Digital mathematics libraries of today

Thierry Bouche

Cellule MathDoc & institut Fourier, Grenoble

Towards a European virtual library in mathematics

ESF Preparatory meeting
Santiago de Compostela, 13th March 2009

Outline

1. The mathematical literature
2. The mathematical corpus
3. The electronic mathematical literature
4. The digital mathematics library
5. Conclusions

The mathematical literature

Stakes

- Mathematical *validated* literature never becomes obsolete
- Old results are not superseded by newer ones: they are their foundation
- It's valid only as a *whole*, building a wide network of references
- It's useful to other sciences in an *asynchronous* fashion
- It must be carefully archived, indexed and preserved
- It must be accessible over the long term

The mathematical literature

Time scale

- Instant preprint circulation (labs, arXiv, email, home pages)
- Actual publication delayed 1-2 years

  Publication's goals: prestige, career, attribution, quality rating... But prominently to secure the version of the work suitable for further reference
- About 50% citations in today's bibliographies are more than 10 years old
- About 25% citations in today's bibliographies are more than 20 years old

The mathematical infrastructure

Components

Thus research in mathematics needs

- Meeting facilities
- Channels for instant dissemination of new results
- A reliable publication system producing the validated reference texts (aka: the mathematical corpus)
- Fast and detailed reviewing/indexing services for efficient discovery
- A good reference library where the genuine texts can be fetched
- And good interconnections between these components

The mathematical infrastructure

The reference library

We thus need a reference library, which should be

- exhaustive
- up-to-date
- well organized
- widely open
- easy to use for non-mathematicians

In the paper realm, this was almost achieved through a network of strong lab's libraries and interlibrary loan, thanks to large union catalogs.

The mathematical corpus

Milestones 1/2

-300 Euclid's *Elements*

1665 Birth of scientific journals (*Journal des sçavans, Philosophical transactions*)

1800 About 200 journals where math articles are published

1810 First math-only journal (*Annales de mathématiques pures et appliquées*, aka *Annales de Gergonne*)

1850 About 1000 mathematical research articles published each year

1950 About 6000 mathematical research articles published each year

The mathematical corpus

Milestones 2/2

1978-1986	TeX
1992	arXiv math preprints, overlay journals
1994	First non-specialized math-only electronic journal it's free (*New York Journal of Mathematics*)
1995	JSTOR digitises 6 English speaking math journals (400 000 pages)
2000	Massive digitisation projects emerge in Europe (ERAM, NUMDAM)
2003	NSF DML planning project

The mathematical corpus

Size

A rough estimate on the size of the whole corpus of written mathematics in the occidental scientific tradition:

- 3 million items were published spanning 100 million pages
- 100,000 new items appear each year
- Less than 1/5 was published before the 20th century
- More than a half after 1950
- A rather stable distribution: 80% journal articles, 10% chapters in collective books, 10% books
- 600 math journals alive
- 10 million pages digitised? 65% of core journals available digitally?

The mathematical corpus

Multilingualism

The mathematical corpus is *multilingual*

- Very old items were written in Greek, Arabic, Latin and are usually read only through translations
- Modern era items were written mostly in the author's mother tongue up to 20th century
- At least English, French, German, Italian have served as *lingua franca* up to the end of 20th century
- French is still alive but steadily loosing influence
- English amounts to more than 75% of the whole corpus...

  and more than 95% of the new items

The mathematical corpus

The impossible catalog

1868 *Jahrbuch über die Fortschritte der Mathematik*

1894 *Répertoire bibliographique des sciences mathématiques* ("valuable" references from 19th century)

1931 *Zentralblatt für Mathematik und ihre Grenzgebiete*

1940 *Mathematical reviews* and AMS classification

1990 Electronic versions (MathSci Disc, CompactMath) and online access (telnet. . .)

1995 Web access (MathSciNet, ZMATH)

2000 Links to original texts

2002 Bibliographies, backward links

2004 mini-DML, DMR, Ulf Rehman's lists

2009 Each registry has its own partial coverage
A huge part of the existing corpus is not indexed, not linked from anywhere, can be found only by chance!

The mathematical E-infrastructure

New components

In the digital realms, the mathematical community has similar needs, which need dedicated infrastructures to take advantage of the new paradigm.

- Meeting facilities: eScience
- Channels for instant dissemination of new results: arXiv, DRIVER
- A reliable publication system producing the validated reference texts: e-publishing platforms
- Fast and detailed reviewing/indexing services for efficient discovery: reviewing databases, LIMES
- A good reference library where the genuine texts can be fetched: a virtual mathematics library
- And good interconnections between these components: standards and protocols

The mathematical E-literature

Disorganization

- Many digital items are duplicated among various providers
- Many paper items are missing a digital counterpart
- Many collections are split accross providers
- Collection holders are very volatile

⟹ Managing an exhaustive and up-to-date access requires zillions of subscriptions, and superhuman monitoring capabilities

The mathematical E-literature

Content providers

- Gallica retrodigitised, public domain (old), free access
- GDZ/DigiZeitschriften retrodigitised, nothing post-2000, free access
- NUMDAM/CEDRAM retrodigitised/e-publishing platform, moving wall
- JSTOR retrodigitised after moving wall, not-for-profit, English only, (expensive) subscription based library service
- Project Euclid retrodigitised and e-publishing platform, not-for-profit, journal level policies
- Oxford University Press retrodigitised/e-publishing platform, no moving wall, English only
- ScienceDirect Elsevier e-publishing platform, retrodigitised content as one optionnal package
- SpringerOnline Springer-Verlag e-publishing platform, retrodigitised content as one optionnal package (English only)
- And very small projects! (Pôlib: 2 old books, MSRI: 28 books from 1 series...)

The mathematical E-literature

Journal accessibility report

Acta math.	~~Mittag-Leffler~~ (1882-2005); Springer (1882-1997), Springer (1997-)
Ann. Math.	~~Euclid, ELibM~~ (2001-); arXiv (2001-2005?); JSTOR (1884-2003); MSP (2008-)?
Bull. LMS	OUP (1865-)
CRAS	Gallica (1835-1965); None (1966-1996); Elsevier (1997-)
Crelle	GDZ (1826-1997); Walter de Gruyter (1998-)
Duke Math. J.	Euclid (1935-1999), Euclid (2000-)
Liouville	Gallica (1836-1935); None (1936-1996); Elsevier (1997-)
Math. Ann.	GDZ (1869-1996); Springer (1869-1996); Springer (1997-)
Pacific J. Math.	Euclid (1951-1996); Albany (1997-2003); MSP (1997-)
Rend. Palermo	None (NUMDAM?) (1887-1941); Springer (1952-)
Théor. nombres Bordeaux	Séminaire : GDZ (1972-1988); Journal : NUMDAM (1989-2005); ELibM (1994-2007); CEDRAM (1989-)

The mathematical E-literature

The digital downside

Electronic media has downsides for scholars and librarians

- Many new access barriers (copyright, licences, DRM)
- No standards for interfaces, file formats, metadata
- No standards for interoperability
- Technology not mature enough for user-dependant access path (IP numbers, URL, DOI, PURL, OpenURL. . .)
- Value is measured by counts (*not* scientific value)

The mathematical E-literature

Needs

In our brave new digital world, doing research based on mathematical results would be much more easy with a database serving the basic features of the reference library, plus e-only add-ons

This means

- A global (distributed) facility dedicated to archive newly published or digitised material
- One up-to-date registry of all available resources
- Mechanisms for interlinking the holdings with existing and future infrastructures
- Seamless navigation accross the whole corpus

Vision

A reference digital mathematics library should asemble as much as possible of the digital mathematical corpus in order to

- preserve it over the long term,
- make it available online
- at reasonable cost,
- in the form of an authoritative and enduring digital collection,
- updated continuously with publisher supplied new content,
- augmented with sophisticated search interfaces and interoperability services,
- developed and curated by a network of institutions

DML architecture

"Available online"

Collections should be

- Cared for and accessed locally

  (digital files preserved physically at each participating institution: *not virtual libraries*)
- Usable, accessible globally

  (though a virtual union catalog, and metadata sharing
  with cooperating services like reviewing databases
  or more general search engines, portals, etc.)

DML architecture

“Reasonable cost”

Business model should be modelled on the current library system:

- Free to patrons
- Free to anyone would be appreciated,

  but not at the risk of loosing the sustainibility
  or reliability of the system
- A reasonable business model is that full texts become freely accessible a while after their publication (aka: moving wall), when the publisher gained a reasonable return on investment

DML architecture

Institutions

Should be

- Scientifically reliable (authoritative)
- Long lasting (enduring)
- Not-for-profit
- Committed to the effort
  (digital legal deposit for mathematical content?)

DML architecture

Keeping up-to-date

The system should be viewed as a backend to the publishing system:
it doesn't aim at replacing it, as it is not meant to produce new content,
but it cannot be reliably run by commercial entities seeking profit, just as our university libraries.

- Publishers should transfer their output rapidly after publication
- It could be remastered in order to generate all formats required by the library operation (archivable unrestricted formats, metadata schemas, etc.)

DML architecture

Augmented metadata

Each provider has its own metadata set and structure.
If we ever want to have a large infrastructure able to cope with everything, we will need

- Specific development at each institution for ingesting and remastering the content it acquires
- General shared procedures for sharing and enhancing data internally

DML architecture

Augmenting metadata

In the worst case, the source file is a collection of images representing graphically a mathematical text: it should be possible to derive automatically a good approximation of the relevant metadata using

- OCR (possibly math-aware)
- Structure recognition
- Metadata capture
- Using relations in order to put the item in proper context
  (deducing more accurate metadata from already existing metadata for items linked or similar to that one)

DML architecture

Policies

General principles for better usability are generally agreed upon:

- Free metadata and navigation
- Eventual open access (moving wall)
- No long-term economic, legal, technical barriers
- No dependance upon viability of any economic agent

Conclusions

- Mathematicians are waiting for a reference digital library
- It should be a distributed collection of physical archives
- With a central access point allowing seamless navigation and integration to existing tools
- It has to be a public service (at least not-for-profit)
- lasting for ever
- But it should keep up-to-date!
- Immediate free access is *not* mandatory
- Eventual open access *is* mandatory