Proceedings of the IATUL Conferences

1997 IATUL Proceedings

Issues Concerning SGML-based Electronic Publishing in Library Settings

Tujia Sonkkila
Helsinki University of Technology Library

Tujia Sonkkila, "Issues Concerning SGML-based Electronic Publishing in Library Settings." *Proceedings of the IATUL Conferences.* Paper 25.
http://docs.lib.purdue.edu/iatul/1997/papers/25

ISSUES CONCERNING SGML-BASED

ELECTRONIC PUBLISHING IN LIBRARY SETTINGS

Sonkkila, Tuija
Helsinki University of Technology Library,
P.O.Box 7000, FIN-02015 HUT, Finland

In terms of academic publishing, Helsinki University of Technology (HUT) is a major producer in Finland. Annually, it publishes hundreds of titles in roughly 200 different scientific publication series. The lifespan of individual publications varies considerably, but one thing is in common: the publishing process is decentralized to the degree that normally it is the writer herself who takes care of the whole publishing process, from tapping the keyboard to storing the print run.

In spring 1996, the Library took the initiative of seeking funding from the Ministry of Education in Finland for a four-year project of HUT electronic publishing. The main goal of the project is to establish a set of technical procedures for electronic publishing of HUT scientific publication series. Another important goal is to increase local understanding and knowledge about the importance of standards in academic publishing in general, and the benefits of SGML in particular. In-house project partners include the Department of Automation and Systems Technology, the Department of Computer Science and Engineering, and the Computing Centre.

An amount of FIM 180K (USD 35K) was granted by the Ministry from its special Information Society Fund . After that, a subsequent FIM 130K (USD 25K) has been recieved. I'd like to point out that this funding covers direct labor costs of the project, special PC software, and running costs such as travel expences. In four years time, additional costs are estimated to go up to FIM 1.000.000 (roughly USD 200K). These include hidden costs like in-house cooperation and consultancy of various kind, as well as clearly visible costs such as purchase and maintenance of SGML database management system and a server.

The project staff consist of one full-time worker, a student of HUT, doing his master's degree in computer science, and a part-time project manager, belonging to the Library staff. During the initial period of the project it was assumed, a bit too optimistically really, that SGML-competent persons would be available for the project to hire, from the day one. What was soon discovered though, was that this was not the case. SGML knowledge seems to be divided between three separate group of people: those who have already their hands full of SGML work, typically in a big high-tech company like Nokia; those with a long university career focusing to the theory of structured documents; and SGML consultants. So, it became obvious that project timetable had to be adjusted to include the learning curve of SGML. Some amount of additional consulting was also found to be necessary.

As of writing this the first protype in the HUT SGML project is approaching its completion and the pilot phase is about to begin in early autumn. During the pilot phase, which is expected to run one-two months, a small number of researchers will be preparing one manuscript of their own selection from start to finish by using the project protype. They have two alternatives for this.

First , the writer may use a native SGML editor. There are a number of them on the market. If the writer wishes, she can use FrameMaker+SGML from Adobe. There are a number of spare licenses available. The main thing is that she - or rather the editor used - has to follow a certain set of rules, called a document-type definition (DTD), of how to construct the document. In other words, the structure of the document has to conform to the DTD chosen.

Second , the writer might like to use a certain MicrosoftWord template file with a pre-defined set of styles for marking different elements, provided by the project.

In the former case , the result is in pure SGML, in the latter, the result has to be converted from the Word format into SGML . The conversion will be made, for the time being, with FrameMaker+SGML. Later on, it will most probably be replaced by a true SGML conversion tool like Balise.

Here I'd like to underline the fact that the conversion is by no means a fully automatic process. Some 60-80% procent of the whole document can be automatically converted, depending on the amount of mathematical formulae, graphics, and tables, which have to be individually dealt with. Fortunately, these are not head-aches of the HUT project only; SGML projects world-wide suffer of them.

As to metadata, there were a couple of sets available, from which Dublic Core Metadata Element Set was chosen, mainly because of the Nordic Metadata Project, the management of which is at the Helsinki University Library, the National Library of Finland. Among other things, this project will provide a free DC > FINMARC converter. FINMARC is the Finnish version of the international MARC bibliographic record format.

Finally, instead of network delivery in SGML format, down-conversion to HTML was thought to be more appropiate at this stage, due to shortage of proper public-domain SGML clients. SGML-to-HTML conversion is done with JADE, a DSSSL engine by James Clark. DSSSL is a SGML-related standard for defining what should be done to individual elements in a given DTD in, say, a conversion instance. JADE reads DSSSL instructions - called a stylesheet - and performs the actual conversion.

For those of you familiar with the issue of whether or not to use an industry-standard DTD , that is, an international standard for a given field of industry, it might be mentioned that document analysis resulted in the choice of constructing an own DTD. HUT publications are structurally reasonably straightforward, quite on the contrary to some DTDs inspected. DocBook DTD for instance, which is an industry-standard DTD for all book-like things, is, probably just for that reason, very big, very confusing, a real monster if I may say so. In addition, this was a good opportunity to learn how to do a DTD.

Future will show if an own DTD was a wise decision or not. SGML analysts tend to emphasize the benefits of industry-standard DTDs though (or subsets of them), particularly in network delivery, because stylesheet construction and maintenance may otherwise become a substantial burden. One month with the before-mentioned DSSSL has proved this to be quite true. Manuals are non-existent, real-life examples are scarce, and the DSSSL standard is not particularly readable. Hopefully though, the project worker grew quite comfortable with the LISP-like language of DSSSL, and work could proceed. Another alternative would have been to write a SGML-to-HTML converter in Perl, which is a Unix-based popular and powerful scripting language.

Project-wise, there are in fact two lessons to be learned from working with free SGML tools like JADE. First, in order to be able to compare the pros and cons of new programs and scripting languages to other, existing alternatives, someone in the project force has to be familiar with a variety of programming languages (working knowledge of the Unix environment is a must, although most of the SGML tools are available for DOS as well). Second, there better be a somewhat conservative - or should I say, realistic - kind of attitude towards new and challenging technical obstacles; not too eager to tackle them all at once, because that might have adventurous results and would lead to time-consuming and frustrating experiences in any case. But too much humbleness is not good either, because that would lead nowhere.

The question of database management is yet to be dealt with. Theoretically, SGML files - which are technically ASCII files - could be queried and retrieved by standard Unix tools added with a number of public-domain SGML applications and a suitable GUI, but that kind of construction would only be a temporary solution.

The HUT SGML project will from now on be divided into four workpackages. First, piloting, which includes an initial training period, followed by regular meetings, helpdesk, and one or two workshops. Second, SGML marketing at HUT; ie making the goals of the project widely known at the university, particularly among the heads of the departments, in the recently appointed Implementation Group for the HUT Information Strategy, and to the soon-to-be-appointed Information Manager at HUT. Third, cooperation with other related projects in Finland, for gaining mutual benefit. And forth, maintenance and improvement of the prototype.

Not mentioned on this list, but bubbling under, is the question of infrastructure, which is a long list of "if's", "how's" and "who's". I have the pleasure - and pain! - to deal with some of these aspects during my post-graduate studies.

As to the deliverables of the project, there is not much to mention yet. There will certainly be a final report at some stage, but long before that I'd hope to be able to get the project Web homepage up-to-date. There is a set of pages already, but unfortunately only in Finnish at the moment, so I don't bother you with a URL. Yet.

It may take quite a while for a new standard to gain acceptance. In this respect, SGML has been no exception. SGML was given the status of an ISO standard in 1986. Before and after that, the principal usage of SGML has taken place in the field of technical documentation, where the benefits of getting different end-products for different clientele from the same, structured SGML database have been obvious. It is only now, in the age of multiple new electronic publishing platforms, when SGML is getting foothold in academic circles as well.

Indeed the future of structured documents looks very promising, thanks to a new, simple dialect of SGML, named XML (Extensible Markup Language). The goal of XML is to enable generic SGML to be served, delivered, and processed on the Web. Precisely for this reason XML is an eagerly awaited novelty. It seems to be the missing link between SGML and HTML, the web-application of SGML.

The SGML standard has been critized, and not without reason, to be hard to understand, too clumsy, that it has odd structures noone uses (and should not use). XML is quite the opposite: stripped off from all oddities of SGML, added with nice new link structures, it surely feels a step into right direction.

To conclude, I'd like to add my two cents on cooperation and on the library as a initiative taker in a publishing project.

Cooperation at university level is never a trivial task, partly because of the amount of time and effort it takes, often without any immediate results. Differences in work culture may be hard obstacles, clashes of interest between organisational units likewise. Nevertheless, cooperation do counts, particularly in publishing, and especially now, when the playground of academic publishing is open for new players. It is quite natural that the Library, which has over twenty years of working experience about why information should be structured and maintained so as to be readily and meaningfully accessible by a computer, has something to give in this matter.

ABBREVATIONS AND ACRONYMS

ASCII	American Standard Code for Information Interchange
DC	Dublin Core
DOS	Disk Operating System
DTD	Document Type Definition
DSSSL	Document Style Semantics and Specification Language
GUI	Graphical User Interface
HTML	Hypertext Markup Language
HUT	Helsinki University of Technology
JADE	JAmes [Clark]'s DSSSL Engine
LISP	List Processing
MARC	Machine-Readable Catalogue

SGML	Standard Generalized Markup Language
URL	Uniform Resource Locator
XML	Extensible Markup Language