Tuija Sonkkila
Helsinki University of Technology Library

Tuija Sonkkila works as a Systems Librarian at the Helsinki University of Technology Library. She has been the Head of the computing services of the Library since 1994. Since 1997 the HUTpubl project management has included in her responsibilities. Ms Sonkkila received her master's degree (M.Soc.Sc) in Social Sciences (majoring in Library and Information Science) from the University of Tampere in 1991. Currently she is working on her licentiate thesis on academic authoring and descriptive markup.

Jörgen Westerling
Remtec Systems Ltd.

Jörgen Westerling works as a Systems Solution Consultant at Remtec Systems Ltd. He received his Master of Science (Eng.) degree in 1998 from Helsinki University of Technology. His Master's Thesis was part of the HUTpubl project, in which he developed an SGML-based publishing process for HUT publications. He is currently the president of SGML User's Group Finland.

Abstract

This paper describes HUTpubl, an SGML-based project conducted by the HUT [Helsinki University of Technology Library]. The goal of the project is to establish a working SGML/XML-based publishing model for HUT scientific publication series. In-house project partners include the Department of Automation and Systems Technology, and the Department of Computer Science and Engineering. HUTpubl is also a member of an umbrella project RAJU (Rakenteinen julkaiseminen yliopistoissa, a newly formed consortium of three Finnish universities). RAJU seeks to define guidelines and to give recommendations for the use of structured documents in higher education in Finland.

The paper elaborates on lessons learned in HUTpubl, both technical and those more closely related to finance and workforce. Findings in other related projects and research activities are also included. Electronic publishing in general is confronted with a multitude of demands and hopes, expressed by users on one hand, and by institutions on the other. One of the key issues concerns long-term availability of digital information. In addition, research findings indicate that users would like to place more detailed full-text information retrieval requests. Due to differing interests, some users focus their attention to figures, others in tables or bibliographies, to name just a few examples. Furthermore, there is a wish to deliver publications on many platforms, which asks for suitable mechanisms of combining information with different sets of output specifications. In all these three cases, the capabilities of today's desktop editors fall short. Yet they are among the most frequently used tools to produce scientific publications.
It is claimed that the answer would lie in the use of structure-oriented editors and descriptive, platform-independent markup. But the move is not a trivial one. One of the first big challenges is the author himself. To what extent is he willing to modify his working habits? Does he accept the possibility of letting someone else define the layout of his work? Another major issue is the publication process. The nature of changes in work-flow are as much organizational as they are technical. There is no easy way to copy existing publishing models from one university to another, as lucrative as it might seem.

1 Introduction

As a publisher of print publications HUT is a big non-commercial one. Annually, the number of individual titles is between 400 and 500, published in over 200 different scientific publication series, locally also called laboratory series. Titles in these series are reports, working papers, theses, dissertations, etc. According to data gathered from the local database TKKJULKAISEE, on average 15-20 % of all publications published by HUT departments between 1994 and 1997 were titles in HUT's own series.

The publishing process at HUT is decentralized. In other words, there is no central publishing unit at HUT, no "HUT University Press", as it were. In this respect, HUT differs from many other Finnish universities. Traditionally also, the twelve departments and their laboratories are very independent in all internal matters such as publishing procedures. The only HUT-wide recommendations concern layout and typographical details of the cover page of the printed document. In 1998, these recommendations were totally revised to reflect the new visual image of HUT. Recommendations were published as an Editors Kit and distributed to every financial unit.

The role of the Library in HUT publishing activities so far has been that of a promoter of standards in metadata creation and use, both in paper publications, and in their electronic equivalents. The Library also acts as a local access point of ISBN numbers. Despite of all effort, the level of standardization has remained low.

In early 1997 the Helsinki University of Technology (HUT) Library launched a project called HUTpubl (URL: http://www.hut.fi/Yksikot/Kirjasto/HUTpubl/), funded by the Ministry of Education in Finland from the special Information Society Program funds. The project had four technical objectives. First, to design a DTD for HUT scientific publication series. Second, to test the usability of the DTD for authoring with a number of real-life examples. Third, to evaluate two different word processors in structured authoring. Fourth, to design a conversion procedure from SGML to HTML format for WWW delivery.

Besides these specific objectives the project targeted to two unwritten, more abstract ones. First, to gather knowledge about SGML implementation issues. Second, to promote the importance of publishing and metadata standards at HUT. As of writing this, the outcome of the project is a pilot
study, a Master's Thesis on the subject, and a small document base. In this paper we will elaborate on lessons learned, present a number of findings on academic authoring, mainly based on current literature, and, in the last section, present the HUTpubl prototype in more detail.

2 Lessons learned

2.1 Make your plans clear

Without project funding from outer sources, today's universities would not be able to experiment in, say, information management issues. But project funding has drawbacks. Probably the most serious one is its impulsiveness and unexpectedness. Therefore, it would be wise to plan for the future right from the start. Present the objectives of your project in simple terms and graphics to key persons in your university, to those whom you will approach the day your project funding dries out. Even though it is extremely difficult, this effort could provide valuable help when the project gets into trouble. Naturally, it will not harm the project otherwise either!

2.2 Know your house

Modern universities are confronted by such entrepreneurial facts of life as result management and market value. This means that, slowly but steadily, universities will begin to take a close look at what they produce, how efficiently, and what kind of impression it generates in the public. One line of production is publishing. Be prepared for this to happen. Ferret out opinion leaders and experts in this field in your university, and try to sell them your ideas, because they will be those who will be asked for advice.

2.3 Be ready to fight for your staff

Academic SGML projects are forced to compete on capable IT staff. How universities succeed in this in average may be read from surveys among university students; the only employers that are even more undesirable than universities are communities! Therefore, it would be wise to try to get part of the permanent staff involved in the project. Otherwise, in the worst case, you may find yourself one day to be the just about the only one in the house interested in the subject.

3 On academic authoring

By definition SGML provides a standardized way to build document repositories that are independent of software and hardware environment. This is a strong argument, both political and economical. In the US, the Task Force on Archiving of Digital Information [1] claims that much of the recent work on the so-called digital libraries has been notably silent on the archival issues. In Europe, the DLM-FORUM organised by the European Commission proposes SGML as the format of structured text [2]. Still,
examples of the use of SGML in academic information management activities are rare. Notably rare they are in the area of scholarly authoring.

As Goldie [3] underlines, the higher "upstream" in the publishing chain SGML is introduced, the bigger the benefits. An interesting question is, how far up in the scholarly authoring chain SGML could be implemented. There exists hardly any literature about researchers as authors of structured text. One reason that comes easily in mind could be the small number of SGML implementations in higher education in general. It should be noted here that the role of SGML in the numerous projects launched and propagated by the TEI [Text Encoding Initiative] in humanities is a central tool for research rather than a publishing tool for documents. Project deliverables themselves are SGML databases, such as text corpuses. But whether any e.g. reports published by these projects are generated through a SGML production line, is not known.

Contrary to clerical staff who use editors as their primary work instrument, and whose authoring tasks might be said to be either pre-defined or otherwise highly expectable, academic authors reporting on their research activities are free actors. They produce text after their own will and interest, with a tool of their own selection. Therefore it is reasonable to believe that they might be rather reluctant to follow guidelines and recommendations in this field, especially if there is no clear evidence that their effort would bring them any good. Nevertheless, researchers do accept guidelines coming from outside, notably from the area of scholarly journals. In article submissions researchers have to follow format descriptions (sometimes quite detailed) expressed by the editorial board.

Goldie [3] notes that the development of SGML tools has been slow, and exceptionally slow has been the development of tools belonging to both ends of the production line: authoring and printing. Recently, Sabarthez [4] has expressed interesting thoughts on how reasons for this might be found in the history of personal computing: the spread of easy-to-use GUBs [graphical user interface], and sophisticated low-end publishing software opened up a conceptual gap between them and mainframe-based SGML systems of that time.

An often used solution on constructing a suitable, "upstream" authoring environment is to use template files with common word processors. Even though there is little empirical research on the actual use of styles in word processors, there is recent evidence [5] that styles are overlooked. Part of the problem is said to be the paper metaphor communicated by WYSIWYG [What You See Is What You Get], because in the paper format the use of styles makes little sense. Styles are thought to be paper output properties only. Like [6] has pointed out, the success of modern WYSIWYG text processing and desktop publishing systems lies within their ability to give instant feedback on how the final document will look like on paper [6]:

*Acting on the feedback, the user then has immediate control. And people*
like having feedback and control
To take this feature away from the author may cause frustration, especially while proofreading [7]. Some have claimed quite categorically [8] that typesetting should not be left to others than the author himself.

4 The HUTpubl process

To be able to enjoy all the advantages of a structured publication process, but at the same time also be able to use the traditional methods of publication production, the model for the HUT process has been developed in parallel with the traditional print-oriented process. The process as a whole thus consists of two parts, the structured one of which was developed in the HUTpubl project.

4.1 Implementation of the process

The most important tasks in development of an SGML-based process are [9]:

- Document and work-flow analysis
- Formal specification and documentation
- Selection of tools
- Implementation and modification
- Testing, training and support

These tasks were more or less followed when developing the process for the HUT publications.

4.1.1 Document analysis

The document analysis was conducted as a study of a sample set of existing publications from some of the most usual series at HUT. The analysis resulted in some one hundred different logical elements. These logical elements were then analysed and a DTD was constructed in the formal specification phase.

4.1.2 Formal specification

When beginning the formal specification phase, the choice was between using an existing DTD or constructing our own. Of these, the second alternative was selected. There were a couple of reasons for this choice. Firstly, the main candidate of the existing DTD:s, DocBook, is so complex, that we felt the task of stripping it down to our needs was a undesired on. Secondly and most importantly, we wanted to gather experience of designing and writing DTD:s. In the design phase, we strived to name elements in accordance with the DocBook, to keep compatibility.

4.1.3 Selection of tools
The tools used in the project was Adobe's FrameMaker+SGML for structured editing, Microsoft Word for unstructured editing, JADE for the SGML to HTML conversion, nsgmls for parsing SGML documents, CVS for version control and PGP for publication authentication.

4.1.4 Testing and training

The HUTpubl process was tested using two writers and one researcher. The test subjects were trained in the basics of SGML and in using the HUTpubl DTD before the test phase was started. The writers had no prior experience using editors for structured information and no experience with SGML.

The result of the testing was that the developed DTD was usable for producing technical publications at HUT with a few corrections. The DTD was modified along the way and the test subjects then used the modified version. Three publications were produced during the test phase.

4.2 A publication prototype

As a part of the process design, a prototype publication was converted from MS Word to SGML-format. The publication was an existing licentiate thesis by Håkan Mitts. The prototype publication was also a test for the complete HUTpubl process, from production to distribution.

References


