A HISTORICAL PERSPECTIVE ON THE EVOLUTION OF CONTROLLED VOCABULARIES IN EUROPE

Celson Lima[1], Alain Zarli[1], Graham Storer[2], Jaime Acevedo-Alvarez [3]

ABSTRACT

The development of Controlled Vocabularies (CV's) - dictionaries, classifications, taxonomies and ontologies - has been the focus of many research initiatives around the world targeting the Construction sector. Bringing experience of several pan-European initiatives over the past decade and more, the authors highlight milestones on the path of evolution of CVs to the present day in terms of development and adoption of results, outline the main obstacles encountered and speculate on the future of CVs in the European Construction sector.

KEY WORDS

Controlled Vocabularies, Ontology, Taxonomy Thesaurus, Dictionary, Classification Systems.

INTRODUCTION

For 40 years the development of Controlled Vocabularies (CV's) such as dictionaries, classifications, taxonomies and the now 'appealing' ontologies has been the focus of research projects and initiatives in Europe and elsewhere. Some of the better known efforts are: ISO12006 (parts 2 & 3), BS6100 and UNICLASS (British Standards), LexiCon (The Netherlands), Barbi (Norway), bcBuildingDefinitions taxonomy (from e-Construct project), ICONDA® terminology (Fraunhofer IRB), e-COGNOS ontology (from e-COGNOS project), Standard Dictionary for Construction (SDC, France). International efforts include SI/SfB, Masterformat, Omniclass and the Canadian Thesaurus.

Even a brief review of the above initiatives and projects reveals how much effort and dedication has been invested in the work, with the aim of preparing the Construction sector for engaging in leading edge advances of semantic-related ICT resources. Preliminary ideas related to developing e-Commerce/e-Business related tools and resources useful to the construction supply chain for publishing catalogues in their own technical and natural languages and becoming actors in the wider European and global arena.

It is perhaps timely to ask some probing questions: After some decades invested in this topic, what is the reality in Europe? Are Controlled Vocabularies in daily use by Construction actors or are they just inspirational works of 'fine' art on the 'research walls'? If Controlled Vocabularies are fully adopted and used on a daily basis, what might be done with them in the future and what trends are now observed in the research area? What are the domains of work where Controlled Vocabularies will likely play an important role? What is the future of e-business and e-commerce related activities in Europe, a very fragmented market where the national (sometimes regional) norms and regulations impose a strict control on products and services construction-oriented? Finally, a challenging question: how can Construction be e-ready for business exploitation of the Semantic Web?

[1] CSTB, Centre Scientifique et Technique du Bâtiment, Route des Lucioles, BP 209, Sophia Antipolis CEDEX, France, Tel: +33 4 93956722, Fax: +33 4 93956733, Email: {celson.lima;alain.zarli}@cstb.fr.

2 GSC, 18 Amersham Hill Drive, High Wycombe, HP13 6QY, United Kingdom,Tel: +44 1494 446300, graham@storer-consulting.fsnet.co.uk

[3] Dipl.-Ing. Architecture, International Cooperation, Fraunhofer-Informationszentrum Raum und Bau IRB, Nobelstrasse 12, 70569 Stuttgart, Germany, Phone +49 711 970 2976, FAX +49 711 970 2599, jaime.acevedo-alvarez@irb.fraunhofer.de

This paper discusses the questions raised above, based on the experience gathered by the authors through their involvement in several European initiatives related to the subject. Section 2 presents the main reasons for development of CVs in Construction (the why). Section 3 presents very briefly a selection of European/International initiatives on this area. Then Section 4 paints a picture of where we are now that informs a speculative and provocative discussion in Section 5, on where we are going *versus* where we could/should be going. Finally, some conclusions close the paper.

REASONS FOR DEVELOPING CONTROLLED VOCABULARIES IN CONSTRUCTION

Why develop controlled vocabularies in fields of activity? What is the need and purpose? The following points help to answer these questions.

- **Vocabularies give names to things that have meaning at a certain level of detail.** Vocabularies provide convenient shorthand for exchanging information. For example, the word "dog" conveys "a domestic carnivorous animal with four legs that typically has a long muzzle, pointed ears, a fur coat, a long fur-covered tail, and whose characteristic call is a bark". It is certainly different to "elephant" or "bicycle". So, if somebody says, "Where is my dog?" we know the kind of thing to look for. But there are many dogs. We could add adjectives like "small", "long", "short-legged", "drooping eared", "German" (which adjectives must have agreed meaning in the dog context) or we could simply use another name "dachshund" or "sausage dog". These need to have agreed meaning, not least because to the English or French the word dachshund is foreign and the other is a descriptive nick-name. The deeper we need to go with meaning to add detail or to differentiate, the more control there needs to be of the language. Between specialists in one discipline there should be precise understanding of words (in this case, zoologists who may even use Latin names) but between non-experts or between different kinds of expert there can be misunderstanding. To illustrate the importance of meaning in a construction related example, what is the difference between a "brick pillar" and a short length of thick wall made from brick? A bricklayer and a cost estimator might use different terms. The answer (in UK at least) is that the difference for estimators is defined by rules related to the objects dimensions.
- **Vocabularies are important to conveying human thought in a concise way, and with precision in a given working context.** Vocabularies must be controlled to avoid the Humpty Dumpty situation depicted in Figure 1. There must be as much precision as possible, although in human exchanges we often say that something is like something else e.g. the dog is like a dachshund but with longer legs. We can then ask questions to refine meaning and (perhaps finally) identify the breed of dog.
- **Controlled vocabularies are even more important to electronic information exchange in any form.** Whilst humans can ask clarifying questions based on their experience and knowledge, computers do not have that as a general capability (though in limited contexts artificial intelligence may enable it). So there needs to be precision built into the language of computer communication used. There is less room for confusion if an item is referred to by an explicit catalogue/part number that *defines* rather than *describes* it to the supplier. But not everything can be conveyed that simply; by a code. Architectural details, a building frame, a plumbing system etc. are usually designed to result in requirements that facilitate choice of components to satisfy the need. So generic types like wall, pump and foundation are used that are then specialised according to properties (such as dimensions, material, colour, strength etc) which themselves must have precise (i.e. agreed) meaning. Although codes can be used to identify components and systems, it

is more convenient that they take the form of the names we humans use. “Pump” not A25GH7, unless we are buying from a catalogue!

Figure 1: Average ± Standard Deviation of Buffer Size Relative to Number of Rolls after 1,000-Iteration

- **Who should control vocabulary for the Construction Industry?** The answer is the practitioners and ICT providers through standardisation processes of one kind or another. The struggle of the past decade or more has been to put in place the national, European, international and cross discipline organisation for that to happen. The Construction Industry is hugely fragmented in terms of disciplines involved, their locations and skills and often undertakes major projects as international consortia. Such consortia are temporary alliances for the duration of a project, perhaps no more than a year or so in a design phase. Project lead times can be exceedingly short with teams from different disciplines (and within disciplines) put together at very short notice. The need for standardised, controlled vocabularies in such a work environment is crucial.
- **Where has the resource come from to cover the cost of vocabulary development and control?** The industry is extremely large (~10% GDP) but it is not noted for its research investment. The small size of the majority (80% +) of companies, their vast number and their distribution inhibits the collaboration necessary. Progress has been mainly through efforts arising from collaborative research projects, initiatives and developments such as those below. Finance for effort has been provided by funding programmes (national & EU) and the participating organisations themselves.

STATE OF THE ART ON CVS FOR CONSTRUCTION IN EUROPE

As previously indicated, a large body of effort has been devoted to the creation and use of CVs around the world. This section briefly summarises a suite of relevant research projects, and pan-European and international initiatives (Figure 2). This panorama is not claimed to be exhaustive but simply aids discussion in the next section.

The story begins with the **CI/SfB** (Construction Index/*Samarbetskommitten for Byggnadsfragor*), a Scandinavian system of classification originally set up in 1959 and specially designed for the construction sector. It is claimed to be in worldwide use for technical and trade literature in the broad construction area. The CI/SfB was used in North America as the basis of the MasterFormat™, which is the standard for specification-writing used in most commercial building design and construction projects.

MasterFormat is a master list of numbers and titles intended for use in the organizing of specifications, and contracting and procurement requirements initially started with 16 *divisions* coded with 5 digits. In order to cope with the changes required by the modern Construction industry, in 2004 MasterFormat was heavily updated; new sections were added

(the initial 16 were extended to 50) and the number codes are now composed of 8 digits instead of the initial 5. MasterFormat targets the standardised communication of projects for all actors involved.

MasterFormat works in harmony with **Uniformat**. Uniformat is an arrangement of construction information based on physical parts of a facility called systems and assemblies. It aims: (i) to achieve consistency in economic evaluation of projects; (ii) to enhance reporting of design program information; and (iii) to promote consistency in filing information for facility management, drawing details and construction market data. Masterformat says *what* the construction item is; Uniformat says *where* it belongs.

ICONDA®Bibliographic began life in the mid 1980's as the database of the International Council for Research and Innovation in Building and Construction (within CIB). Since then, ICONDA® has found a key role in various information products. Fraunhofer IRB, the Information Centre for Planning and Building of the Fraunhofer-Gesellschaft, presently coordinates maintenance of the database and its marketing. Today, ICONDA® is a supra-national organisation incorporating content provided by 1 supranational and 23 national organizations in 14 countries worldwide. Access to ICONDA® based products will be facilitated by a multilingual terminology of around 100 000 terms in English, German, French, and Spanish. Expansion to Italian, Lithuanian, Romanian and Slovenian languages is underway. Principal terminology sources are the INIST Vocabulary, the Canadian Thesauri and ICONDA® own vocabulary.

BS6100 (British Standard 6100), dating from the 1990's, is a glossary of terminology used in the UK Construction sector. Its role was to provide *a comprehensive list of terms that will promote better understanding between various sections of the construction industry, facilitate trade and provide better tools for improving handling of information*. BS6100 was merged with **UNICLASS** (the Unified Classification for the Construction Industry published in 1997 as a substitute for the widely accepted but increasingly out-dated CI/SfB). UNICLASS is a construction-specific information classification system that covers information generated from all phases of a construction project. It is structured with a faceted classification system rather than a hierarchical one.

ISO 12006[4] (parts 2 & 3) came from another direction of concern: the International Organisation for Standardisation. ISO was also targeting the development of standard CVs for the Construction sector in a world-wide scale. On one hand, ISO12006-2 targets the definition of a model for classification systems (it is not a classification system in itself); it sets out an approach whereby particular classification systems that meet *regional or national* requirements can be developed according to a *common international* approach. On the other hand, ISO 12006-3 defines a schema for a taxonomy model, providing the ability to define concepts by means of properties, to group concepts, and to define relationships between concepts. Objects, collections and relationships are the basic entities of the model.

The ISO foundation work was adopted and used by some institutions around the world. Among them, we can cite Stabu (Netherlands), Edibatec (France), and the Norwegian construction industry, which respectively started their own implementations of ISO-based tools, namely the **LexiCon, SDC, and BARBI**. In other words, they are separate implementations of dictionaries that comply with the specification given in ISO 12006-3.

Next we refer to the **IAI** (International Alliance for Interoperability) and its Industry Foundation Classes (IFC). The IFC model has been progressively developed by the IAI since

[4] Both standards were officially released in 2001, after the normal process through standardisation (PAS-Publicly Available Specification, and DIS-Draft International Standard), although they started to be be used before then.

1995 through several releases implemented in software for data exchange and sharing across applications. Since the IFC.2x release (October 2000), a 'core' of the model has been protected against change and formally accepted as ISO PAS 16739 in November 2002 under the external "harvesting" procedures of ISO TC184/SC4. IFC is the IAI's vehicle to implement the Building Information Model (BIM) concept, aiming to increase productivity of design, construction and maintenance operations in the building life cycle.

The IFC model is rooted in approaches originally developed within the work of **STEP** (ISO TC184/SC4); especially in the development of the ISO 10303 series of standards. In particular, IFC has adopted and/or adapted certain parts of the STEP standards including: formal specification language EXPRESS from ISO 10303 part 11; encoding of files for data exchange is undertaken using ISO 10303 part 21; and schema from ISO 10303resource standards such as parts 41, 42, 43 and 46. Despite the fact that from a "semantic perspective", the IFC model per se cannot be considered as ontology/taxonomy, part of it has been used to support reasoning and to exchange meaningful pieces of information among different software tools.

Collaborative EU Projects

There have also been many project-based European efforts including: eConstruct (Lima et al., 2003), e-COGNOS (El Diraby et al., 2005), CEN/ISSS eConstruction series of Workshops (Böhms et al., 2004), FUNSIEC (Lima et al., 2005), CONNIE (Cerovsek et al., 2006), and the on-going SEAMLESS (Lima et al., 2006) project.

eConstruct developed the *Building and Construction eXtensible mark-up Language* (*bcXML*), which supports the eBusiness communication process needed between clients, architects and engineers, suppliers and contractors for the (e)procurement of products, components, and services. The *bcBuildingDefinitions*, the taxonomy developed by eConstruct to show the power of bcXML, contains nearly 3 000 terms *specifically related to doors*, expressed in six languages[5]. Such taxonomy can be instantiated to create catalogue content or the actual requirements (queries) and solutions (answers) messages.

The **e-COGNOS** developed a KM-oriented software infrastructure enabled by a semantic engine: an ontology server (and its respective ontology). The ontology focuses on construction concepts related to e-COGNOS main objective: **consistent knowledge representation of construction knowledge items**. The e-COGNOS ontology is composed of two taxonomies (concepts and relations). The taxonomy of concepts is grounded in the IFC model which forms its higher levels, according to the following foundation statement: *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 Domains) and according to certain conditions (Technical Topics).*

The **CEN/ISSS eConstruction** series of workshops worked towards the formulation and embodiment of identified required semantic themes. This initiative recognised that it is not really possible to propose standardised *Semantic Resources* (SRs – i.e. ontologies, taxonomies, dictionaries, thesauri, and the related resources) for the construction sector but that it was possible to provide key principles that organisations could follow after deciding to use SRs to support their business activities. This initiative emphasised the need to take into account two key factors, namely *purpose* and *application* areas, when considering development and/or use of SRs.

FUNSIEC worked with the following questions: Is it possible to establish semantic links (mappings) between different SRs? The answer is yes, it is possible, which was demonstrated

---

[5] : English, French, Dutch, German, Norwegian, and Greeklish (Greek language written with Latin characters). Additional information about e-Construct can be found at (Lima 2003).

through the results of such projects: the OSIECS Kernel, and both OSIECS meta-model/model. The former is a software tool built to identify and propose semantic mappings between two SRs. OSIECS meta-model/model are the mapping tables produced by the OSIECS Kernel.

CONNIE tackled the exploitation of multi-lingual content in norms and regulations for the European Construction sector. It produced a software infrastructure to help organise, index, classify and use (in a pan-European way), the contents (regulation/norms) available within the CONNIE environment. The CONNIE infrastructure relies strongly on use of CVs to index and share multi-lingual content in an efficient way.

SEAMLESS targets the deployment of a seamless infrastructure to help SMEs to participate more easily in the e-business world (i.e. providing e-services to support business needs, such as procurement, production follow-up, etc.). The SEAMLESS infrastructure is sector-independent, but to demonstrate its potential two vertical sectors were used: Textile and Construction. The knowledge-related side of SEAMLESS is based on a hierarchy of ontologies covering three levels of representation, namely: the global level (the whole SEAMLESS environment), the mediator level (the intermediate level providing a mapping between the global level and the SMEs), and the local levels (the lowest level where the SMEs use their 'personalised' CVs). In order to support the operation of the SEAMLESS environment, a sector-specific hierarchy of ontologies is being developed.

Currently, North America is attempting to bring classification within a single, multi-facetted approach called **Omniclass**, which started under the name of Overall Construction Classification System (OCCS) in 2001 but renamed to Omniclass in 2002. It is based on ISO12006-2 as a framework and it uses MasterFormat for work results, UniFormat for elements, and Electronic Product Information Cooperation (EPIC) for structuring products. The first version of Omniclass 1.0 was officially released in March 2006.

Last but not least, the picture is completed by the Canadian thesaurus. This is a bi-lingual thesaurus specifically created to represent construction terms in English and French. The enrichment of this thesaurus has been re-launched and new developments and improvements are expected to be released in the near future.

WHERE WE ARE NOW AND WHERE WE COULD/SHOULD GO NEXT

Works and initiatives described in the previous section allow us to say that we are in a good position but we have not achieved the goal. Companies are not yet sufficiently capitalising on the results provided by the research world and standardisation still needs to find its place in this arena. However, we should not be pessimistic; very good work has been produced and solid results are now available. Education and awareness are the key words behind what needs to be done in order to push things forward.

The assessment of the results produced by FUNSIEC emphasised the importance of education (in the wide sense) of Construction practitioners regarding the use of semantic resources. CEN/ISSS eConstruction workshop suggested similar. Education here means providing *good practice examples to the end users* (companies and individuals) showing how they can benefit from the use of CVs in their daily business, how they can expand their capabilities and increase potential in terms of their markets, and what the tangible benefits/improvements of CVs can be to them. And we need to find an everyday vocabulary to use in talking about such matters with construction practitioners! Examples may be that language.

The authors continue to work in this field and the latest experience shows that although good and powerful contributions are already in place, every time new research starts, people want to develop afresh, 'to show innovation', rather than to extend and refine the valuable

legacy of past work. This is a very natural human behaviour and difficult to change, influenced by research programmes that call for "innovative, groundbreaking" work. For instance, in a new IST project[6], where CVs are required, the development team has found arguments to justify the development of 'yet another' ontology editor and a new tool to produce semantic mappings. Acknowledging the past and the extent to which we might reuse and build upon it seems to be difficult for us. Thus to those who came on board IAI in the late 1990's the IFC's were a newly discovered phenomenon that generated interest. However, those involved in the detailed development acknowledge the contribution that earlier work and thinking in STEP made.

Therefore, most researchers want to plant a flag in new continents and not be explorers, mappers and exploiters of already discovered ones. Yet exploitation is what will benefit the industry. That is where the prizes should be. A bigger and better mouse-trap will not be of benefit unless existing ones have actually been taken out of their boxes and tried in practice! We ignore the considered recommendations from standards-related initiatives, like the CEN/ISSS eConstruction workshops, that spoke of the need to 'analyse what is available', 'reuse current results', and generally learn from the past – but these are considered to be "second class" activities. Emphasis is placed on being revolutionary rather than evolutionary; that is what attracts prestige in research.

However, the picture is not totally bleak with results lost in a research black hole. Business initiatives, even supported by less advanced solutions, are pushing things forward. IFCs have catalysed the adoption of the BIM concept and, considering the fact that part of the IFC model can be used as an ontology, new experiments on the area have been launched and we can expect solid results very soon. Other more modest but very useful initiatives (e.g. ICONDA® family, CSTB products) are making money using dedicated CVs. This is more than enough for some but not for others. Hence ICONDA® is working to enhance their 'semantic side' to something more modern, supported by new technologies and CVs. For instance, the ICONDA® agency has made agreements with several new countries around the world in order to enrich its terminology; CSTB is underway with an internal project to extend the capabilities of dictionaries and taxonomies supporting the search processes of content-based products, such as their CD-REEF and I-REEF. Both examples are also following very closely standards-related initiatives aiming to capitalise on them.

CONCLUSIONS

Communication is about exchanging signals. Humans are able to use words, body gestures, images, etc.. Jargon is usually used inside any given community and those not belonging to that community *will* have difficulty understanding and communicating. If we are to be clear and unambiguous, therefore, we must 'control' the vocabulary we are using in communication. Only parties fully knowing the words and their meanings are equipped to engage in risk-free communication. When it comes to computer-based communication, clear meaning is even more crucial since computers cannot establish dialogues in order to elucidate 'what is meant by that'. The conceptual approach to handle this situation often relies on the adoption of formal CVs (as much as possible) which help define the universe of discourse (the working context) of those involved in the communication process.

Several examples can be found around the world, coming from very different initiatives ranging from industry collaborations to feasibility projects funded by research programmes. Results are emerging; education is gaining an improved status on the European scene for several reasons, including European policies, straight businesses efficiency benefit, and

[6] Authors are intentionally avoiding to identify the initiative, for obvious reasons.

general evolution of the area. LexiCon and BARBI (two implementations of ISO 12006) have joined forces; IFCs are becoming *the* standard supporting the inevitable BIM concept; International Framework Dictionary (IFD) is attracting broad attention, and governments have published policies that directly or indirectly enforce the adoption of shared CVs and semantic-related resources. This is the future path we are called to – with no acceptable excuses and no other choice in moving forward.

Recalling the words of McGuiness (with adaptation to our terminology in parentheses), *an ontology* (or CV) *is required when there is a need to communicate/exchange* (transfer and/or share) *various sorts of information where the meaning is fundamental. Ontology* (CV) *is also useful when the reuse of existing knowledge is required. From a non-exhaustive list of uses, an ontology* (CV) *can be used for simple kinds of consistency checking, interoperability support, validation and verification testing, configuration support, help to perform structured, comparative and customised search as well as to exploit generalisation/specialisation of information* (Mcguiness 2002).

This means whenever we need to communicate precisely, the vocabulary must be controlled, our jargon must be shared and meaningful, and our semantics must be refined for the sake of the communication process. This is the mission behind the development and use of CVs in the Construction sector. This is the justification for proposing, developing, and assessing CVs. This is the quest that keeps the authors of this paper involved in this field. Results are still in their infancy, but they are promising and exciting, and hold the key to the future exploitation of semantic vision in the industry.

REFERENCES

Lima C, Stephens J, Böhms M. The bcXML: Supporting eCommerce and Knowledge Management in the construction industry. Itcon Journal, v. 8, p. 293-308, 2003.

El-Diraby T, Lima C, Fiès B. Domain Taxonomy for Construction Concepts: Toward a Formal Ontology for Construction Knowledge, Journal of Computing in Civil Engineering, Vol. 19, No. 4, October 2005, pp. 394-406.

Böhms M, Lima C, Storer G, Wix J. Framework for Future Construction ICT, International Journal of Design, Sciences & Technology, Volume 11, Number 2, 2004, p. 153-162, editor: Dr. Reza Beheshti.

Lima C, Silva C, Sousa P; Pimentão JP, Duc CL. Interoperability among Semantic Resources in Construction: Is it Feasible? In proceedings of CIB / W78 22nd Conference on Information Technology in Construction, p. 285-292, ISBN 3-86005-478-3, CIB Publication No. 304, Dresden, Germany, July 2005.

Cerovsek T, Gudnason G, Lima C. CONNIE - Construction News and Information Electronically. Joint International Conference on Computing and Decision Making in Civil and Building Engineering, Montreal, Canada, 14 - 16 June 2006.,

Lima C, Bonfatti F, Sancho S, Yurchyshyna A. Towards an Ontology-enabled Approach Helping SMEs to Access the Single European Electronic Market, In proceedings of the 13th ISPE International Conference on Concurrent Engineering: Research and Applications, 18 - 22 September, 2006, Nice, France.

Mcguiness D. Ontologies Come of Age, In Dieter Fensel, J im Hendler, Henry Lieberman, and Wolfgang Wahlster, editors. Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. MIT Press, 2002.