A Quick Scan on possibilities for automatic metadata generation

Frank Benneker

A Quick Scan on possibilities for automatic metadata generation

Frank Benneker

October 2006

Learning objects in practice 4

samenwerken aan vernieuwing van het Hoger Onderwijs

DU Digitale Universiteit

Colophon

A Quick Scan on possibilities for automatic metadata generation

Learning objects in practice 4

Stichting Digitale Universiteit
Oudenoord 340, NL-3513 EX Utrecht, The Netherlands
P.O. Box 182, NL-3500 AD Utrecht
Telephone +31 - 30 - 238 8671
Fax +31 - 30 - 238 8673
e-mail buro@digiuni.nl

Author

Frank Benneker

Copyright

Date

October 2006

Distribution

The series Learning objects in practice is available from the following recognized websites:

- www.du.nl/leerobjecten
- dspace.ou.nl
- www.sco-kohnstamminstituut.uva.nl
- www.hbo-kennisbank.nl.
- elearning.surf.nl

Contents

Introduction

To developers of learning materials, filling in a set of metadata is an altruistic deed, from which only others, and not they themselves, will profit. The only benefit for developers is that, if others also dutifully fill in their metadata, they will be able, some time, to benefit from those efforts. Under these circumstances, people will only be willing to provide the barest essentials. Therefore, the set of metadata categories should be as small as possible, if possible multiple-choice lists should be available, and the tool with which metadata are filled in should be as user-friendly as possible. Ease of use is not only related to well-known characteristics such as an attractive user interface. Filling in metadata becomes much more pleasant if data that have been provided during the proposal and definition phase of a project can be re-used as metadata. More generally, if the provision of metadata is integrated into the whole learning object development process, this will improve the quality of the resulting metadata set. Instruments for generating metadata descriptions will have to be integrated into the learning object development process that leads to the storage and management of materials developed. This is the best way to guarantee the integrity of the materials and the metadata attached to them. Integration will also facilitate the development of search engines with which developers will search the learning object repository (LOR).

Large amounts of information can be made available in several ways. For example, in Google-like search engines the user can search simply by typing in words into a free search field. These search engines do have more advanced search options, which allow for the input of specific key words (for example only results in a specific language).

In separating the simple and more advanced search option, several varied strategies can be used, for example a wizard-like dialogue that guides the user step-by-step towards a smaller set of results, or the use of free text fields, or the use of fields containing a check list with a restricted number of options etcetera. Although search strategies are beyond the scope of this quick scan, there is a direct connection to metadata. Metadata quality determines the quality of the search results found. Even Google depends on the quality of its own metadata. Putting effort into a good metadata process and into the metadata quality is one of the most important prerequisites for a successful LOR and for the re-use of learning objects in general.

The Dutch Digital University and learning object metadata

Making metadata go away: hiding everything but the benefits is the title of a paper by Duval & Hodgins (2004) with which I agree whole-heartedly. To me, being a relative newcomer to the discussion on learning object metadata (I am professionally involved since 2000), the discussions sometimes appears to be endless and not oriented towards concrete and usable results. This quick scan aims at being a small step forwards towards concrete solutions. Our aim is not to start the discussion all over again, but rather to guide the Dutch Digital University (DU) and to build on good work that has been done by others.

Attaching metadata to learning objects according to the DU guideline is a complex and time-consuming process. If the number of learning objects to be developed is small, one can still keep track of time investment, yet with larger amounts of objects, attaching metadata will require more and more resources.

The creation of a standard learning object metadata record according to the DU guideline will soon take one hour. Suppose that we wish to describe an existing collection of 1500 learning objects according to the DU guideline, then this soon takes one person-year of work. The conclusion is

justified that extensively manually attaching metadata to large collections of learning objects is a very difficult, if not impossible, scaling process. Yet, metadata are an indispensable element in recording essential characteristics that should make possible the application of and search for learning objects. Alternatives to the manual (human) metadata process are more than desirable. One alternative is to reduce the amount of metadata that has to be put in manually by a human by applying 'smart' software that can generate metadata automatically.

The aim of this quick scan is to investigate whether such 'smart' software exists and whether this software is usable in the Digital University context. We do not aspire to test this software extensively and to come up with a detailed comparison of these products. This quick scan is a short exploration and not the end point. It just displays the opportunities for simplifying the attachment of metadata to learning objects. This quick scan' aim is to contribute to a closer collaboration between the diverse organizations in The Netherlands that are setting up repositories of learning materials and who are struggling with the metadata working process.

The starting point of this quick scan is the DU metadata guideline (Werkgroep DU-metadata richtlijn, 2004), which is based on the Dutch translation of IEEE LOM.

A short analysis

What is automatic generation of learning object metadata about? This section describes some techniques and concepts that play a role in this process.

Metadata extraction and automatic classification

It is important to make a distinction between metadata extraction and automatic classification. These are two different perspectives on automatic metadata generation.

- Extraction
  Based on certain algorithms, information is extracted from documents and subsumed under the relevant metadata fields. Examples include language recognition and key word assignment. Within this approach, workable solutions are available. The technology applied is easily accessible and utilizes common and general concepts from computer technology.
- Classification
  A tool for automatic classification has relevant domain knowledge at its disposal. This domain knowledge is utilized in the analysis of the object at hand, for example a learning object about the heart can be classified using medical domain knowledge. This type of tool results in a domain specific classification. An important observation is that these tools first have to go through a training using a set of documents that together provide a good description of a specific domain. Thus at the beginning of the process, investment is high. Benefits are most clearly present in the case of attaching metadata to a large number of objects. The technology applied originates from research on artificial intelligence.

Who creates the metadata?

Liddy (2005) describes three main scenarios for metadata creation:

1. metadata is produced by metadata specialists, a job position found in library organizations.
2. metadata is produced by the learning object owner/developer.
3. metadata is automatically generated by software tools.

The first scenario is mentioned by Hermans and De Vries (2006). They add to these scenarios the situation in which no metadata have been prescribed, and in which annotations and ratings by users play an important role.

In actual practice, a combination of Liddy's three scenarios is often found. For example, in their metadata guideline, the Digital University recommends a combination of scenarios (1) and (2). Liddy subscribes to the viewpoint that manually attaching metadata is costly and labour-intensive[1].

Usable versus perfect metadata

At several occasions, Huijsen, Grootveld, Brussee, Setten & Porskamp (2005) - report on automatic metadata generation by the Telematica Institute – correctly warns that one cannot rely solely on automatically generated metadata[2]. Everyone knows the famous examples of how to fool Google. For example, a search using the Dutch phrase "raar kapsel" ('funny hair style'; dd. January 10, 2006) will lead one to the homepage of the Dutch prime minister. A critical note is warranted. The goal of perfect metadata generation, which results in the same metadata set in every situation, is an unrealistic goal.

The techniques that we are looking for are being applied in a restricted domain (education) and are being applied to learning objects that are developed by instructional designers to the best of their knowledge. The context is well-defined and the users of the metadata tools are known. In my opinion, this implies that the development requirements for metadata extraction and classification tools are much more easily met than in a situation in which a generic tool has to be built for any object in an unknown context. Relying solely on technology is unwise, yet things do not have to be made more complicated than they are.

The composition of learning objects

As a rule, learning objects include more than one type of material. They are not composed of only text or only images. The average learning object consists of a combination of digital materials, each with its own characteristics and its own possibilities and impossibilities for automatic metadata generation. With automatic metadata generation (extraction and classification) the possibilities offered by the source material are a major concern.

- Textual materials
  Documents mainly consisting of text offer the best opportunities for automatic metadata generation. These opportunities are built on wide experience and expertise. It is important to note that most algorithms have been developed for texts in English. There are not that many good algorithms for the Dutch language. We did not investigate the opportunities for other languages. For the Dutch language, further research is needed in collaboration with a target group that wants to use the results.

- Multimedial materials
  A growing proportion of learning objects are fragments consisting of video, images or sound. With these types of materials, automatic metadata generation is much less developed than with textual materials. Moreover, results are not good enough yet. The generation process proceeds in several steps, for example speech within a video or sound fragment is converted into a text document[3] – currently, this first step works quite well. This semi-finished product is the basis for metadata generation. Algorithms for pattern recognition are used as well. Unfortunately, these techniques are not (yet) available for large-scale application. Metadata is available in most scenario's when learning objects are used in an educational setting. Learning Objects need

[1] Liddy [2005], "However, manual metadata assignments, like cataloging is a labor-intensive and costly function, requiring special knowledge and training."

[2] Huijsen et al.: "This difference between the human readable and separate metadata will not only lead to embarrassing disclosures, it also means, unfortunately, that we cannot rely solely…."(translated from Dutch).

[3] See for example http:///www.blinkx.com

metadata to be of use in real word setting. Good examples are the pictures taken by a digital camera. In most cases, metadata will be automatically attached to the picture (using the EXIF format[4]). Converting a metadata description in format x into format y (from EXIF into LOM) is a relatively simple process. Something similar has been managed in the Digital University project Rechten Online ('Law Online'), in which materials in EML format have been converted into IMS QTI format. These conversion algorithms must be developed separately for specific domains and specific goals.

- Compound materials (complex objects)
  In many cases learning objects are composed of several elements, for example a text, an exercise and some illustrations. Metadata can be attached to each of these components. An example is the complex learning object described in Poortman and Sloep 2006. Clearly much could be gained if the elements within these types of collections (learning object) could inherit the relevant metadata (re-use). During this quick scan I haven't yet come across good solutions. Further research is needed, also because of the possibilities of combining inheritance with the operationalisation of contextual and profile information.

Where is the extractable metadata located?

In a number of papers, Duval and Hodgins describe the possibilities for and the concepts behind metadata extraction from learning objects. The main idea is some sort of self-description which is described in the relevant fields of a LOM record. They consider this an important research domain. Their paper *A LOM Research Agenda* (Duval & Hodgins, 2003) describes the ways in which metadata can be generated (semi-)automatically.

- Metadata extraction from learning objects
  Much of the information that one wishes to store in the metadata is present in the object itself. Several techniques exist for extracting this information from the object, for example algorithms for language recognition and making summaries, pattern recognition with images. With web pages, HTML scraping techniques exist and ActiveX components can be used to extract information from MS Office documents. Office documents contain several metadata elements (author, date, version, etcetera). The metadata are built on profile information that is stored in for example MS Word. One requirement for extraction is that the profile information is correct.

- Metadata inheritance, using metadata of related learning objects
  Learning objects are often part of a larger collection of objects. The most obvious solution is inheritance of shared metadata (re-use of metadata).

- Contextual information and profile information
  Several information systems in institutions, for example online study guides, directory services and HRM systems, contain much relevant information on persons and courses (modules). Making this information operational in the metadata workflow, such that this information can be automatically incorporated into the metadata, can yield profit. The AMG framework gives some examples of how such a network of metadata delivery can operate.

In addition to these descriptive metadata, Duval and Hodgins introduce the concept of "social recommendation" for capturing those data on learning objects that are of special interest[5] to people looking for usable learning objects. These kind of metadata is added after use in a continuous

[4] See http://www.exif.org/

[5] Duval and Hodgins (2004), "we should be able to make use of information about how the learning object helped the user and organization to achieve the goal in affective and efficient way".

process. One very famous example is Amazon[6]. These 'feedback' metadata cannot at all be generated automatically as the metadata reflects personal interpretation and appreciation. Some components of it can be generated automatically. A repository can support these feedback mechanisms, for example by making available profile information (name, background, etcetera) the moment a person logs in; this information can be fed into a feedback form automatically. Process information such as version and data can be generated automatically as well. The AMG framework that is being developed in Leuven is a concrete realization of the proposals that have been made by Duval and Hodgins.

Development in the Netherlands and Belgium

The automatic generation of metadata and classification of learning objects is a topic on the agenda of several organizations in the Netherlands and Belgium. A lot of research on this technology is also being done by researchers in other countries. The quick scan will describe three initiatives in the Netherlands and Belgium because of their potential use and cooperation for the Digital University.

The research done by SURFnet

In 2005 SURFnet commissioned research into categorization software. The study was executed by Quo Vide. The main goal of this study was to improve the quality of the SURFnet search engine by adding software that is able to generate metadata automatically and to classify sources automatically. This study is characterized by a thorough analysis of the market of the available commercial solutions. They observed the fact that commercial solutions are expensive. Yearly costs varied from 50.000 euros to one million euros. Products in this cost range are beyond the reach of average educational institutions.

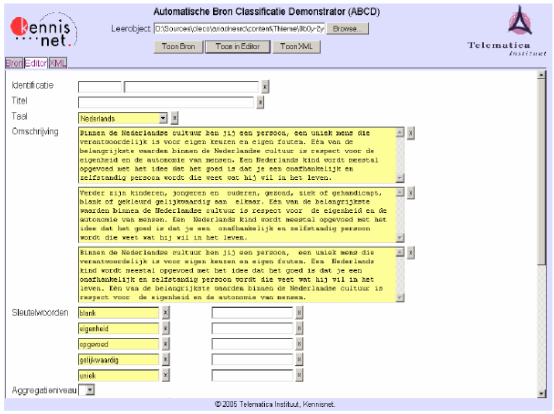
In the final report a detailed description is given of the various technologies that are used and the operation of several algorithms is briefly described. The Telematica Institute took in their research that they carried out for Kennisnet most elements for their domain description directly from the SURFnet research. It is beyond the focus of the quick scan to research this topic in depth, but it is an very interesting topic. The quick scan has a few references in the literature list to the several final reports on the website of SURFnet. The main conclusion of Quo Vide is based on a proof of concept with three vendors, which have been selected from a long list of possible vendors (Appendix 1).

> In view of the results of the proof of concept it is not possible to form a clear picture of the status of the technology of automatic metadata classification at this moment. Unfortunately the proof of concept failed to give a good insight into the quality of the automatic classification solution from various vendors. It has become clear that from an economic point of view automatic classification doesn't provide the necessary surplus value for the SURFnet FAST search engine. The costs for implementation, training and maintenance of the categorization software are probable higher than the possible benefits.
> *Conclusions of the SURFnet research produced by Quo Vide (December 2005) (Translated from Dutch)*

[6] http://www.amazon.com

The ABCD Demonstrator

The Telematica Institute has executed a research for Kennisnet on the possibilities of automatic classification. The scope of this research is discussed in a different part in the quick scan. An important part of the Kennisnet assignment is the development of a demonstrator, the ABCD demonstrator. This demonstrator shows that it is possible with limited resources and time to develop a practicable solution that supports the process of generating metadata. A fully functional automatic solution isn't realized but the tool provided acceptable results based on the concept of suggestion. This concept works with a suggestion that made it possible for metadata entry to be based on a number of predefined metadata values for the relevant metadata fields. The metadata must be verified by a person and changed if needed. The ABCD demonstrator is unfortunately not available for general use and an inquiry with Kennisnet has not delivered an answer yet if Kennisnet will make tool available to other communities for further research and development.



Figuur 13: Automatisch geëxtraheerde metadata in de demo metadata-editor.

Automatic Metadata Generation (AMG)

A research group of the Catholic University of Leuven has been active in the field of automatic metadata generation for several years. They have developed a conceptual framework that describes the different aspects and concepts that play a role in the process of automatic metadata generation.

This metadata framework is the basis for a software solution that is developed in Leuven to actually generate metadata automatically. The metadata is extracted from learning objects using extraction techniques in combination with information about the context in which the learning object is developed and used. This results in a metadata record. On a limited scale algorithms are used that make use of characteristics such as the language in which the object is written. In the AMG framework there is no use of classification algorithms like the ones that are used by the ABCD demonstrator. The AMG software is not tested in this quick scan. A quick test is done with an online provided demo module on the website (http://ariadne.cs.kuleuven.ac.be/amg) of the AMG project. This module offers possibilities for a limited category of objects to generate metadata automatically in several metadata schema's like IEEE LOM and Dublin Core. They have used the software among other things to describe the metadata of learning objects extracted from Blackboard and deliver them with a repository. In the spring of 2006 a completely rewritten version of the software is made available. The software is available for other communities on the AMG website.
The AMG software is promising and deserves more research in a future project.

Alternatives to generating metadata records

Letting developers fill in metadata and describe educational materials is not the only meaningful way to make learning material available. The main goal of this quick scan is not to provide full

analysis of the problem area but to provide a snapshot. A short characterization of some alternatives for a metadata record (description) offers perhaps some clues for further research to alternatives of searching and finding learning objects.

The Google Approach

Metadata as a LOM record is not the only way to make a learning object findable. The technology used by e.g. Google and Yahoo to index objects and allow them to be found is not part of the research in this Quick scan. (e.g. a Google search engine for Learning Objects). In a follow-up study it is advisable to research these possibilities. A future experiment that might attract interest is the option to make the learning objects collected by LOREnet available through scholar.google.com.

Social tagging as an alternative approach

The classical way in which a metadata workflow is organised is characterized by a strong a-priori formalism. All possible actions are written down and elaborated in such a way that a univocal and correct metadata record is generated. In actual practice this workflow is sometimes considered as restrictive and a burden. New developments in the domain of social software offer an alternative for this strong formalism. One still creates metadata but it is based on personal intuition and insights. The philosophy behind such a social network is the fact that one describes one's own objects and in the process of describing makes use of a consensus on the concepts and terminology that is being used. The consensus is laid down by accepted concepts in everyday communication between people. A picture of a dog will be 99 out 100 times being labelled by a tag that has something to do with a dog or a dog species. Metadata in a social network is not in pursuit of 100% accuracy and is not based on formal agreements but is a reflection of what is considered an adequate description by a user group.

Such a collection of tags that evolves in a network of users is called a folksonomy. This is an alternative to the formal taxonomies that are in principal developed by information specialists and library organizations.

Who owns the metadata?

In the discussion about metadata it is often overlooked that correct metadata records represent a commercial value. Metadata is the way to find the objects and metadata is this sense a key factor of a successful repository. Therefore choices have to be made in regard to the ownership and the right to use metadata. 'Who has access to metadata' and 'what is one allowed to do' are important questions. Examples from another practice are the collections of e-mail addresses and profile information as an instrument for direct marketing. A second (Dutch) example is the AH-Bonus card, which registers every item purchased at AH by its owner (AH is the biggest supermarket in the Netherlands). It is important that if we develop a network of reusable learning objects we not only take care of the copyrights on the object but also take care of the rights on the use of the metadata.

Key articles and reports

This quick scan makes use of a number of sources. Four of them are the backbone of the quick scan. They hold quite interesting information for further research.

Huijsen e.a. (2005). Automatische Classificatie Eindrapport, Telematica instituut.

The Telematica Institute has executed a study for the Dutch Kennisnet organisation on the possibilities of automatic classification. This report gives an insight into the options and problems of the present status of the technology of automatic classification. The report is characterized by a thorough and classical approach of the problem area. Several technologies and strategies are

briefly discussed. An important conclusion is the fact that there are only a few suitable algorithms for the Dutch language available. This research is built on the research that is done on behalf of SURFnet on commercial and open-source solutions.

To build an actual demonstrator was another part of the project besides the domain analysis. The demonstrator makes use of the metadata search profile developed by Kennisnet. The demonstrator makes use of elements developed by the work done in Leuven on the AMG framework. The end result is promising and deservers further development. In my opinion is desirable that the DU, SURF and Kennisnet combine resources and expertise in a follow-up. A very useful element of the report is table in which the options of each metadata field are discussed. The metadata schema of Kennisnet is similar to the schema used by the DU.

A critical comment is in my opinion the conservative approach to the problem area. The classical approach and thoroughness in which one deals with the metadata process is found in the way that an innovative approach are negatively looked upon and that the weak points of the alternatives are pointed out instead of the positive contributions that are possible.

Duval & Hodgins (2003). A LOM Research Agenda
Erik Duval and Wayne Hodgins are veterans in the development of the LOM, the learning object metadata standard. A development that started in the mid 90's and which results are used more and more by a growing community. The LOM is considered a successful standard for applying metadata to learning objects. Duval and Hodgins do not consider the LOM as a goal in itself but as a first step towards serious re-use of learning materials. It is not a exiting step but a necessary one. In this research paper sixteen topics are covered. Each topic is a possible research topic. The objective is to mark the next phase in the process of re-use of learning materials. The potential research topics range from the development of suitable taxonomies to business models. Topic number 8 covers the domain of automatic metadata generation. In my opinion, this article shows the way to the kind of discussion that is essential to the implementation of learning object metadata. This article has many interesting ideas and concepts that cry out for further research and discussion. A point of critique is the fact that focus is mostly on the early adopters and that it takes huge steps through the arena of learning objects.

Vandoolaeghe & Van Isterdael (2005). Metadata voor leerobjecten in een digitale leeromgeving
This article (in Dutch) is an excellent description of the domain of metadata for learning objects. A number of relevant discussions, open ends and other metadata related issues have found their way to this article. All elements are discussed in a clear way. In short it is a very useful introduction for everybody that wants to know more about metadata for learning objects and for those who are looking for perspective on recent developments and it shows the relevant relations between several developments.

Cardinaels, Meire & Duval (2005). Automating Metadata Generation: The Simple Indexing Interface
This article provides a good description of the possibilities of the AMG Framework that is developed in Leuven. The article presents an interesting case study, which shows one possible way of describing the material that is stored in the local Blackboard implementation with the help of the AMG framework. This article describes in short the essential elements and techniques that are a part of the AMG framework. The article is a good starting point for further research into the AMG tools in a follow-up project.

Recommendations for a follow-up

A quick scan doesn't provided a detailed plan for the future, a quick scan provides some assistance in developing new projects and for a follow-up research.

What are we looking for?

In a number of studies a case is made for leaving the royal road and sacrificing the demand for completeness for a pursuit of practicability at a reasonable cost. The advice is not to aim at perfection but to find the balance between extensive use of software techniques and the effort that people put into the classification of learning objects.

Fully automated metadata extracting and classification is at this moment not realistic but the support of the existing metadata workflow is achievable. This support will on the first place be an optimization of the workflow, e.g. smart use of templates and secondly to use software tools that will provide a suggestion for some of the metadata fields of the compulsory metadata fields of DU metadata set.

The AMG framework of Leuven deserves further research, e.g. by performing tests with available learning objects from the DU collection. The AMG framework is an important candidate for the metadata generation in the DU development projects. This might result in a network of (web-) services that DU member institutions could use in their metadata workflow.

A cautious step for a common approach

The DU project VAMP is a follow-up which will use the results presented in this Quick scan. In April 2006 a workshop is organised in which several stakeholders from the Dutch higher education landscape were represented. The goal of the workshop was to establish a common approach between the different stakeholders and to develop in cooperation a toolset for the automatic generation of metadata. The university of Leuven was also present at the workshop. This university offered their cooperation in developing the toolset based on the AGM framework which was developed at the university in Leuven.

Folksonomies

Metadata as a LOM record is not the only way to make a learning object findable. The technology used by e.g. Google and Yahoo to index and allow objects to be find is not part of the research in this Quick scan. (e.g. a Google search engine for Learning Objects). In follow-up studies it is advisable to look at these approaches in searching for learning objects. The DU project Social Networking with Learning Objects will investigate in which role folksonomies could play in getting more positive results when looking for learning objects

Who owns the metadata?

Investigate the choices that are an option for the DU in relation with the ownership and copyright of the metadata of learning objects. This metadata represents an economic asset. A discussion on these issues could have a positive effect on the use and re-use of learning objects. It is important that the DU takes a clear position on these issues in respect to their products. Who owns the metadata and which are the limitations on the use of the metadata?

References

Cited works

Cardinaels, K., Meire, M., & Duval, E. (2005). Automatic metadata generation: the simple indexing interface. *Proceedings of the 14th international conference on World Wide Web, May 10-14, 2005, Chiba, Japan* (pp. 548 – 556). New York: ACM Press.

Deken, J.J.E., & Wijland, M.W.P.J. van (2005, 18 augustus). *Resultaat Onderzoek naar Categorisatie Software voor SURFnet; Samenvatting – Rapport* [Results of the study into categorisation software for SURFnet: Summary] (JD/2005/SURFNET/003). Noord-Scharwoude, The Netherlands: Quo Vide. Available at http://www.surfnet.nl/publicaties/surfworks2005/indi-2005-008-12.pdf.

Deken, J.J.E., & Wijland, M.W.P.J. van (2005, 20 december). *Resultaat Proof-of-concept Automatische classificatie SURFnet; publieke Samenvatting* [Results proof-of-concept automatic classification SURFnet: public summary] (JD/2005/SURFNET/007). Alkmaar, The Netherlands: Quo Vide. Available at http://www.surfnet.nl/publicaties/surfworks2005/indi-2005-012-30.pdf.

Duval, E., & Hodgins, W. (2004). Making metadata go away: “Hiding everything but the benefits”. In W. Jianzhong (Ed.), *DC-2004: Proceedings of the International Conference on Dublin Core and Metadata Applications* (pp. 29-35).

Duval, E., & Hodgins, W. (2003). A LOM Research Agenda. *Proceedings WWW2003, May 2003 Budapest.*

Hermans, H., & Vries, F. de, (2006). *Organizational scenarios for the use of learning objects* (Learning objects in practice 2). Utrecht, The Netherlands: Stichting Digitale Universiteit. Available at http://www.du.nl/leerobjecten.

Huijsen, W., Grootveld, M., Brussee, R., Setten, M. van, & Porskamp, P. (2005, december). *Automatische Classificatie; eindrapport* [Automatic classification; final report]. Enschede, The Netherlands: Telematica Instituut.

Liddy, E. (2005). Metadata: A promising Solution. *Educause review May/June 2005*.

Poortman, S., & Sloep, P. (2006). Educational models: A case study into transferability of didactical structure in a complex learning object (Learning objects in practice 3). Utrecht, The Netherlands: Stichting Digitale Universiteit. Available at http://www.du.nl/leerobjecten.

Vandoolaeghe, F., & Van Isterdael. W. (2005). *Metadata voor leerobjecten in een digitale leeromgeving* [Metadata for learning objects in a digital learning environment]. Universiteit Antwerpen.

Werkgroep DU-metadata richtlijn (2004). *Werken met metadata in DU-projecten* [Using metadata in DU-projects](Deel 1: Handleiding en Deel 2: Bijlagen [Part 1: Manual and Part 2: Appendices]) (Versie 1.1, kenmerk ELO.DEL.2300/2301). Utrecht: Stichting Digitale Universiteit. Available at http://www.du.nl/digiuni//download/temp/ELO.DEL.2300.werkenmetmetadatainDUprojectenhandleiding.pdf and http://www.du.nl/digiuni//download/temp/ELO.DEL.2301.werkenmetmetadatainDUprojectenbijlagen.pdf.

Yilmazel, O., Finneran, C.M., & Liddy, E.D. (2004). Metaextract: An NLP System to Automatically Assign Metadata. *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries* (pp. 241-242).

Further reading

Cardinaels, K., Duval, E., en Olivié, H. (2002). Issues in Automatic Learning Object Indexation, In P. Barker & S. Rebelsky (Eds.), *Proceedings of ED-MEDIA World Conference on Educational Multimedia, Hypermedia & Telecommunications* (pp. 239-240).

Greenberg, J. (2004). Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications. *Journal of Internet Cataloging, 6*(4), 59-82.

Liddy, E.D., Allen, E., Harwell, S., Corieri, S., Yilmazel, O., Ercan Ozgencil, N., et al. (2002). Automatic Metadata Generation & Evaluation. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, August 2002, Tampere, Finland* (pp. 401-402). New York: ACM Press.

Metros, S.E. (2005, July/August). Learning Objects: A rose by Any Other Name …. *EDUCAUSE Review, 40*(4), 12–13. Available at http://www.educause.edu/apps/er/erm05/erm05410.asp.

Ochoa, X., Cardinaels, K., Meire, M., & Duval, E. (2005). Frameworks for the Automatic Indexation of Learning Management Systems Content into Learning Object Repositories. *Proceedings of EDMEDIA 2005, World Conference on Educational Multimedia, Hypermedia & Telecommunications, Montreal, Canada* (pp. 1407-1414). Chesapeake, VA: AACE. Available at http://ariadne.cs.kuleuven.ac.be/amg/publicationsFiles/paperAMG2.doc.

Patton, M., Reynolds, D., Choudhury, G. S., & DiLauro, T. (2004, November). Toward a Metadata Generation Framework: A case study at Johns Hopkins University. *D-LIB magazine, 10*(11).

Weibel, S.L. (2005, July/August). Border Crossings: Reflections on a Decade of Metadata Consensus Building. *D-LIB magazine, 11*(7/8).

Links

Automatic metadata generation. Website. Available at http://ariadne.cs.kuleuven.ac.be/amg.

Appendix 1: List of vendors in the SURFnet study

Besides a literature study Quo Vide sent out a Request For Information (RFI) to the following vendors:

1. Teragram Categorizer
2. Inxight Categorizer
3. Wordmap
4. Nstein
5. Autonomy
6. ClearForest
7. Convera
8. Verity K2
9. Triple Hop Technologies
10. Engenium
11. Data Harmony
12. Entrieva (voorheen Semio)
13. SmartLogik
14. Temis
15. Kofax - Mohomine Classifier
16. Recommind
17. yellow brix
18. Entopia
19. Collexis
20. Irion
21. Interwoven

Source: Deken and Wijland (2005).

Appendix 2: A list of several open source software tools and algorithms for the generation of metadata

An important set of open source software tools is developed by the University of California. The toolset with the name IVia is to be found on the website: http://ivia.ucr.edu/. The AMG framework is among the users of this tool. The metadata extraction tool is capable of extracting metadata for the following fields of a metadata record:
(The algorithms are limited to the English language)

- Titles
- Descriptions
- Keyphrases
- INFOMINE Categories (requires model)
- Language
- Media Type (i.e. MIME Type)
- Creator, contributor, and publisher
- Library of Congress Classification (requires model)
- LCSH
- LCC outlines

The ABCD-demontrator uses several open source solutions besides the one that are provided by the AMG framework.

Language recognition is often based on the NGram algorithm. An open source Java-implementation of this algorithm is available on: http://sourceforge.net/projects/ngramj

To generate a summary the ABCD-demonstaror uses a software program written by Mark Watson, KBTextmaster, http://www.markwatson.com/opensource/
The program is changed to fit the needs of the Kennisnet project.

One tool to extract metadata that is not yet mentioned is MetaExtract of Syracuse University (Elizabeth Liddy). This tool is based on principles developed in the domain of natural language processing (NLP). An article by Yilmazel and others on MetaExtract (Yilmazel, Finneran & Liddy, 2004) briefly describes the possibilities of this approach. During the Quick scan I didn't succeed in finding additional information on MetaExtract.