Digital Mathematics Library
Report of the Technical Standards Working Group

Date May 18, 2003

Thierry Bouche, Ulf Rehmann

Cochairs:

Thierry Bouche Grenoble	thierry.bouche@ujf-grenoble.fr
Ulf Rehmann Bielefeld	rehmann@mathematik.uni-bielefeld.de

Members:

Pierre Berard	Pierre.Berard@ujf-grenoble.fr
Jon Borwein	jborwein@cecm.sfu.ca
Keith Dennis	dennis@rkd.math.cornell.edu
Michael Doob	mdoob@cc.UManitoba.CA

Contents

This document gives technical recommendations to ensure the integration of digitized mathematical documents in a uniform "Digital Mathematics Library" (DML).

Since the digitized documents will be produced by various projects, possibly applying different methods and technologies, these recommendations define general technical standards in order to make the DML as a whole easily accessible, usable, maintainable, and sustainable.

A digitization project requires several procedures. The most critical tasks are the *scanning and archiving processes*, which are substantial for the quality and longevity of the data to be preserved. The scanning part requires most of the work, it cannot easily be repeated and should therefore be performed with greatest care.

Other tasks, like enhancing the data by OCR layers[1], annotations, metadata, and web links, could be either postponed or possibly redone, if later on more advanced technology becomes available.
The actual file formats or implementations mentioned here are presented as examples, which, at the time of this writing, can be used in order to achieve the proposed standards.

1 Scanning Quality

Recommendation:

- Scan quality:
  600 dpi bitonal minimum quality level
  300 dpi bitonal/gray-scale is discouraged

- In special cases and in the long run, higher resolutions, gray-scale, or even color may be more suitable.

  Explanation: Special projects may consider higher standards. For instance, manuscripts or old printed documents would certainly profit from gray-scale scanning or higher resolution. The general rule of thumb is that no significant data should be lost at this stage. A human eye must be able to read the scanned pages as easily as the original. Any further process will rely upon this scan quality. If storage volume and production costs permit, the higher the standards, the better the longevity of the data.

  The recommendations for scanning quality are compliant with - or higher than - the scanning standards given by the Digital Library Federation [5].

- Obvious flaws of the printing like skewed printing areas should be corrected during the scanning process.

- The printing area of each page should be positioned at the same place for all pages of a given object, possibly reflecting the differences for "right" and "left" pages. Page jumpings, rotations, varying margins and dimensions of images are discouraged.

  Note: a possible choice could be the approach chosen by Gallica [9]: Always put the text into the minimum ISO A format into which it fits.*

2 Archiving Formats

Recommendation:

[1] OCR = Optical Character Recognition

- Scanned raw data should be stored in an open format, possibly ASCII based, (to ensure longevity and sustainability) under lossless compression using public compression standards.

  Example formats are: pnm (portable anymap) [16], TIFF (Tagged Image File Format) [19]; example compression schemes are CCITT G4 [4] for bitonal, LZW [12], ZIP [21] for gray, color)

It is proposed that the archived raw data be made publicly available whenever possible. This will support distribution and hence longevity of the primary data. The more copies that exist, the greater will be the likelihood that data will survive unforeseen events.
At least the archived raw data should be made easily available for re-engineering when enhancements of any type are possible.

---

3 File Name and URL Conventions

The following should be guaranteed:

- unique and meaningful names for all files, displaying some basic information about its content.

  Example: NUMDAM [13] uses a file naming scheme like

  for journal volumes: journal-acronym_year_series_volume_issue

  for articles: volume-id_first-page_order

  (different fields separated by '_'). This scheme allows to assign to any possible NUMDAM file (at logical units level: serial, volume, issue, article) a unique ID, making it robust and general.

  Of course, conventions for file name length have to be obeyed.

- file name conventions should be made public for each DML server.

  Explanation: Since there will be numerous single DML projects, and since it seems too optimistic to expect that there will be a unique naming scheme for all projects/servers, a prefixing method identifying the project/server is recommended.

- stable URLs for all documents.

- stable and possibly uniform appearance of web pages for all servers (possibly organized in a Math-Net like manner) (ordering scheme governed by MSC 2000 [1]),

- uniform access techniques for all documents,

---

4 Delivery Formats

Recommendation:
The delivery formats should allow:

- mixed image/text data (multi-layered page structures), not only showing the scanned image, but also carrying (hidden) text layers for OCR text, links and other annotations.

  Explanation: (hidden) annotations like OCR-ed versions of the text should be possible and included to make the documents searchable, Internet links should be possible to let references point to reviews and/or to their location in the DML, and resolve other references, internal and external.

- multi-page faithful images of the scanned raw data

  Explanation: "Multi-page" is required to allow convenient access to a whole document, i.e., an article or a book (see the section "Download Units" below). "Faithful" means: the delivery data should be a lossless compression of the raw scanned data.

- fast downloading, even for large documents (books)

- random access to single pages of documents

  Explanation: It should be possible to look, for example, at the reference page at the end of a 300 page document without downloading all the preceding 299 pages.

- quick rendering, fast sequential page flipping, convenient zooming

  Explanation: This makes the use of the documents convenient and easy. E.g., zooming allows the recognition of small indices and other tiny details.

The delivered files should contain

- Searchable OCR text layers basically in ASCII, non ASCII letters like accents, diaereses in unicode encoding.

- Links to Mathematical Reviews (MR) [10], Zentralblatt für Mathematik (ZBL) [20], Jahrbuch über die Fortschritte der Mathematik, and other similar sources (e.g., there are (or have been) other such services - the Russian Refrativni Zhur. Mat. & the contemporary of the JFM (merged with it in 1933) - Revue Semestrielle des Publications Mathemétiques).

- no "garbage text" (e.g., from unrecognized material like formulae)

Further Explanations, Examples:
When choosing a file format, it is important to check that knowledge and implementation of this format is sufficiently stable among digitization and software vendors, or that a sufficient free software community supports it. Proprietary formats, only supported by fragile start-ups and whose conversion or management depend on very specific or lossy software, is to be avoided.

Two file formats currently conform to the requirements above: DjVu [6] , PDF [15].

5 Download Units

- Depending on the manuscript, different download units are recommended:

	primary:	*secondary:*
Journals:	single articles	volumes
Books:	whole book	chapters

  Comment: The choice between "primary" and "secondary" may depend on the delivery format.

- Browsing through tables of content is desirable.
- Single pages as the *only* download unit is discouraged.
- Download of page ranges is desirable.

6 Server Techniques

- The web servers should support random access to single pages (see the delivery formats section).

  Explanation: As a single file per article or book may yield very large files (there are many fundamental papers of 300 and more pages, e.g.), it is expected that the delivery format allows an automatic page by page delivery through the web server to enable random access to single pages without decomposing the whole file.

  It should also be possible to point directly to a given page in the file (e.g., when the request comes after a word search in the plain text, e.g.), or to a given location (something similar to http://..../book#chapter.3). Highlighting a given word or text portion should also be possible.

Hints concerning the given format examples:

- for DjVu, "indirect documents" should be delivered (this can be achieved by setting up the server appropriately, at least for apache).
- for PDF, "byte optimized" (or "linearized)" files should be delivered, server should be configured for "byte serving".

7 Further recommendations

It is suggested to set up public servers for

- format conversions,
- performing OCR,
- automatic supply of metadata for an article (using Dublin Core [8], Open Archive [14], or similar encodings),
- uploading digitized material to the DML,
- registry of all ongoing projects, keeping track of ongoing/completed/planned digitizing projects and allowing mathematicians, librarians and other interested users to propose further material for registration,
- scan servers: a scan server should be a place with good scanning equipment, where people can send their paper material to in order to get it scanned at high quality.

Reason: Setting up the DML is a task for many people and will last 10–15 years or longer. Any individual or institutional contribution of digitizations therefore should be welcome. Individuals should be encouraged and enabled to help.

In order to enable many contributors to provide digitized material in a sufficiently high quality, it is necessary to provide public tools to transform the material into the right format, which is sometimes technically demanding, and to provide text layers by OCR (this should be optimized for the language the manuscript is written in, therefore it would be good to have public servers for the various language areas). Also, it should be easy for contributors to provide the scanned material with (elementary) metadata such as MSC, keywords and phrases on Dublin Core and/or Open Archive basis.

In principle, this technology will be an advantage for any scientific discipline (as well as for more general areas of electronic literature, so the suggestion of a set of servers like this as a basic archiving infrastructure might help to convince funding agencies to give support for DML projects).

Public format conversion servers could also contribute to solve the long term archiving problem, since they provide a dynamic tool for achieving this.

Of course, all these servers should be able to handle mass data upload (script driven), as well as individual files.

The Digitization Projects should:

- Use stable URLs and stable interfaces.
- Offer exportable records for monographs and serials in standard formats to all libraries for their online catalogues. These books and journals should be on the “library shelves” of every library in the world!
- Offer exportable records for journal articles in standard form to the databases of Mathematical Reviews and Zentralblatt für Mathematik.

- Offer records for reference linking to MR and ZBL when the references have been identified.

All Libraries should:

- Add electronic records for all these freely available journals and books to their online catalogue.
- Notify users: Post notices in journal sections of the library to alert users to the fact that certain journals are also online. Perhaps even mention if they are searchable or have reference linking.

The databases MR and ZBL should:

- Add all journal articles from the digitization projects which are not already listed; link to all.
- Add the citation information from the projects to their databases.

Steps for the DML to take:

- Keep an up-to-date listing of the math digitization projects.
- Keep an up-to-date listing of mathematics items and status of digitization.
- Maintain a volunteer network similar to Project Gutenberg [17].
- Collect and disseminate information between the projects.
- Keep lists of digitization vendors (quality, prices, etc.).
- Keep statistics on production: Per page costs, numbers of references, etc,
- Maintain information on and develop software tools:
- Keep track of current best commercial and free tools: OCR; Tools to work with TIFF, PDF, DjVu, etc.
- Coordinate the development and distribute software tools:
  – Reference extracting software (easier for complete journals),
  – Matching software for reference linking
  – Software to convert searchable PDF to searchable DjVu,
  – Software to put references directly into PDF, DjVu files.
- Establish servers analogous to the Any2DjVu Server [2] for conversion purposes: OCR; TIFF to PDF, DjVu; link location & insertion

8 Remarks

A kind of prototype server for special file conversion and OCRing is the Any2DjVu server: `http://any2djvu.djvuzone.org/` [2]
A prototype for a metadata server is the MathNet MMM (Mathematics Metadata Markup) server:
`http://www.mathematik.uni-osnabrueck.de/cgi-bin/MMM3.1.cgi` [11]
Both servers work via Web masks, but can also be driven (for mass production) by LWP scripts [3].

References

[1] 2000 Mathematics Subject Classification
`http://www.ams.org/msc/`

[2] Any2DjVu Server
`http://any2djvu.djvuzone.org/`

[3] Sean M. Burke, Perl & LWP, 244 p., O'Reilly 2002
`http://www.oreilly.com/catalog/perllwp/`

[4] Comité Consultatif International Téléphonique et Télégraphique Group 4 (CCITT G4) Protocol
`http://isp.webopedia.com/TERM/C/CCITT.html`

[5] Digital Library Federation
`http://www.diglib.org/standards/presreformatsum.htm`

[6] DjVuZone, The Technology for Scanned Documents on the Web
`http://www.djvuzone.org/`

[7] Specification of DjVu Image Compression Format, Version of 1999/04/29.
`http://djvu.sourceforge.net/specs/`

[8] Dublin Core Metadata Initivative
`http://dublincore.org/`

[9] Gallica
`http://gallica.bnf.fr/`

[10] Mathematical Reviews (MR)
`http://www.ams.org/mathscinet`

[11] Mathematics Metadata Markup Server
`http://www.mathematik.uni-osnabrueck.de/cgi-bin/MMM3.1.cgi`

[12] LZW Data Compression by Mark Nelson Dr. Dobb's Journal October, 1989
`http://dogma.net/markn/articles/lzw/lzw.htm`

[13] NUMDAM: Numérisation de documents anciens mathématiques
`http://www.numdam.org/`

[14] Open Archive Initiative
`http://www.openarchives.org/`

[15] Portable Document Format (PDF)
`http://www.adobe.com/products/acrobat/adobepdf.html`

[16] PNM Format
`http://netghost.narod.ru/gff/graphics/summary/pbm.htm`

[17] Project Gutenberg
`http://promo.net/pg/`

[18] SourceForge Net
`http://sourceforge.net/`

[19] TIFF, Tagged Image File Format
`http://home.earthlink.net/ ritter/tiff/`, see also RFC1314.

[20] Zentralblatt für Mathematik (ZBL)
`http://www.emis.de/ZMATH/`

[21] ZIP, GZIP file format specification version 4.3, RFC 1952
`http://www.faqs.org/rfcs/rfc1952.html`