Towards a Corporate Semantic Web Approach in Designing Learning Systems. Review of the TRIAL SOLUTION Project

Michel BUFFA*, Sylvain DEHORS**, Catherine FARON-ZUCKER*, Peter SANDER*
**MAINLINE, I3S, UNSA*
930 route des Colles, bât ESSI, BP 145, 06930 Sophia Antipolis cedex, France
{buffa,faron,sander}@essi.fr
***ACACIA, INRIA Sophia Antipolis*
2004 route des Lucioles, BP 93, 06902 Sophia Antipolis cedex, France
Sylvain.Dehors@sophia.inria.fr

Abstract. The TRIAL SOLUTION EU project focused on the publication of personalized electronic documents based on existing scientific books. Its general approach consists in slicing electronic books into elementary learning resources and annotating them with metadata enabling the retrieval of resources by a semantic search. The annotated resources are published into a repository available for teachers or students to produce personalised teaching or learning materials with delivery tools. In this paper we give an overview of the project, emphasizing the authoring tool we have developed to annotate the learning resources, and we review it by the light of the Semantic Web.

Introduction

TRIAL SOLUTION EU project [1] focuses on the publication of personalized electronic documents based on existing scientific books. Designing new electronic documents from scratch is very expensive and print-oriented authors rarely possess the required competency to perform such a specialized task. On the other hand, most recent regular printed books have a digital format, e.g. Latex or Microsoft Word, that can help in automating the publication of personalised electronic documents from printed books. The benefits of the approach based on re-engineering existing materials have been described in [1] and the process itself is detailed in [2]. This was the starting point of the TRIAL SOLUTION project.

The general approach consists in slicing electronic books into elementary learning resources and re-engineering these resources by refining the slices and annotating the resources with metadata on : content, didactic features and interoperability interfacing; this in order to enable intelligent retrieval. The annotated resources are then published into an online repository available for teachers and students to produce personalised documents, or just to find relevant learning materials.

The TRIAL SOLUTION platform integrates three main services: the automatic extraction and annotation of learning resources from electronic books, the re-engineering of the repository of learning resources and the retrieval of learning resources based on their annotations. Our contribution to the project is the authoring tool for re-engineering the learning resources by improving on their initial slicing and by adding metadata to them. In

[1] TRIAL SOLUTION stands for *Tools for Reusable, Integrated, Adaptable Learning - Systems/standards for Open Learning Using Tested, Interoperable Objects and Networking* It was funded by the EU as part of the IST Program within the EU's Fifth RTD Framework Program. http://www.trial-solution.de/

the next section we will give more details on the book enhancement and learning resource extraction and annotation processes. In section 2 we then review the TRIAL SOLUTION project by emphasizing the similarity of its approach with Semantic Web approaches of automatic annotation of learning resources based on ontologies.

1. Extraction and Annotation of Learning Resources: the Trial Solution Approach

1.1. Automatic Extraction of Learning Resources from Books

The first step of the TRIAL SOLUTION process consists in automatically disaggregating a textual document into a set of slices, using a tool from the SIT enterprise[2] that works with structured format documents such as LaTeX or well-styled Word documents. It extracts an XML table of contents where each entry corresponds to a slice, i.e. a learning resource. Details on the DTD can be found on the project site. Sections, chapters, tables, figures, examples are some of the easiest slices that the tool can identify automatically, and course documents using a standard style can be sliced more efficiently.

The extracted learning resources are automatically annotated with metadata about the author of the original book, the semantic content of the resources, and the relationships between resources in the original book. Hyperlinks or *"see also"* sentences are used to determine relationships while keywords help in classifying some resource contents. The Splitter looks for keywords and sentences specified in a thesaurus for the book's domain. It is based on a Thesaurus Management tool and a Key Phrase Assignment tool. The first tool checks for consistency of the thesaurus and the second one manages a collection of mathematical key phrases, extracted from a standard textbook on mathematics [3].

However, this automatic process is almost always insufficient and may even produce some wrong guesses, as stated in [2]. The extracted learning resources have to be re-engineered using our authoring tool.

1.2. Re-engineering Learning Resources

We have developed the TRIAL SOLUTION re-engineering tool to allow re-engineers to fix and improve the automatically-produced base of annotated and interconnected learning resources. Experiences on book enhancement using this tool are described in [2]. Re-engineers must have good knowledge of the original book content, and are either the authors themselves or (more often) teachers using the book as course material.

Our re-engineering tool is a client of the learning resource server. It talks to the server using a custom protocol where both the requests and the values returned by the server are based on an XML encoding defined in the TRIAL SOLUTION DTD.

When connected to a repository of learning resources automatically extracted from electronic books, our re-engineering tool enables (1) assigning the resources a title, (2) editing their contents, (3) editing the tree structure of the whole repository: resources can be split, merged, deleted; sub-resources can be created or reorganized, and (4) editing the metadata associated to the resources. We distinguish between three kinds of metadata. *Types* represent the pedagogical role of the resource contents, eg, a definition, a theorem, etc. *Keywords* specify the topics that the resource contents address. *Relations* with other resources like "*references*", "*requires*", etc. build up a semantic network between resources.

Our re-engineering tool also enables editing the thesaurus upon which the metadata are built. The thesaurus can hold various relationships between keywords: subsumption, synonymy and relatedness, and a description can be specified for each keyword.

---

[2] Infotech Slicing Technology (SIT) GmbH, http://www.slicing-infotech.de/de/index.php

2.1. Related Work

A similar automatic slicing approach has been applied in the SeLeNe [4] project where documents follow the DocBook DTD. In the domain of Intelligent Tutoring Systems (ITS) and Adaptive Hypermedia Systems (AHS), where using small bits of educational content is a very common practice, the more-common approach is rather to author the content specifically through authoring tools [5].

Other projects like Ariadne [6] or UBP [7] have proposed "learning object repositories" (LOR) where resources are stored and annotated with standard metadata but no assumption is made on the origin of the resources, especially whether they have been authored on purpose or come from existing documents. The OLR [8] repository uses RDF for storing the annotations and takes advantage of the chapter structure of the global course to inherit annotations. All these LORs are provided with annotation tools that help filling in the metadata.

To sum up, the TRIAL SOLUTION re-engineering tool adopts the same strategy as LORs annotation tools, but the specificity of our approach relies in the coupling of a manual annotation process and a preliminary automatic phase of resource extraction and annotation.

2. Review of Trial Solution

2.2. Web Standards

The Trial Solution project is compliant with some major open standards such as the IMS Content Packaging[3] for modelling objects that have to be exchanged between the different tools of the platform.

Metadata for describing the resources are compliant with the Dublin Core Metadata[4], the IMS Learning Resource Metadata[5] and the LOM Learning Object Metadata[6] standards, but deviate when forced to deal with new type of material [9]. During the project, it was necessary to develop our own metadata specification due to shortcomings in the then-available standard specifications.

However the technology has evolved considerably since the design stage of the project. Web services have come into their own and would now be employed for the client-server exchange of complex objects. The semantic web languages RDF(S) and OWL would now best suit the expression of the learning resource metadata.

2.2. Towards a Corporate Semantic Web Approach

TRIAL SOLUTION fits right within the scope of the semantic web. We have built a repository of learning resources which are meant to be shared and reused. They can be efficiently retrieved by means of semantic annotations about their domains, their pedagogical roles and the structural and semantic relationships between them. At the end of the project we argue for a homogenous representation of all these metadata in RDF annotations based on ontologies of concepts and relations about the resource structure, domain, and author's pedagogy. In addition to homogeneity for sharing and reusing, this language provides a much richer expressivity than simple keywords for describing resources.

---

[3] IMS Content Packaging Specification: http://www.imsproject.org/content/packaging/index.html
[4] Dublin Core Metadata Element Set, Version 1.1: http://dublincore.org/documents/dces/
[5] IMS Learning Resource Metadata: http://www.imsglobal.org/metadata/index.html
[6] IEEE Standard for Learning Object Metadata (LOM): http://ltsc.ieee.org/wg12/

Another conclusion we draw from TRIAL SOLUTION is about the population and annotation of a learning resource repository. The experience convinced us that manual annotation of resources is overwhelming for teachers or other scientists when facing a large amount of resources. We aim at automating this process as much as possible; the extraction of knowledge from structured format documents is quite conclusive and could be improved. We tend towards an approach of knowledge extraction from both textual documents and their authors. The acquisition, of the relationships between the author's pedagogical organisation and the layout of the document, will help an automatic interpretation of the layout. By doing so, the annotations about pedagogical features will be much enriched. Here we speak for minimizing domain dependent annotation in advantage of pedagogical annotation. This approach is quite feasible within learning resource repositories about specific domains: in this case some information about the domain becomes implicit.

Finally, in the TRIAL SOLUTION project, the end-user goals were quite general, which made the annotation task difficult: when composing a personalized document, he/she may search for any resources for anything at all. More specific learning scenarii and profiles should improve the adequacy between the annotation contents and the end-user requests.

Conclusion

The Semantic Web was only emerging at the design stage of the TRIAL SOLUTION project. However the TRIAL SOLUTION approach fits right within the scope of the semantic web. We built a repository of annotated learning resources for sharing and reusing. The resources are annotated based on an ontology, with knowledge about their contents, pedagogical features and relationships among them.

At the end of the project, we turn towards a corporate semantic web approach in designing learning systems, where (1) repositories are domain dependent, (2) the document authors are involved in the learning resource acquisition process and (3) user profiles and pedagogical scenarii are taken into account in the annotation contents. We are currently involved in the design of a corporate semantic web for learning experiencing this approach.

References

[1] I.Dahn. Slicing Book Technology – Providing Online Support for Textbooks, in Proc. of the International Conference on Distant Education (ICDE'2001), Dusseldorf, Germany, 2001.

[2] I. Dahn, M. Armbruster, U. Furbach, G. Schwabe. Slicing Books – The Authors' Perspective, in R. Bromme, E. Stahl (eds.): *Writing Hypertext and Learning: Conceptual and Empirical Approaches*, Pergamon, 2001.

[3] Bronstein, Semendjajew, Musiol, Mühlig, *Taschenbuch der Mathematik,* Verlag Harri Deutsch, 2000.

[4] P.Rigaux and N.Spyratos, Metada Management and Learning Object Composition in a Self e-Learning Network, in Proc. of Workshop on Information Search, Integration and Personalization, Japan, 2003.

[5] Brusilovsky, P. Developing adaptive educational hypermedia systems: From design models to authoring tools. In T. Murray, S. Blessing and S. Ainsworth (eds.): *Authoring Tools for Advanced Technology Learning Environment*. Dordrecht: Kluwer Academic Publishers, pp. 377-409, 2003.

[6] E. Duval, E. Forte, K. Cardinaels, B. Verhoeven, R. Van Durm, K. Hendrikx, M. Wentland-Forte, N. Ebel, M. Macowicz, K. Warkentyne, and F. Haenni, The ARIADNE Knowledge Pool System, in Communications of the ACM 44 (2001), no. 5, pp. 73-78.

[7] Simon B., Quemada J., A Reflection of Metadata Standards based on Reference Scenarios, 2002, http://www.wu-wien.ac.at/usr/wi/bsimon/publikationen/SimonQuemada-ReflectionOnMetadataStandards.pdf

[8] J. Brase, W. Nejdl, Ontologies and Metadata for eLearning, Handbook on Ontologies, pp. 579-598, Springer Verlag, 2003.

[9] W. Lenski, E. Wette-Roch. The Trial-Solution Approach to Document Re-use, in Proc. of Workshop on Electronic Media in Mathematics, Coimbra, Portugal, 2001.