Workflow in the Digital Mathematics Library Project

How Mathematics is Stored and Retrieved

Petr Sojka

Masaryk University in Brno, Czech Republic
sojka@fi.muni.cz http://www.fi.muni.cz/usr/sojka/

Abstract. This paper is a progress report of the retrodigitization project of the Czech Digital Mathematics Library, DML CZ. We are aiming to digitize valuable mathematical journals and books (250,000 pages) published in the Czech and Slovak Republics, and make them publicly available in digital form. We describe here the project work-flow: the key concept is a gradual enhancement of the digital material by 'knowledge enhancing' filters applied to the markup-rich XML data.

1 Introduction

The proposed WDML, World Digital Mathematics Library, [2] in a digital and easily accessible form is a dream of mathematicians over the world. In addition to the projects Cornell funded by NSF, EMANI: electronic mathematical archiving network, NUMDAM: Numérisation de documents anciens mathématiques, German digital research library funded by German Research Foundation and realized at Göttinger Digitalisierungs Zentrum, a DML-CZ: Czech Digital Mathematical Library project for the retrospective digitization of library materials of mathematical journals and books published in in the Czech and Slovak Republics (200–300,000 pages in total) is being realized [8]. We identify the following main steps in retrodigital document processing:

acquisition document acquisition, preparation, copyright issues handling;
scanning document scanning, main metadata entering, scanning checks;
image processing main OCR, image enhancements;
semantic processing document markup enhancement, semantic processing, document classification, citation linking, document clustering, indexing;
presentation visualization techniques of the document repository, digital library web portal, interfaces to other services and search engines.

In this overall architecture of processing, the *raw data* is transformed into *information* and ultimatelly, *knowledge*. We workflow is built on *extensible*, *open* formats (XML), while the key data processing on *extensible*, *open source* tools that gradually enhance and enrich the scanned data into a mathematical knowledge library of a new type. The best current practices of previous projects (NUMDAM, JSTOR and the Digitization of the Otto Encyclopædia [7]) are being followed so as not to reinvent the wheel.

The structure of this paper is as follows: Section 2 describes the scanning phase and digital storage issues. We describe the technique of gradual markup enrichment in Section 3. The closing remarks in Section 4 deal with the organization, presentation and delivery of the digitized material.

2 Scanning and Image Processing

Scanning at 600 DPI and 4-bit depth is taking place in Jenštejn near Prague, in the Digitization Center of the Library of Academy of Sciences using two A2 book scanners Zeutschel OS 7000. The original TIFF files are archived before entering the production line. The scanning process itself is inexpensive—pure scanning costs are reported at about ten percent of the whole page processing price [2] thanks to the high degree of automation.

With the Book Restorer software, dozens of image corrections are being made in automated overnight jobs. These include: adjustments to the histogram, binarization, deskew, despeckle, geometrical correction and lighting correction (1D & 2D). Then the data are fed into Sirius, which was developed for the National Library in Prague by Elsyst Engineering. The main metadata (page numbers) are entered and page numbers scanned. The system allows export in XML format with links to the normalized images.

From Image to Text with Visual Markup Documents consisting of a host of images are not much useful unless a full text layer has been created by optical character recognition (OCR) techniques. The OCR program, ABBYY Finereader/PDF Transformer, can provide the text of a document with *visual markup* as a text-under-image searchable PDF or HTML. The text is encoded in Unicode. The links between the textual and visual layers of a document are stored and preserved for further document processing, as XML encoded data.

3 Document Markup Enhancement

Scanned text with *visual markup* has to be converted into a *logical markup* to enable high precision search techniques. This is a difficult step, often ambiguous, reverting the process of typesetting. To face that, as most mathematics papers are typeset with the TEX engine, TEX typesetting rules and fonts have to be taken into account so as to enrich document objects with a *structural markup* in MathML. As MathML (XML namespace), allows for the storage of both *presentation/visual* (e.g. TEX) and *logical/content markup*, it is an ideal format for storing both layers of information. We are currently testing the Infty system [9] for its structural recongnition of mathematics, and if successfull, it will be merged with FineReader API.

Structure Markup Enrichment The structure markup of document paper can be flat (marking only the key parts of a paper such as title, author, abstract), or it can be more detailed. The level of detail can be specified or enforced by a *Document Type Definition*, DTD, or similar formalisms such as XML Schema, which allow even more detailed type checking.

For many markup tasks, regular expressions can be developed and used [7]. Even more demanding tasks such as the identification and markup of bibliographic entries in a document are performed by smart regular expressions matching various citation styles as shown in CiteSeer [3] or ACM Digital Library projects. Similarly, eXstyles by Inera allows the definition of thousands of rules for the semi-automated editing of

poorly marked-up (Word) files. In our workflow, we have sets of regular expressions to locate the main metadata (title, authors, etc.) in the scanned OCR.

The main mathematical databases of reviews and abstracts as Zentrallblatt MATH and MathSciNet already contain most needed metadata (at least for journals)—they are collected as the first iteration of metadata.

An application for metadata editing is being developed at the Institute of Computer Science, Masaryk University, which pairs scanned images with OCR texts, trying to locate available metadata against them. A metadata copyeditor sees all these data in WYSIWYG form, and makes the necessary editing of mandatory metadata fields. After this step, both the metadata and the OCR text with partial XML markup (e.g.bibliographic entries) are saved.

There are numerous ways that lead to markup enhancement, so for every document object, several versions of markup richness should be stored in a repository, allowing the building of pipes of programs to further enhance the markup. Today's tools and technologies of *corpora management* allow for the effective handling of a corpus, the size of the entire published mathematical literature to date.

The problem of *language identification* is an example of a markup enrichment filter. A relatively easy task is identifying the language of document chunks (paragraphs). It is easily extensible, because only bigram statistics of a new language have to be computed. This way, the source language tag for every paragraph or sentence can be added automatically.

Data Storage, Indexing, Semantic Processing All digital documents should be efficiently and effectively stored in a *digital repository system* to ensure

- a platform for digital data enhancement processing,
- the long-term unambiguous identification and preservation of digital material,
- open access-friendly *digital rights management* (DRM).

Currently, the data are loaded into the open source system Kramerius. Various modules for it are being developed (indexing, PDF generation on the fly).

The software developed at the NLPlab at the Faculty of Informatics, Masaryk University has superior scaling capabilities for indexing, querying, browsing and statistical computations of textual corpora: the corpus manager Manatee and it's graphical user interface Bonito. This system is capable of efficient storage, indexing and processing of billions of words. It will be used for a full-text search across the entire repository.

Today, *natural language processing* technologies make it possible to process digital documents not only on the level of syntax, but semantic processing is also needed to achieve the ideal of the *Semantic Web* for which many problems have yet to be solved. XML languages, such a RDFS in RDF or DAML+OIL, provide a means of machine interpretable document semantics, compatible with our model of linked document layers based on XML.

Mathematics Subject Classification Scheme (MSC) was compiled by the Editorial Offices of Mathematical Reviews (MR) and Zentralblatt MATH (Zbl) and is widely accepted by the publishers of mathematical journals. These are the main sources leading towards the creation of a taxonomy of mathematics (referred to as *ontology* in semantic web jargon) to which documents will be linked.

4 Identification, Visualization and Dissemination

All the document object parts accumulated by intelligent document processing should be stored and linked together. It is essential to have an *open access interface* at least to the archive metadata available for citation indexes and search engines. It should also be accompanied by a *document object identifier* (DOI) or persistent URL (PURL). The expanding use of DOI for scientific data is evident, and the DOI system is on track to becoming an ISO standard.

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) will be used to allow search engines and people on the Web to harvest stored metadata.

DjVU For document delivery, tagged text-under-image format PDF (or PDF/X for printing) such as a rich media container can be *generated* from primary sources, containing layers of scholar's interests. In addition, DjVU format in it's searchable form will serve for image low-bandwidth dissemination as well.

The success of tools based on the TouchGraph engine [6] such as Amazon or Google browsers in the style of WebODAV [1] inspired us to use a similar approach for handling digital library visualization and presentation [4]. Metadata, classification links and relations are stored in the KAON database [5] as an RDFS. It has been verified that the amount of metadata of the whole mathematics literature worth archiving (estimated at about only 50 million pages) is achievable and could be visualized on a modest workstation, even with a rich set of RDF data and document descriptions.

Closing Remarks We have designed an architecture of, and a methodology for building a fully fledged mathematics archive. Semantic enhancements filtering, metadata linking and visualization play a major rôle in the architecture. The research has been supported by the Czech National Programme *Information Society*, Grants No. 1ET208050401, 1ET200190513 and 1ET100300419.

References

1. Mao Lin Huang, Peter Eades, and Robert F. Cohen. WebOFDAV: Navigating and Visualizing the Web On-line with Animated Context Swapping. In *Proceedings of the 7th International WWW Conference*, pages 638–642, 1998.
2. Allyn Jackson. The Digital Mathematics Library. *Notices of the AMS*, 50(4):918–923, 2003.
3. Steve Lawrence, C. Lee Giles, and Kurt Bollacker. Digital Libraries and Autonomous Citation Indexing. *IEEE Computer*, 32(6):67–71, 1999.
4. Zuzana Nevěřilová and Petr Sojka. XML-Based Flexible Visualisation of Networks: Visual Browser, 2005. Submitted.
5. Daniel Oberle, Raphael Volz, Boris Motik, and Steffen Staab. An extensible ontology software environment. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 311–333. Springer-Verlag, 2004.
6. Alexander Shapiro. TouchGraph LLC at SourceForge, 2005.
7. Petr Sojka. Publishing Encyclopædia with Acrobat using TEX. In *Towards the Information-Rich Society. Proceedings of the ICCC/IFIP conference Electronic publishing '98*, pages 217–222, Budapest, Hungary, April 1998. ICCC Press.
8. Petr Sojka. From Scanned Image to Knowledge Sharing. pages 664–672, Graz, Austria, June 2005. Know-Center in coop. with Graz Uni, Joanneum Research and Springer Pub. Co.
9. Masakazu Suzuki, Fumikazu Tamari, Ryoji Fukuda, Seiichi Uchida, and Toshihiro Kanahori. INFTY — An integrated OCR system for mathematical documents. pages 95–104, Grenoble, France, 2003. ACM.