Metadata in E-Learning: Automatic Extraction and Reuse

Michael Sonntag[1]

The concept of metadata is not new, but currently its use is rather restricted. One of the areas it is actually employed, although usually only to a small extent, is E-Learning. Reasons for this slow adoption are the additional work required for adding it and a lack of applications providing direct advantages to end users. Some approaches to ameliorate both problems are presented in this paper together with applications realizing them: Automatic extraction of metadata from learning resources and the use of metadata to create an index, roadmaps and cross-references between learning units.

1 Motivation

One problem of E-Learning is finding appropriate courses. As the investment in a completely new course is higher compared to creating a conventional course, this aspect is very important. Through the last years a large number of widely differing courses was developed using many different learning systems, pedagogical approaches, content formats and topics. However, relying on current search engine technology to find a matching one for a specific need is rather difficult: Only full text search is supported. This does not work with some course formats, e.g. binary encodings or when compressed. Also, it is useful only when searching for certain topics. Retrieving courses in a specific language might be successful, but when requesting an educational level (e.g. an undergraduate course) or for a certain didactical approach (self-organized learning), this doesn't work at all. A solution for this could be methods introduced by the semantic web, especially the use of ontologies and the annotation of material with metadata. Through explicit marking with metadata matching courses (or elements thereof) can be found more easily and with better precision. However, even with these methods, metadata poses problems (see also [4]):

1. Search engine support: General search engines do not yet support metadata. However, repositories or dedicated E-Learning search engines employ them, although these might be hard to find themselves. This problem can be relied upon to change over time, however.

[1] Institute for Information Processing and Microprocessor Technology (FIM), Johannes Kepler University Linz, Altenbergerstr. 69, A-4040 Linz, Austria. E-Mail: sonntag@fim.uni-linz.ac.at

2. Missing metadata: When creating a course, authors rarely add this additional information. Some of the reasons are that its benefit is not immediately visible and does not improve the initial use of the material (no added value *now*, only for potential later reuse). Moreover, additional work is required by using an (often different or cumbersome) editor for inserting or composing the metadata information at the different elements (the whole course, individual parts like lessons or resources, etc.). This is an area where the individual author is required for improvement and where a competitive advantage can be gained through findability and ease of assessment of suitability as well integration.
3. Lots/Differing standards: Several metadata standards/specifications exist [10], and even though many are rather similar (most are, like IMS [7] and ARIADNE [2], based on LOM [11]), they differ in details (both naming and structure). They also exist in several versions already (e.g. IMS: 5th version). Making matters worse, they are based on vocabularies often unknown to authors who should annotate their materials according to them. One example is the learning resource type (IMS): The exact difference between "diagram", "figure", and "graph" can be difficult for non-native speakers or those unfamiliar with the specification. Another issue is the format of the metadata: Whether RDF or "pure" XML is much disputed (e.g. [15] and [13]).

This paper will present some approaches to ameliorate these problems in chapter two and focus on two of the strategies shown, reuse for other areas and metadata as help for certification in different forms, in the following chapters. Chapter five presents some tools implementing several of the approaches. The paper ends with some conclusions.

2 Reducing problems of metadata

One way to reduce the additional work required for adding metadata to an otherwise finished course is automatic extraction. Some metadata can be extracted from other information (e.g. the content) or converted from other metadata formats, obviating the need for explicit (re-)annotation. Examples for exact derivation are the language of the content, its structure and type. Other elements can only be extracted with a lesser degree of certainty or depending on the format/content. Examples for these are the title or keywords. Keywords e.g. can be extracted perfectly if specified in file properties of documents (e.g. Microsoft OLE 2 Compound Document format or PDF), very well when explicitly marked as such (on separate line and preceded by "Keywords:"), but only very roughly from plain content text trough automatic and unsupervised text analysis.

Another way to enlarge the use of metadata is tool support for authoring. Providing additional information (and especially examples) in easy-to-use tools for adding metadata according to different standards as well as support for conversion between them would be very helpful. Especially the latter is a difficult field because of the fragmentation of standards and their often subtly different meanings, so just syntactically converting it to another structure (e.g. renaming the content or moving it into added elements) will sometimes be not enough but require additional changes to the content (selecting a different value).

The reluctance to invest the additional work for adding metadata could be overcome when focusing not only on searching (which is often a rather remote use for developers), but also on other uses, especially such providing immediate benefits during the initial use, and especially for singular courses.

3 (Re-) Use of metadata for other areas

Reuse of metadata is in my opinion the best incentive for adding it, as immediate benefits are gained. Some ideas for this are presented, which partly have already been implemented (see below).

3.1 Automatic index creation

Creating an index for learning material is a lot of work, especially as there is usually little or no support from the editor. Text processors mostly support this, but for other elements (graphics, animations, sound files, etc.) this is unavailable. Integration is also lacking: At most an index for several individual files of the same type can be created, but none spanning a whole course, which in most cases consists of many different kinds of content. An index provides an alternative and learner-dependent (instead of defined by the teacher) approach of navigation for the course content, which is especially important in the area of continuous learning ([21]), where quickly locating parts of interest is desired. If the complete course is assembled from several independent parts, it also strengthens the integration. As the parts are independent, there are no cross-references or links between them. An index is at least a first approach to integrating them more fully by allowing finding related issues in different parts of the course.

However, if metadata is available for the individual elements (and depending on its quantity and quality) an index can be created automatically. Keywords of all the learning material elements are collected and arranged together with links to the content. As in a full index, duplicates are merged to a single entry with several references. To avoid problems with slightly different words (the courses need not be completely homogenous, e.g. when assembled from parts of different authors), this can be supported by taking into account the language of the keywords and using an appropriate stemming algorithm. For grouping similar keywords therefore only the word stem should be used. This introduces some uncertainty, as different words (or when incorrectly guessing a non-specified language) could result in identical stems. Therefore when using this approach, the full keywords should be listed alongside the references.

A drawback of this "reuse" is that this is a meta-index in the sense that it only links to individual learning resources as a whole, but not to locations within them. Implementing this is difficult as it would depend on both the actual structuring of the content (its file type) and the navigation on the specific software used to display the resource. Quality therefore depends on the content's granularity: Many small and annotated resources yield a large and useful index, while a few large documents result in few entries leading to a single large document, which is less advantageous.

Subentries are possible but probably not very accurate unless the content is very finely partitioned (so keywords apply only to small elements). Here the same methods as for collapsing similar entries could be used. However, it would depend on subentries starting with a longer identical substring or word; different but related words could only be grouped together with the use of an extensive ontology containing both words under consideration. As such ontologies are currently rare or only available for small and isolated areas, this is at the moment not a very promising approach, but could be of help in the future.

3.2 Relations between elements (automatic cross references)

When learning material is created by assembling individual parts from several sources, cross references are missing (each part must be self-contained; e.g. a requirement for submanifests in IMS). However, links between parts referencing related topics or additional information are essential parts of web-based training: One reason for the Internet's success is its abundance of cross-links in addition to often strict hierarchical navigation, allowing pursuing topics independently of the main and pre-given structure.

This also facilitates explorative learning, where learners take a more active role by selecting their individual way through the material which need not be hierarchical. With a history showing visited and especially unvisited areas (a rather simple task if the LMS supports adaptation, but otherwise quite difficult) even completeness of learning is achievable. These cross connections can be added either manually through some effort (identifying for each part all the other elements which might be of interest), or automatically based on metadata (see [17] for a system using explicitly provided metadata and an ontology). Links can be derived from several elements of metadata:

- Keywords: Learning units with similar keywords focus on similar topics and can therefore be connected and marked as "related". This produces a list of related elements which can also be enriched with additional metadata (e.g. the referenced items title or description). This is especially useful and working well if the same content is presented in different ways, e.g. units providing a textual, a graphical, an audio or video description of the same content or one in a different language.
- Explicit relation data: E.g. the LOM standard contains several predefined relation types. Especially the "references" and "requires" elements could be used for creating links to other elements (or checking for completeness). This probably only works for parts, which were at least potentially intended for combination, else this annotation would not be contained. As they were explicitly created, no margin for errors remains so this should be used whenever present.
- Classification: If elements are described in detail by their position in a subject area, this can also serve to connect material. In common hierarchical and numerical classifications even a measure for the distance is possible by taking into account the difference in numbers and hierarchy steps between the two items. This allows better annotation of the links and e.g. presentation in different degrees of importance (location, color, size, …). An example is the ACM classification (already converted into an ontology [18]), which could be used directly for teaching modules classified accordingly at each subpart.

The same restrictions as with automatic index creation apply: References can only originate at and target complete units. As reuse will mostly combine numerous small units and rather less frequently group a few big items, this is of less concern here. Another drawback is that their annotation (i.e. the type and meaning of the reference) is only weakly ascertained unless it is based on explicit relation information.

Both for creating links according to keywords and classifications extended ontologies about the topic area could be helpful for ascertaining the degree of similarity and the kind of relation.

3.3 Annotated course lists

Information about courses in listings of those available is often separated from the course itself: Data (topic, prerequisites, technical requirements, etc.) must be entered separately and is stored independently, resulting often in inaccurate data and duplication of effort for entry and update. For E-Learning courses these could be partially extracted from metadata of the learning material (if present there). Through this the problem can be ameliorated and metadata reused.

However, not all required information is contained therein (e.g. the LOM metadata standard does not provide information about the room of the course, schedules for different activities, or the teacher), as usually only the material itself is annotated, excluding organizational data. This information is commonly part of the learning management systems (LMS) instead, which is responsible for lifecycle and presentation of courses. Here also specifications exist, but these are far less universally used than those for metadata or content aggregation.

3.4 Personalized course delivery

Some parts of the metadata could be used for personalizing course delivery, e.g. again keywords and classifications, or other metadata like interactivity type/level, end user role or difficulty. However, this provides only a small base for adaptation and additional specific metadata is required for comprehensive personalization (see e.g. [3] for such an extended system).

4 Metadata and certifications

Metadata can improve comparability between courses and curriculums: A detailed structured description of the content as well as the method of delivery and pedagogical models employed is contained. But not only materials can be compared, also examinations and to some extent even results. When a course according to certain material (including associated tests) has been finished, its metadata describes the proficiencies of the students and can be seen as describing the certificate issues based on this in detail. This allows easier comparison of results between courses with similar content not only according to the verbal description of the course but also automatically through comparing metadata of the course or its individual elements, while avoiding the need to introduce separate assessment metadata (e.g. [6]; probably better results but requiring again more metadata entry and management). This however works only on one of the two elements of certification: contents and achievement with respect to the content items. Only the former is comparable, while although the latter can be measured (e.g. percentage of questions answered correctly), comparing it between courses is difficult. A measure for the difficulty of a course is defined in metadata standards, but it doesn't necessarily apply to the examination. Additionally, the assessment standard need not be identical between different authors, institutions or target groups. A specification for "Reusable Definition of Competency or Educational Objective" by IMS [8] exists, however it only defines an unstructured set of independent statements. Compared to a complete ontology with a hierarchy (competencies consisting of smaller parts or alternatives) this is probably only slightly useful. However, if this (or any similar information) is integrated into metadata it would improve the method outlined above considerably, as comparisons could then be made with exactly the same accuracy as used for the definition of the competencies.

Metadata can also be useful when comparing materials to decide which one to use for a course (or for inclusion in larger material; after it has been found), although often established standards are insufficient. E.g. in Austria the ministry of education has recently adopted its own extensions to the LOM metadata standard ([1]). Every electronic supplemental for officially approved schoolbooks (or completely electronic teaching material) must be annotated according to this standard, resulting in at least some annotation of electronic material. This intends to allow easier comparison and selection between them by teachers. The specification is based on LOM (using RDF binding; a pure XML binding was developed at this institute) and defines additional elements and some additional values for existing elements (e.g. contributor roles: sponsor, publisher, author, editor, creator). It is modeled after the "application profile" pattern ([5]), consisting of several namespaces optimized for local applications.

Additional elements of importance in this context are:

- School type, educational level and curriculum coverage: This defines for which level of competence and which target group the material is appropriate and includes taxonomies respectively exact content definitions. As assessment standards are also defined in these categories (although not included in the specification), comparison of results is then possible reliably. The only variations remaining are from the individual assessment of the teacher (which is unavoidable).
- Certifications: In Austria all school books (and electronic material) must be approved before they can be used in schools (usually several approved books/courses exist from which the teacher can then select according to her/his preferences). This certification status is included as metadata, providing a different kind of certification missing in other standards. It is stored in plain text and cryptographically unsigned, however, and therefore relies on a trusted server (as is the case in Austria). This is a serious drawback as it prevents the passing on and reuse of modules with reliable certification. If a new course is assembled from pieces (or whole units) of approved content, it must be certified in detail and completely again, instead of only checking the scope of the compilation or any additional elements which are not certified themselves (or excerpted from certified base material).
- Relation to schoolbook: Some material is intended as a supplement to conventional books, other as replacements. For the former the relation with the book must be specified in detail: Is it the same con-

tent as the book, additional examples, additional content or does it provide more details? Also, to which chapters/sections in the book an entity is related must be specified. If the book is known to the person comparing two students, results in electronic tests can then be compared with a high degree of certainty and even to students learning with the physical book only. Still, this is a very restricted use of classification for assessment as it relies on a single reference element, the physical "base" book.

Tool support for automatic comparison of metadata between two units or courses is currently missing. Still, manual comparison of results of students or courses can be eased if metadata is available.

5 Practical example: Extraction and use of metadata

At the institute several tools for furthering the use of metadata have been implemented and other ones are currently in development. These span the "lifecycle" of metadata from extraction/creation to use.

5.1 Creation

Metadata can be added to learning material either manually or through automatic extraction. For the latter a tool was developed. It extracts the content language, the title and keywords from several file formats. Supported are Microsoft files (Word, PowerPoint, Excel), PDF, CPS-Manifests (according to LOM final draft and IMS specifications in versions 1.1 – 1.2.2), HTML files (as META tags or as contained Dublin Core data; referenced external data is not extracted) and plain text files. The software can be easily extended to support more languages or to extract and add different metadata. Extraction relies primarily on metadata already contained in the different formats in some kind, but as fallback also extraction from plain content text is implemented and used if other methods fail.

Language detection is based either on directly contained language specification, but in contrast to the other elements this is rather rare (e.g. the Microsoft formats do not provide this in their metadata). Therefore an approach using stop word lists is used. Stop words are words which are very common in a certain language but possess little meaning (e.g. articles like "the"). These lists are commonly used by search engines to filer out unimportant words from queries. Here the exact opposite approach is taken: Everything *but* stop-words is filtered out. The list with the most matching words is assumed to be from the same language as the text. Testing showed this to be very reliable if the input is a complete text and not just a list of phrases, not too short and consists of a single language (individual foreign words are no problem: these are almost never stop words).

Keyword and title extraction is mainly based on existing metadata, except for plain text. There it should not be used as a final result but rather a starting point for human revision: A relatively simple custom algorithm is used. All words excluding stop words are counted in their base form (word stemming algorithms specific to the language used). Afterwards the top 5% words are discarded, as they are probably of less importance (very common words, although not stop words), as well as all words with a relative occurrence below 80%. All keywords are however required to occur at least four times. These values result from tests of the algorithm but are only rules of thumb; they can be changed easily. This algorithm can also be replaced completely by another (e.g. commercial) implementation without changing the code.

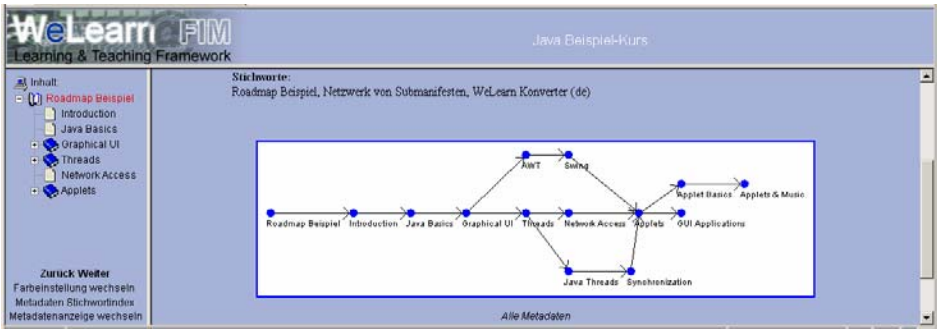
Input for the tool is a complete manifest and the base directory for the actual content files. It then extracts the metadata from all files referenced and integrates it into a copy of the manifest in the same standard that is already used (IMS Version 1.2.2 as default if no metadata was present previously). The software can be used either stand-alone with a command line interface or be integrated into other software, as planned for the new version of WeLearn.

It must be noted that metadata extraction is used here in the sense of "creating new metadata from content", not sharing already existing metadata in a defined format (e.g. [14]), which is the next step.

5.2 Offline presentation

Most E-Learning standards are geared towards online and dynamic delivery. Sometimes the content should also be provided offline, e.g. on CD-ROMs, with most of the functionality still available but without installing any software (other offline viewers usually contain an embedded web server to install, e.g. the SCORM conformance test suite [19], or the Reload SCORM Player [16]). For this application a converter was developed ([9]) producing different views (Applet, DHTML and simple HTML; for catering to different user groups and browsers), which employs metadata in several ways:

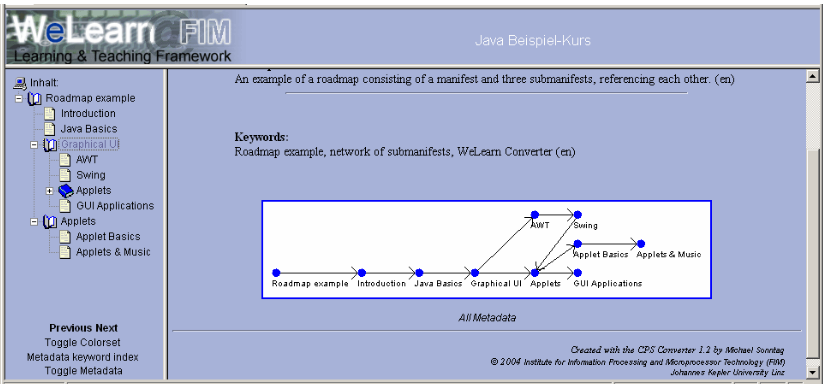
- Integration into output: Any metadata item can be individually accessed and rendered on the pages created through an XPath expression referencing it. The provided examples employ this to display German or English description and keywords depending on the actual template used.
- Creating a course roadmap: Based on the aggregation level of the content items as well as submanifests, a graphical view of the course is created [12]. This is an alternative to the main navigation; it need not be a tree, but can also be a network, depending on the actual content (see *Figure 1*).



Stichworte:
Roadmap Beispiel, Netzwerk von Submanifesten, WeLearn Konverter (de)

Figure 1: Example of full roadmap (German template)

- Filtering out certain elements: The content can be filtered upon conversion to only show items of a certain classification. This allows creating smaller courses, e.g. for certain target groups, from a larger course without internal changes. Basing them on a single main unit avoids problems when updating content, as different versions are automatically updated on the next conversion (see *Figure 2*).



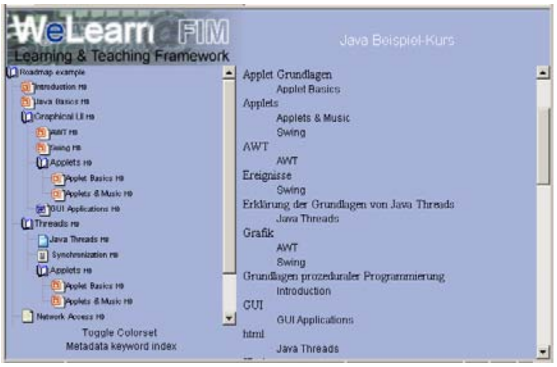
An example of a roadmap consisting of a manifest and three submanifests, referencing each other. (en)

Keywords:
Roadmap example, network of submanifests, WeLearn Converter (en)

Created with the CPS Converter 1.2 by Michael Sonntag
© 2004 Institute for Information Processing and Microprocessor Technology (FIM)
Johannes Kepler University Linz

Figure 2: Example of filtered roadmap (English template)

- Content index: An index of the content is created fully automatically, based on metadata keywords. These are linked directly to the individual items they refer to (see *Figure 3*). Depending on the navigation used, this is also connected back to the navigation structure (both Applet and DHTML navigations support this feature).



Applet Grundlagen
Applet Basics
Applets
Applets & Music
Swing
AWT
AWT
Ereignisse
Swing
Erklärung der Grundlagen von Java Threads
Java Threads
Grafik
AWT
Swing
Grundlagen prozeduraler Programmierung
Introduction
GUI
GUI Applications
html
Java Threads

Figure 3: Index page

5.3 Online use

For the next version of the institute's Online Learning Platform (OLP) called WeLearn (currently in development), metadata is employed more extensively. Already implemented is providing learners with a personalized page according to their interests (derived from the keywords associated with the materials they visit and derived e.g. from the text of their posts in forums). Base data is extracted either from metadata provided by the author in the course material, or (if missing) through the tool described above. Planned is finding learners with similar interests, which improves cooperative learning [20]. Another use is creating an image of the current position of other learners within the learning material in the vicinity. This is based on the roadmap presented above with additional graphical indications: coloring nodes for current location and number of near students and additional textual information.

Producing cross-connections between items within a course and between courses is currently under consideration. As the content in a LMS might be large, comparing all data with each other requires too much time with often little chance of success. This would be a task specifically suited to agents as it requires a larger degree of intelligence. Also, integration into the systems user interface is yet unclear.

6 Conclusions

When metadata is actually available for a larger amount of resources (either through manual addition or automatic derivation) search engines will be more inclined to actually support it. In the meantime, other benefits can be used instantly: Metadata need not only be used to show a description of the content but can serve different other purposes. One example is adaptation to the actual user in several forms. This starts from rather trivial selection of equal content according to the media type (text or video) and ranges over identifying similar or related content and providing additional navigation structures like an index up to automatically deriving interests of users and selecting matching content.

Metadata is a very important element especially for learning content and learning management systems, as this is data to be reused often (and possibly in different contexts). It is also naturally associated with other data like different lessons, information for the teacher, or exercises. What is currently still missing, or available only in rudimentary form, is information on connections with other metadata. Here the natural barrier is that only data known in advance (i.e. existing and known courses) can be explicitly referenced. Therefore methods for automatically creating such cross-connections are of prime importance. The methods and tools presented here can serve as an initial stepping stone for approaching this goal.

Acknowledgment

This paper is a result of the project "Integrating Agents into Teleteaching-Webportals" sponsored by the FWF of Austria (Fund for the support of scientific research; Project number P15947-N04).

Literature

[1] Austrian Metadata specification for electronic Teaching-/Learning materials, Version 1.32 from 12.1.2004 http://elearning.bildung.at/statisch/bmbwk/de/elearning/metadatenmodellversion1_3_2.pdf

[2] ARIADNE Educational Metadata Recommendation: http://www.ariadne-eu.org/en/publications/metadata/index.html

[3] CONLAN, O., HOCKEMEYER, C., WADE, V., ALBERT, D.: Metadata Driven Approaches to Facilitate Adaptivity in Personalized eLearning Systems. Journal of the Japanese Society for Information and Systems in Education, 2003, http://www.cs.tcd.ie/Owen.Conlan/publications/JSISEv1.23_Conlan.pdf

[4] CURRIER, S.: Metadata Quality in e-Learning: Garbage In – Garbage Out? http://www.cetis.ac.uk/content2/20040402013222

[5] HEERY, R., PATEL, M.: Application Profiles: Mixing and matching metadata schemas. Ariadne, Issue 25 (24.9.2000) http://www.ariadne.ac.uk /issue25/app-profiles/intro.html

[6] HSIEH, C., SHIH, T. K., CHANG, W., KO, W.: Feedback and Analysis from Assessment Metadata in E-Learning. In: Proceedings of the 17th International Conference on Advanced Information Networking and Applications (AINA'03). http://www.mine.tku.edu.tw/scorm/hsiehct_assessment.pdf

[7] IMS Meta-data specification: http://www.imsglobal.org/metadata/index.cfm

[8] IMS Reusable Definition of Competency or Educational Objective: http://www.imsglobal.org/competencies/index.cfm

[9] LOIDL-REISINGER, S., SONNTAG, M.: Using metadata in creating offline views of e-learning content. In: AUER, M. E., AUER, U. (Eds.): Interactive Computer Aided Learning. Kassel: Kassel university press 2003

[10] LOIDL-REISINGER, S., PARAMYTHIS, A.: Distance Education - A Battlefied for Standards; In: SZÜCS, A., WAGNER; E., TSOLAKIDIS, C.: The Quality Dialogue. Integrating Quality Cultures in Flexible, Distance and eLearning; Proceedings of the 2003 EDEN Annual Conference, Rhodes (Greece)

[11] IEEE WG12: Learning Object Metadata. http://ltsc.ieee.org/wg12/

[12] MÜHLBACHER J. R., SONNTAG, M.: Roadmaps - Navigational Aids and Tools for Reusing Content in Distance Education. In: SZÜCZ, A., WAGNER, E., TSOLAKIDIS, C.: The Quality Dialogue. Integrating Quality Cultures in Flexible, Distance and eLearning. Proceedings of the 2003 EDEN Annual Conference, Rhodes, 16-18.6.2003

[13] NILSSON, M., PALMER, M., NAEVE, A.: Semantic Web Metadata for e-Learning – Some Architectural Guidelines. In: Proceedings of the 11th World Wide Web Conference. Royal Institute Of Technology, Stockholm, 2002.

[14] Open Archives Initiative. http://www.openarchives.org/

[15] QU, C., NEJDL, W.: Integrating XQuery-enabled SCORM XML Metadata Repositories into an RDF-based E-Learning P2P Network. Educational Technology & Society, 7 (2), 2004 http://ifets.ieee.org/periodical/7_2/8.pdf

[16] RELOAD: Reusable eLearning Object Authoring & Delivery: http://www.reload.ac.uk/

[17] RONCHETTI, M., GIULIANI, A., SAINI, P.: De-Fragmenting Knowledge: Using Metadata for Interconnecting Courses. http://eprints.biblio.unitn.it/archive/00000407/

[18] SAINI, P., RONCHETTI, M.: Deriving ontology-based metadata for E-Learning from the ACM computing curricula. http://eprints.biblio.unitn.it/archive/00000405/01/017.pdf

[19] Advanced Distributed Learning Initiative: SCORM http://www.adlnet.org/

[20] SONNTAG, M., LOIDL-REISINGER, S.: Cooperative Agent-Supported Learning with WeLearn. In: CHROUST, G., HOFER; C. (Eds.): Euromicro 2003. New Waves in System Architecture. Proceedings of the 29th Euromicro Conference. Los Alamitos, IEEE Computer Society 2003

[21] ZANG, D., ZHAO, J. L., ZHO, L., NUNAMAKER, J. F.: Can E-Learning replace classroom learning? Communications of the ACM 5 (47) May 2004