Digitalization of Otto Encyclopædia

Petr Sojka
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
E-mail: sojka@informatics.muni.cz

Abstract

An experience from the process of adding logical markup to visually tagged scanned data is presented. Method of gradual markup enhancement is outlined. 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 29 000 page digitization project to create an electronic encyclopædia is described. Problems faced in applying Adobe's Acrobat technology for encyclopædia publishing are discussed.

Abstrakt: V příspěvku jsou shrnuty zkušenosti z digitalizace dvou projektů– digitalizace Ottova slovníku naučného (téměř 29 tisíce stran) a Ottova slovníku naučného nové doby (11 tisíc stran). Je popsán poloautomatický postup postupné konverze naskenovaných dat do tvaru vhodného pro přesazbu, který obsahuje informace o struktuře a o logických částech textu. Je navržena metoda navigace v takýchto rozsáhlých encyklopediích, reflektující různé pohledy uživatele na data. Zmíněny jsou také problémy způsobené použitím formátu PDF a technologie Adobe Acrobat.

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

Klíčová slova: digitalizace, digitální replika, navigace, hypertext, Acrobat, CD-ROM, encyklopedie, Jan Otto

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

1 Motivation

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

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 the project of a digital encyclopædia CD-ROM production.

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

2 Otto encyclopædia digitization project

We were participating in the design and retypesetting phases of a project aimed on creation of digital replica of Czech encyclopædia "Ottův slovník naučný" [5] (OSN; 28 volumes, almost 29 000 pages, about 170 000 entries, 4888 illustrations) published in 1888–1908. In the text authoring 1046 specialists were involved. The encyclopædia is about four times bigger than Nouveau Larousse illustré published during 1896–1907 in France.

The republishing project has been realized during 1996–1998 to reserve the informations collected for future generations. Similar project of digitization of another Czech encyclopædia—"Ottův slovník naučný nové doby" [6] (OSNND, 12 volumes, about 11 000 pages) has been just finished, using the experience from work on OSN. OSN and OSNND are the biggest Czech encyclopædiæ ever published and are considered as the most valuable sources of general information and knowledge useful not only for historicians.

As a format for electronic delivery of digitized encyclopædia Adobe's Portable Document Format [1] (PDF) was chosen, because achieving of highest visual fidelity with respect to the original was a high priority. This format has been successfully used for WWW and electronic journal publishing [8]. Several free readers of PDF documents are available (Acrobat Reader, xpdf, Ghostview or DocuReader [10]).

Full scanning and retypesetting of 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 experience with.

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

3 Navigation methods—document views

The reader of huge electronic document needs navigation help. We distinguish several types of reader's "view" of a document. There is logical structure of 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. in multicolumn layout. This represents the *physical view* on a document. A third point of view may be based on the reader's particular interests—he wants to collect related information about particular topic of interest and relations between document parts and 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—localisation of an item in the huge bookmark tree using the mouse is nearly impossible; new method for logical structure visualisation has to be used. Characters used in Acrobat bookmarks are restricted to `PDFDocEncoding` [1] only, which is not sufficient (there are not all characters used in the Czech language available).

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

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

Let's take an example of our encyclopædia. The tree of a document structure (28 volumes with about 5 000 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 by just l mouse clicking.

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 its proper order. As printing is usually demanded by user for pages of interest and one to one reproduction is preferred to allow readers to have the look and feel of the original pages of interest. Adding navigation buttons and areas to get an 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 on 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 in doing that kind of document classification on Internet).

In our application we have identified the need to generate "see also" links based on a computed measure of "similarity" based on semantic map of the Czech language using an approach motivated by [7]. 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 tree-structured semantic net of an encyclopædia—semantic view. It can be visualized in the same way as 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." *Donald Knuth*

4 Production

Usage of Adobe's Acrobat Capture program was found inappropriate because customisation 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 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 batch mode of processing. Some transformations needed human intervention; several interactive special purpose programs were used for instance for spellchecking (mainly as Visual Basic macros) and for markup validation and correction.

This semiautomatic transformation has lead to fully featured tagged data embodying both logical and visual document 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

Typesetting system TEX was chosen for retypesetting in distributed, heterogeneous 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 got stuck. The very first attempts to use WYSIWYG typesetting programs were disastrous. We benefitted from large CTAN (Comprehensive TEX Archive Network) database of free fonts, special purpose macros and programs.

The standard way of producing PDF from TEX is going via 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 Thế Thành [2, 3]—were done, allowing for very compact PDF files. It also allows high degree of reuse of document parts in an object oriented manner of PDF. pdfTEX, however, still lacks support of some PDF features like bitmap font support that we needed in our application; we ended finally using Distiller leaving pdfTEX as option for the next release.

TEX macros allowed full automatization of typesetting of navigational document (more than 9000 screen pages with alphabetically sorted keyword index), as well as cross references between the document cores in several PDF files. The process of creation of Acrobat's articles was fully automated too. This has been accomplished with 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 showed that bottom up approach to get fully tagged data for retypesetting of large volumes of text is feasible. The whole OSN encyclopædia fits on a single CD-ROM and 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 the user convenience. This made second CD-ROM with index and the like. CD-ROM has got very positive feedback from its readers. High degree of automation proved possible; manual work is necessary, though, to meet all requirements. Special attention has to be given to searching and navigation tools, allowing reader's different document views and digestion. Using document design for paper edition for electronic document delivery is problematic, but can be partially eliminated by offering different document views to the user, taking into account 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 variable letter width technique [4] applied to multiple master fonts to achieve more uniform greyness in the text than in the original print.

6 Acknowledgement

We acknowledge enthusiasm, consideration and care of all people we met participating our bit in the design of above mentioned projects. Support for preparation of this report under the grant CEZ:J07/98:143300003 is acknowledged.

References

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

[3] Hàn Thế Thành. pdfTEX distribution, available from `ftp://ftp.cstug.cz/pub/tex/local/cstug/pdftex`.

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

[5] Jan Otto. Ottův slovník naučný. 28 volumes, Prague, 1888–1908.

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

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

[8] 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.

[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`.