MathDoc and the Electronic Publishing of Mathematics

Elizabeth CHERHAL-CLEVERLY
Cellule MathDoc
Université Joseph Fourier, Grenoble
Elizabeth.Cherhal@ujf-grenoble.fr

Laure HEÏGÉAS
Cellule MathDoc
Université Joseph Fourier, Grenoble
Laure.Heigeas@ujf-grenoble.fr

Abstract

France has a long tradition in the publication of mathematics. The very first "mathematics only" journal in the world, the *Annales de Gergonne* was published from 1810 to 1831 and several of the foremost current mathematical journals are published in France today. This paper will develop the original work of the "MathDoc" team to make accessible these and other journals, and more generally to promote the electronic publishing of mathematics.
MathDoc is a small multidisciplinary team, supported by Université Joseph Fourier Grenoble and CNRS[1], whose competences range through mathematics, computer science, databases and Information Retrieval to documentation and librarianship.
MathDoc is devoted to providing access to mathematical literature in the broad sense.
Examples will be given of different applications and projects in the field of electronic publishing and digital libraries.

- The NUMDAM archive giving access to over 8000 retro digitised articles:
  We shall present the editorial and technical choices of the archive, and develop some of its more interesting features such as its hyperlink network.
- Front Math for Gallica:
  The French National Library (BnF) has a great many digitised math documents, in particular the complete works of many remarkable mathematicians. These can be found on BnF's digital library website.[2] However, the granularity of the BnF metadata is not sufficient to easily find individual papers. The Gallica front-end will be presented.
- The « pôle des revues de mathématiques », mathematical journals portal, started in 2005:
  The aim of this project is to provide a complete electronic publishing infrastructure to several academically published French (and European) mathematical journals, and manage their websites. The main workflow functionalities will be described, and focus will be made on the originality of mathematics publishing as opposed to publishing other sciences.
- Seamless web access to digital or digitised math documents (mini-DML):
  We will describe the aims of the mini-DML project, the required infrastructure, and the technology for providing a one-stop search site for all digital mathematical literature available online.

1 The MathDoc team

The "Cellule de Coordination Documentaire Nationale pour les mathématiques" (Cellule MathDoc) is a small team, supported by CNRS and Grenoble 1 University. The team was created in 1995 by Professors Pierre Bérard and Laurent Guillopé, with an aim to providing access to mathematical documentation, giving assistance to mathematical libraries, and working to further a European infrastructure for the Zentralblatt-MATH database. The team, although small in numbers at the time (three members of staff at first), quickly showed its competence in many areas of mathematical documentation. For instance, MathDoc developed the web interface and search engine to Zentralblatt-MATH. Other projects in the digital library field were set up, such as the merged mathematical library catalogues, the national index of online preprints and theses from 1998 onwards and the NUMDAM project from 2000 onwards. The staff increased, as the number of projects grew, and there are currently seven people

[1] Centre National pour la Recherche Scientifique
[2] http://gallica.bnf.fr

working at MathDoc besides the two mathematicians, who decide on the general orientations, there are: one assistant, three computer scientists and three librarian/information scientists. MathDoc's activities range from library applications to digital libraries, and now to electronic publishing. Four recent examples of these will be described below.

2 The NUMDAM digital archive

NUMDAM (Numérisation d'Anciens Documents Mathématiques or digitization of (old) mathematical documents) is an archive containing the entire production of ten[3] French academic mathematical journals from the first number to year 2000. At the time of writing the archive contains over 8300 articles. Efforts are being made to keep the archive up to date, past the year 2000. The collection is to be increased in the coming year(s) by more journals and seminars. The principles underlying NUMDAM are integrity, interactivity and freedom of access after a varying moving wall. Integrity because the entire backrun has been digitzed, every page scanned at high resolution, and the page format reproduced, interactivity due to detailed structured metadata which allows to search not only over the normal metadata, but also full text and cited references and freedom of access because, as the funding going into NUMDAM is entirely public, the collections are made freely accessible after a few years. A moving wall, which varies from 0 to 10 years, protects the publishers' current production. NUMDAM has brought added value and better visibility to the journals, and in no cases have subscriptions fallen.

The main features of NUMDAM are:

- Access to the articles through browsing or searching
- A "Complete record" compliant with the "best practice" statement made by the committee for electronic communications of the International Mathematical Union. A full bibliographic reference, together with abstract, reference list and hyperlinks is freely available, even if the full text of the article is behind the moving wall.
- The full text is available for more than 90% of the collection (after the journal dependant moving wall)
- Download unit: the download unit is the full article, as opposed to separate pages. Download formats are indirect DjVu and linearised PDF, to allow for quicker (page at a time) downloads.
- The "dual" interface: Metadata is presented in html with links, but the download unit is a faithful image of the full text.
- Stable URLs: Every article has a stable and comprehensible URL.

One of the most interesting features in NUMDAM is its hyperlink network: as well as the "normal" sort of links found in most databases such as author's name, journal volume, erratum to original, original to erratum, etc. NUMDAM has, because of its good metadata and some automated processes, made links from each article to its review (Math Reviews (MR), Zentralblatt (ZM), Jahrbuch (JFM)…). As the metadata for each article contains the list of references, an original pattern matching program finds the reviews in the aforementioned databases also for the bibliographic references and makes links there. The bibliographic references are also linked to an article inside NUMDAM, and vice versa (an article is linked to everything that cites it). Statistics for 7899 articles give: 4970 MR links, 5388 ZM links, 2031 JFM links. 4877 articles contain formalized reference lists, amounting to 78413 cited items. Of these 75% have a link to ZM, 66% to MR. There are less than 100 erratum relations, but more than 5000 direct links to NUMDAM.

A recent feature in NUMDAM is its enhanced search engine. When searching full text, links are proposed to the page numbers containing the words (and not only to the file). There is also an option to search expressions present in the same page. NUMDAM has also an OAI-PMH[4] server, and a "cloaked" interface for web crawlers to provide better indexing. One can find links to NUMDAM from MR, ZM, Google, Yahoo and many OAI aggregators. On the technical level, NUMDAM is based on XML metadata containing unicode characters as far as possible, and the EDBM software for the web interface. To automatically reconstruct the hyperlink network, the database is rebuilt whenever a new collection is added. Most features of NUMDAM will be also present in the mathematical journals portal presented later on in this paper.

[3] Annales de l'institut Fourier (1949->), Annales de l'institut Henri Poincaré (1930-1964), Annales mathématiques Blaise Pascal (1994->), Annales scientifiques de l'École normale supérieure (1864->), Annales de l'université de Grenoble (1945-1948), Bulletin de la SMF (1872->), Journées Équations aux dérivées partielles (1974->), Mémoires de la SMF (1964->), Publications mathématiques de l'IHÉS (1959->), Annales de la Faculté des Sciences de Toulouse (1887->)

[4] http://www.numdam.org/oai

3 Front Math for Gallica

NUMDAM is not alone in providing digitised mathematical documents. The French National Library (BnF) has digitised a great many valuable mathematical documents. A partnership between BnF and MathDoc exists for the "concerted digitisation of mathematics" and stipulates that no duplicates should be digitised. If NUMDAM digitises a collection, BnF will not do so, and vice versa.

BnF has a very important digital library called Gallica[5] in whuich one can find online the complete works of many remarkable mathematicians such as Cauchy, Abel, Jacobi, Fourier and Laplace. However, the granularity of the BnF metadata is not sufficient to easily find individual papers. MathDoc is building a front end for the user, in order to make Gallica's valuable resources findable, not only via MathDoc, but also in the mini-DML project described below.

MathDoc will, in the end, provide "article level" metadata for some journals[6] digitised by BnF, and for the contents of the aforementioned complete works. The metadata will also be given back to BnF. As in NUMDAM, metadata is in XML and the front end displays the XML for user-friendly browsing, and points links to the appropriate part of the digitised files on Gallica. Each "catalogued" item will display a full record, and a stable URL for third parties (including mini-DML below) to point to.

4 The "pôle des revues de mathématiques » (mathematical journals portal)

4.1 Background and purpose

As stated in the introduction to this paper, France produces many first class mathematical journals. Some of these are "home" published: either inside academic institutions or by learned societies. These journals are all supported by CNRS (National Centre for Scientific Research), and their means of production are minimal. CNRS has decided to federate its support in setting up an advanced electronic publishing platform, thus enabling a better online visibility of their current production.

As NUMDAM already provides a high quality of digitisation and distribution for old paper volumes, MathDoc has been put in charge of developing a portal for the current issues. MathDoc can be considered as a partner to the journals and the service is designed to address the needs of low-cost independent journals in order to enhance their productivity and visibility on the web. The journals will benefit from the technology and features of NUMDAM, and some more. The service will include browsing, searching, cross linking, and also archiving and maintenance of all the articles.

4.2 Related journals and economic model

Currently, the journals concerned by MathDoc's portal are *Annales de l'Institut Fourier* and *Bulletin de la Société Mathématique de France*. It is likely that other academic journals such as *Annales de la Faculté des Sciences de Toulouse* will soon also be included. The economic model will at first remain the same as the existing one, that is the subscription to the paper will imply the subscription to the electronic version. As this portal has public funding, a wide and democratic access to the electronic articles is desirable. As in NUMDAM, the journals and MathDoc will agree on a moving-wall (all the electronic issues older than a given number of years will become free). This appears to be a fair compromise between journals' economic constraint and open access. This economic model is to move in time according to the wishes of the publishing partners, such as pay-per-view, electronic subscription only, or "package subscription".

4.3 Publishing workflow and MathDoc's offered service

The publishing flow is composed of three parts : production, online publication and archiving. All the scientific production and validation remains under the responsibility of the publishers. MathDoc deals with the online publication and archiving. First, the publishers will provide MathDoc with three formats for each article published in a given issue : the original TeX/LaTeX file written by the author and amended to fit the standards of the journal, the PDF file generated from TeX/LaTeX and the metadata in XML. MathDoc will help the journals to provide these three formats by advising and developing technical solutions, such as setting up a common LaTeX format and appropriate software, or developing extraction, conversion or input software for metadata. In this way, the

[5] http://gallica.bnf.fr

[6] Journal de Mathématiques Pures et Appliquées, Comptes Rendus de l'Académie des Sciences, bulletin des Sciences Mathématiques

production will be more normalized and efficient. The online publication consists in setting up a web site well-identified to the journal's image. From this site the user can browse through all the issues or launch a search within the journal or the whole portal, and include the NUMDAM archive in his/her search. In fact, this site will have the same advanced functionalities as the NUMDAM site, such as full-text searching. Each article has a record which contains a link to the full text in PDF and all the user relevant informations extracted from the metadata. Links will be generated for each bibliographical reference cited in the article. These links will point to the portal/NUMDAM and/or well known bibliographical databases, ZentralBlatt, Math Reviews and Jahrbuch. Lastly, there will be more metadata than in NUMDAM: the abstract, the keywords and the Mathematics Subject Classification for instance.
Generally a publisher is looking forward rather than backwards, preparing the next issue of a journal for instance, and is not so concerned about maintaining and archiving the published volumes. Traditionally libraries have been in charge of keeping archives of paper volumes. By assuring this service for electronic volumes, MathDoc will offer an interesting facility to the publishers and to their readers. Everything relative to a given volume/issue/article will be archived: the original TeX/LaTeX file and its background, the displayable PDF file and its updated versions, the XML metadata.

4.4 Conclusion

To conclude, MathDoc will provide to independent journals a high quality online publishing facility. Without sacrificing the journal's image, the portal will facilitate all the parts of the publishing flow by unifying the workflow and by using open and well documented formats and standards.
Finally, the portal will aim to be as visible as possible to the outer world, via links from ZentralBlatt and MathSciNet databases, exposition of clear metadata via OAI and indexation by web crawlers such as Google Scholar.

5 The mini-DML project

The acronym DML refers to "Digital Math Libray", meaning a collection of all digitally available mathematical documents. Mini refers to the fact that currently only some of the literature is available, and even less has metadata available for harvesting. Mini-DML's goal is to offer unified indexation of all mathematical documents available in digital format.
Mini-DML has no ambition to provide added value, such as reviews, or other features to be found in the traditional reviewing databases. The idea is to use OAI-PMH technology to collect (and possibly expose) simple metadata about digital collections, then provide access to it through a one stop search site. At the time of writing, mini-DML contains metadata from sources such as ArXiv, NUMDAM, project Euclid, the Gallica front end, etc…. Currently, mini-DML has an XML schema[7] on which it is based, and propositions are being discussed with other partners providing access to important digital collections to improve the metadata schema.
If this project goes forward in the proposed way, after a little while it should be possible to build a comprehensive DML database, to which requests could be addressed in order to successfully resolve matches for any cited reference string from any published (or even unpublished) paper.

References

Bouche, Thierry. – NUMérisation de Documents Anciens Mathématiques, in Actes des Journées, La Numérisation des Collections, Atelier – Journée d'études organisée par le laboratoire Reconnaissance des Formes et Visions de l'INSA, dans le cadre de l'Institut des Sciences du Document Numérique (ISDN) Rhône–Alpes, 25 juin 2001

Bouche , Thierry. - NUMDAM and other digitisation activities at MathDoc, CEIC (Committee on Electronic Information and Communication), Grenoble, 11 mars 2005

Cherhal, Elizabeth. - "Numérisation/conservation partagée : la collaboration BnF/Cellule MathDoc ; pistes de travail pour de nouveaux services". Table ronde, Journées des pôles associés, BnF, Paris,. 1 Juillet 2004

MathDoc: see http://www-mathdoc.ujf-grenoble.fr

NUMDAM: see http://www.numdam.org

OAI : see http://www.openarchives.org

Project Euclid: see http://projecteuclid.org

[7] http://www.numdam.org/OAI/minidml.xsd