Computing Ontology Creation

Krassen Stefanov, Kornelia Todorova

Abstract: *In this paper an approach for the development of an Ontology for the domain of Computing Education is presented. This approach was applied in the FP5 IST Project DIOGENE - A Training Web Broker for ICT Professionals. I am outlining our work on the Computing Ontology creation, and am giving some guidelines and hints for further usage of the Ontology.*

Key words: *Computer Systems and Technologies, Information and Communication Technologies, Computing Education, Ontology, Metadata*

INTRODUCTION

Diogene is an EC funded project under the 5th Framework Programme – Information Society Technologies (contract IST-2001-33358). The main project objective is to design, implement and evaluate with real users an innovative training Web brokering environment for ICT individual training (but based upon a domain-independent platform) able to support learners during the whole cycle of the training, from the definition of objectives to the assessment of results through the construction of custom self-adaptive courses.
Six high quality courses will be implemented by project partners: *Object oriented analysis and design based on UML*, *XML*, *Professional engineering process improvement training*, *Increasing Organisational Performance with the Balanced IT Scorecard*, *Digital images for multimedia*, and *Php language for dynamic web pages*.

The definition of a knowledge representation methodology able to describe Learning Objects in a machine-understandable way and a learner model able to formally profile learners in terms of actual acquired knowledge, learning preferences and styles, will be undertaken. On its base a set of software tools for Learning Object knowledge modelling to be used during the course arrangement phase will be implemented. Then a Main Ontology covering the Computing Education domain will be created, and a Metadata indexed set of high quality courses will be arranged.

In this paper we will describe the process of Computing education Ontology creation. First of all, we will start with the explanation of the term Ontology, and the chosen format for the knowledge representation.

Approaches for Ontology creation – DIOGENE case study

The name "ontology" comes from Greek philosophy and means "the study of the nature of being". The term is used in the domain of Knowledge Representation "to categorize the kinds of things existing". The aim is to fix a common vocabulary of terms able to describe as much knowledge about the world as possible from given domain, and to subdivide this knowledge in a coherent class hierarchy, so as to create a shared knowledge representation language. Usually an Ontology is composed of (some of) the following: classes of objects, a vocabulary of terms (instances), and various relations between terms and classes. Depending of the chosen formalism for expressing these things, we distinguish DAML, OIL, SHOE, and other types of Ontologies [5, 6]. For the purposes of the DIOGENE project the SHOE was chosen as a representation formalism.

The Computing Education Ontology created in DIOGENE consists of one class (Computing Education field), a lot of instances (terms from the Computing domain), and a pre-defined set of three relations:

- *HP* (Has Part): *HP (x, y_1, y_2, ..., y_n)* means that the concept *x* is composed of the concepts y_1, y_2, ..., y_n , that is to say: to learn *x* it is necessary to learn y_1 and y_2, and, ..., and y_n.
- *R* (Requires): *R* (*x, y*) means that to learn *x* it is necessary to have already learnt *y*. This relation poses a constraint on the Domain Concepts' order in a given Learning Path.
- *SO* (Suggested Order): *SO* (*x, y*) means that it is *preferable* to learn *x* and *y* in this order. Note that also this relation poses a constraint on the DCs' order but now it is not necessary to learn *y* if we are interested only in *x*.

Relations between the instances are declared using "slots".

Some important features of the chosen relations are:

- Each concept can be split only by one Has Part relation
- The hierarchy defined by these three relations should not contain loops
- The lowest-level concepts are intended to be realised by corresponding Learning Objects, providing the conceptual link between the Ontology and underlying Metadata level

Each concept definition includes:

- A name – a word or short phrase used by the ontology creators to refer to the concept.
- A description – a textual definition of each concept, in English. This is derived from a dictionary or thesaurus.

Both attributes can be used in documenting the ontology, for example, in clarifying the process of ontology creation, and in assisting in the manual attachment of Learning Objects to concepts. They are important elements in the automatic attachment of Learning Objects to the ontology, enabling information retrieval or natural language processing techniques to be used.

There are three strategies for the creation of an Ontology: top-down, bottom-up, and a combination of the two. There is no best approach - it is depending of the Ontology creator's preferences and abilities, as well as on the availability of Learning Objects to be used as a starting point.

Ontology creation is iterative process of modelling the given domain, by choosing the most important concepts and identifying the most relevant relations between them. The intended usage of the Ontology is the main guiding principle during the design and modelling process.

There are various tools that can assist in this process. The most used such tools (also referred to as Ontology editors) are: Protégé 2000, OntoEdit, OilEd, WebODE, Ontolingua, Ontosaurus [2, 4, 7, 8, 9, 10].

Analysis of tools for Ontology creation

We performed an in-depth analysis of all available Ontology editors. Some important issues which arose from this analysis are:

- There is no suitable ontology editor based on SHOE specification.
- Three "best candidates" were studied more carefully: Protege, OilEd and WebODE.
- Finally Protégé was chosen as the best fit for DIOGENE.

For the purposes of the DIOGENE Project we chose Protégé 2000, developed by the Knowledge Modelling Group at Stanford University. It provides an integrated knowledge-base editing environment and an extensible architecture for the creation of customised knowledge-based tools. It has been developed using a plug-in architecture, where new services can be added easily to the environment. It conforms to the Open Knowledge Base Connectivity (OKBC) protocol for accessing knowledge bases stored in knowledge representation systems.

The tool accesses classes, instances, slots and applications through a uniform GUI which enables convenient co-editing between these elements. It is designed to guide developers and domain experts through the process of system development, allowing the reuse of domain ontologies and problem-solving methods, thereby shortening the development time required. This is an iterative process, with cycles of revision to various components of the system.

In order to create the Computing Education Ontology, we first start with investigating all available sources, and the results are presented in the next chapter.

Analysis of existing sources

The field of Information and Communication Technologies is a broader term, which is also often referred to as Computing, Computer Science, Computer Systems and technologies, Informatics, etc. We studied a lot of sources - existing electronic and printed classifications, thesaurus, ontologies, etc. Below we shortly describe the sources of main interest.

ACM Computing Classification System (CCS)

The 1998 version of the ACM CCS scheme, valid also today, is the key resource in the ICT area. ACM's first classification system for the computing field was published in 1964. Then, in 1982, the ACM published an entirely new system. New versions based on the 1982 system followed, in 1983, 1987, 1991, and now 1998 [1].

The full classification scheme involves three concepts the four-level tree (containing three coded levels and a fourth uncoded level), General Terms, and implicit subject descriptors.

The Mathematics Subject Classification (MSC)

The Mathematics Subject Classification [11] is used to categorize items covered by the two reviewing databases, Mathematical Reviews and Zentralblatt MATH. The Mathematics Subject Classification is broken down into over 5 000 two-digit, three-digit, and five-digit classifications, each corresponding to a discipline of mathematics (e.g., 11 = Number theory; 11B = Sequences and sets; 11B05 = Density, gaps, topology).

The current classification system, 2000 Mathematics Subject Classification, is a revision of the 1991 Mathematics Subject Classification. It is the result of a collaborative effort by the editors of Mathematical Reviews and Zentralblatt MATH to update the classification.

UNESCO Thesaurus

The *UNESCO Thesaurus* [13] is a controlled vocabulary developed by the United Nations Educational, Scientific and Cultural Organisation which includes subject terms for the following areas of knowledge: education, science, culture, social and human sciences, information and communication, and politics, law and economics. The *UNESCO Thesaurus* allows subject terms to be expressed consistently, with increasing specificity, and in relation to other subjects. It can be used to facilitate subject indexing in libraries, archives and similar institutions.

As in other subject thesauri, the terms in the *UNESCO Thesaurus* are linked together by three types of relationships: hierarchical, associative, equivalence. It was first published in 1977. A second edition was issued in 1995. The *UNESCO Thesaurus* database can be browsed in two ways: alphabetically and hierarchically (by area of knowledge and by micro-thesaurus). The *UNESCO Thesaurus* is available in paper and digital formats.

ASIS Thesaurus of Information Science

They are the most powerful Internet thesaurus available. They include Thesaurus Navigator (TN), established in 1989 to provide advanced technologies for information storage and retrieval. TN displays terminology from thesauri built with the Thesaurus Construction System (TCS).

IEEE Web Thesaurus

They are used mainly for the IEEE Keyword search.

INSPEC Thesaurus

INSPEC is produced by the Institution of Electrical Engineers and is regarded as the premier database for access to the world's leading scientific and technical literature in physics, electrical engineering, electronics, communications, control engineering, computers and computing and information technology.

Thesaurus Computer Science

One of the best thesaurus in ICT field is in German - Thesaurus Computer Science (alphabetisch) [12].

Internet Portals and their Search Interfaces

AVEL Sustainability Knowledge Network [3] is a portal and brokerage service for engineers, other professionals and researchers concerned with sustainable systems. It is also a resource for students in senior secondary and tertiary education. It is using part of the ACM CCS scheme.

Other portals in computing or computer science mainly are offerings from commercial players (IBM, Microsoft etc) pointing to their own products or research. The more generic services such as Yahoo and Top20 (http://www.top20computerscience.com/) also have offerings but not really useful as a source for ontology creation.

The subject taxonomies/hierarchies now used in Yahoo, AltaVista, and the Open Directory Project indexes (etc.) appear to have been created ad hoc, and appear to change a lot

over time. Presumably a large staff is needed to evolve the classification schemes as new categories become relevant to users.

Proposed solution

Our choice is to use the ACM CCS as a main source for Computing Education Ontology. We also propose a set of changes to the ACM CCS in order to be practically used as Ontology and to follow the chosen Ontology syntax.

First of all, we slightly broaden the interpretation of the CSS as a tree, and think about it as a DAG (directed acyclic graph). The roots of the DAG are the 11 first-level nodes of CCS (marked with letters from B to L), which actually means that at the beginning our Ontology is a DAG consisting of 11 different trees. We can also permit "joining" of these trees by allowing some cross-linking, but only if this will not cause a loop.

For example, the following is a correct Ontology: HP(X,Y) and HP(Z,Y). As you can see, Y is "son" of both X and Z, and you have not a single root.

We also propose not to include in the Computing Ontology the special nodes "General" and "Miscellaneous", appearing at each level in the CCS.

The special opinion is given to the concepts with the same name, existing in the different parts of the CSS. We adopted the following solution:

- if the concepts are exactly the same, to make a cross-link;
- if concepts are not exactly the same, to use a renaming scheme.

The main idea which can enable the use of the Computing Ontology according to the DIOGENE specifications (as outlined in the internal document "OG - A Guide to Ontology Creation"), is that for each new course offered through Diogene, the necessary adjustment (change, enlargement) of the ICT Ontology is necessary. In order to do so, the course authors should prepare a list of all main concepts taught by their course, and to link these concepts with the relations as explained in the above mentioned document.

The relation "Has Part" have to be used for the definition of the hierarchy of the concepts, in order for them to fit to the CCS trees. The other two relations, "Requires" and "Suggested Order" have to be used to define the order of the different sub-trees (different parts of the DAG).

For each new sub-tree of concepts, created for every specific course, it should be decided at what place(s) in the Ontology DAG it should be included. For a single course more than one such sub-tree can be created.

Another important decision was to shift from SHOE to DAML+OIL knowledge representation specification, which partly was forced by the analysis of our first experiments with the Ontology creation. As a consequence we have to develop a new Protégé Plug-in for customising it to the needs of our own Ontology.

We have further to decide about the procedure how our Computing Education Ontology should be enlarged and changed in the future, in order to preserve the completeness and the correctness of the Ontology.

CONCLUSIONS AND FUTURE WORK

We are just starting to work on the Computing Education Ontology creation. We hope that during the next two years we will be able to improve the quality of the prepared material – either the concepts and links between them, as well as increasing the number of new courses which terminology to be covered.

REFERENCES

[1] ACM Computing Classification System (CCS). (http://www.acm.org/class/1998/overview.html and http://www.acm.org/class/)

[2] J. C. Apirez, O. Corcho, M. Fernandez-Lopez and A. Gomez-Perez, 2001. WebODE: a Scalable Workbench for Ontological Engineering. K-CAP'01, Canada. (http://mkbeem.elibel.tm.fr/paper/KCAP2001-35.pdf)

[3] AVEL Sustainability Knowledge Network (http://avel.edu.au/)

[4] S. Bechhofer, I. Horrocks, C. Goble and R. Stevens. OilEd: a Reason-able Ontology Editor for Editor for the Semantic Web. (http://www.cs.man.ac.uk/~horrocks/Publications/ download/2001/oiled-dl.pdf)

[5] D. Fensel, F. van Harmelen, I. Horrocks: *OIL: A Standard Proposal for the Semantic Web*. Deliverable 0 in the European IST project OnToKnowledge. (http://www.ontoknowledge.org/oil/downl/otk.del02.pdf).

[6] J. Heflin, J. Hendler, S. Luke. *SHOE: A Knowledge Representation Language for Internet Applications*. Technical Report CS-TR-4078 (UMIACS TR-99-71). 1999. (http://www.cs.umd.edu/projects/plus/SHOE/pubs/techrpt99.pdf).

[7] Stanford University Knowledge Systems Laboratory. Ontolingua. (http://www.ksl.stanford.edu/software/ontolingua/)

[8] Stanford University. Planning a Protégé 2000 Project. *Protégé 2000 User Guide.* (http://protege.stanford.edu/doc/users_guide/)

[9] Stanford University. Protégé 2000. (http://protege.stanford.edu)

[10] Y. Sure, M. Erdmann, J. Angele, S. Staab, R. Studer, and D. Wenke. OntoEdit: Collaborative Ontology Development for the Semantic Web. (http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/ 2002_iswc_ontoedit.pdf)

[11] The Mathematics Subject Classification (MSC - http://e-math.ams.org/msc/)

[12] Thesaurus Computer Science (alphabetisch). (http://www.inf.fu-berlin.de/bib/thesaurus-cs.html)

[13] UNESCO Thesaurus (http://www.ulcc.ac.uk/unesco/)

ABOUT THE AUTHORS

Assoc.Prof. Dr. Krassen Stefanov, Department of Information Technologies, Faculty of Mathematics and Informatics, University of Sofia "St. Kliment Ohridski", Phone: +359 2 8161 511 or +359 2 8656 157, E-mail: krassen@fmi.uni-sofia.bg.

PhD.Stud. Kornelia Todorova, Department of Information Technologies, Faculty of Mathematics and Informatics, University of Sofia "St. Kliment Ohridski", Phone: +359 2 8161 511 or +359 2 8656 157, E-mail: cornelia@fmi.uni-sofia.bg.