Cahiers GUTenberg

AN EXPERIENCE FROM A DIGITIZATION PROJECT

Petr Sojka

Cahiers GUTenberg, nº 28-29 (1998), p. 276-282.

<http://cahiers.gutenberg.eu.org/fitem?id=CG_1998___28-29_276_0>

An Experience from a Digitization Project

Petr Sojka

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
Email: sojka@informatics.muni.cz

Abstract. *An experience from the process of adding logical markup to visually tagged scanned data is presented. The method of gradual markup enhancement is shown. Methods of navigation in a large hypertext document based on typesetting from logical markup are suggested—physical, logical and semantic user views. Their application on a 28 000 page project to create an electronic encyclopædia is described and problems faced when using Adobe's Acrobat technology for publishing are discussed.*

Keywords: digital replica, markup, navigation, hypertext, Acrobat, CD-ROM, encyclopædia,

"Go forth and create masterpieces of electronic publishing art."

1. Introduction

Current computer technologies allow easy production and delivery of huge documents like encyclopædias on a CD-ROM for public, everyday use on cheap hardware. Digitizing such multivolume books in hypertext form is a challenge.

Users accustomed to using paper editions for years often reject electronic media if their look and feel is totally different from the printed version. The task is therefore to digitize the work exactly as was published on paper but with the add-ons such as searching, navigation, etc. Clever substitution of intuitive navigation based on physical book handling is essential for conservative users.

In this paper we describe our ideas about navigation in such huge documents, and experience we gained during our participation in a project for the production of a digital encyclopædia CD-ROM.

"You have done something that you are excited about." *Leslie Lamport*
"Historia magistra vitæ."

2. Otto encyclopædia project

We were participating in the design and retypesetting phases of a project aimed at the creation of a digital replica of Czech encyclopædia "Ottův slovník naučný" (OSN) [4] (28 volumes, 27 789 pages, 139 418 entries, 4888 illustrations) published in 1888–1908. The project has been realized during 1996–1998. Currently, a similar project of digitization of another Czech encyclopædia—"Ottův slovník naučný nové doby" [3] (OSNND)—is being undertaken, using the experience from work on OSN. OSN and OSNND are the biggest Czech encyclopædiæ ever published and are one of the most valuable sources of general information and knowledge, useful not only for historians.

As a format for electronic delivery of a digitized encyclopædia Adobe's Portable Document Format [1] (PDF) was chosen, because achieving highest visual fidelity with respect to the original was a high priority. This format has been successfully used for WWW and electronic journal publishing [6]. Several free readers of PDF documents are available (Acrobat Reader, xpdf or DocuReader [10]).

Full scanning and retypesetting of the encyclopædia was needed because the original was typeset using hot type and no original electronic data were available. We faced several design and production problems whose solutions we want to share.

"By indirections find directions out." *Shakespeare (Hamlet, act 2, sc. 1, l.66)*
"Put yourself in the reader's place!" *Don Knuth*

3. Navigation methods—document views

The reader of a huge electronic document needs navigation help. We distinguish several types of reader's "view" of a document. There is logical structure to a document—the *logical view*. The reader wants to inspect or read an electronic document on computer screen bearing in mind its physical structure—e.g. multicolumn layout. This represents the *physical view* of a document. A third point of view may be based on the reader's particular interests—he wants to collect related information about a particular topic of interest and relations between document parts and to visualize them; we call this view the *semantic view*.

3.1. The logical view—expressing logical structure

PDF format offers several possibilities for navigation: Acrobat bookmarks for the logical view, articles for the physical view and hypertext links for switching between various physical or logical views. However, technical problems appear in practice, for instance limitations imposed by media used in document transmission like physical dimensions and resolution of a user computer screen. Acrobat bookmarks are of problematic use for a *highly* structured document—localization of an item in the huge bookmark tree using the mouse is nearly impossible; a new method for logical structure visualization has to be used. Characters used in Acrobat bookmarks are restricted to PDFDoc encoding only, which is not sufficient (not all characters used in the Czech language are available).

Current possibilities of intelligent searching within PDF are restricted too, with no support for Czech in Acrobat 3.

The logical view has to be presented in another way. A portable and effective structure for navigation of a document can be mechanically generated by the following approach.

Let's take the example of our encyclopædia. A tree of a document structure (28 volumes with about 5000 entries each) was balanced to have minimum depth l and every node had at least n successors. For $N = 140\,000$ entries we can manage to have $l = 4$ and $n \approx 20$ (the l-th square root of N).

For each node (entry in the encyclopædia) a hypertext-sensitive navigation link was generated and typeset such that each one n-th part of a navigation page on the screen described a subtree—in our case of alphabetically ordered encyclopædia entries it was the first entry in a subtree. This method allows the user to jump to a particular page with just l mouse clicks.

3.2. The physical view—expressing document layout

Acrobat Articles proved to be very intuitive and handy for navigation in our multicolumn application. The reader would also (using Adobe's Access application) have the articles *spoken* to him, in proper order. As printing is usually demanded by the user for pages of interest and one-to-one reproduction is preferred to give readers the look and feel of the original pages of interest, nothing more can be done at this point of view. Adding navigation buttons and areas to get information on actual user position is a must, but basic support is already built into most PDF readers (together with zooming, various types of fitting actual page on screen, thumbnails).

3.3. The semantic view—expressing relations

Another point of view of the document is based on a semantic map of document notions and their relations. Information can be structured and sorted with the semantics of the articles in mind. (Yahoo attempts to do this kind of document classification on the Internet.)

In our application we have identified the need to generate "see also" links based on a computed measure of "similarity" based on a semantic map of the Czech language using an approach motivated by [5]. We could then make the "navigation document" using this semantic net. These links will create an equivalence relation over the encyclopædia entries and applying this approach recursively we can get a tree-structured semantic net of an encyclopædia—the semantic view. It can be visualized in the same way as the logical structure tree.

"Data cannot be used at a finer grain than it is marked up at." *R. Jelliffe*
"The ability to handle lots of cases is Computer Science's strength and weakness." *Don Knuth*

4. Production

Use of Adobe's Acrobat Capture program was found inappropriate because customization proved to be inefficient or impossible, and because of problems with handling different alphabets and the need for high-level markup for navigation and searching. Thus retypesetting from scanned data with low-level (visual) markup was needed.

4.1. Going from visual to logical markup

Special markup formal languages were developed for tagging scanned data and their further processing. From tags that were inserted during the scanning process several document transformations were applied to get richer and logical markup. Most transformations were done automatically, using pattern matching tools (mainly based on regular expressions, carefully selected and *debugged* substitutions). UNIX tools `sed`, `awk` and `perl` were heavily used in the batch mode of processing. Some transformations needed human intervention; several interactive special-purpose programs were used for instance for spelling error corrections (mainly as Visual Basic macros) and markup validation and correction.

This semiautomatic transformation has led to fully featured data tagging, embodying both logical and visual structure. This bottom-up approach proved

effective and successful. About 40 000 hypertext links were added semiautomatically using just syntax and morphology information on words.

4.2. Retypesetting and PDF generation

The typesetting system TEX was chosen for retypesetting in a distributed, heterogenous environment. This proved to be a good choice; we are deeply convinced that with a program that is not open and available with sources we would sooner or later have got stuck.[1] We benefited from a large CTAN database of free fonts, special-purpose macros and programs.

The standard way of producing PDF from TEX is via a DVI file to Postscript by `dvips` and then to PDF via Adobe Distiller. Experiments with a modification of TEX which is able to produce PDF directly—pdfTEX program by Hàn The Thanh [7, 8]—were done, allowing for very compact PDF files. It also allows a high degree of reuse of document parts in an object-oriented manner of PDF. pdfTEX, however, still lacks the support of some PDF features like bitmap font support that we needed in our application; we ended up using Distiller, leaving pdfTEX as an option for next release.

TEX macros allowed full automation of typesetting of a logical navigation document (more than 9000 screen pages), as well as cross references between the document cores in several PDF files, and automatic creation of Acrobat's articles. This has been accomplished with a pdfmark mechanism for passing information from high-level TEX markup via DVI and Postscript to PDF via Distiller.

"We are all apprentices in a craft where no-one ever becomes a master."
Ernest Hemingway

5. Conclusion and future work

Our experience from participation in the project[2] showed that the bottom-up approach to get fully tagged data for retypesetting of large volumes of text is feasible. A high degree of automation proved possible; although manual work is necessary to meet all requirements. Special attention has to be given to searching and navigation tools, allowing reader's different document views and digestion. Using the document design for paper edition for electronic document

1 The very first attempts to use WYSIWYG typesetting programs were disastrous.

2 The whole OSN encyclopædia fits on a single CD-ROM and the first version is out. As searching in Czech within PDF is not possible, texts were exported and indexed by Verity's tools as a separate application with links to the PDF version for user convenience. This made a second CD-ROM with index and the like.

delivery is problematic, but can be partially eliminated by offering different document views to the user, taking into account the quite different transportation medium—computer screen.

New methods remain to be developed for automatic semantic-view generation. Unicode [9] support in Acrobat to allow PDF-based Czech searching is awaited for the next OSN encyclopædia regeneration. We are about to experiment and test the variable letter-width technique [2] applied to multiple master fonts to achieve more uniform greyness in the text than in the original print.

6. Acknowledgements

We acknowledge the enthusiasm, consideration and care of all people we met during our participation in the design of the above-mentioned project.

Bibliography

[1] Tim Bienz, Richard Cohn, and James R. Meehan. *Portable Document Format Reference Manual, Version 1.1.* Addison-Wesley, Reading, MA, USA, 1996. Version 1.2 of manual is available in electronic form from `http://www.adobe.com/prodindex/acrobat/adobepdf.html`.

[2] Miroslava Misáková. Typography of Quality in Computer Typesetting (in Czech). *Master's Thesis*, Masaryk University Brno, Czech Republic, December 1997.

[3] B. Němec et al. Ottův slovník naučný nové doby. 12 volumes, Prague, 1930–1943.

[4] J. Otto. Ottův slovník naučný. 28 volumes, Prague, 1888–1908.

[5] Gerard Salton, Chris Buckley, and James Allan. Automatic Structuring of Text Files. *Electronic Publishing*, 5(1):1–7, March 1992.

[6] Philip N. Smith, David F. Brailsford, David R. Evans, Leon Harrison, Steve G. Probets, and Peter E. Sutton. Journal Publishing with Acrobat: the CAJUN Project. *Electronic Publishing*, 6(4):481–493, December 1993.

[7] Hàn The Thanh, Petr Sojka, and Jiří Zlatuška. The Joy of TEX2PDF—Acrobatics with an Alternative to DVI Format. *TUGboat*, 17(3):244–251, July 1996.

[8] Hàn The Thanh. pdfTEX distribution, available from `ftp://ftp.cstug.cz/pub/tex/local/cstug/pdftex`

[9] Unicode Consortium. *The Unicode Standard*, Version 2.0. Addison-Wesley, 1996. see also `http://www.unicode.org/unicode/standard/standard.html`

[10] Zeon Corporation. DocuReader 2.0. `http://www.zeon.com.tw/dreader.htm`