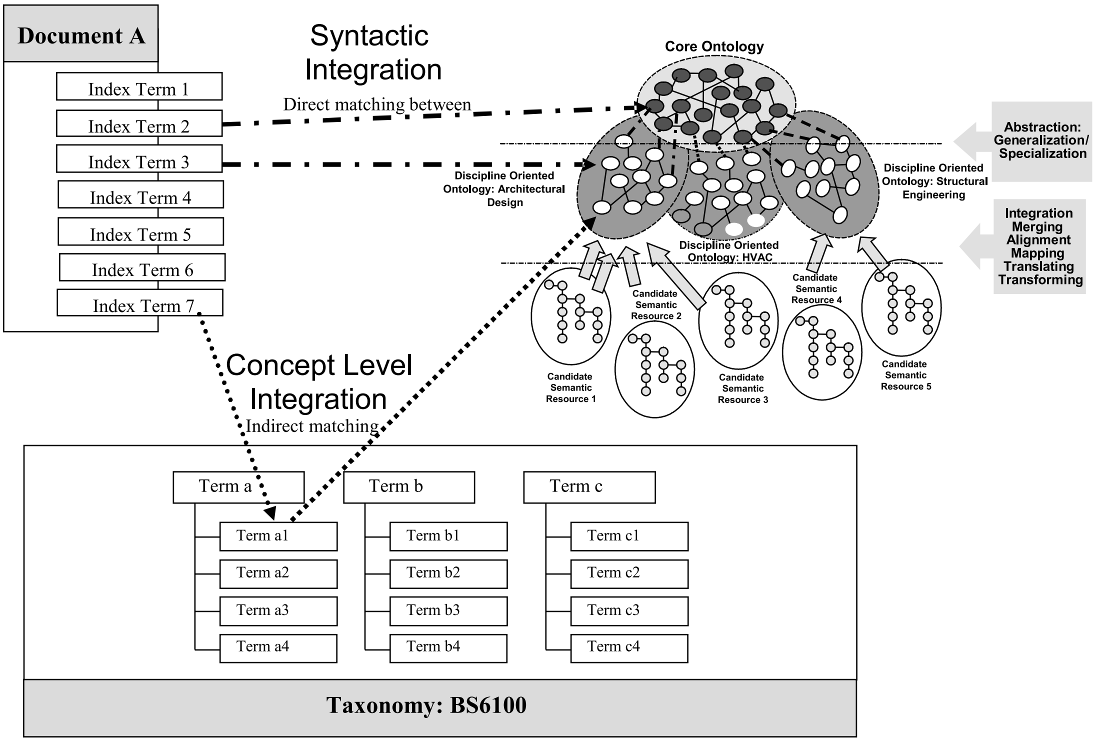

Figure 5 Method for concept integration making use of a pivotal semantic resource

the given discipline. The following formula, known as 'tf-idf' (Salton and Buckley, 1988; Baeza-Yates and Ribeiro-Neto, 1999), is used:

$$W_{i,j} = f_{i,j} \times idf_i \tag{1}$$

where $W_{i,j}$ represents the quantified weight that a term t_i has over the document d_j; $f_{i,j}$ represents the normalized occurrence of a term t_i in a document d_j, and is calculated using Equation (2):

$$f_{i,j} = \frac{\mathrm{freq}_{i,j}}{\max_{\text{for all terms in document}} \mathrm{freq}_{\text{term},j}} \tag{2}$$

where $\mathrm{freq}_{i,j}$ represents the number of times the term t_i is mentioned in the document d_j; $\max_{\text{for all terms in document}} \mathrm{freq}_{\text{term},j}$ computes the maximum over all terms that are mentioned in the text of document d_j; idf_i represents the inverse of the frequency of a term t_i among the documents in the entire knowledge base, and is expressed as shown in Equation (3):

$$idf_i = \log \frac{N}{n_i} \tag{3}$$

where N is the total number of documents in the knowledge base, and n_i the number of documents in which the term t_i appears. The intuition behind the measure of $W_{i,j}$ is motivated by the fact that the best terms for inclusion in the ontology are those featured in certain individual documents, capable of distinguishing them from the remainder of the collection. This implies that the best terms should have high term frequencies, but low overall collection frequencies. The term importance is therefore obtained by using the product of the term frequency and the inverse document frequency (Salton and Bukley, 1988).

The following hypotheses apply when comparing concepts from the ontology with the extracted key words from the documents:

- If an ontological concept and a document key word have the same name or a common stem, then they can be considered as semantically equivalent. This is motivated by the nature of the construction jargon where variations of the same noun are used in various contextual situations. This is best illustrated by 'Architecture' where variations of the concept exist, including '*architectural*' used to express an '*architectural* point of view' and '*architectural* solution'. While, when looked at in isolation, the extracted key word 'architectural' will be treated as semantically equivalent to the ontology concept 'architecture', the non-elementary index term 'architectural solution' (as explained earlier) will in fact be considered as a potential new concept for the ontology, to be validated by knowledge experts.

- If an ontological concept and a document key word have distinct names, both concepts are looked up in the BS6100 glossary. The algorithm checks first if both concepts are related through a specialization/generalization mechanism. In the affirmative, the key word is added as such to the ontology in accordance with the identified relationship, and is marked as 'new', which means that it will only be made officially part of the ontology once approved by the relevant knowledge expert(s). If the ontological concept and the document key word are not related through a specialization/generalization relationship, then the quantified semantic weight of the index term (Equation 1) is used as a criterion to decide whether it should be added as a concept to the ontology. At this stage, only index terms with a relatively high semantic weight were retained and added to the ontology with no relationship to other concepts defined at this stage.

In order to devise a threshold for the term weighting factor, an experiment was conducted on the 'Architecture domain' ontology, whereby domain experts were asked to analyze the retrieved concepts. A collection of 56 documents was used, out of which 480 concepts were extracted and ordered by the decreasing value of their highest associated term weighting factor (computed over each document of the related document collection). Out of the 480 extracted concepts, 447 concepts were new and not included in the initial existing taxonomy in Architecture. The three selected architect domain experts for the experiment analyzed and retained the top 468 concepts arguing that the remaining 17 concepts were not semantically relevant. This revealed that below the value of 0.22 of the term weighting factor, related concepts were either semantically equivalent to the ones defined above this value, or were deemed not worth including in the ontology. This figure (0.22) was then used as a threshold for retaining the concepts of the three other domain ontologies.

Finally, the properties of the new ontological concepts are either refined or defined for new concepts that have been integrated into the ontology (each property defined in the concept will receive, if applicable, ranges, restrictions, etc.). This task is undertaken manually by the knowledge experts.

Step 4: ontology concept relationship building

The next step in the methodology includes building the relationships connecting the concepts, including those that have not been retained in the previous stage. Concept relationships can be induced by patterns of co-occurrence within documents. As described in Baeza-Yates and Ribeiro-Neto (1999), overall relationships are usually of a hierarchical nature and most often involve associations between 'broader' and 'narrower' related terms. Broader term and narrower term relationships define a specialization hierarchy, where the broader term is associated with a class and its related narrower terms are associated with specialized instances of the class. While broader and narrower terms relationships can be defined automatically, dealing with related term relationships is much harder. One reason for this is that these relationships depend on the specific context and the particular needs of the group of users, and are thus difficult to establish without the knowledge provided by specialists. We therefore distinguish three main types of relationships:

- Generalization/Specialisation Relationship (e.g. Wall can be specialized into separation wall, structural wall, Loadbearing Separation Wall).
- Composition/Aggregation Relationship (e.g. Door is an aggregation of a Frame, a Handle, etc.).

- Semantic relationship between concepts (e.g. a Beam supports a Slab, and a Beam is supported by a Column).

The last two categories above are addressed in this step. The process is semi-automated in that relations are first identified automatically. Contributions from knowledge specialists are then requested to qualify and define the identified relations. In order to assess the relevance of relationships between concepts, an approach that factors the number of co-occurrences of concepts with their proximity in the text is adopted. This is known as the 'Metric Clusters' method (Baeza-Yates and Ribeiro-Neto, 1999) (Equation 4). This proceeds by factoring the distance between two terms in the computation of their correlation factor. The assumption being that terms which occur in the same sentence seem more correlated than terms that appear far away.

$$C_{\mathrm{u,v}} = \sum_{t_i \in V(S_u)} \sum_{t_j \in V(S_v)} \frac{1}{r(t_i,\ t_j)} \tag{4}$$

The distance $r(t_i, t_j)$ between two key words t_i and t_j is given by the number of words between them in the same document. $V(S_u)$ and $V(S_v)$ represent the sets of key words that have S_u and S_v as their respective stems. In order to simplify the correlation factor given in Equation 4, it was decided not to take into account the different syntactic variations of concepts within the text, and instead use Equation 5, where $r(t_u, t_v)$ represents the minimum distance (in terms of the number of separating words) between concepts t_u and t_v in any single document.

$$C_{\mathrm{u,v}} = \frac{1}{\mathrm{Min}[r(t_u, t_v)]} \tag{5}$$

The following exception should be noted: in the case where the minimum distance between two terms is null, the correlation factor will return to '∞'. When this exception arises, the correlated terms are considered as candidates to form a composite term. The assumption taken in the computation of Equation 4 is that the closer the correlation factor is to 1, the stronger the correlation between the two terms is likely to be. Based on the correlation factors, index terms are linked to relevant concepts from the ontology with a blind relationship. In order to establish a threshold for the correlation factor, it was decided that only terms co-occurring within the same sentence should be considered. An analysis of a sample set of documents showed the average size of a sentence to be 12 words. $C_{\mathrm{u,v}}$ has therefore been given a threshold of 0.1 (maximum distance between two terms co-occurring in a 12-words sentence). The decision to retain the average size of a sentence was motivated by the results of the experiment, as the threshold resulting from the use of a '12-words sentence' helped identify a limited but most of the time justified number of relationships.

The domain knowledge experts have the responsibility of validating the newly integrated index terms as well as their given names, and then defining all the concept associations that do not belong to the generalization/specialization category. First, these relationships were established at a high level within the Core Ontology, and then subsequent efforts have established relationships at lower levels within the discipline ontologies.

Four domains have been selected to develop sub-ontologies, namely Architecture, HVAC (Heating, Ventilation and Air Conditioning), Electricity, and Structure. These were chosen on the basis of the availability of technical documentation provided by the participating end-user organizations. A wide range of documents was used for each selected domain, including Full Specification Documents, related technical regulations, Bills of quantity and internal company technical documents. Given the project-oriented nature of the industry, these were selected across several projects so as to maximize the chances of including a broad range of the relevant concepts and terminologies used. Table 2 summarizes the useful information used to construct the various domain ontologies.

Table 2 Figures in relation to the ontology building approach

Ontology domain	Sub-ontology architecture	Sub-ontology HVAC	Sub-ontology electricity	Sub-ontology structure
Number of relevant concepts in (initial) selected taxonomy	33	31	19	13
Number of documents used to identify relevant terms and relationships	56	45	43	47
Number of extracted new ontological concepts from document set	447	125	99	175
Number of manually retained ontological concepts (non-overlapping or semantically equivalent)	435	111	75	145
Percentage of manually retained concepts (from extracted new ones)	97.3%	88.8%	75.7%	82.8%
Total number of concepts in sub-ontology	468	142	94	158
Percentage of newly (manually validated) added concepts	93%	78%	80%	91%
Number of automatically identified relationships	573	151	128	211
Number of manually recommended and retained relationships	445	125	78	149
Percentage of manually recommended and retained relationships	78%	83%	61%	71%

The use of discipline documents to identify ontological concepts and relationships was revealed to be the right strategy to construct the discipline sub-ontologies as illustrated in Table 2. Compared to the number of concepts found in the initial taxonomies, the use of documents helped identify 93, 78, 80 and 91% of the total number of concepts in the Architecture, HVAC, Electricity, and Structure sub-ontologies, respectively. As far as the relationships are concerned, the approach helped identify:

- the relations that exist between concepts found in the initial taxonomies (which only included relations of type 'IS-A'). For instance, IfcColumn, IfcBeam and IfcSlab are concepts that existed in the initial taxonomy as a specialization of Ifc Building Element. Relationships between these concepts have now been enriched in the Architecture sub-ontology by the following new semantic links: Beam *supports* Slab; Column *supports* Beam;
- new relations brought by the inclusion of a substantial number of new concepts. For instance, the new concept Opening has generated a number of new relations such as Opening *is performed on* Wall; Window or Door *is fitted on* Opening.

While the initial core ontology included 71 concepts, it has now been expanded and enriched in the light of the newly added concepts in the various sub-ontologies. The core ontology includes 114 concepts in its current version. A comprehensive version of the Core Ontology is illustrated in Figure 6. Also, it is worth noting that while the concepts in the initial taxonomies identified the core product components within each discipline, the use of documents helped identify concepts that provide a level of completeness, which makes the sub-ontologies self-sufficient to conduct end-users' business processes.

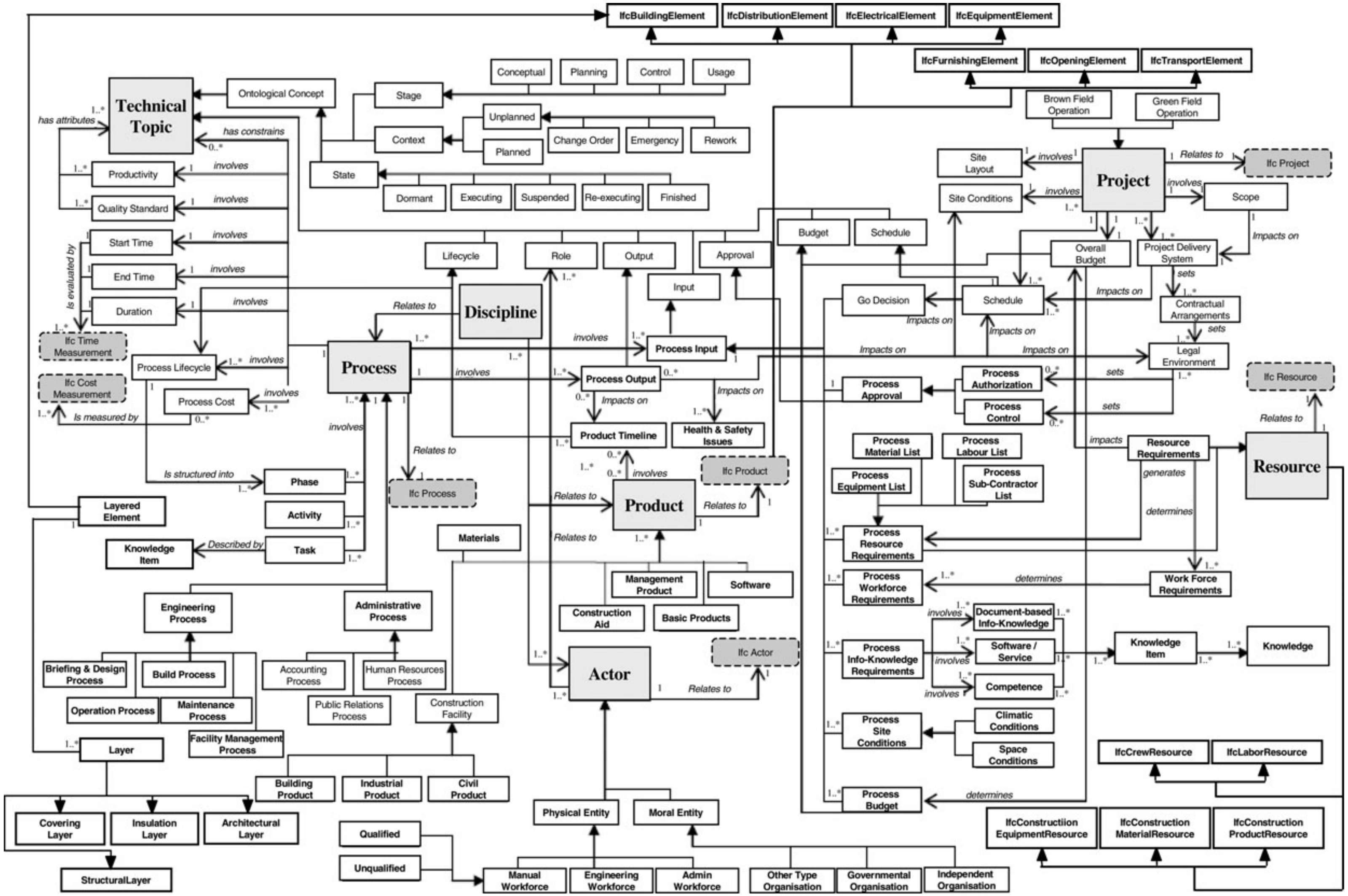

Figure 6 Illustration of the eCognos Core Ontology

7 Ontology maintenance and validation

7.1 Testing and validating the ontology

The research involved four organizations that each selected a knowledge-intensive business unit in order to test and validate the ontology developed to date. The ontology was used through a set of knowledge management dedicated services delivered as part of the eCognos platform (Wetherill *et al*., 2002). The testing of the ontology aimed at:

(a) establishing the completeness of the ontology in terms of conceptualizing the targeted disciplines;
(b) assessing the relevance of the concepts and relationships; and
(c) evaluating the usefulness and advantages afforded by the use of an ontology for knowledge management purposes, as opposed to traditional approaches based on key words.

The main services of the eCognos platform that were used to validate the ontology include: Knowledge Extraction, Knowledge Indexing, and Knowledge Searching services. The eCognos platform generates a knowledge representation (KR) for every single knowledge item (KI) managed by the system. The latter includes any text-based electronic object that holds construction-related semantics. When adding a new KI in the knowledge base, the knowledge extractor service is responsible for extracting the best relevant key words from the KI and calculating their statistical weights. This set of key words/weights is referred to as a semantic vector, and is submitted to the ontology service in order to get the best ontological representation of the KI. An example of a semantic vector is illustrated hereafter:

{(airlock,0), (balcony,0), (beam,0.05), (covering,0.1), (duct,0.5), (hindge,0.02), (lift well, 0.2), (shaft,0.3), (tendons,0), ...}.

The dimension of the vector is determined by the number of concepts in the ontology.

A KR is then built for the KI and stored in the eCognos repository. It is worth highlighting that this process is independent from the one illustrated in Figure 1. While the latter is used to develop the ontology, the former illustrates examples of the use of ontology to generate ontological knowledge representations of KIs used on a daily basis by construction practitioners. Also, when searching for knowledge, the system submits the user's query in order to get the best ontological representation of the resulting semantic vector. The search service was used in particular to enable users to perform searches across knowledge items, including documents, and to allow them to retrieve those knowledge items and their associated representations based on the concepts used in the ontology.

Two main metrics were used to evaluate the eCognos ontology:

(a) performance evaluation in terms of the response time delivered by the use of the ontology; and
(b) the retrieval performance in terms of relevance of the retrieved document set through the use of ontology.

Four sample document sets (representing the four selected case studies) comprising about 12 000 documents provided by the project construction end-users were used. The exact number per case study is given in Table 4. These were gathered from several recently completed projects. Fifteen queries per case study have been formulated by the discipline experts involved in the research. These were based on their own information needs experiences on projects, traditionally performed manually or using ad hoc search facilities. Table 3 illustrates the queries used in the 'architecture' case study.

The experts were asked to identify manually the document set matching each formulated query (within the context of their discipline). While the identification process was achieved in a matter of days, this did reflect (timescale wise) the current information and document search practices in the construction industry. These identified relevant document sets have then been used as a basis

Table 3 Queries used in the 'architecture' case study

Find documents that address sustainability issues in design
Locate all documents that reference health and safety regulations
Find all detailed specification documents that relate to thermal insulation
Find all documents providing detailed specifications of lifts
Locate documentation related to earthwork
Locate quality control documents
Find all documents that refer to quantity surveying work
Find all documents that refer to the client's brief
Locate all documents that prescribe doors' fire resistance
Find documents that describe space furnishings
Find all documents that refer to air conditioning
Locate all documents that refer to pre-cast concrete
Find documents that prescribe the use of asphalt in buildings
Locate all detailed specifications that refer to pipelaying work

Table 4 Performance evaluation of the ontology

Evaluation of the ontology			Evaluation without ontology by simple use of index terms		Evaluation with ontology	
Case study	Number of reference documents	Number of formulated queries	Average recall	Average precision	Average recall	Average precision
Case study 1 'health and safety'	11 789	15	0.30	0.06	0.55	0.45
Case study 2 'electrical domain'	11 913	15	0.25	0.10	0.60	0.50
Case study 3 'structure domain'	11 073	15	0.15	0.05	0.50	0.40
Case study 4 'architecture'	11 834	15	0.45	0.10	0.70	0.55

to compute the average Recall (the fraction of the retrieved relevant documents) and Precision measures (the fraction of the retrieved documents, which is relevant). These have been quantified in two scenarios:

(a) through the simple use of index terms; and
(b) through the use of the developed ontology (as illustrated in Table 4).

In the first scenario, a full-text summarization of the documents and queries was preformed, while in the second scenario the summarization relied exclusively on the concepts of the ontology. The discipline experts that generated the queries in the first place and identified the relevant document sets have been involved in the relevance assessment work.

In terms of response time, the difference between the two approaches (with and without the use of ontology) was comparable, and hence is not reported in the paper. The difference was in the retrieval performance provided by the ontology as illustrated in Table 4. The description of the field trials and testing scenarios are beyond the scope of this paper [related information can be found in (eCognos Consortium, 2003)]. The main results of the evaluation are summarized below:

- The issue of the comprehensiveness of the ontology was raised, as a number of concepts were reported missing in some of the discipline ontologies. While acknowledging the potential limitations of automated IR techniques, this suggests either that the selected term-weighting threshold was too high or that the scope and coverage of the document corpus used to build the ontology supplied by the participating companies were not comprehensive enough. This can be addressed by lowering the term-weighting threshold in consultation with domain experts while enlarging the documentary corpus with further documents drawn from projects that exhibit distinct characteristics in terms of project type, construction system, choice of materials and geographical location.

- The assumption adopted earlier, whereby most construction semantics is expressed through elementary or non-elementary concepts composed exclusively of nouns, was challenged through the implementation process as we did encounter concepts that referred to verbs such as 'Reinforced Concrete. While Stage 1 failed to detect such concepts, domain specialists, when asked to qualify the relationship between related elementary concepts, have ultimately identified these. In this particular example, 'Reinforced Concrete' is created by objectifying the relationship between 'Concrete' and 'Steel Bar'. Based on the four targeted domain ontologies illustrated in Table 2, the inclusion of verbs when combining words (nouns and verbs) co-occurring with a null syntactic distance into a single indexing component triples the number of retained concepts (from an initial number of 846–2342). Based on the experience of this first iteration, only a small number of concepts involving verbs as elementary or non-elementary concepts (exactly 9) were reported missing. The inclusion of verbs would have put an additional burden on knowledge experts, as these would have to analyze the 1496 additional concepts to retain only a few. It is therefore argued that while this can cause potential sources of concept loss, knowledge experts' time saving gained by restricting index terms to elementary or composite nouns justifies the choice of this option. It is also believed that the latter stages of the approach, in particular Stage 4, can allow these to be identified and created.
- Despite reporting missing concepts, the search functionality with ontology outperformed the traditional full-text search approach as more relevant document subsets have been retrieved. The knowledge representation technique, based on ontology, provides a more accurate user and machine interpretable summarization of documents, as illustrated by the field trial results with significant improvements in the precision and recall factors.
- The use of integrated services articulated around a common in-house ontology promotes the wide adoption of common standards and the sharing of a common understanding of terms and concepts. However, subtle variations of the semantics of certain terms may exist across companies. This suggests that discipline ontologies in the construction sector may need further refinement to be adapted to the norms and values of an organization.
- Some business processes involve the use of concepts from more than one discipline ontology. While this is supported by the eCognos methodology through the generic core ontology, it triggers another issue in relation to supporting dynamic views or perspectives that involve concepts drawn from more than one discipline ontology.

The validation of the methodology stresses the importance of the incremental and iterative approach to building the ontology. The following section describes the environment used to maintain the eCognos ontology.

7.2 *Managing the ontology*

Use of the eCognos ontology involves two types of user, namely the ontology manager and the ontology users. The ontology manager is responsible for the management of the ontology. As such, he/she has the authority to update the structure of the ontology by adding/removing ontological concepts, adding/removing relationships between concepts and adding/removing concept properties with the assistance of and validation from relevant knowledge experts. The ontology users are not allowed to directly change the ontology structure, but their expertise will be used to 'refine and enrich' the ontology. They can suggest the inclusion of new entities, which will be evaluated by the manager through consultation with relevant knowledge experts.

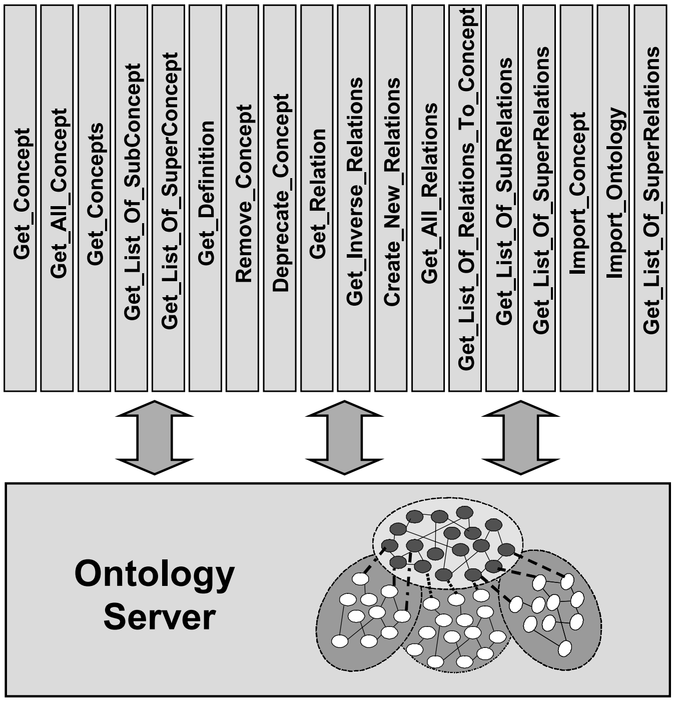

Figure 7 Ontology server

The approach adopted in order to handle the changes suggested by the ontology users involves storing the suggested changes as ‘quarantined items’ that must be evaluated and approved (or rejected) by the ontology manager. This avoids undesirable changes in the ontology structure, while providing a pragmatic mechanism to handle the changes requested by the end-users. Additionally, the entities/properties removed are, in fact, marked as ‘deprecated’. Such a strategy allows compatibility with future search operations that still use the deprecated entities.

The ontology is being used in two different ways. On one hand, to support an advanced search process (the user’s query is not based on key words; rather it is based on ontological instances fully based on the eCognos ontology) and, on the other hand, to support the KM-related services of the eCognos platform.

This required the development of a full service aimed at managing access to, and the retrieval of, concepts from the ontology. This is referred to as the ontology server, and is dedicated to servicing requests to the ontology (Figure 7). The ontology server is accessible through a dedicated API supporting eCognos, as well as any other eCognos-compatible KM system. A dedicated interface was developed and used by the ontology manager to maintain the ontology (Rezgui, 2006).

8 Conclusion

The paper presented an ontology development methodology to support the knowledge management needs of practitioners in the construction sector. A layered and modular approach was adopted to structure and develop the ontology. This was justified by the fragmented nature of the sector, organized into a variety of disciplines. The uniqueness of the methodology is illustrated by the combination of the following distinctive features:

- The modular structure of the ontology: the ontology takes into account the fragmented nature of the construction sector and its organization into established disciplines. It, therefore, mirrors the discipline-oriented nature of the industry.

- The support for the multiple interpretations of concepts across disciplines: the construction disciplines have their own norms and values, reflected in the development of dedicated semantic resources. The proposed ontology architecture is modular, with sub-ontologies dedicated to each established discipline, federated by a core ontology defined at a higher level comprising generic concepts applicable across sectors and enabling reconciliation of different terminologies used across disciplines.
- The collaborative nature of the ontology development process: It supports parallel development of the various ontology modules while promoting their integration through the core ontology.
- The iterative nature of the ontology development process: given the large-scale dimension of the construction industry and its project-oriented nature, an iterative process that proceeds by refining and extending the ontology over time helps converge towards a complete and true conceptualization of the domain.
- The ontology development approach: this is semi-automated and relies on discipline-oriented documentary corpuses to identify concepts and relationships using tf-idf and metric clusters techniques, which are then validated by human experts. It proceeds by re-using and building on existing semantic resources, improved and enriched with additional concepts drawn from discipline-oriented documentary corpuses.

The ontology developed to date is far from being complete, and will probably never be, as an ontology should be viewed as a living system. Similar efforts, such as the ISO standard for the exchange of product model data (STEP, 1994) project and its application to various industry sectors (including manufacturing), have taken almost a decade or longer to come to fruition. The eCognos ontology is by no means different. The issue of the existence of a unique ontology for an entire sector remains open. This suggests that while the eCognos Core Ontology forms a robust basis for interoperability across the discipline-oriented ontologies, the latter will need adaptation and refining when deployed into an organization and used on projects. Another issue that was raised is that related to the adoption of user-specific views or perspectives on the global ontology. In fact, in many instances, some actors might be required as part of their job to deal with more than one discipline ontology to conduct a task. This necessitates some flexible mechanisms that can enable the rapid combination of two or more discipline ontologies into a single view/perspective. This constitutes one of the author's future research objectives.

Acknowledgments

The author would like to thank the anonymous reviewers for their helpful comments, suggestions and recommendations that have substantially improved the paper. The author would also like to acknowledge the financial support of the European Commission under the eContent (FUNSIEC) and IST (eCognos: IST-2000-28671) Programmes.

References

Aussenac-Gilles, N., Biébow, B. and Szulman, S. 2001 Revisiting ontology design: a methodology based on corpus analysis. In *12th International Conference in Knowledge Engineering and Knowledge Management, (EKAW'00), Lecture Notes in Artificial Intelligence*, vol. 1937. Berlin: Springer, pp. 396–408.

Baeza-Yates, R. and Ribeiro-Neto, B. 1999 *Modern Information Retrieval*. New York, NY, USA: Addison Wesley.

Bechhofer, S., Horrocks, I., Goble, C. and Stevens, R. 2001 OilEd: a reason-able ontology editor for the Semantic Web. In *Joint German/Austrian conference on Artificial Intelligence (KI01), Lecture Notes in Artificial Intelligence*, vol. 2174. Berlin: Springer, pp. 396–408.

Bernaras, A., Laresgoiti, I. and Corera, J. 1996 Building and reusing ontologies for electrical network applications. In *Proceedings of the European Conference on Artificial Intelligence (ECAI 96)*, Budapest, Hungary: von Neumann Society of Computer Science, pp. 298–302.

Blazquez, M., Fernandez, M., Garcia-Pinar, J. M. and Gomez-Perez, A. 1998 Building ontologies at the knowledge level using the ontology design environment. In Gaines, B. R. and Musen, M. A. (eds.), *Proceedings of the Knowledge Acquisition Workshop (KAW98)*, Banff, Alberta, Canada, 18–23 April 1998. Available at http://ksi.cpsc.ucalgary.ca/KAW/KAW98/KAW98Proc.html; http://ksi.cpsc.ucalgary.ca/KAW/KAW98/blazquez/

Broglio, J., Callan, J. P., Croft, W. B. and Nachbar, D. W. 1995 Document retrieval and routing using the INQUERY system. In Harman, D. K. (ed.), *Overview of the Third Retrieval Conference (TREC-3)*. Gaithersburg, USA: NIST Special Publication 500–225, pp. 29–38.

BS6100 1992 Glossary of building and civil engineering terms, British Standards Institution. http://bsonline.techindex.co.uk

Cimiano, P. and Volker, J. 2005 Text2Onto: a framework for ontology learning and data-driven change discovery. In *Proceedings of the 10th International Conference on applications of natural language to data bases, (NLDB'05)*, Alicante (Spain), *LNCS* 3513, Springer, pp. 227–238.

CoMMA Consortium 2002 Corporate memory management through agents. In Cunningham, P. and Cunningham, M. (eds.) *Proceedings of E-Work & E-Business 2002*, Madrid: IOS Press, pp. 383–406.

Corcho, O., Fernando-Lopez, M. and Gomez-Perez, A. 2003 Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data and Knowledge Engineering*, **46**, 41–64.

Cristani, M. and Cuel, R. 2005 A survey on ontology creation methodologies. *International Journal on Semantic Web and Information Systems* **1**(2), 49–69.

eCognos Consortium 2002 *eCognos Ontology*, Project Deliverable D3.6, France: CSTB.

eCognos Consortium 2003 *eCognos Field Trials*, Project Deliverable D4.2, France: Derbi.

eConstruct 2001 Final edition of the bcXML Specifications. Available at http://www.econstruct.org/6-public/bcxml_cd/publicdeliverables/d103_v2.pdf

Euzenat, J. 1996 Corporative memory through cooperative creation of knowledge bases and hyper-documents. In *Proceedings of the 10th Knowledge Acquisition for Knowledge-Based Systems Workshop (KAW96)*, pp. 1–18.

Euzenat, J. 2004 An API for ontology alignment. In Hiroshima, J. P., van Harmelen, F., McIlraith, S. and Plexousakis, D. (eds.), *Proceedings of the 3rd conference on international semantic web conference (ISWC)*, 2004, vol. 3298, Heidelberg (DE): Springer Verlag, pp. 698–712.

Farquhar, A., Fikes, R. and Rice, J. 1996 The ontolingua server: a tool for collaborative ontology construction. In *Proceedings of the 10th Knowledge Acquisition for Knowledge-Based Systems Workshop (KAW96) Ban*, pp. 44.1–44.19. Avaliable at http://ksi.cpsc.ucalgary.ca/KAW/KAW96/KAW96Proc.html; http://ksi.cpsc.ucalgary.ca/KAW/KAW96/farquhar/farquhar.html

Fernandez-Lopez, M., Gomez-Perez, A., Pazos-Sierra, A. and Pazos-Sierra, J. 1999 Building a chemical ontology using METHONTOLOGY and the ontology design environment. *IEEE Intelligent Systems & their applications* **4**(1), 37–46.

Ganter, B. and Stumme, G. 2003 Creation and merging of ontology top-levels. In *Proceedings of the 11th International Conference on Conceptual Structures (ICCS'03), LNCS* 2746. Dresden, Germany, July 21–25, 2003. Springer, pp. 131–145.

Gomez-Perez, A. 2001 Evaluation of ontologies. *International Journal of Intelligent Systems* **16**(3), 10:1–10:36.

Gruber, T. 1994 Towards principles for the design of ontologies used for knowledge sharing. *International Journal of Human Computer Studies* **43**(5/6), 907–928.

Grüninger, M. and Fox, M. S. 1995 Methodology for the design and evaluation of ontologies. In *Workshop on Basic Ontological Issues in Knowledge Sharing,* August 19–20 Montreal: AAAI Press.

Guarino, N., Masolo, N. and Vetere, G. 1999 OntoSeek: contentbased access to the web. *IEEE Intelligent Systems* **14**(3), 70–80.

Guarino, N. and Welty, C. 2000 Supporting ontological analysis of taxonomic relationships. In *Proceedings of the 19th International Conference on Conceptual Modeling (ER'00), Lecture Notes in Computer Science*, vol. 1920. Berlin: Springer, pp. 210–224.

Haava, H. M. 2004 Semi-automatic method to ontology design by using FCA. In *Proceedings of the CLA 2004 International Workshop on Concept Lattices and their Applications*, Ostrava, Czech Republic, September 23–24 (CEUR Workshop Proceedings), Vol. 110. Available at http://www.CEUR-WS.org

Hahn, U. and Schulz, S. 2003 Towards a broad-coverage biomedical ontology based on description logics. *Pacific Symposium on Biocomputing* **8**, 577–588.

Holsapple, C. W. and Joshi, K. D. 2002 A collaborative approach to ontology design *Communications of the ACM* **45**(2), 42–47.

Horrocks, I. and van Harmelen, F. 2001 Reference description of the DAML-OIL (March 2001) ontology markup language, Technical Report. Available from http://www.daml.org/2001/03/reference.html

IAI 2006 International alliance for interoperability. IAI Web Site. Available at http://www.iai-international.org, accessed July 15, 2006.

ISO 12006-2 2001 Building construction: organization of information about construction works (Part 2: framework for classification of information). International Standards Organisation.

Kalfoglou, Y. and Robertson, D. 1999 Use of formal ontologies to support error checking in specifications. In Fensel, D. and Studer, R. (eds.), *Proceedings of the 11th European Workshop on Knowledge Acquisition, Modeling and Management (EKAW'99), Lecture Notes in Artificial Intelligence*, vol. 1621. Berlin: Springer, pp. 207–224.

Kalyanpur, A., Parsia, B., Sirin, E. and Hendler, J. 2005 Debugging unsatisfiable classes in OWL ontologies. *Journal of Web Semantics* **3**(4), 268–293.

Kayed, A. and Colomb, R. M. 2002 Extracting ontological concepts for tendering conceptual structures. *Data and Knowledge Engineering* **40**, 71–89.

Kietz, J. U., Maedche, A. and Volz, R. 2000 A method for semi-automatic ontology acquisition from a corporate intranet. In *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management (EKAW'00), Workshop on Ontologies and Texts*, CEUR Proceedings, Juan-Les-Pins, **51**, 4.14.14.

Klein, M. 2001 Combining and relating ontologies: an analysis of problems and solutions. In Gomez-Perez, A., Gruninger, M., Stuckenschmidt, H. and Uschold, M. (eds.), *Workshop on Ontologies and Information Sharing, International Joint Conferences on Artificial Intelligence (IJCAI'01)*, Seattle, USA, 4–5 August.

Klein, M., Fensel, D., Kiryakov, A. and Ognyanov, D. 2002 Ontology versioning and change detection on the Web. In Gomez-Perez, A. and Benjamins, V. R. (eds.), *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02), Lecture Notes in Artificial Intelligence*, vol. 2473. Berlin: Springer, pp. 247–259.

Lenat, D. B. and Guha, R. V. 1990 *Building large knowledge-based systems: representation and inference in the Cyc project*. Boston: Addison-Wesley.

Lopez, M. F. 1999 Overview of methodologies for building ontologies. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI-99), workshop on ontologies and problem-solving methods*, Stockholm, August 2, 1999.

Lowe, E. J. 1995 Ontology. In Honderich, T. (ed.), *The Oxford Companion to Philosophy*. Available at http://www.xrefer.com/entry.jsp?xrefid=553052

McGuiness, D. L., Fikes, R., Rice, J. and Wilder, S. 2000 An environment for merging and testing large ontologies. In *Seventh International Conference on Principles of Knowledge Representation and Reasoning 5KR 2000*, Breckenridge, Colorado, USA, San Francisco, California: Morgan Kaufmann Publishers, pp. 12–15.

Neches, R., Fikes, R. E., Finin, T., Gruber, T. R., Senator, T. and Swartout, W. R. 1991 Enabling Technology for Knowledge Sharing. *AI Magazine* **12**(3),36–56.

Noy, N. F. and Klein, M. 2002 Ontology evolution: not the same as schema evolution, Technical Report SMI-2002-0926, Stanford.

Noy, N. F. and Musen, M. A. 2000 PROMPT: Algorithm and tool for automated ontology merging and alignment. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI'00)*, Austin: AAAI Press.

Noy, N. F., Sintek, M., Decker, S., Crubézy, M., Fergerson, R. W. and Musen, M. A. 2001 Creating semantic web contents with Protégé-2000. *IEEE Intelligent Systems* **16**(2), 60–71.

Omelayenko, B. 2001 Syntactic-level ontology integration rules for e-commerce. In *Proceedings of the 14th FLAIRS Conference (FLAIRS-2001)*, Key West, Florida, May 21–23, 2001. Menlo Park, California: AAAI press.

Partridge, C. 2002 The role of ontology in integrating semantically heterogeneous databases, Technical Report 05/02, LADSEB-CNR, Padova, Italy, June 2002. http://www.loa-cnr.it/Papers/ladseb_tr05-02.pdf

Pinto, H. S. and Martins, J. P. 2000 Reusing ontologies. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI 2000) Spring Symposium Series, Workshop on Bringing Knowledge to Business Processes, SS-00-03*. Menlo Park, California: AAAI Press, pp. 77–84.

Rezgui, Y. 2001 Review of information and knowledge management practices state of the art in the construction industry. *Knowledge Engineering Review* **16**(3), 241–254.

Rezgui, Y. 2006 Ontology-centered knowledge management using information retrieval techniques. *Journal of Computing in Civil Engineering* **20**(4), 261–270.

Rezgui, Y. and Zarli, A. 2006 paving the way to the vision of digital construction: a strategic roadmap. *Journal of Construction Engineering and Management* **132**(7), 767–776.

Rezgui, Y. 2007a Exploring virtual team-working effectiveness in the construction sector. *Interacting with Computers* **19**(1), 96–112.

Rezgui, Y. 2007b Knowledge systems and value creation: an action research investigation. *Industrial Management and Data Systems (Emerald)* **107**(2), 166–182.

Salton, G. and Buckley, C. 1988 Term weighting approaches in automatic retrieval. *Information Processing and Management* **24**(5), 513–523.

Staab, S., Schnurr, H. P., Studer, R. and Sure, Y. 2001. Knowledge processes and ontologies. *IEEE Intelligent Systems* **16**(1), 26–34.

STEP 1994. Overview: STEP (ISO 10303) Product Data Representation and Exchange initiative, TC 184/SC4, http://www.tc184-sc4.org/SC4_Open/SC4_Work_Products_Documents/STEP_(10303)/, last accessed 05/04/2005.

Stojanovic, L. and Motik, B. 2002 Ontology evolution with ontology. In *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management (EKAW02) Workshop on Evaluation of Ontology based Tools (EON2002)*, CEUR Workshop Proceedings, Siguenza, vol. 62, pp. 53–62.

Studer, R., Benjamins, V. and Fensel, D. 1998 Knowledge engineering: principles and methods. *IEEE Transactions on Data and Knowledge Engineering* **25**, 161–197.

Stumme, G. and Maedche, A. 2001 Ontology merging for federated ontologies on the semantic web. In: Gomez-Perez, A., Gruninger, M., Stuckenschmidt, H. and Uschold, M. (eds.), *Workshop on Ontologies and Information Sharing, International Joint Conferences on Artificial Intelligence (IJCAI'01)*, Seattle, USA, 4–5 August 2001.

Sugumaran, V. and Storey, V. C. 2002 Ontologies for conceptual modeling: their creation, use and management. *Data and Knowledge Engineering* **42**, 251–271.

Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R. and Wenke, D. 2002 OntoEdit: collaborative ontology engineering for the semantic web. In *Proceedings of the First International Semantic Web Conference (ISWC'02), Lecture Notes in Computer Science*, vol. 2342. Berlin, Springer, pp. 221–235.

Swartout, A. T. 1999. Ontologies. *IEEE Intelligent Systems and Their Applications* **14**(1), 18–19.

Swartout, B., Ramesh, P., Knight, K. and Russ, T. 1997 Toward distributed use of large-scale ontologies, aaai symposium on ontological engineering, Stanford, 1997.

Tomovic, A., Janicic, P. and Keselj, V. 2006 n-Gram-based classification and unsupervised hierarchical clustering of genome sequences. *Computer Methods And Programs In Biomedicine* **81**, 137–153.

Uschold, M. and King, M. 1995 Towards a Methodology for Building Ontologies. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI95), Workshop on Basic Ontological Issues in Knowledge Sharing*, Montreal, 1995.

Welty, C. and Guarino, N. 2001 Supporting ontological analysis of taxonomic relationships. *Data and Knowledge Engineering* **39**(1), 51–74.

Wetherill, M., Rezgui, Y., Lima, C. and Zarli, A. 2002 Knowledge management for the construction industry: the e-cognos project. *Journal of Information Technology in Construction* **7**, 183–196, http://www.itcon.org/2002/12