From Scanned Image to Knowledge Sharing

Formats and Technologies in the Digital Mathematics Library Project

Petr Sojka

(Masaryk University in Brno, Czech Republic
`sojka@fi.muni.cz`)

Abstract The main obstacle to easy accessing the vast amount of knowledge is the fact that they are not available in well-designed, standard, fully indexed electronic form, together with detailed metadata and full-text search capabilities.

This paper is a case study of design issues in a subproject of WDML (World Digital Mathematics Library) aimed at digitizing valuable mathematical journals and books published in the Czech and Slovak Republics, to make them publicly available in digital form. We discuss here the design of the work-flow aiming at having mathematical knowledge stored in digital library. The key concept is a gradual enhancement of the digital material by 'knowledge enhancing' filters applied to the markup-rich XML data.

Key Words: digital library; metadata handling; semantics of mathematical documents; knowledge management; digitization; MathML; visualization; portal-systems; repositories of knowledge; DML-CZ

Category: H.3.7, H.5, H.3, H.4

> We dreamed of making the incredible breadth of information that librarians so lovingly organize searchable online.
> —Larry Page, founder of Google

1 Introduction

The main obstacle to easy accessing the vast amount of knowledge is the fact that they are not available in well-designed, standard, fully indexed electronic form, together with detailed metadata and full-text search capabilities. A vast amount of valuable material remains in paper-only format. It is clear that the exploitation of information published on paper is severely limited by the fact that most of the books and journals, and even teaching materials, stay in stone libraries only, while academics have found that online publications have much greater impact [Lawrence, 2001].

Google has recently offered to pay $150 million for the digitization of Harvard, Stanford, Oxford and University of Michigan libraries, plus the New York Public Library, and to have them indexed in the Google Print project[1]. This may have far-reaching implications, some of which are drawbacks:

- handling rare and local publications not stored in those libraries is not an issue;

– nothing is known about handling *semantics* and *mathematics*—up to now word form indexing with sophisticated PageRank algorithm is used, without using more intelligent *semantic web* technologies.

Several initiatives are currently undertaken to build WDML, World Digital Mathematics Library[2], [Jackson, 2003] and there are already several finished projects (Cornell[3] funded by NSF; DIEPER[4]—Digitized European Periodicals) and the following ongoing digitization projects:

EMANI electronic mathematical archiving network[5],

NUMDAM Numérisation de documents anciens mathématiques[6],

German digital research library funded by German Research Foundation and realized at Göttinger Digitalisierungs Zentrum[7], and

DML-CZ Czech Digital Mathematical Library.

The key to the wide utilization of the knowledge contained in all documents created in the above projects, as well as in projects handling *born-digital* documents (JSTOR[8], CiteSeer[9] etc.) is the format and structure of its digital form. The effective creation and exchange of metadata and interfaces for handling text and graphics scanned from paper, together with the methods and algorithms for marking, classifying and delivering these terabytes or pentabytes of data are crucial problems to be solved when preparing digital libraries of research and teaching materials.

The structure of this paper follows the problems tackled on the long path from scanned image of a page to the scholar benefitting from the knowledge described therein. In [Section 2], the objectives of the DML-CZ project and its main phases are described. [Section 3] describes the scanning phase and digital storage issues. We describe the technique of gradual markup enrichment in the [Section 4]. [Section 5] discusses current methods of semantic processing of digital data. Closing remarks in [Section 6] deal with the organization, presentation and delivery of digitized material.

2   DML-CZ

DML-CZ (2005–2009) is a project for the retrospective digitization of library materials of mathematical journals and books published in Central Europe (in the Czech and Slovak Republics). We have identified the main steps of 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 document repository, digital library web portal, interfaces to other services and search engines for the semantic database document.

In this overall architecture of processing the *raw data* is transformed into *information* and finally, *knowledge*. We build this workflow on *extensible*, *open* formats (XML), and key data processing on *extensible*, *open source* tools that gradually enhance and enrich the scanned data into information and finally a mathematical knowledge library of a new type.

3 Scanning and Image Processing

The best current practices[10] of previous projects (WDML, JSTOR, Digitization of Otto Encyclopædia [Sojka, 1998]) are being followed so as not to reinvent the wheel. The scanning process is cheap—scanning costs are reported at about ten percent of the whole page processing price [Jackson, 2003]. TIFF, PDF or DjVU[11] formats may serve for image archival and dissemination.

3.1 From Image to Text with Visual Markup

Documents consisting of a host of images are not useful without a full text layer created by optical character recognition (OCR) techniques. OCR programs such as ABBYY[12] Finereader/PDF Transformer can provide the text of a document with *visual markup* as text-under-image searchable PDF or HTML. The text should be encoded in Unicode, and the viewing tools should be able to use the free fonts from the STIX[13] project, that cover the comprehensive set of fonts used in mathematics and other scientific literature. Linking between the textual and visual layers of a document should be stored and preserved for further document processing.

3.2 Data Storage

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).

Latest reports [Smith et al., 2004] show that the DSpace™ open source system seems to be meeting our demands and is an ideal candidate for effective, safe and long-term storage of and access to digital data. Another option is to use home-grown software above BerkeleyDB database[14].

Superior scaling capabilities has the software developed at the NLPlab at the Faculty of Informatics, Masaryk University for indexing, querying, browsing and statistics computations of textual corpora: corpus manager Manatee[15] and it's graphical user interface Bonito[16]. This system is capable of efficient storage and processing of hundreds of million words forms.

4 Document Markup Enhancement

Extensible markup language (XML) is widely accepted as the format of choice for the future millennium. Scanned text with *visual markup* has to be converted into *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 [Knuth, 1986], TEX typesetting rules and fonts [Padovani, 2003] have to be taken into account to enrich document objects with a *structural markup* in MathML[17]. 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.

4.1 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[18], which allow even more detailed type checking.

There are successful examples of digital archives with a structure rich markup such as PubMed Central[19], the free digital archive of biomedical and life sciences journal literature. They have developed a DTD suite for journal archiving and exchange, together with tools[20] to handle the tagged documents.

For many markup tasks standard regular expressions can be developed and used [Sojka, 1998]. 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 Cite-Seer [Lawrence et al., 1999] or ACM Digital Library[21] projects.

There are numerous ways leading 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[22] and technologies of *corpora management* allow for the effective handling of a corpus the size of the entire published mathematical literature to date. Methods of structure identification from OCR imaging have also been developed [Taghva et al., 1998], together with interactive spelling correction of OCR errors [Taghva and Stofsky, 2001].

The problem of *language identification* is an example of a markup enrichment filter. A relatively easy task is document chunks (paragraph) language identification [Dunning, 1994]. It is easily extensible, because for a new language 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.

4.2   Born-Digital Documents

The final form of born-digital and retro-digitized documents should ideally be the same to allow uniform access and handling schemes for scholarly use.

Most new mathematical documents are being prepared in some TeX macro-package as LaTeX or AMSLaTeX. Most publishers of scientific literature have adapted their workflow so that rich-structure SGML or XML versions are created as part of their publishing workflow for long-term storage [Bazargan, 2004]: it is reported that the translation filters are written using free software only.

As part of MoGWLI[23] (Mathematics on the Web: Get It by Logic and Interfaces) project, several tools for the automatization of MathML documents processing have been developed.

Project helm[24] (Hypertextual Electronic Library) presents a *content-centric* architectural design of an electronic library that allows its inner detailed structure of mathematical documents to be employed.

5   Semantic Processing

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 *Semantic Web*[25] for which many problems will have to be solved. XML languages, such a RDFS [RDF, 2004a] in RDF [RDF, 2004b] or DAML+OIL[26], provide a means of machine interpretable document semantics, compatible with our model of linked document layers based on XML.

5.1   Document Classification and Clustering

Mathematics Subject Classification Scheme (MSC[27]) was compiled by the Editorial Offices of Mathematical Reviews (MR) and Zentralblatt MATH (Zbl) and

is widely accepted by the publishers of mathematical journals. Currently, the MSC2000 version is used. These are the main sources towards the creation of a taxonomy of mathematics (in semantic web jargon called *ontology*) to which documents will be linked.

There are technologies for document clustering, indexing and retrieval such as Vivismo[28] or Verity[29]. These allow an automatic hierarchical structuring of documents in the Yahoo style, or a computation of document similarities [Salton et al., 1994].

6 Identification, Indexing, Visualization and Dissemination

All the document object parts accumulated by intelligent document processing should be stored and linked together. The process of building smaller (university, government departments[30]) *information bridges* can be seen on the Internet. Having an *open access interface* at least to the archive metadata available for citation indexes and search engines is essential. It should also be accompanied by a *document object identifier* (DOI[31]) or persistent URL (PURL). The expanding use of DOI for scientific data [Paskin, 2004] is evident, and the DOI system is on track to becoming an ISO standard.

A major benefit to the user is when paper references in bibliography lists are hyperlinked within the archive (as in CiteSeer, or Google Scholar[32]) or the whole Internet. The main review and abstracts mathematical databases as Zentrallblatt MATH[33] and MathSciNet[34] are trying to achieve that goal. Algorithms to do autonomous citation indexing have already been developed [Lawrence et al., 1999], but a vast collection of metadata of the papers is needed.

The Open Archives Initiative Protocol for Metadata Harvesting[35] (OAI-PMH) will be used to allow search engines and people on the Web to harvest stored data.

6.1 Presentation

For document delivery, tagged text-under-image format PDF (or PDF/X[36] for printing) such as a rich media container can be *generated* from primary sources, containing layers of scholar's interests. It may be superseded by another format in the future—Adobe is pushing a new version of PDF specification [Adobe Systems Incorporated, 2004] in about every eighteen months, so it is essential to have all material stored in an open and extensible way together with tools that can handle them (e.g. XML+XSLT), and generate a delivery format on demand, in accordance with the actual standards and user preferences. For an even wider dissemination of documents, their compliance with Section 508 of the Rehabilitation Act[37] is a good strategy.

6.2   Visualization

The main bottleneck in electronic exchanges of knowledge is at the recipients end. 'Lost in hyperspace' fear is widespread among web users; it is estimated that only one percent of papers are actually read cover to cover. The importance of the organization and presentation of information in digital library systems is underestimated [Dreher et al., 2004].

Methods for visualization and navigation [Tufte, 1990] in steadily growing digital data on the Internet are badly needed [Geroimenko and Chen, 2003]. The success of tools based on the TouchGraph engine [Shapiro, 2005] such as Amazon[38] or Google browsers[39] in the style of WebODAV [Huang et al., 1998] inspired us to use a similar approach for handling digital library visualization and presentation [Nevěřilová and Sojka, 2005]. Metadata and classification links and relations are stored in KAON database [Oberle et al., 2004] as 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 moderate workstation, even with a rich set of RDF data and document descriptions.

> We should experiment; we should try out new things;
> we should tinker with technology and find better ways to communicate.
> —John Ewing [Ewing, 2002]

7   Closing Remarks

We have designed an architecture of and methodology for building a fully fledged mathematics archive. Semantic enhancements filtering, metadata linking and visualization play a major rôle in the architecture. We have discussed using current standards in several digital document processing steps with extensive use of current XML technologies. We argue for a wide use of extensible formats and open source software for the implementation of core functionality of a digital document repository of research and educational materials.

Acknowledgement

References

[RDF, 2004a] (2004a). RDF Vocabulary Description Language 1.0: RDF Schema. http://www.w3.org/TR/rdf-schema/.
[RDF, 2004b] (2004b). RDF/XML Syntax Specification. http://www.w3.org/TR/rdf-syntax-grammar/.

[Adobe Systems Incorporated, 2004] Adobe Systems Incorporated (2004). *PDF Reference: Adobe Portable Document Format Version 1.6.* Adobe Press, fifth edition.

[Bazargan, 2004] Bazargan, K. (2004). LaTeX to MathML and back: A case study of Elsevier journals. In *Proceedings of PracticalTEX2004.* TUG.

[Dreher et al., 2004] Dreher, H., Krottmaier, H., and Maurer, H. (2004). What we Expect from Digital Libraries. *Journal of Universal Computer Science*, 10(9):1110–1122. http://www.jucs.org/jucs_10_9/what_we_expect_from.

[Dunning, 1994] Dunning, T. (1994). Statistical identification of language. Technical Report MCCS 94-273, New Mexico State University, Computing Research Lab.

[Ewing, 2002] Ewing, J. (2002). Predicting the Future of Scholarly Publishing. pages 52–58. Springer-Verlag.

[Geroimenko and Chen, 2003] Geroimenko, V. and Chen, C. (2003). *Visualizing the Semantic Web: XML-Based Internet and Information Visualization.* Springer-Verlag.

[Huang et al., 1998] Huang, M. L., Eades, P., and Cohen, R. F. (1998). WebOFDAV: Navigating and Visualizing the Web On-line with Animated Context Swapping. In *Proceedings of the 7th International WWW Conference*, pages 638–642.

[Jackson, 2003] Jackson, A. (2003). The Digital Mathematics Library. *Notices of the AMS*, 50(4):918–923.

[Knuth, 1986] Knuth, D. E. (1986). *The TEXbook*, volume A of *Computers and Typesetting.* Addison-Wesley, Reading, MA, USA.

[Lawrence, 2001] Lawrence, S. (2001). Online or invisible? *Nature*, 411(6837):521.

[Lawrence et al., 1999] Lawrence, S., Giles, C. L., and Bollacker, K. (1999). Digital Libraries and Autonomous Citation Indexing. *IEEE Computer*, 32(6):67–71.

[Nevěřilová and Sojka, 2005] Nevěřilová, Z. and Sojka, P. (2005). XML-Based Flexible Visualisation of Networks: Visual Browser. Submitted.

[Oberle et al., 2004] Oberle, D., Volz, R., Motik, B., and Staab, S. (2004). An extensible ontology software environment. In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, International Handbooks on Information Systems, chapter III, pages 311–333. Springer-Verlag.

[Padovani, 2003] Padovani, L. (2003). *MathML Formatting.* PhD thesis, University of Bologna.

[Paskin, 2004] Paskin, N. (2004). Digital Object Identifiers for scientific data. In *Proceedings of 19th International CODATA Conference*, Berlin.

[Salton et al., 1994] Salton, G., Allan, J., and Buckley, C. (1994). Automatic structuring and retrieval of large text files. *Communications of the Association for Computing Machinery*, 37(2).

[Shapiro, 2005] Shapiro, A. (2005). TouchGraph LLC at SourceForge.

[Smith et al., 2004] Smith, M., Rodgers, R., Walker, J., and Tansley, R. (2004). DSpace: A Year in the Life of an Open Source Digital Repository System. In Heery, R. and Lyon, L., editors, *Proceedings of ECDL 2004, LNCS 3232*, pages 38–44. Springer-Verlag.

[Sojka, 1998] Sojka, P. (1998). 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. ICCC Press.

[Taghva et al., 1998] Taghva, K., Condit, A., and Borsack, J. (1998). Autotag: A Tool for Creating Structured Document Collections from Printed Materials. In Hersch, R. D., André, J., and Brown, H., editors, *Lecture Notes in Computer Science 1375*, pages 420–431, Berlin, Heidelberg. Springer-Verlag.

[Taghva and Stofsky, 2001] Taghva, K. and Stofsky, E. (2001). OCRSpell: an interactive spelling correction system for OCR errors in text. *IJDAR*, 3(3):125–137.

[Tufte, 1990] Tufte, E. (1990). *Envisioning Information.* Graphics Press.

Sojka P.: From Scanned Image to Knowledge Sharing ...

Links

[1]http://print.google.com

[2]http://www.wdml.org

[3]http://www.library.cornell.edu/dmlib/

[4]http://dieper.aib.uni-linz.ac.at/

[5]http://www.emani.org

[6]http://www.numdam.org/

[7]http://gdz.sub.uni-goettingen.de/en/index.html

[8]http://jstor.org

[9]http://citeseer.ist.psu.edu

[10]http://www.ceic.math.ca/Publications/Recommendations/BPs.pdf

[11]http://www.djvuzone.org

[12]http://www.abbyy.com/

[13]http://www.stixfonts.org/

[14]www.sleepycat.com/

[15]http://www.textforge.cz

[16]http://nlp.fi.muni.cz/projekty/bonito/

[17]http://www.w3.org/Math/

[18]http://www.w3.org/XML/Schema

[19]http://www.pubmedcentral.nih.gov/

[20]http://dtd.nlm.nih.gov/tools/

[21]http://portal.acm.org/dl.cfm

[22]http://www.textforge.cz/products.html

[23]http://www.mowgli.cs.unibo.it

[24]http://helm.cs.unibo.it/

[25]http://www.w3.org/DesignIssues/Semantic.html

[26]http://www.daml.org/language/

[27]http://www.ams.org/msc/

[28]http://vivismo.com

[29]http://verity.com

[30]http://www.osti.gov/bridge/

[31]http://www.doi.org

[32]http://scholar.google.com

[33]http://www.zentralblatt-math.org/MATH/home

[34]http://www.ams.org/mathscinet

[35]http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

[36]http://www.pdfx3.org/

[37]http://www.section508.gov/

[38]http://www.touchgraph.com/TGAmazonBrowser.html

[39]http://www.touchgraph.com/TGGoogleBrowser.html