Michal Růžička

Automated Processing of TeX-Typeset Articles for a Digital Library

In: Petr Sojka (ed.): Towards Digital Mathematics Library. Birmingham, United Kingdom, July 27th, 2008. Masaryk University, Brno, 2008. pp. 167--176.

Persistent URL: http://dml.cz/dmlcz/702533

Automated Processing of TEX-Typeset Articles for a Digital Library

Michal Růžička

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

Abstract. Experience in setting up a comprehensive journal processing system based on the TEX typesetting system with the CEDRAMworkflow is described, following the example of the Archivum Mathematicum journal. The system automates the preparation of issues and simultaneously generates the materials needed for the Czech Digital Mathematics Library project (DML-CZ). The second part of the article describes the process of transformation of archival born-digital articles into a DML-CZ-suitable format.

Key words: TEX, publishing system, digital mathematical library, DML-CZ, metadata

1 Introduction

Since 2005, a digital mathematics library has been under development in the Czech Republic. The goal of the Czech Digital Mathematics Library project (DML-CZ) [1,2,3,4,5] is the preservation in digital form of the contents of the major part of mathematical literature ever published in the Czech lands, and to provide free access to the digital content and bibliographical data. [6]

From a viewpoint of the contents of the digital library, there are three main periods of time that must be addressed within the digital library project.

1. A retro-digitization period — Documents that are available only in paper format and must be digitized for the needs of a digital library.
2. A retro-born-digital period — Documents that are already born-digital but have been made without awareness of the digital library. The format of these documents is often not suitable for the needs of the digital library.
3. A born-digital period — Documents that are born-digital and made in such a way as to meet the needs of both the publisher and the digital library.

This article discusses the processing of the Archivum Mathematicum journal [7] in order to acquire both the retro-born-digital and born-digital data needed for the DML-CZ project.

Petr Sojka (editor): DML 2008, Towards Digital Mathematics Library, pp. 167–176.
 ISBN 978-80-210-4658-0

2 The Retro-Born-Digital Period of the Archivum Mathematicum Journal

The Archivum Mathematicum journal has been published using $\mathcal{AMS}$-TEX and LATEX since 1992. During this period, there have been several changes in style files and the initial mixture of the $\mathcal{AMS}$-TEX and the LATEX sources nearly became a LATEX `amsart.cls` monoculture.

Since 1992, there have also been some changes in the editorial staff of the journal. Consequently, it has not been possible to collect all the source codes for all the issues, something that further complicated the whole task.

2.1 Extraction of Bibliographical Metadata

It was especially necessary to collect bibliographical metadata for the DML-CZ project, more specifically, the list of references from every article that included one. Further metadata about the articles and issues were already available from other sources.

Differences Between $\mathcal{AMS}$-TEX and LATEX Sources. It has already been mentioned that the format of the source codes of the articles was not homogeneous and varied not only from issue to issue but also among the articles within one issue. In general, there were two major formats (each containing roughly 50% of the articles) — articles written using $\mathcal{AMS}$-TEX and articles written using LATEX document class `amsart.cls`. Over the course of time, there was a trend towards the latter.

In addition to the necessity for a slightly different process of metadata extraction, there was one major difference between the two groups of articles — $\mathcal{AMS}$-TEX has a group of logical macros for bibliography marking, so that it was possible to preserve semantic information of all bibliography items even on the output.[1] This contrasts with the LATEX `thebibliography` environment, which contains visual but not logical markup.

Preprocessing of Articles. The internal format of DML-CZ metadata is XML. Therefore it was desirable to extract metadata from the original TEX format directly into XML.

A very good tool for transforming LATEX documents into XML is the Tralics program [8,9]. Tralics is a *LATEX* translator. It was therefore necessary to perform some preprocessing of the $\mathcal{AMS}$-TEX articles. Inasmuch as only article bibliographies were extracted, the LATEX articles were also preprocessed in a similar way in order to prepare input files in the LATEX format containing only the bibliography.

[1] Regrettably, not all authors used these macros properly and a non-negligible portion of the $\mathcal{AMS}$-TEX articles had items such as publisher, year of publication, and so on marked using a common `\paperinfo` macro without further structuring.

For both $\mathcal{AMS}$-TEX and LATEX articles, scripts (in the language of the `ex` program[2] in that case) were prepared. Those scripts transform the source code of a regular $\mathcal{AMS}$-TEX/LATEX article into a minimal LATEX document that is ready for further processing by Tralics. The workflow can be seen in Figure 1. The following is an example of a minimal LATEX document derived from a $\mathcal{AMS}$-TEX article:

```
\documentclass{archivum}
\begin{document}
  \Refs
  \ref\key1\by Gancarzewicz, J., Michor P. W.\paper Natural...
  \endref
  \ref\key2\by Zajtz, A.\paper On the order of natural...
  \endref
  ...
  \endRefs
\end{document}
```

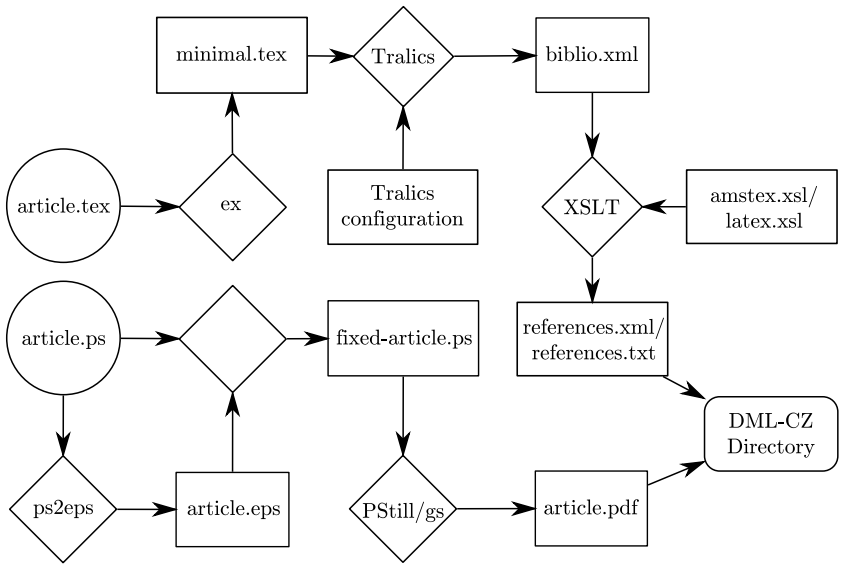

Fig. 1. Schema of the Archivum Mathematicum retro-born-digital period workflow

[2] The `ex` program is a part of the widespread `vim` text editor.

Conversion of LaTeX Sources into XML by Tralics. The minimal LaTeX document mentioned above is ready for further processing by Tralics. It was again necessary to prepare two different configurations for the $\mathcal{AMS}$-TeX and LaTeX sets of bibliographical macros. These configuration files instructed Tralics in the translation of input TeX macros into output XML document.

In order to make the Tralics configuration as simple as possible, the Tralics configuration files were made in such a way as to produce 'neutral' XML output containing just logically marked bibliographical data reflecting the original $\mathcal{AMS}$-TeX markup (in the case of articles originally written in the $\mathcal{AMS}$-TeX language).

The Tralics configuration files contained a special definition of the $\mathcal{AMS}$-TeX bibliographical macros using Tralics-specific commands. The bibliographical macros defined in this way took their arguments and enclosed them by an XML element of the name of the original macro in the output.

The translation of the 'neutral' XML files into the desired final XML format was performed using XSLT (see Figure 1 on the previous page). The following is an example of an output XML file:

```
<?xml version="1.0" encoding="UTF-8"?>
<references>
  <reference id="1">
    <prefix>[1]</prefix>
    <title>Natural...</title>
    <authors>Gancarzewicz, J., Michor P. W.</authors>
    ...
  </reference>
  <reference id="2">
    <prefix>[2]</prefix>
    <title>On the order of natural...</title>
    <authors>Zajtz, A.</authors>
    ...
  </reference>
...
</references>
```

The articles originally written in the LaTeX language did not contain any logical markup. The 'neutral' XML produced by Tralics therefore reflected visual markup rather than semantic structure. Thus XSLT was made to produce only a plain text output with minimal markup — every bibliographical record was divided into 'authors', 'title' and 'suffix' fields. As visual markup varied slightly between different authors and articles, this method was not firm enough and further human checking was necessary.

2.2 Conversion of Articles from PostScript into PDF

The DML-CZ digital library requires not only metadata about the articles but also the articles themselves in electronic form. Due to the changes in style files

and incomplete source codes, it was not possible to recompile all the articles. Even a small change in the output is strongly incompatible with the purposes of a digital library.

Fortunately, nearly all the articles of the retro-born-digital period were available as PostScript files. However, the form of these files was not fully suitable to the needs of the digital library. Desirable final format of the articles was PDF.

Automated Modifications of the PostScript Files. The first problem of the PostScript files was their BoundingBox — the smallest axis-aligned rectangle that entirely encloses all elements on the page. The PostScript files were incorrect in terms of both the BoundingBox and the paper format information. The position of the text on the page was also incorrect.

The sheer number of such articles dictated that the whole process of PostScript correction be automated. The BoundingBox of each PostScript file was detected by the `ps2eps` utility from the standard TEX Live distribution [10] and fixed within the PostScript file. Using the real BoundingBox value it was also possible to calculate a correct position for the text on the page. See Figure 1 on page 169.

Substitution of Bitmap Fonts. The second problem with the PostScript files involved embedded low-resolution bitmap fonts, which are unfavourable to the future needs of digital library users.

Bitmap fonts with a fixed resolution (300 DPI in this case) are appropriate to use at that particular resolution. However, compared with outline fonts, the visual quality of bitmap fonts tends to be poor when scaled or otherwise transformed. Nowadays, publications are printed using much higher resolutions, so low-resolution bitmap fonts are less suitable than outline ones. Moreover, publications in a digital library are very often read on a computer screen. Computer screens usually have much lower resolution than 300 DPI and electronic publications are frequently scaled on the screen. Thus, bitmap fonts are not appropriate even for this purpose. Therefore, several ways of exchanging original bitmap fonts for their outline alternatives were investigated.

All the PostScript files were made with the `dvips` program. There have been several changes in font embedding since 1992. Bitmap fonts with a resolution of 300 DPI were embedded in the older articles and outline fonts in the newer ones.

Some methods of font substitution are mentioned in [11]. However, the FixFont program [12] mentioned in the article did not succeed. Moreover, there was no helpful error message. The FontRep Adobe Acrobat plug-in [12], which was mentioned in [11] as well, is unavailable from the plug-in homepage and there is no contact information for its author.

Finally, the PStill program [13] was partially successful. PStill is able to substitute bitmap fonts in a dvips-produced PostScript file as a part of the conversion of this file into PDF. However, PStill depends on the presence of

the names of the fonts used in the comments in the PostScript code. Older versions of the `dvips` program did not include these comments. Therefore, not all PostScript files containing bitmap fonts could be substituted. The remainder of the articles were translated by the well-known GhostScript program-suite. See Figure 1 on page 169.

3 The Born-Digital Period of the Archivum Mathematicum Journal

The Czech Digital Mathematics Library is going to archive not only retro-digitalized and retro-born-digital publications but also new publications. Therefore, the preparation of a publishing system able to generate the materials needed for the digital library is required.

Nowadays, publishers often use some kind of an automated document workflow [14], frequently based on XML. XML formats allow users to separate visual representation of data from content. This is important not only for database publishing but also for deriving document metadata.

The TEX typesetting system is good at the separation of format from content as well. Thus a publishing workflow could easily be based on TEX and/or XML [15]. A combination of the TEX system and XML has also been chosen for the new processing system of the Archivum Mathematicum journal.

3.1 CEDRAM Base

The system of the Archivum Mathematicum journal is based on a system [16,17] used for the French CEDRAM project [18]. Journals collated around the CEDRAM project use a common processing system for the preparation of issues. This system is based on the TEX typesetting system and the Tralics LATEX to XML translator.

Both these components are also used in the new Archivum Mathematicum system. One part of the metadata for the DML-CZ project is the list of references from every article that included one. To avoid loosing semantic information, as did the retro-born-digital articles written using the LATEX `thebibliography` environment, the BibTex program is used for typesetting bibliographies.

Preserving the structure in bibliographical metadata allows us to provide functions such as searching in particular fields in bibliographies and grouping publications written by the same authors, as well as generating web pages with different visual representation for different parts of a bibliographical record. The structure of the bibliographical metadata corresponds to the structure of BibTex bibliography files. As BibTex is used for article preparation it is a very natural way to structure bibliographical metadata.

The Tralics program is a really important part of the system. The conjunction of Tralics and BibTex bibliographies is used for the generation of article metadata in an XML format. The Tralics ability to use BibTex bibliographical databases directly makes the generation of metadata considerably simpler.

3.2 Workflow of an Issue

From the user's point of view, the new system is not too different from the 'traditional' way of issue preparation. The `cedram.cls` classfile used is based on the `amsart.cls` classfile with only a slightly extended set of user macros. The `amsart.cls` origin of the CEDRAM classfile and only minimal changes in macro set made the transfer of the Archivum Mathematicum journal even simpler, in view of the fact that the `amsart.cls` file had been used in the past. The preparation of the articles is nearly the same as in the past and practically all the new actions are processed automatically.

The base of every issue under control of the new system is a 'driving' file and a set of independent articles in separated directories (see Figure 2 on the following page). The following is an example of a driving file:

```
\documentclass[AM,english,RedoBibTeX,Volume,Couverture,XML]{cedram}
% volume number, issue number, month, year
\IssueInfo{44}{1}{}{2008}
\SetFirstPage{1}

\begin{document}
\makefront
\articles
  \includearticle{article1}
  \includearticle{article2}
  ...
\makeback
\end{document}
```

All of the processing is driven directly from the classfile using the TEX `\write18` feature. This TEX command allows the user to carry out ordinal system commands directly from the TEX source code.[3] In this way, article metadata are translated into XML directly using Tralics (Step 2b in Figure 2 on the next page).

The compilation of the 'driving' file in the `pdflatex` program (Step 1 in Figure 2 on the following page) starts a huge set of automated actions. In general, all the articles are compiled independently. The compilation produces a PDF of each article (Step 2a) and these are subsequently merged into the final issue PDF file (Step 3). The cover of the issue in PDF format, which also contains an automatically generated table of contents, is produced as well (Step 4). A side-benefit of this compilation is the creation of metadata (Step 2b).

A big advantage of this way of processing articles is the complete isolation of each article. Any unwanted interference is eliminated.

[3] The `\write18` feature has to be explicitly enabled by specifying `-shell-escape` or similar option on the TEX command-line.

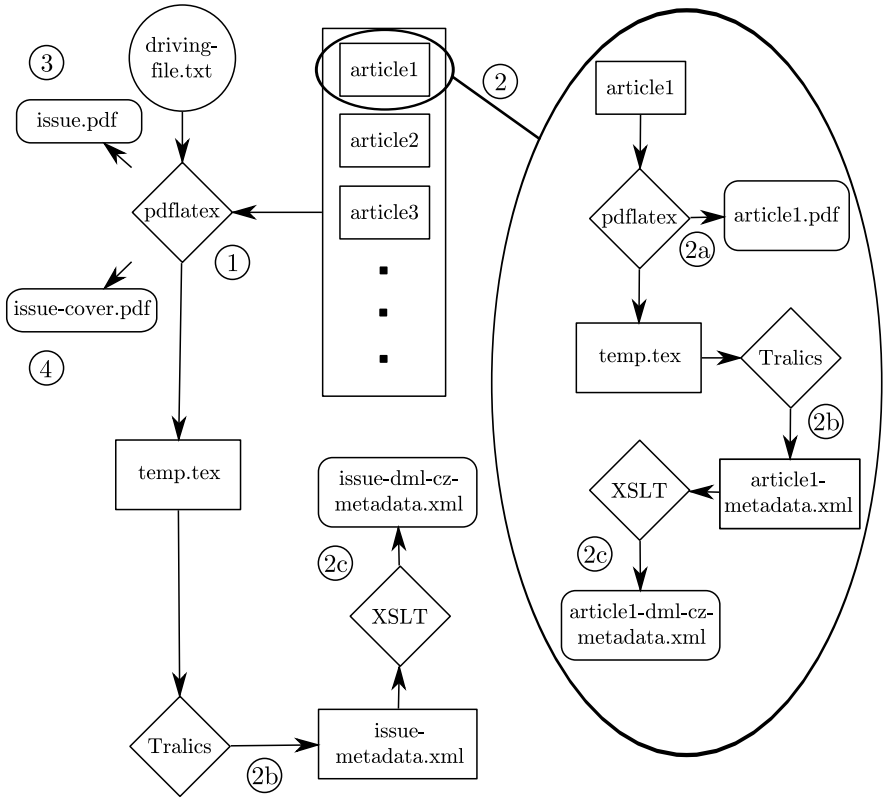

Fig. 2. Schema of the Archivum Mathematicum born-digital period workflow

3.3 Extensions to the CEDRAM Workflow

The new processing system of the Archivum Mathematicum journal contains some extensions to the CEDRAM workflow. The compilation of the issue and further actions are driven by the standard Unix `make` program using Makefiles.

Metadata on both the whole issue and every article included are automatically generated as a set of XML files by Tralics during the compilation of documents (Step 2b in Figure 2). The format of these XML files is nearly the same as in the CEDRAM workflow. The final XML files in DML-CZ format are subsequently created by XSLT (Step 2c).

The DML-CZ XML files, all the articles in the PDF format, and the plain text of these articles are subsequently packed into a common ZIP archive. This archive can be sent to the appropriate person directly by e-mail with the proper Makefile target.

The system also creates an electronic version of the Archivum Mathematicum journal. Web pages containing information about the articles are made

automatically using the issue metadata. These web pages and the articles in both the PDF and the PostScript format are then saved in a separate directory.

Furthermore, the system contains minor supportive tools such as the creation of a database review form and the generation of lists of 'suspicious' (unnaturally long) words in the source codes of the articles, which helps to reveal typing errors.

4 Conclusion

Since 2008, the Archivum Mathematicum journal has been using the new publishing system. The first issue of this journal was published using this system, and some other journals are considering using the system. The final goal is the preparation of comprehensive journal processing system based on the TEX typesetting system that would automate the preparation of issues and simultaneously generate the materials needed for the DML-CZ project.

In addition to the documents for print and metadata for the digital library, the system also generates supporting outputs for editorial staff. The web pages of the journal electronic edition are automatically generated as well as the database review form.

The transformation of the articles of the retro-born-digital period was another part of the project. Bibliographical metadata of the 1992–2007 period have been extracted and translated into DML-CZ metadata-rich format. The articles were automatically processed, corrected and saved in the form needed for the DML-CZ project.

Acknowledgement This research was supported by grant reg. no. 1ET200190513 of the Academy of Sciences of the Czech Republic.

References

1. Sojka, P.: From Scanned Image to Knowledge Sharing. In Tochtermann, K., Maurer, H., eds.: Proceedings of I-KNOW '05: Fifth International Conference on Knowledge Management, Graz, Austria, Know-Center in coop. with Graz Uni, Joanneum Research and Springer Pub. Co. (June 2005) 664–672.
2. Bartošek, M., Lhoták, M., Rákosník, J., Sojka, P., Šárfy, M.: DML-CZ: The Objectives and the First Steps. In Borwein, J., Rocha, E. M., Rodrigues, J. F., eds.: CMDE 2006: Communicating Mathematics in the Digital Era. A. K. Peters, MA, USA (2008) 69–79.
3. Sojka, P., Panák, R., Mudrák, T.: Optical Character Recognition of Mathematical Texts in the DML-CZ Project. Technical report, Masaryk University, Brno (September 2006) presented at CMDE 2006 conference in Aveiro, Portugal.
4. Bartošek, M., Krejčíř, V.: Jak se dělá digitální matematická knihovna. In Sborník konference AKP 2007, Liberec, Czech Republic (2007). Available from WWW: `http://dml.muni.cz/docs/akp2007-sbornik.pdf`.
5. Czech Digital Mathematics Library [online]. [cit. 2008-05-30]. Available from WWW: `http://dml.cz/`.

6. Czech Digital Mathematics Library: About DML-CZ [online]. [cit. 2008-06-22]. Available from WWW: http://dml.cz/about/.
7. Archivum Mathematicum [online]. Masaryk University, Brno. Last modified 14 May 2008 [cit. 2008-05-18]. Available from WWW: http://www.emis.de/journals/AM/.
8. Grimm, J.: Tralics, a LaTeX to XML Translator. In Proceedings of EuroTeX, TUGboat **24**(3) (2003) 377–388.
9. Tralics: a LaTeX to XML translator [online]. Last modified $Date: 2008/05/13 09:32:16 $ [cit. 2008-05-18]. Available from WWW: http://www-sop.inria.fr/apics/tralics/.
10. TeX Live@TeX Live [online]. $Date: 2008/05/17 00:21:31 $ [cit. 2008-05-25]. Available from WWW: http://www.tug.org/texlive/.
11. Probets, S., Brailsford, D.: Substituting outline fonts for bitmap fonts in archived PDF files. Software-Practice and Experience. **33**(9) (2003) 885–899.
12. Research — Fonts [online]. [cit. 2008-05-25]. Available from WWW: http://www.eprg.org/research/fonts/.
13. Siegert, F.: PStill: ...generate, reprocess, normalize and extract content for PDF, EPS and PS. [online]. [cit. 2008-05-25]. Available from WWW: http://www.pstill.com/.
14. Krell, H.: What's New With Springer Production? [online]. [cit. 2008-05-28]. Available from WWW: http://www.springer.com/societies?SGWID=0-40801-12-481803-0.
15. Interview of Kaveh Bazargan and CV Radhakrishnan -- co-directors of River Valley Technologies [online]. Interview completed 2006-09-20, $Date: 2006/06/28 23:30:02 $ [cit. 2008-05-28]. Available from WWW: http://www.tug.org/interviews/interview-files/river-valley.html.
16. Bouche, T.: A pdfLaTeX-based automated journal production system. In Proceedings of EuroTeX 2006, TUGboat **27**(1) (2006) 45–50.
17. Bouche, T.: CEDRICS: When CEDRAM Meets Tralics. (2008) In: Sojka Petr (editor): *DML 2008 – Towards Digital Mathematics Library*, Birmingham, UK, July 27th, 2008, pp. 153–165.
18. Centre de diffusion de revues académiques mathématiques [Center for diffusion of mathematic journals] [online]. [cit. 2008-05-25]. Available from WWW: http://www.cedram.org/.