pm — Institut für Print und Medientechnik der TU Chemnitz
[Institute of Print and Media Technology at Chemnitz Technical University]
Reichenhainer Str. 70 – 09126 Chemnitz, Germany
Tel: +49 (0) 371-531-8091 Fax: -3780
arved.huebler@mbv.tu-chemnitz.de — http://www.tu-chemnitz.de/pm

PROF. DR. ARVED C. HÜBLER

4th ICCC/IFIP Conference on Electronic Publishing
August, 17-19th 2000 at Kaliningrad/Svetlogorsk, Russia.

# Improving Individual Book Publishing Concepts with XML Schema

Arved C. Hübler, Klaus Kreulich
Chemnitz Technical University, Institute for Print- and Media Technology
http://www.tu-chemnitz.de/pm

## Introduction

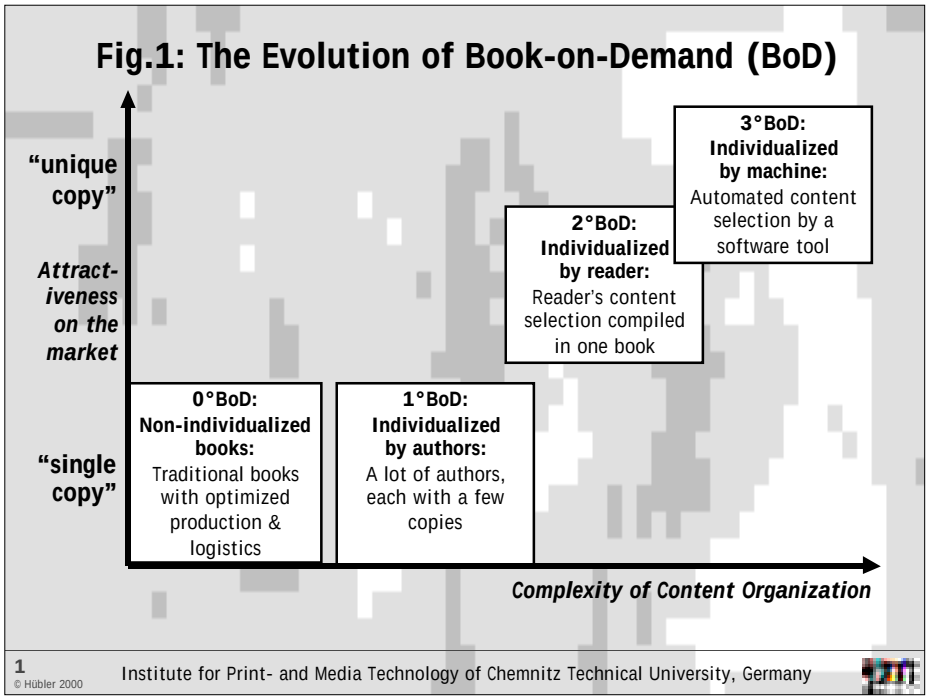
The traditional way of publishing books is a one-dimensional procedure. An author creates content, believing that somebody may be interested in it. A publisher invests money in his book project, financing the production, the marketing and the distribution to the bookstores. The readers select the books in the bookstore under a wide range of decision criteria's. At the end of the chain, the public libraries save the book, storing the information as cultural heritage. Since years with the growing of digital printing technology, Internet and data based publishing also a new type of book production is in the discussions: Book-on-Demand (BoD). Especially suppliers of the new enabling technologies promote this new type of book.

The argumentation for this book publishing strategy is guided very strong by the traditional way of book production and of traditional book concepts itself. Core feature of this BoD concept is the masterless print process of digital presses. Because of this, the make ready times for printing decline to zero, the print production costs of one copy are the same as multi copy production. Since more then ten years this technical capability is used as Print-on-Demand PoD in invoice printing and other applications. Now also the bindery becomes "on-Demand", the inline production of a single brochure is possible on a low but acceptable quality stage. But the extension from PoD to BoD did not change the book market in a larger way. Only some specific branches occurred where BoD could take some significant parts of the market share. But today's BoD does not cast any doubt on the traditional book itself.

The main error of these first BoD concepts is that they ignore real cost structure of the books. Typically for a book for 20$ in the bookstore, only 0,17$ must been paid for the printing, and all technical costs including materials are below 2,50$ [1]. Because of this



© pm 04.07.2000




reason, real cost savings in production and logistics are not so high and only in some special applications BoD is attractive.

We have to take another important fact into account: There are only a few authors writing books without any commercial interest and with the goal to sell only two or three copies. The success story of a book is, either in non-material or in commercial categories, to spread it. From this point of view, BoD is a disaster for the author and publisher, both can not think about individual books economically. BoD solved them only their second problem, the production. But the first problem, how to reach the reader who needs the book, remains.

Consistently, today's BoD is not really successful. We call this concepts "1st stage BoD (1°BoD)", whereby the books are characterized by new production concepts and new output technologies (PoD, but also E-Books and CD-ROMs), but traditional business models and content handling.

## The True Individual Book

New concepts of individual books have to look for the reader's interest. The reader does not need individual books with the meaning that he could select between a huge number of uninteresting contents from unknown authors. He likes to have individual selection of highly significant content so he gets the comprehensive opportunity to find answers for his questions. Content selectivity in this way is something totally different to the 1°BoD concept. It is a recompilation of books by mixing different content in a meaningful order, controlled by the individual interest of the reader. We call this concept the "2nd stage BoD (2°BoD)" or Dynamic Book. The far away vision for this new type of book is an automated content selection, which might been the "3rd stage BoD (3°BoD)" or the Generic Book [2] (see Fig.1).



© pm 04.07.2000

Today there are some examples of this kind of books realized. Commercially personalized children books are known, where the player's names are eligible by ordering the copy. Also prototypes in different research applications are reported, e.g. an individualized study guides, which is under development at our institute [3,4]. But there are several serious problems for a general realization of 2°BoD products, either in the structural but also in the technical sphere. So the access to the content from different sources may be possible via Internet, but the copyright management, version consistency and similar problems must been solved first.

At the technical side, with the XML technology a base is given for a principle new approach in defining document structures. But up to now, the existing XML (and former SGML) applications are oriented at the traditional book concept with it's semantic hierarchy and linear understanding of textual reception. But for real individual books, a separation in one non linear structure supporting the content selection and one highly linear structure, supporting the assembling process for the output media, is necessary. In Fig.2 the *selection manager* and the *layout manager* are functional software tools for this different tasks. One important additional function is the transformation of the non-linear content structure to the linear output structure, which may be organized with a *structure manager* tool. To implement a full 2°BoD Workflow, additional components are necessary to solve specific transformation tasks.

XML may be powerful enough to support these requirements of 2°BoD products. But today's common DTD's do not have the necessary functionality for this.

### Today's Book-Oriented DTD's are limited

From the various uses of XML or SGML in the field of automated document management a

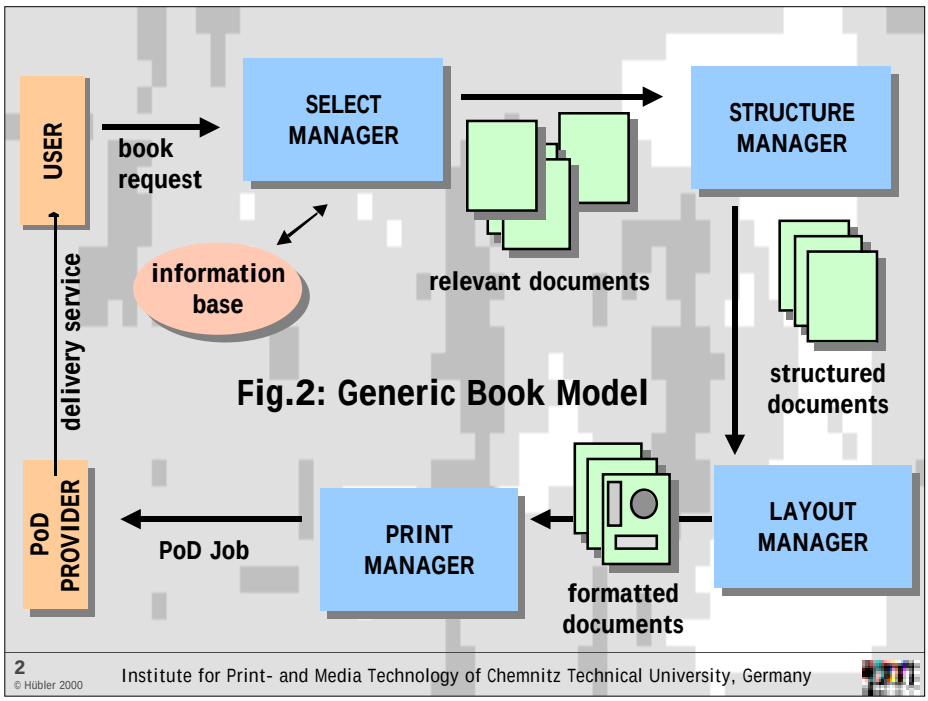


Fig.2: Generic Book Model



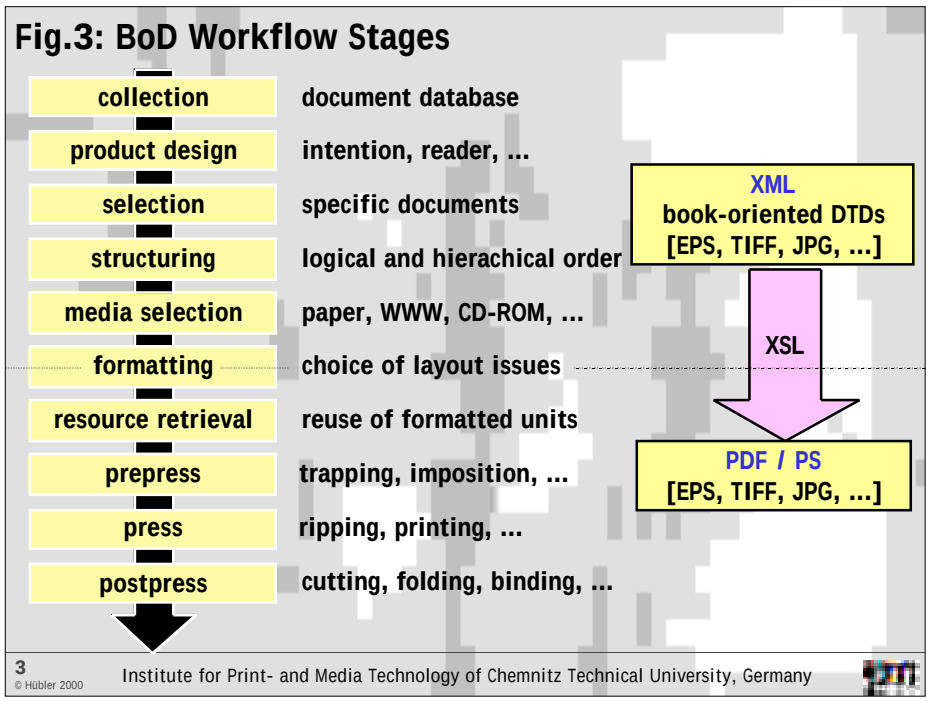
© pm 04.07.2000

lot of DTD's have evolved since the passing of the SGML standard in 1986. In the meantime, some of these DTD's have established themselves as official or unofficial standards. Some prominent book oriented DTD's are ISO 12083 DTD, DocBook DTD, Text Encoding Initiative (TEI) DTD, MIL-STD-38784 DTD, HTML and XHTML and as newest OEB DTD, a DTD for content representation on E-Book media.

In the development of BoD applications these DTDs can be used as the basis for modeling textual structures. Thereby, however, the universal weaknesses of DTD's have to be considered. A serious disadvantage of DTD's is that they only offer very inexact data types. At the end of a hierarchic element definition stands a #PCDATA element. This means that the content of the element can consist of any number of symbols. Additionally, the useable XML data types for attributes are weak. In practice this means that data often cannot be checked, as regards correct contents, by the standard means of an XML parser. In an automatic finishing process, especially in the communication between different business partners, such as a PoD provider and a publisher, expendable measures of protection have to be made in order to extract the meant digits. In other words, in the case of DTD's the burden of adding program logic to deal with unspecified data falls on the developer.

Another disadvantage of DTD's is the insufficient support to reuse content models. DTD's offer the mechanism of the parameter entities, that is a comfortable way to use strings repeatedly. However, a logic link between two elements with different names but similar content models is not possible. The above listed DTD's provide some subtle solutions for customizing them in particular application relevant contexts. However, all these solutions are based on entities, wherefore processing programs generally have to be adjusted to the customized elements. With the assistance of the type concept, XML Schemas offer new opportunities.

As indicated above, the completion of a production process and the realization of a busi-






© pm 04.07.2000

ness are parts of the BoD Workflow (see Fig. 3). Data concerning the ordering process and the print and postpress workflow are to be taken under consideration. In this context, the DTD's above are to be complemented by suitable DTD's of other usage's or newly developed DTD's. For a versatile application, a modularization of the entire BoD DTD's into a content share, a control share, a printing share, a postpress workflow share etc. would be useful. The existing mechanisms for modularizing DTD's are rudimentary. XML Schemas mean improvements in this area, as well.

## Benefits of XML Schema [5]

To overcome the shortcomings of DTD's the W3C chartered a new Working Group that is concerned with the development of an XML Schema Recommendation. The current working draft contains fundamental approaches to improve a data management with XML. Overall XML Schema simplifies the automatic processing of data and documents. The interface between XML documents and databases is continually simplified and so the handling of dynamic data is supported.

One of the fundamental improvements compared to DTD's is a uniform syntax for document instances and document definitions; XML Schemas are XML Instances of the W3C Schema definition themselves. The processing of XML Schemas and XML Instances by the same software tools is possible. An automated processing of XML Schemas can be done on a much higher level than that of DTD's. In addition, less expenditure for the data management is necessary. The administration of the data as well as the programming expenditure is eased.

An important extension of the existing DTD properties is the opportunity to use strong typed data. XML Schema offers, similar to a modern computer language, an extensive quantity of implicit data types. Beyond it, within XML Schemas deduced or completely redefined types can be declared.

A further important property of XML Schemas is the support of Namespaces, with which the reuse of entire XML Vocabularies and also individual structure definitions is made easier. The Namespace concept was developed independently from XML Schemas, but in connection with XML Schemas for the first time a workable option for use is possible.

The reuse of Schemas or parts of them is supported by an object-oriented approach in type definition. Like the categories of an object oriented computer language can inherit their properties from a basis category, the transition of properties between hierarchic types of XML Schemas is facilitated. The so-called deduced types can have exactly the same properties as their basis types, but restrictions or extensions can also modify them. From the novelties and improvements that XML Schema offers various uses of BoD applications can be deduced.

## Conclusions

The definition of the term "book" is changing strongly from the old author and publisher oriented 1°BoD model to reader oriented 2°BoD approaches. This means, that the traditional concept of a linear textual content structure may be expanded by the demand of the reader's selection procedures in a dynamic content base. The older ideas of hypertext will get a new stimulus, not as an alternative for the traditional book, but as an integrated pre-stage in the book content assembly. The XML technology of today is not as



© pm 04.07.2000

that powerful as to cover all the necessities related to these 2°BoD document types. XML schemas could be a further milestone on the road to future book technology. An enhancement of XML functionality to support hypertextual semantics and the transformation of these structures to a linear content representation must be one of the next steps in XML development.

**References**

**[1]** Hübler, Arved C. in Kipphan, Helmut (Ed.): Handbuch der Printmedien — Springer Verlag, Berlin Heidelberg, 2000 (Part 1.9) / will be published in English in 2000

**[2]** K. Kreulich: The Generic Book as an Application of Intelligent Information Retrieval-Systems, Abstracts of the 22nd Annual Conference of the German Society for Classification, Dresden, March 1998, S.65

**[3]** Kreulich, Klaus; Stasch, Eckhard; Hübler, Arved: value-added Digital Libraries Services: Individualized Crossmedia Output with XML Documents — 10th VALA Conference, Melbourne, 16. -18.2.2000

**[4]** Study-Guide-on-Demand, Project of the Institute for Print- and Media Technology at Chemnitz Technical University, Germany. http://www.pm.tu-chemnitz.de/sf/

**[5]** Hübler, Arved C.; Kreulich, Klaus: Modules for an XML Schema in the Book-on-Demand Process — XML 2000 Europe Conference, Paris 14.6.2000



© pm 04.07.2000