

From Semantic Document Annotation to Global Search Facilities for Personalised Study Programmes

German Nemirovskij, Sandra Rapp
Albstadt-Sigmaringen University
Business and Computer Science
nemirovskij@hs-albsig.de, rappsand@hs-albsig.de

Eberhard Heuel
FernUniversität in Hagen
Information Systems and Databases
eberhard.heuel@fernuni-hagen.de

ABSTRACT

The vision of a common international education space, including universities, infrastructure, and services raised in the last decade. It implicates the composition of personalised study curricula consisting of modules offered by different universities and colleges. In this scenario there is a strong need for efficient services enabling individual search and comparison of study programmes and modules.

The SWAPS (Semantic Web Approach for Personalisation of Study) project, presented in this paper, is focussed on the development of a dedicated search system, based on Semantic Web technologies.

The SWAPS system facilitates automated semantic annotation of programme and module descriptions available on the web. The extracted semantics are stored in machine readable form offering a realistic perspective for efficient processing of semantic search and comparison of programme and module descriptions. This is an important step towards creating a global search and brokerage system for personalised study programmes.

KEYWORDS

Semantic Web, Semantic Annotation, Semantic Search, Personalised Study, Study Module, Study Programme

1. INTRODUCTION

The world wide educational systems underwent significant changes over the past three decades. As a result of social and cultural developments, students have got a lot of new options for designing their personal study career. A large variety of study programmes became available on every level of personal qualification, from undergraduate programmes to special post-doctoral and further education programmes offered for highly qualified professionals. Furthermore, study programmes have become modular, giving students the chance of more mobility during their studies. Organisational and legal efforts have been taken to guarantee high compatibility of study programmes and modules, e.g. the so called Bologna Process in Europe [3]. At the same time new instructional methods (such as blended learning [4]) and infrastructures exploiting digital communication media and networks enable students to pursue their studies in a location independent way.

These processes of change in the education systems will go on. The student of the future will compose study modules offered by different educational institutions to a highly personalised curriculum. According to this vision the role of educational institutions, such as universities, will change fundamentally [14]. They are to be considered as components of a common, integrated education space. Their traditional tasks like offering lectures, seminars, materials, and infrastructure for education will be supplemented by new tasks, e.g. the management of personalized educational programmes.

Unfortunately, the realisation of this scenario is still hindered by the lack of comfortable and easily available technologies supporting students, docents and education managers. The most important technology in this context is the support for search and comparison of study programmes and modules. Such service should help current and prospective students to search for suitable study programmes and single modules as well as to compose personal curricula by expressing their multi-faceted requirements. Managers and docents should be supported when assessing awards of students gained during previous studies, e.g. during a college

year abroad. Educational institutions and faculties may use the service for search of potential cooperation partners. The development of a web portal offering a graphical user interface as well as a collection of web services for search and comparison of study programmes and modules, based on Semantic Web technologies, is the objective of the SWAPS (Semantic Web Approach for Personalisation of Study) project.

The focal idea of the project is to offer users a large range of highly detailed and most recently updated programme and module descriptions by using descriptions already available online. This approach first time mentioned in [23] contains a lot of non-trivial requirements in view of the heterogeneity of these descriptions concerning structure, languages and vocabularies used. The SWAPS project tackles these requirements by exploiting the semantics of programme and module descriptions. Related documents are annotated and indexed with reference to an ontology describing the system of terms and relations used in the domain of interest. The same ontology is used in the queries set users' search operations. Thus, we achieve a highly efficient semantic search and comparison of module descriptions.

After discussing related state-of-the-art work in section 2 of this paper, we will present four characteristic use cases for the SWAPS system. The structure of programme and module descriptions to be annotated is discussed in section 4. A detailed description of the annotation workflow is shown in section 5 while section 6 gives some conclusive remarks and an outlook to future work.

2. RELATED WORK

Launched under the European Framework Programme 5, the R&E project *Cuber* [5] delivered one of the most important issues concerning a higher level of flexibility in European study programmes. The project's result was the CUBER system. It offers two interfaces: one for the users who are searching for study modules and another one for the university managers and docents, supporting them in submitting the metadata which describe a course [14].

The system revealed two problems: 1. A large amount of metadata needed for the description of only one study module leads to motivational problems among docents or module/programme managers. 2. The discrepancy between the accuracy and exhaustiveness of users' search queries needed to achieve good results and users' often scarce knowledge about the search domain makes the interaction of user and system extremely complex. The interaction interface is overloaded and the requested efforts to complete the search process are high.

In order to overcome the first problem, we suggest the automation of submission achieved by automated annotation of published documents which describe study programmes and modules. In general the annotation technology is explained well in [12]. In [11] is proposed an approach for annotation of documents including tabular structured descriptions of technical product features. As long as study module and programmes descriptions have a similar tabular structure (we see a lot of parallels between this approach and the one suggested in this paper) we adopted some ideas stated in this paper for our annotation approach. Besides that, [1] shows an approach for annotation of HTML documents based on Cascading Style Sheets. We used this experience while programming the annotators for extraction of tabular structured information from HTML documents.

The second problem (high interaction complexity) is tackled through implementing semantic search and comparison that use information about relations between concepts within the search domain [9]. The result of "semantic search" [7] is not based on the occurrence of search terms in a document as it works in the "lexical search", instead it is based on the analysis of relationships between *concepts* [21] building the user's query and *concepts* used in the documents. Relationships within an entire corpus of concepts used in all related documents are modelled by ontologies [21], i.e. hierarchically organised graphs. The use of this additional information simplifies the searching process and makes the semantic search significantly more efficient than the lexical search.

Thus, the satisfying search results may be achieved not by enabling users to extensively specify the properties of study programmes and modules, such as the subject of study, the learning contents and outcomes, the teaching/learning method, the terms of time, the location of study events, the price level, etc., but rather by analysing relations between properties already indexed and comparatively short user queries.

Semantic search and semantic annotation are increasingly applied for management of digital learning resources as it is shown in [6]. [17] describes the semantic annotation of learning resources using features of

the document layout. The most recent approaches concerning personalised syllabi for learning resources which are available through digital libraries [19] suggest the standardisation of syllabi representation in order to facilitate the personalised access to learning resources. [18] presents a concept aiming at annotation of syllabi to facilitate semantic search and access to learning resources stored in a digital library through the Semantic Web. Though there are significant differences between management/usage of digital learning resources on the one hand and study programs/modules on the other hand, we have widely exploited the experience of the projects described above.

3. SWAPS USE CASES

The SWAPS system is basically built for two user groups: 1) students who are beginning or proceeding their studies and are looking for information about programmes, modules and courses globally offered by educational institutions 2) docents and managers of educational institutions who are looking for support in assessing students' performance and awards gained in previous studies at other institutions. The following four typical use cases will be supported by the SWAPS system:

1. A student wants the system to generate a personal curriculum for bachelor/master/integrated studies. In this case a possibly complete specification of the student's personal profile and requirements is needed, including among others a list of subjects studied in the past, the awards gained in the past, possibly her/his learning style such as holist/serialist [20], the profession s/he plans to practise after the studies, a list of subjects s/he wants to study, time terms for the planned studies, preferred location, etc. As a result the system generates a set of personal curricula matching the student's query best.

2. A student looks for a particular study module/course. In a Google-like style s/he specifies the most important characteristics of the item to be searched for, e.g. subject and location. The system generates a list of items found.

3. A docent wants to figure out if the programmes/modules/courses attended by a student in the past comply with the prerequisites required in the programme/module/course by the own institution. The docent selects the programmes/modules/courses attended by the student in the past and the programme/module/course the student applies for and fires the query. If the systems' answer is negative, it specifies the particular subjects/credits the student is missing to be accepted as a participant.

4. A docent wants to figure out if the module/course attended by a student at a different institution may substitute a module/course offered by the own institution. If the answer of the system is negative, it specifies the particular subjects/credits that are missed in the module/course attended by the student previously.

4. SEMANTIC ANNOTATION OF STUDY PROGRAMME AND MODULE DESCRIPTIONS

To make the search/comparison results useful and attractive for the users, detailed descriptions of a large range of study programmes and modules must be available. The experience of previous projects [14] shows, however, that it is not realistic to expect that docents and educational managers create suitable dedicated descriptions for the purpose of search systems and that this fact actually builds a strong barrier on the way to an efficient search system. This is why we decided to reuse already existing descriptions of study programmes published on the web, extracting the attributes of modules and programmes described in the documents, storing these attributes in the SWAPS ontology and using them later for processing the incoming search queries.

As mentioned above, European universities undergo the so called Bologna process [3] aiming at the compatibility of study programmes and the fostering of intra-European student exchange. One of the important items of the Bologna process is the subdivision of study programmes into formally and fully described modules. In most cases their descriptions are available online. The descriptions of study modules of the Durham University (Department of Computer Science) include, for example, the following fields: *Type, Level, Credits, Availability, Module Cap, Location, Prerequisites, Corequisites, Aims, Content, Learning Outcomes, Modes of Teaching, Teaching Methods and Contact Hours, Summative Assessment and Formative Assessment* (figure 1).

In the United States you usually find the subdivision of degree programmes (such as master and bachelor programmes) into courses. The quality of course descriptions available online varies a lot. In many cases, however, full descriptions are available from the faculties. The most exhaustive course descriptions are provided, e.g., by MIT [15].

In Australia courses (the term is equivalent to study programmes) are subdivided in subjects which correspond to European study modules and to courses in the United States. Course handbooks (similar to programme handbooks of British universities) include descriptions of subjects which are usually well structured. You find a sample of well structured and exhaustive descriptions at the University of Melbourne (Faculty of Science) [10].

Module Description

Department: COMPUTER SCIENCE

COMP2092: SOFTWARE ENGINEERING (40 CREDITS)

Type	Open	Level	2	Credits	40	Availability	Available in 2006/07	Module Cap	None.	Location	Durham
-------------	------	--------------	---	----------------	----	---------------------	----------------------	-------------------	-------	-----------------	--------

Prerequisites

- Programming and Data Structures (COMP1082) OR Introduction to Programming (COMP1011).

Corequisites

- None.

Excluded Combination of Modules

- None.

Aims

- Have gained a detailed understanding of the phases of the software development lifecycle and know the best practice and problems that are associated with each of the individual phases.

Content

- Project management.
- Human computer interaction.
- Software requirements analysis.
- Software design.
- Software testing and software quality assurance.
- Advanced software lifecycles.
- Business and professional issues.

Learning Outcomes

Subject-specific Knowledge:

- Have gained a first hand experience of software development that is realistic and applicable to software development in industry.
- Have an appreciation of the problems facing the software development industry in terms of the software development process and general project management.
- Able to describe and analyse how each of the issues within software engineering interrelate.
- Have gained additional knowledge of the problems faced in 'real world computing' through representatives from industrial software development companies.

Figure 1: Module description “Comp2092: Software Engineering”, Durham University, Department of Computer Science [8]

The study module and programme descriptions are usually available in HTML or PDF formats. They are made to be read by humans and are not machine readable. However the information enclosed in the documents can be extracted and restored in a machine readable form by means of semantic annotation., e.g. by marking up particular text chunks representing information classes (concepts) , like *Person*, *Language*, *Country*, etc., which descriptions in a formal machine readable form are available.

By now, the fully automatic annotation has not been developed. The state of the art is the semi-automatic annotation: After an annotation is made by machines it is overviewed and adjusted by humans [12]. Yet the existing approaches show a rather high automation level if they are applied to documents using strongly restricted vocabulary and grammar, e.g. if the structure or layout are annotated as described in [1], [11], and in [17].

Well-defined structure and restricted grammar of programme and module descriptions are essential prerequisites to make these documents a perfect source for highly automated semantic annotation and indexing. Among 100 evaluated descriptions of study programmes and modules 90 were published in a tabular form, had not more than 40 fields, and a common vocabulary consisting of less than 900 terms. Furthermore, names of module or programme attributes, such as *Content*, *Learning Outcomes* or

Prerequisites are strictly separated from attribute values in most of descriptions (see figure 1). This regularity enables the SWAPS system to annotate the pairs *attribute-name/attribute-value* in a highly automated way.

The annotation results are used for the document indexing needed later to answer to the search queries, and for the extension of the SWAPS knowledge base, an instantiated ontology needed for semantic document annotation. This knowledge base contains all the attribute classes and class instances, e.g. *attribute-name/attribute-value* pairs, which have ever been annotated by the SWAPS system. Whenever an unknown pair is found, it is manually assigned to an appropriate attribute type and stored in the knowledge base. We call the workflow of attribute extraction followed by the indexing of documents the *SWAPS Annotation/Indexing Pipeline*. This pipeline is described in the following section.

5. ANNOTATION/INDEXING PIPELINE

In the current version of the SWAPS system, the annotation and indexing of programme and module descriptions is processed through several steps integrated into the annotation/indexing pipeline (figure 2), which more detailed description is given in [22]. While developing the pipeline we exploited the experiences gained in an approach for the extraction of product data from web pages [11]. As product descriptions mostly have a tabular layout, their annotation processing is similar to that needed for tabular structured module descriptions. However, product attribute values like *length*, *width*, *contrast* or *colour* appear elementary and “simple” compared to complex values of module or programme attributes like *Content*, *Learning Outcomes* or *Prerequisites*. The need for an extended approach results in the SWAPS pipeline.

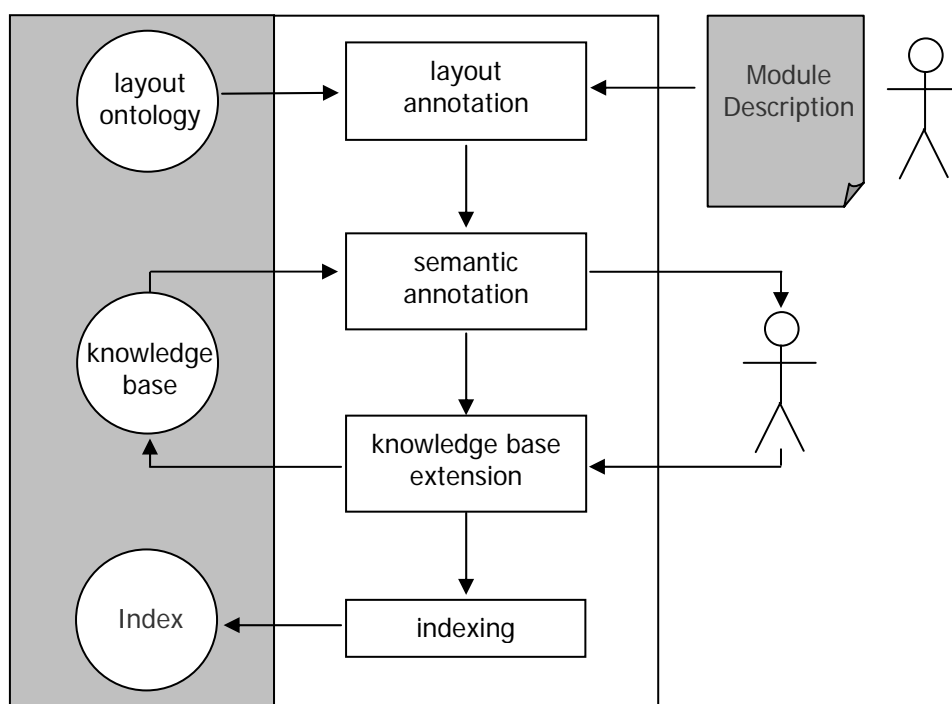


Figure 2: SWAPS Annotation/Indexing Pipeline

Layout annotation: In the first phase of the pipeline a docent or a module manager registers a module by submitting the URL of the description available online via a special web interface. The annotation of the document layout is carried out. This is a fully automated process. The goal of this step is to extract the *attribute-name/attribute-value* pairs. For a certain number of formats and structures the appropriate annotation methods are already available, e.g. an approach for the layout annotation of CSS based web pages is presented in [1]. The most simple layout annotation case is that of `<dd>` and `<dl>` HTML-tags used to mark up the attribute names and values respectively. In the *SWAPS Pipeline* several layout annotators using

different strategies are invoked in concurrence. The best result achieved is considered as the input for the further pipeline processing.

In the pipeline draft the extraction of *attribute-name/attribute-value* pairs is named as *layout annotation*. On this stage no attempt is made to analyse the semantic (meaning) of the attributes (this is the task of the next phase). The layout annotation is rather focused on the attribute localisation. During layout annotation it does not yet make any difference if a course, a module, or a programme description is annotated. The currently processed description is annotated as an instance of the class *EducationalUnit*.

The result of the layout annotation is the instantiated layout ontology containing the class descriptions as well as the class instances identified in the document being annotated. Figure 3 shows a fragment of the layout annotation result for the module description presented in figure 1. The content of table headers, e.g. denoted <th>, is annotated as instances of the *AttributeName* class. The content of table data fields, e.g. denoted as <td>, is annotated as instances of the *AttributeValue* class. Neither the language used, nor the syntax, nor the semantics of the annotated text blocks are analysed as yet.

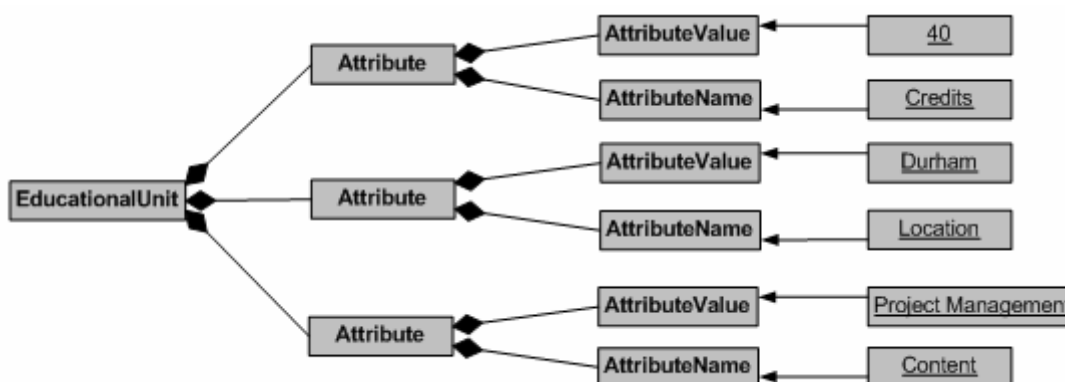


Figure 3: A fragment of layout annotation of the module description shown in figure 1

Semantic annotation: Semantic extraction is the scope of the following pipeline phase “semantic annotation” (see figure 2). Attributes are now classified according to their semantic values and represented through “meaningful” classes (concepts) like *LocationAttribute*, *CreditsAttribute*, *ContentAttribute*, etc. (see figure 4).

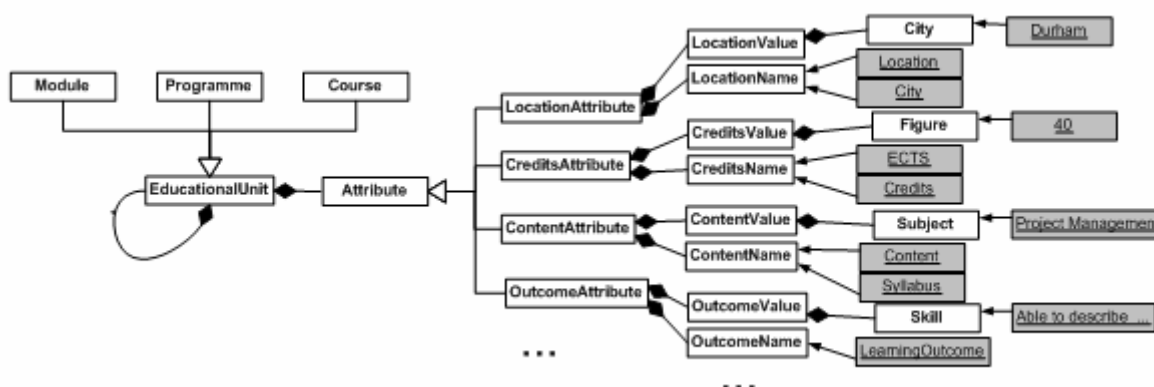


Figure 4: SWAPS knowledge base (entities of instances are shown in grey colour)

To achieve this result attribute names identified during layout annotation are compared with the names of class instances stored in the knowledge base. If matches are found, the “meaningless” *attribute-name/attribute-value* pairs are annotated as instances of “meaningful” attribute classes.

Furthermore in order to increase the search/comparison efficiency the SWAPS system attempts to infer the semantics of attribute values. For this reason attribute values are annotated using common concepts like *Figure*, *Person*, *Skill*, *City*, etc. (figure 4). With this step the semantic annotation process is completed.

At the end of the semantic annotation phase the educational unit which was annotated is located in the hierarchy of educational units (see the left part of figure 4). An educational unit may be a *Programme*, a *Module* or a *Course* and may enclose other educational units. A study programme may aggregate study modules and courses, modules may aggregate courses. Course is an elementary component in this hierarchy. It may not aggregate any other components as it is shown in figure 5.

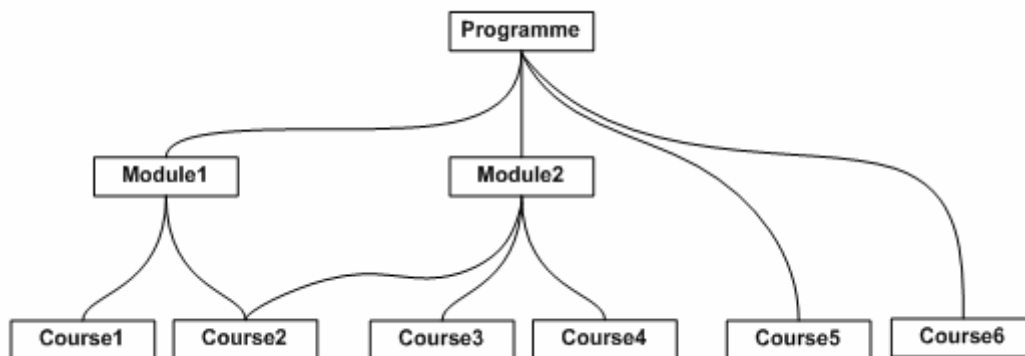


Figure 5: Hierarchy of educational units

Knowledge Base extension: In this system quality of machine annotation depends on the quality of the knowledge base. The more classes and class instances the knowledge base encloses, the more precise and differentiated is the annotation result. Thus the continuous extension of the knowledge base is an important requirement for the SWAPS system. It is met by involving the submitter of programme and module descriptions into the extension process.

The machine generated annotation is presented to the submitter in form of a text with mark-ups. A similar graphic interface is described in [12]. The submitter is able to check, change or supplement the annotation result manually. S/he can annotate items which were not automatically identified by assigning them to existing classes. Furthermore the submitter can add new classes to the knowledge base and use them for further annotation.

Indexing: Indexing is the last phase of the SWAPS pipeline. The study programme and module descriptions submitted by authorised users are considered to be reliable and should be available for the processing of search queries. For this reason the terms occurring in attribute values, attribute class identifiers and the identifier of the educational unit described in the document are stored in the SWAPS index, a data structure used for processing of semantic search. To understand the role of the index we give a short description of the search workflow: After a search query is fired by the user it is inspected for the attribute names stored in the knowledge base, e.g. “Content”, “Location”, “Credits”. In case of success the search domain in the second part of the search is significantly reduced. In the second part, the query is inspected for the attribute values. If there are correspondences between attribute values stored in the index and those mentioned in the query, the tuple consisting of an attribute value, an attribute class identifier and an identifier of the correspondent educational unit is added to the search result.

6. CONCLUSION AND PERSPECTIVES

The SWAPS system is one step on the way towards an integrated world-wide educational space including universities, services, and a common distributed infrastructure. A search system based on semantic web technologies as presented in this paper is assumed to be an important part of this infrastructure. It facilitates highly automated annotation of study programme and module descriptions available online. The extracted attributes of programmes and modules are used for indexing these descriptions and for semantic search, supporting students in designing their personal curricula over the whole range of available programmes and modules.

By now, the kernel components of the SWAPS systems (the search engine, the interface for the submission of descriptions and the annotation tool) have been basically developed. For system deployment on a large scale we need a critical mass of module and programme descriptions. The large number of related

descriptions to be found on the web, however, can not be assumed to be reliable and up-to-date. That is why they can only be used for the extension of the knowledge base but not for semantic search issues. To guarantee the reliability of information, we set up a procedure for submission of these descriptions by docents or by faculty staff.

To this concern we are looking for cooperation with-faculties and departments of universities. Starting within a local network of German universities, we wish to extend partnership and cooperation to international partners in the near future.

7. REFERENCES

- [1] S. Behhofer, S. Harper, D.Lunn: SADie: Semantic Annotation for Accessibility, *The Semantic Web – ISWC 2006*, Springer, Berlin, 2006, pp 101-115
- [2] N.J. Belkin, G. Muresan, Measuring Web Search Effectiveness: Rutgers at Interactive TREC, *WWW2004 Conference Workshop*, 2004
- [3] Bologna Declaration Communiqué of the meeting of European Ministers in charge of Higher Education, Prague, May 19th 2001, <http://ec.europa.eu/education/policies/educ/bologna/bologna.pdf>, (20.02.2007)
- [4] C. J. Bonk, C. R. Graham, *The Handbook of Blended Learning: Global Perspectives, Local Designs*, Pfeiffer Wiley, 2005
- [5] Cuber, Websites of the Project, <http://www.cuber.net> (20.02.2007).
- [6] S. Dehors, et al., Semi-automated Semantic Annotation of Learning Resources by Identifying Layout Features, *SW-EL*, Kaohsiung, 2005
- [7] L. Ding et al., Search on the Semantic Web, *IEEE Computer*, 10(38), 2005, pp. 62-96
- [8] Durham University, Department: Computer Science, Module COMP2092: SOFTWARE ENGINEERING, http://www.dur.ac.uk/faculty.handbook/module_description.php?module_code=COMP2092 (April 4.4.2007)
- [9] R. Guha, R. McCool, E. Miller, Semantic Search, *WWW2003*, Budapest, May 20-24, 2003
- [10] Guide to Courses, Faculty of Science, The University of Melbourne, <http://www.unimelb.edu.au/HB/facs/SCL.html>, (20.02.2007)
- [11] W. Holzinger, et al., Using Ontologies for Extracting of Product Features from Web Page, *The Semantic Web – ISWC 2006*, Springer, Berlin, 2006, pp 286-299
- [12] A. Kiryakov, et al., Semantic Annotation, Indexing, and Retrieval, *Elsevier's Journal of Web Semantics*, Vol. 2, Issue (1), 2005.
- [13] R.R. Korfhage, *Information Storage and Retrieval*, Wiley, 1997
- [14] B.J. Krämer, CUBER für die Ausbildung à la carte, in: G. Simonis (ed.), *Lernort Virtuelle Universität*, Verlag Leske & Budrich, 2003
- [15] MIT Subject Listing & Schedule, Spring 2006-07, <http://student.mit.edu/@9041442.26459/catalog/index.cgi> (20.02.2007)
- [16] K. X. S. de Souza, et al., Visualization of ontologies through hypertrees, *Proceedings of CLIHC*, Rio de Janeiro, 2003, pp. 251-255
- [17] D. Taibi, et al., A Semantic Search Engine for Learning Resources, *m-ICTE2005, Recent Research Development in Learning Technologies*, 2005
- [18] M. Tungare, et al., "Towards a Syllabus Repository for Computer Science Courses", *Full paper accepted for ACM SIGCSE 2007*, Covington, March 2007
- [19] M. Tungare, et al., "Towards a Standardized Representation of Syllabi to Facilitate Sharing and Personalization of Digital Library Content", *Proceedings of SW-EL workshop 2006*, Dublin, June 2006
- [20] G. Pask, , "Learning strategies, teaching strategies, and conceptual or learning style", *Learning strategies and learning styles* ch.4 pp.83-100, R.R.Schmeck (ed.), New York: Plenum, 1988
- [21] M. Ehrig et al., "Similarity for Ontologies - a Comprehensive Framework", *Proceedings of 13th European Conference on Information Systems*, May 2005
- [22] G. Nemirovskij et al., "Semantic Search and Annotation of Programme and Module Descriptions for Personalisation of Study", Submitted to ED-MEDIA 2007, Vancouver, Canada, 2007
- [23] G. Nemirovskij et al., "SWAPS: Semantic Web Approach for Personalisation of Study", Submitted to ICALT 2007, Niigata, Japan, 2007