Online Personalization of Textbook Material – Results of the European Project Trial-Solution

Ingo Dahn[1]

University Koblenz-Landau, Knowledge Media Institute, Universitätsstr. 1. D-56070 Koblenz, dahn@uni-koblenz.de

Abstract: The European project Trial-Solution (see http://www.trial-solution.de) applies Slicing Book Technology to provide online personalization services for teaching materials extracted from pre-existing textbooks. The talk will present the tools and the workflow established in the project for the creation and usage of a library of more than 25,000 learning objects. The roles of metadata, open standards and artificial intelligence technology in the project will be highlighted.

Introduction

Usage of documents has a long tradition in human history. Documents are the standard form in which knowledge is communicated among humans. When knowledge is requested, it is everyday practice that humans go to a library or a web search service and retrieve a document – and only then they start retrieving the knowledge they need from this document. Humans have developed beautiful designs and sophisticated techniques like typesetting to ease working with documents.

Computers can proceed differently. They do not have a sense for the beauty of a book. Simple and clear structures are much more helpful for them, even if this leads to a much larger mass of data which they have to process. They can retrieve the knowledge they need from databases or from raw experimental data. They can also quickly scan thousands of web pages and extract information from these pages – provided this information has been encoded in a computer friendly form!

Basic Technology

The central idea behind the Trial-Solution project is to transform existing documents as much as is necessary to support the user in assembling the information from these documents into new documents ready-made for the actual purpose. Thus the project supports a transition from a document-centric culture to an information-centric culture. This transition poses immediately two challenges.

1. Computers cannot understand the information in existing documents
2. The information in existing documents makes sense only in its document context.

The Trial-Solution project handles these problems with Slicing Book Technology. This technology consists of the following steps (see Dahn, 2001)

1. Decompose existing books into semantic units
2. Add a knowledge base of meta data
3. Design an intelligent advisory system that uses these meta data
4. Compose personalized documents tailored to the learner‘s current needs – on the fly

Slicing Book Technology was presented at the LearnTec Conference 2000 (Dahn, 2000) and at the AERA Conference 2000 (Dahn, 2001). The first book using this technology (Wolter & Dahn 2000) appeared in 2000 using the SIT-Reader – a server software from the company Slicing Information Technology in Berlin (http://www.slicing-infotech.de).

The European project Trial-Solution was launched in February 2000 and completed in April 2003. It investigates the potential of Slicing Book Technology for the combination of semantic units from different sources.

Working with pre-existing content is cost-efficient. Universities and publishers have a lot of high quality documents that can be re-used. The Trial-Solution project has found that the necessary re-engineering for preparing the documents for new online personalization services can be done without involving the authors, except for discussing the final result or the handling of errors that re-engineers may find in the books. Within the Trial-Solution project most of the re-engineering work was done at TU Chemnitz by students who had mastered the subject, supervised by academic staff.

Selecting the Domain

The project selected undergraduate mathematics as its knowledge domain. One reason for this selection was the fact that undergraduate mathematical knowledge is widely used in many disciplines, notably in engineering and in the natural sciences. Another reason was, that mathematical documents tend to have a clear structure that would be easy to analyze.

Note that this analysis had the objective to isolate semantic units and should not be spoiled by layout issues.

[1] Work funded in part by the European Commission under Contract IST-1999-11397

It is only over the last five years – in connection with the hype of XML as a metalingua franca for the exchange of data – that it became generally acknowledged that a clear separation of content and presentation helps to serve humans as well as computers in making use of the available information. We hasten to mention that this is not quite correct. Mathematicians have partly separated content and presentation already since around 1990 when they started to use broadly LaTeX to write their journal papers and books. This was possible due to a high degree of standardization of the language of mathematical formulas. Thus Mathematicians and publishers of Mathematical texts have gained a lot of experience in managing their knowledge in a way that is equally friendly to computers and humans. This emphasizes that studying the properties of mathematical documents can lead to methods and guidelines which are of high interest also for other fields.

In fact Slicing Book Technology as the basis of the project is not specific for mathematics. It expects only well-structured documents which occur in many fields. Besides good teaching materials we mention here legal and technical documents as examples.

A disadvantage – from the point of view of the project – of using LaTeX is that it gives authors a large set of tools to shape their documents, in fact authors and editors have a complete programming language at their disposal. In the numerous documents handled in the project we found a large variety of applications of these tools. Therefore, considerable efforts had to be made to develop the Trial-Solution Tools to be so flexible that all these varieties could be handled.

Another disadvantage of the use of undergraduate mathematics as the subject is that this is usually taught to beginner students. It turned out, that these students are not yet familiar with the use of information as raw material from which new knowledge can be gained. In fact it is an important educational objective to achieve such a creative and flexible use of information and to overcome the beginner's attitude which expects to find everything in a single place which has to be provided by the professor.

The techniques and tools developed in the project are well suited to achieve this educational objective, but their usage has to be explained to beginner students.

The Trial-Solution Tool Suite

In order to demonstrate cost efficiency and to handle 19 pre-existing documents with altogether more than 5,000 pages, the project had to establish a workflow leading smoothly from the original document to the personalized documents which are delivered to the end user. Also the tools for this workflow had to be developed within the project.

The Trial-Solution Tool Suite was designed in a modular way. The system is composed of five modules which communicate only through narrow, well-defined interfaces. We emphasize that this communication is performed by exchanging data and metadata in an open XML-specified format. The tools developed in the project are

- The Splitter
- The Metadata Tool
- The Metadata Server
- The Authoring Tool
- The Delivery Tool

The workflow developed in the Trial-Solution project consists of four steps, each of which is supported by a specific tool.

In the first step a document is automatically decomposed into semantic units. This process, called slicing, also adds basic metadata to the slices. The tool developed and used for this process is the Splitter, developed by the partner Slicing Information Technology. The Splitter is configured manually to understand the intended slicing granularity and the way in which the author of a particular book uses phrases or formatting to express the content structure, content types or key phrases. Then the Splitter creates automatically the basic package structures that are further transformed by the other tools. These packages consist of a collection of files containing the content of the slices, an XML manifest describing the structure of the sliced document and the metadata assigned to the slices and another XML document describing the thesaurus to be used for this book.

Once a thesaurus for a certain topic is available, like that produced in the project, it can easily be joined with the thesaurus produced by the Splitter in order to support a more controlled work in the subsequent steps of the workflow.

The next step assigns automatically key phrases to (groups of) slices. The software used for this step has been developed by the Dutch partner CWI. It analyses the content of the slices, adds new key phrases to the thesaurus and adds pointers from slices to the assigned key phrases to the manifest XML file.

This revised package is imported into the Trial-Solution Reengineering Tool. Its purpose is to revise the structure of the sliced book and the metadata assigned in it. It supports also the extension of the thesaurus by adding new key phrases.

Unlike the previous tools (the Splitter and the Automated Key Phrase Assignment System) the Reengineering Tool is intended to be used by domain experts with little technical expertise. Therefore it is web based and has a graphical user interface.

In fact, within the Trial-Solution project two versions of the Reengineering Tool have been in use. The first,

known as “Koblenzer Buchmacher”, was developed by the author and existed in large parts before the project. It has been only slightly extended within the project to allow concurrent work of multiple users and to support simultaneous operations on groups of slices. Experience with this tool has been also used to draft the specification of the official Trial-Solution Reengineering Tool which then has been implemented at the University Koblenz-Landau and at the University of Nice Sophia Antipolis. This new tool is java based, has an improved graphical user interface and uses a Postgres database for safe storage and retrieval of data and metadata.

These Reengineering Tools, hosted in Koblenz, have been used mostly by a group of students at TU Chemnitz who did the bulk of the re-engineering work supervised by staff personnel.

The fourth step, after the manual revision of the document structure and of the assigned metadata, is the integration of the newly re-engineered book into the project library. The most essential part in this step is the integration of the thesaurus of the book with the thesaurus used by the books already in the library.

Since the slicing book re-engineer is working with very fine grained material, it is frequently necessary to define new key phrases. Also the author of the original work has often already assigned key phrases which are obviously a valuable good to be preserved. In this process it is inevitable that key phrases are duplicated or that they appear in slight linguistic variations. This is achieved by a tool for the maintenance of the project thesaurus. This tool, also known as the Trial-Solution Metadata Server, has been developed by Fachinformationszentrum Karlsruhe and Trinity College Dublin. It has automated components, for example to detect duplicates and linguistic variants, and interactive components, for example to replace key phrases by more standard conformant versions.

The thesaurus has by the end of the Trial-Solution Project almost 16,000 entries for the field of mathematics. Some of these entries are bilingual in German and English. The Trial-Solution Thesaurus is considered as an important result of the project.

The only tool of concern for the end user is the Delivery Tool. This tool is a web server that delivers personalized teaching materials for the field of undergraduate mathematics. Personalization here means the selection of content slices and formatting according to the current needs of the user. It does not include changes to the content which could be seen as a violation of the author’s rights. Delivery is done online on the fly within seconds directly to the reader’s desktop. Changing requests and modifying delivered documents is easy. The typographic quality of the delivered documents meets current publishing standards as far as this can be achieved in a fully automatic way. Personal annotations can be added to parts of documents. These annotations as well as personalized compilations of content can be shared with others. Each user can compose and store various collections of content on the server. He/she can rearrange and administer these collections.

The most innovative feature of the system is the automated advice the user gets in order to compose meaningful documents that take the intended purpose and the user’s knowledge into account. So, for example, the user can select topics or pieces of content and request to collect the information that is necessary to solve a particular exercise or to prepare for an exam on this content. The system can on request deliver related examples or applications, even if they are contained in other books of the Trial-Solution Library. The content which is necessary to understand a particular part can be added automatically.

The Automated Knowledge Management inside the Delivery Tool uses an automated inferencing system developed at the Artificial Intelligence Group of the University Koblenz-Landau. This system applies declarative rules defining the intended documents to metadata on content and to information about the user’s knowledge and interests in order to recommend content slices for delivery.

With mouse click the user can inform the Trial-Solution Delivery Tool that he knows a particular piece of content. Then the Automated Knowledge Management will automatically complete and adapt the internal user model. This will prevent the system from delivering to the user content which is already known, even from parts which he/she never has inspected. For a more refined adaptation of the user model, involving the human tutors of the student, the system has been equipped with an experimental open interface with an external assessment system. This has been tested in connection with the widely used Learning Management System WebCT.

Investigation of user needs has shown that users mostly prefer reading from paper with respect to reading from screen, however also appreciate the availability of electronic versions for rapid access, retrieval and searching. The Trial-Solution Delivery Tool supports both. It keeps the original formatting that the author has used for the original printed book as much as possible and enables the reader to print the (frequently small) delivered documents for reading. Even further, an experimental open connection to a print on demand system has been built into the system and used in practice during the evaluation phase.

Of course, the Trial-Solution Delivery Tool has advanced search capabilities. It offers a full text search as well as a search for key phrases. Complex search requests can be handled using so-called regular expressions. When multiple matching key phrases are found, assistance is offered to restrict the search further. All types of search can be restricted to specific types of content. So it is, for example, possible to search (say in preparation for an exam) for exercises on a particular topic, add automatically the content leading from already know facts to these exercises and then to select

from this collection only the included examples for print in order to get a quick and concrete survey of what is in the books for the exam. This requires altogether 5 mouse clicks in the Trial-Solution Delivery Tool.

The Trial-Solution Delivery Tool has been developed into a stable system that has been successfully used also with large real-life audiences. Nevertheless it has a prototypical character and a number of possibilities for improvements. These concern especially the reactivity of the system and the clarity of the user interface. The lessons learned from this development are now considered in the further development of the Reader software from Slicing Information Technology.

Some Methodological Considerations

The project has devoted a great deal of attention to the design of its metadata system. In a first step existing metadata specifications like IEEE LOM, ADL/SCORM and IMS have been considered. These were found not to be fully adequate for the needs of the project.

Special problems in this respect have occurred with the handling of key phrases which play an important role in the project. The aforementioned specifications allow key phrases only as strings in a concrete language. These strings are directly written into the metadata assigned to a learning object. This approach makes it hard to maintain the thesaurus and the key phrase assignment.

To overcome these limitations, this approach has been changed as follows. The thesaurus is stored in a separate XML file. When a key phrase is assigned to a slice, only a reference pointing to the key phrase in the thesaurus is stored with the slice. In this way, changes in the thesaurus have to be made only in the thesaurus file. Then they are automatically valid at all places where the respective key phrases have been assigned.

A second advantage of this approach is the possibility to overlay a structure to the thesaurus and to consider related key phrases, synonyms, key phrases defining subconcepts and to identify instances of the same concept as key phrases in different languages.

Another new feature that was introduced in the project for the handling of metadata is metadata inheritance. This means, that metadata can be assigned to groups of slices, for example to all slices of a section. Then these metadata are in fact assigned to the item of the organization of the sliced book which represents this group and they are inherited downwards as if they have been assigned to all slices in this group. This is a very efficient way of metadata assignment.

In the design phase of the project it was planned to allow also upward inheritance of metadata, i.e. inheritance of metadata from individual slices to their sections, chapters, etc. This upward inheritance information turned out to be inefficient to process. Moreover there have been only very few uses of this. Therefore this possibility was cancelled in a later stage.

Allowing downward metadata inheritance has the consequence that in order to see all the metadata that are valid for a particular slice all the metadata attached to its "parents" (section, chapter, document) must be considered too. To guarantee this, the project has introduced the principle that whenever a slice is exported from a sliced document, all metadata of its parents must be exported too. In this way, for example, the rights-relevant information about the author of a book must be kept only at the node representing the complete book and it will be automatically applied to all slices of the book. Even further – this approach makes it possible to respect the author's rights even if personalized sliced books are used in a further step as raw material for the composition of further books together with slices from other sliced books. The project has discussed this in detail and found that this provides all information to respect the rights of the author, the publisher and of the persons who compiled a particular sliced book.

The most demanding reengineering task in the project is the assignment of relations between slices. Relations are used by the Trial-Solution Delivery Tool to search for slices which should be suggested for reading to the user. One outcome of our research is that the kind of relations to be used and their individual definition clearly depends on their later purpose in a delivery tool. This explains why different related projects come up with specific metadata specifications which are most appropriate for the services they intend to deliver.

So within Trial-Solution we identified the need for five different types of relations. In other usage scenarios other types of relations may be needed. In sliced books relations are set between individual slices or between groups of slices. In the latter case they are usually inherited by all members of these groups.

Other projects, for example German project "Teachware On Demand", consider only relations between topics (key phrases) and map these relations to relations between the slices that have assigned these key phrases. While this approach greatly reduces the number of relations that have to be assigned (since there are less key phrases than slices), it is also less precise than the approach taken in Trial-Solution. For example it could not differentiate the relations between two problems and their respective solutions if both concern the same topic. Therefore we selected the more demanding approach in order to serve the user with more specific documents.

References

Dahn, Ingo (2000): Symbiose von Buch und Internet, Proc. Learntec 2000, Karlsruhe, p. 551-558

Dahn, Ingo (2001): Automatic Texbook Construction and Web Delivery in the 21st Century, J. of Structural Learning and Intelligent Systems, vol. 14(4), 2001, pp. 401-413

Wolter, Helmut, Dahn, Ingo (2000): Analysis Individuell, Springer-Verlag, Heidelberg