

The NUMDAM program

Thierry Bouche

Cellule MathDoc & institut Fourier, Grenoble

MSRI workshop, April 16th 2005, Berkeley

Presentation
○○○○○○○○○

Features
○○○

Open questions

Outline

- 1 Presentation
 - Definition
 - Overview
 - Motivations
 - Principles
 - Copyright model
 - Collections
 - NUMDAM future?
- 2 Features
 - Main features
 - Main features (links)
 - Main features (most recent)
- 3 Open questions

Definition

Digitise for archiving and delivery the backrun of academic mathematical journals

Archiving : Integral scanning from the first page up to the last one of each volume, including covers, plates, unbound leaves, ads, etc.

Delivery : One multipage file per article, access through tables of contents browsing or searching, freely downloadable after a variable moving wall.

Overview

- NUMDAM was conceived five years ago by MathDoc's directors: Pierre Bérard and Laurent Guillopé.
- MathDoc is a small CNRS-UJF unit providing services to the European mathematicians and librarians (Zentralblatt online, LIMES, electronic catalogues, ...)
- Only public funds so far.
- MathDoc owns no documents itself: NUMDAM is meant as a service to other parties, under the auspices of the French maths community.
- *Not* a new edition.

Motivations

- Mathematical literature never becomes obsolete.
- It's useful to other sciences in an *asynchronous* fashion.
- It's valid only when considered as a *whole*, building a network of (international) references.
- Isolated academic journals starve against commercial concentration.
- The digital format makes it possible to keep the archive alive, with easy access and hyperlinks.

Principles – Integrity

- Full backrun, no editorial choice.
- Scan of every page at high resolution (600 dpi) black & white (text), grey or colour when applicable.
- Page format reproduced.
- Unicode metadata.
- Clean separation between faithful articles' image and added metadata (PDF/DjVu article vs. navigation HTML stuff).

Principles –Interactivity

- Detailed structured metadata captured:
basic bibliographical data plus full (literary) text
and cited references.
- A hyperlink network places the article in proper context:
bibliographies, reviews, errata, etc.

Copyright model

- Electronic version under the title's owner control.
- Authors are asked to give their (exclusive) electronic copyright
to them if still alive
(preferred: academic institution).
- We contract with the copyright owner to allow access at
www.numdam.org.
- Clear identification of the originating journal
(logo, links, first page).
- Moving wall agreed by both parties
(mean: 5, min: 0, max: 10);
free access after the moving wall as a counterpart
to public funding.

Currently online

Title	Period	Owner	Volumes	Pages	Articles
<i>Ann. inst. Fourier</i>	1949	Assoc. A.I.F.	156	51 054	1 811
<i>Ann. I.H.P.</i> [†]	1930-1964	I.H.P.	71	5 860	147
<i>Ann. math. Blaise-Pascal</i>	1994-2002	Labo/UBP	19	2 554	165
<i>Ann. Sci. École norm. sup.</i>	1864	É.N.S.*	295	68 898	1 867
<i>Ann. univ. Grenoble</i> [†]	1945-1948	UJF	3	1 006	47
<i>Bull. Soc. math. France</i>	1872	S.M.F.	167	45 774	2 608
<i>Mém. Soc. math. France</i>	1964	S.M.F.	134	18 118	396
<i>Journées É.D.P.</i>	1974	C.N.R.S.	31	5 976	514
<i>Publ. math. I.H.É.S.</i>	1959	I.H.É.S.*	92	17 424	344
* Contract with a commercial publisher.				216 664	7 899

Forthcoming

Title	Period	Owner	Volumes	Pages	Articles
<i>Ann. Fac. Sci. Toulouse</i>	1887-2000	U.P.S.	207	36 052	1 035
Sém. Prob. Strasbourg	1967-2002	Labo IRMA*	37	17 352	1 254
Séminaires I.H.P. [†]	1953-1985	??	146	20 000	1 800
Sém. Bourbaki	1948-2000	Assoc. N. B.*	44	17 000	893
<i>Ann. Gergonne</i>	1810-1831	P.D.	22	8 000	935
<i>Ann. I.H.P. sér. A</i>	1964-2000	I.H.P.*	64	25 000	1 125
<i>Ann. I.H.P. sér. B</i>	1964-2000	I.H.P.*	39	20 000	936
<i>Ann. I.H.P. sér. C</i>	1985-2000	I.H.P.*	16	10 000	368
<i>Rev. Stat. appl.</i>	1953-2000	SFS	52	21 890	1 080
<i>Ann. Fac. sci. Univ. Clermont</i>	1962-1993	UBP	37	5 000	
<i>Ann. Scuola Norm. Sup. Pisa</i>	1871-2001	SNS	88		
<i>Compositio Math.</i>	1935-1996	Fund. C.	70	40 000	
...					

Future developments

- Older, multidisciplinary journals?
- Books, Ph.D. thesis & other monographs?
- Manuscripts, rare items (Bourbaki archives...)?
- Collaborations with other partners (Regional/European funding, more journals from other countries)?
- Real time integration of current metadata from live journals?

Main features

- Access to the articles through browsing or searching.
- “Full entry” compliant with CEIC best practice (full bibliographic reference + abstract + bibliography + links freely available).
- Full text available for more than 90% of the collection (after journal-dependent moving wall).
- Download unit: full articles (with first page added).
- Download formats (indirect DjVu & linearised PDF) allow for page-by-page downloading.
- Dual nature of the interface: HTML metadata with links vs. multipage faithful image of the full text as download unit.

Links

As many useful links as possible:

- Same author, same journal volume.
- Full text (multipage image+hidden text) when available.
- Reviewing service when matched (MR, ZM, JFM, SPS, ...)
- Errata/original article when necessary.
- From a cited reference to its reviews (MR, ZM, JFM).
- Or to its actual location (NUMDAM, many more in a near future thanks to the mini-DML).
- To (NUMDAM) articles that cite the given article.

For 7 899 articles, we have 4 970 MR ids, 5 388 ZM ids and 2 031 JFM ids; less than 100 erratum relations.

4 877 of them have a formalised reference list, amounting for 78 413 cited items.

79% have a ZM id, 66% an MR id, while only 2% have a JFM id.

More than 5 600 direct NUMDAM links so far.

Recent features

- Enhanced search engine
(returns linked page numbers for words in full text;
option to search expressions present on the same page).
- Full catalogue available through OAI-PMH server.
- Browsing interface “cloaked” for Web crawlers for better indexing where scarce metadata is available.
- Links to NUMDAM from MathSciNet, Zentralblatt, RBSM, Google, Yahoo, OAI agregators.

Open questions

- **Шафаревич** = Šafarevič = Safarevic ≠ Shafarevich?!
- How to obtain a unified treatment of maths in titles, abstracts, bibliographies?
- How to enhance access for non-mathematicians? (Formula searching, vocabulary matching, content indexing. . .)
- Should we dig into the running text for references? (formalised bibliographies are seldom present since the 1930s, standard since WWII only. . .)
- Do we prefer no link rather than few fuzzy ones?
- Will there be one day a comprehensive DML database which could resolve all “near” matches for any cited reference string from any published paper?

Addendum: Related projects at MathDoc

- 4 ● MathDoc: other related projects
 - Outer digitised material
 - Gallica frontend
 - The mini-DML project

Outer digitised material

- LiNum : Livres numérisés mathématiques.
2577 freely accessible books, 651 digitised but copyrighted, provided by large digitisation centers : Gallica (Paris), *Digital Math Books Collection* (Cornell), *Historical Math Collection* (Ann Arbor), *Mathematica* (Göttingen), *Biblioteka Wirtualna Matematyki* (Warsaw), etc.
- Le *Répertoire bibliographique des sciences mathématiques* (1894-1912).
Collaboration Gallica, Paris (scan of the cards), laboratoire de philosophie et d'histoire des sciences, Nancy (structured keyboarding of the cards) and MathDoc, Grenoble (database, indexing, online interface).
- The *Journal de mathématiques pures et appliquées*, aka *Journal de Liouville* (1836-1932) : detailed cataloguing and indexation of the volumes digitised by Gallica.

Gallica frontend

The BNF's server Gallica has a huge amount of valuable mathematics that are somewhat hidden by weak metadata policy. Cellule MathDoc is building a user frontend so that Gallica's resources will be mini-DML compliant:

- Standalone browsing/searching interface for "opaque" works such as journals (JMPA, CRAS, BSM) and collected works of important mathematicians.
- One full record per item (reprinted or original article) so that third parties (including our mini-DML) can link them.

The mini-DML project

The mini-DML:

- Unified indexation of articles available in digital format, taking advantage on the general dissemination of XML/OAI-PMH technology.
- With special emphasis on long-run journals for which a large amount of material never made it into recent review databases (JFM, ZM, MR)...
- But also digitally available texts that are seldom explicitly referenced (reprints—possibly in collected works, e.g. from Gallica, preprints—arXiv, current issues, ...)

Thank you!