EuDML

The EuDML project

and the future of access to the mathematical literature

Thierry Bouche

Cellule MathDoc,
Université Joseph-Fourier (Grenoble 1)

Congreso de la RSME
El futuro de la literatura matemática
Palacio de Congreso, Ávila, 3 de Febrero de 2011

Outline

The mathematical literature

Facets

What is *the* mathematical literature?

Many document sets for different features!

Preprints and other volatile material enable *current* research and *fast* communication between colleagues

Formal publishing of refereed results is *slow* but *needed* to keep the system's entropy low, with high quality standards and reliability

Archiving and providing long term access to all mathematical records is *critical* as mathematical results last for ever and can be used by any science at any time after their publication

These distinctions are essential to the future of *mathematics* thus must not be forgotten in the handling future mathematical literature

The mathematical literature

Facets

What is *the* mathematical literature?
Many document sets for different features!

Preprints and other volatile material enable *current* research and *fast* communication between colleagues

Formal publishing of refereed results is *slow* but *needed* to keep the system's entropy low, with high quality standards and reliability

Archiving and providing long term access to all mathematical records is *critical* as mathematical results last for ever and can be used by any science at any time after their publication

These distinctions are essential to the future of *mathematics* thus must not be forgotten in the handling future mathematical literature

The mathematical literature

Facets

What is *the* mathematical literature?
Many document sets for different features!

Preprints and other volatile material enable *current* research and *fast* communication between colleagues

Formal publishing of refereed results is *slow* but *needed* to keep the system's entropy low, with high quality standards and reliability

Archiving and providing long term access to all mathematical records is *critical* as mathematical results last for ever and can be used by any science at any time after their publication

These distinctions are essential to the future of *mathematics* thus must not be forgotten in the handling future mathematical literature

The mathematical literature

Facets

What is *the* mathematical literature?
Many document sets for different features!

Preprints and other volatile material enable *current* research and *fast* communication between colleagues

Formal publishing of refereed results is *slow* but *needed* to keep the system's entropy low, with high quality standards and reliability

Archiving and providing long term access to all mathematical records is *critical* as mathematical results last for ever and can be used by any science at any time after their publication

- These facets are all equally useful: *complementary* indeed
- Each one needs a dedicated infrastructure
- and a different set of actors and skills

These distinctions are essential to the future of *mathematics*
thus must not be forgotten in the handling future mathematical literature

The mathematical literature

Facets

What is *the* mathematical literature?
Many document sets for different features!

Preprints and other volatile material enable *current* research and *fast* communication between colleagues

Formal publishing of refereed results is *slow* but *needed* to keep the system's entropy low, with high quality standards and reliability

Archiving and providing long term access to all mathematical records is *critical* as mathematical results last for ever and can be used by any science at any time after their publication

These distinctions are essential to the future of *mathematics*
thus must not be forgotten in the handling future mathematical literature

The mathematical literature

This complexity is not shared by all sciences

- When publishing flow is rapid, no need for preprints
- When the novelty is entirely contained in the description of a certain structure: once public no other publication is required (or possible)
- In some areas publishing very high volumes, most of the papers have no value six months after publishing, archiving is scientifically meaningful for a very small subset of the published material, the rest only useful to historians

General solutions developed for other sciences do not fit math!

The mathematical literature

This complexity is not shared by all sciences

- When publishing flow is rapid, no need for preprints
- When the novelty is entirely contained in the description of a certain structure: once public no other publication is required (or possible)
- In some areas publishing very high volumes, most of the papers have no value six months after publishing, archiving is scientifically meaningful for a very small subset of the published material, the rest only useful to historians

General solutions developed for other sciences do not fit math!

The mathematical literature

Preprints

Open archives allow:

- *cheap* instant circulation of new results
- iterative enhancements
- versionning, possibly up to the postprint (or withdrawal...)
- good referencing
- stable URLs for sharing references in collaborative work

And can be (ab)used for:

- Current research information systems
- Green open access
- Permanent archiving?

The mathematical literature

Preprints

Open archives allow:

- *cheap* instant circulation of new results
- iterative enhancements
- versionning, possibly up to the postprint (or withdrawal...)
- good referencing
- stable URLs for sharing references in collaborative work

And can be (ab)used for:

- Current research information systems
- Green open access
- Permanent archiving?

The mathematical literature

Publishing

Publishing in math

- is slow (1-2 years)
- is not the typical way the specialists are made aware of a new result
- provides the version of the work suitable for further reference (including possible *errata*)
- provides work's *attribution* and *date stamp* (with more authority than open archives)
- provides quality rating through publication's reputation, selectivity, prestige...
- was and *should stay* expensive (penalty copy)

The mathematical literature

Publishing Landscape

Big diversity of stakeholders, no dominant business model

- About 600 live core journals published by
  - commercial publishers (big to small),
  - learned societies,
  - university presses,
  - math department
  - or even research groups

  All have an electronic edition?
- 2000 serials *with* math articles
- Books are important items (not so much electronic. . .)
- Many small publishers publish first quality material.
  They are not so open to open access

The mathematical literature

Case studies

Three among the best journals in the world

Annals of Math. edited by a laboratory (Princeton/IAS).
Has been produced internally, has enjoyed full open access at arXiv ("overlay journal" 2001-2005), ELibM, then at project Euclid (2001-2008). Now archived at JSTOR (5 years embargo), charged electronic edition at MSP

Publ. Math. IHES edited by an independent institution.
Has been produced internally with help from PUF (-2000). Now under contract with Springer, archives at NUMDAM (up to 2007 with 5 years moving wall)

Inventiones Math. Springer commercial journal.
Open access retrodigitised archives at GDZ (up to 1996)

The mathematical literature

Library

The libraries's main functions are:

Selection	of collections by subject, document type *not necessarily* formally published: theses, reports, preprints... some grey literature needs to be archived or made accessible to patrons (e.g. Nash, Perelman...)
Acquisition	of actual documents (books, journals, files...)
Cataloguing	(capture, produce, import, enhance metadata)
Archiving	documents, files, metadata
Preservation	of the collections to ensure their long term physical persistance
Access	maintained for patrons

The future mathematical literature

Hopes & fears

- Going electronic *should* be a wonderful asset for opening new ways of interacting with the mathematical corpus beyond old boundaries
- Almost every new mathematical document has a digital counterpart, myriads of older ones too.
- Many functions have been migrating to e-only for the best in terms of user experience (preprints into Open archives/Institutional repositories, reviewing journals to databases, journals. . .)
- Some won't be so easy (books)
- Mathematicians are in many universities the only scholars opposed to e-only subscriptions and e-books
- Which is mainly due to doubts on duration of the digital medium
- They don't question duration of paper, which is *wrong!*
- Adiós Codex (wiki-style "collaborative skywriting") ?

The future mathematical literature

Wishes

- For preprints: keep it simple, keep it free!
- For publishing: keep it professional, efficient, affordable, driven by "sustainable" best practices rather than by short term profit
- For libraries:
  - A global (distributed) facility dedicated to archive newly published or digitised material
  - An up-to-date registry of all available resources
  - Mechanisms for interlinking the holdings with existing and future infrastructures
  - Seamless navigation across the whole corpus
  - Instant and perpetual access

The future mathematical literature

The digital downside

Electronic media has downsides for scholars and librarians

- Mainstream publishing is not adapted to mathematical content...
- No standards for interfaces, file formats, etc.
- Many new access barriers (copyright, licenses, DRMs)
- Costs increase!
- "Value" is measured by counts (*not* scientific value)

The future mathematical literature

Disorganization

- Many paper items are missing a digital counterpart, *but*
- Many digital items are duplicated among various providers, *while*
- Many collections are split across providers, *and*
- Collection holders are very volatile

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

The future mathematical literature

Risks

- Return to privately owned scientific libraries!
- Digitise when trendy, sell the product, drop the archive when it generates not enough profit anymore
- Sophisticated referencing for highest visibility of a subset of the math produced, which would be accessible to the richest happy few
- Universal (charged) access instead of real (eventual) open access.

1. Three facets of the mathematical literature
2. The future mathematical literature
3. The EuDML project
   - Vision
   - The project
   - Consortium
   - One access point
   - A distributed archive
   - Eventual open access
   - Innovation
   - First results
4. EuDML content: state-of-the-art
5. EuDML network
6. Conclusion

The European Digital Mathematics Library

(Eu)DML Vision

The Digital Mathematics Library should assemble **as much as possible** of the digital mathematical corpus in order to

- help **preserving** it over the long term,
- make it **available online**
- possibly after some embargo period (**eventual open access**),
- in the form of an **authoritative** and **enduring** digital collection,
- **growing** continuously with publisher supplied new content,
- **augmented** with sophisticated search interfaces and interoperability services,
- developed and curated by a network of **institutions**

$\Longrightarrow$ EuDML, pilot implementation with content from 12 European partners

The European Digital Mathematics Library

(Eu)DML Vision

The Digital Mathematics Library should assemble **as much as possible** of the digital mathematical corpus in order to

- help **preserving** it over the long term,
- make it **available online**
- possibly after some embargo period (**eventual open access**),
- in the form of an **authoritative** and **enduring** digital collection,
- **growing** continuously with publisher supplied new content,
- **augmented** with sophisticated search interfaces and interoperability services,
- developed and curated by a network of **institutions**

$\Longrightarrow$ **EuDML**, pilot implementation with content from 12 European partners

EuDML | The EUROPEAN DIGITAL MATHEMATICS LIBRARY

The EuDML project

Objectives	Pilot implementation of an innovative one-stop shop for mathematical content from 11 European institutions
Consortium	$12 + 1^2$ partners, $1 + 1^2$ associated partners
Profile	3 years (01/02/2010-31/01/2013), 488 PM, European funding up to: 1.6 M€
Content	**235,000 items; 2 600 000 pages**
Rétrodigitised	NUMDAM, Gallica, DML-PL, GDZ, SPM/BNP, HDML, DML-CZ, DML-E, RusDML.
Born digital	BulDML, CEDRAM, DML-PL, EDPS, ELibM, DML-CZ, DML-E

The EDML Consortium

IST	**Overall management and technical coordination** Instituto Superior Técnico (Lisbonne, Portugal)
UJF/CMD	**Scientific coordination** Université Joseph-Fourier: MathDoc (Grenoble)
CNRS/CMD	Centre national de la recherche scientifique: MathDoc (France)
UB	University of Birmingham: Computer Science Dpt. (Royaume Uni)
FIZ	Fachinformationszentrum: Zentralblatt (Karlsruhe, Allemagne)
MU	Masarykova univerzita: Informatique (Brno, République tchèque)
ICM	University of Warsaw: ICM (Pologne)
CSIC	Consejo superior de investigaciones cientificas: IEDCYT (Madrid, Espagne)
EDPS	Édition Diffusion Presse Sciences (Paris, France)
USC	Universidade de Santiago de Compostela: Instituto de Matemáticas (Espagne)
IMI-BAS	Institute of Mathematics and Informatics, BAS (Sofia, Bulgarie)
IMAS	Matematicky Ustav Av Cr V.V.I. (Prague, République tchèque)
IU	Ionian University: Informatics Dpt. (Corfou, Grèce)
MML	Made Media UK (Birmingham, Royaume Uni)
	~
EMS	European Mathematical Society
SUBGoe	Göttingen University library (Germany)

The European Digital Mathematics Library

A global gateway

One access point

For users A website with personal work spaces, allowing to search and browse the collections

For systems A batch lookup for turning citations into links

Benefits

- Easier discovery of mathematical items
- Higher visibility
- Unique lookup for reference linking
- More value to new material as references lead *somewhere*

The European Digital Mathematics Library

A distributed archive

A network of institutions

- Leverage the existing network of local (European) DML projects
- Kind of opt-in legal deposit of mathematical knowledge
- A distributed archive for full texts
- Reliable not-for-profit scientific institutions for long term preservation

Benefits

- Third party curation of full texts
- Long term availability of the corpus
- No preservation hassle on each content producer
- Publishers concentrate on creating new products
  (EuDML takes care of the published content to maintain it over the long term)

The European Digital Mathematics Library

Eventual open access

The moving wall approach

- Once a publisher has completed a year's output, an archival copy (metadata and full text) is transfered to the relevant institution
- It is ingested and indexed
- New items appear in EuDML central database
- Full texts links point to the publisher's platform where access is controlled
- After a while, the local copy of the full text is readily available as well

Benefits

- Full text secured for preservation and long term access
- More texts with low market value available as open access
- Basic mathematical knowldege, foundation of current science and enabling new research, much more widely available

The European Digital Mathematics Library

Innovation

We will investigate new technology

- MathML metadata (extracted from images by math OCR, converted from TEX sources, or extracted from PDF)
- Math search
- Accessible math
- Deep interlinking
- Compute items categorisation and similarity

Benefits

- Advances in Mathematical knowledge management
- Innovative discovery mechanisms
- Testbed for new user interaction with the corpus
- Deliver readily deployable technological solutions
- Deliver enhanced metadata to providers

The European Digital Mathematics Library

First results

Some of first year's outcome

- The content census
- Prague's *Workshop with content providers* in October 2010 where potential partners and oponents were invited to share views on our objectives and policies:
  - Eventual open access (moving wall)
  - Rapid transfer of publisher's content for better indexing
  - Distributed archiving through a network of reference institutions

  No definitive *No*s
  (but no enthusiastic *yes* from profit-oriented publishers either!)
- First public release expected **July 2011**

The European Digital Mathematics Library

First results

Some more of first year's outcome

- A usability study of 4 DML websites
- The EuDML metadata schema (170,000+ items converted)
- First internal demo
- First release of some enhancing tools such as:
  - Add MathML version to TeX metadata,
  - Add english keywords or MSC from Zentralblatt MATH
  - Recompress PDF

EuDML content

Collections

EuDML partners: digital libraries

BulDML	3 journals
DML-CZ	11 journals, 6 proceeding series, 35 books
DML-E	22 journals
DML-PL	10 journals, 4 series of books
HDML	8 journals, 29 conference proceedings, 20 books
NUMDAM	30 journals, 29 seminars, 270 Doct. Th., 1 series of monographs
SPM/BNP	1 journal

EuDML partners: publishing platforms

CEDRAM	10 periodicals
ELibM	91 journals
EDP Sciences	7 journals

EuDML content

Collections

EuDML partner: reviewing database

ZMATH 3 million reviews

EuDML associated partners and collections

Gallica-Math 1 journal, 98 books

GDZ-Math 42 journals, 1531 monographs, 294 multi-volume works

RusDML 11 journals

EuDML content

Next collections?

EuDML future partners?

BDIM	*processing* 1 journal
eLib SANU	9 Journals
eLib MATF	150 books, 354 Doct. Th.
SwissDML	4 journals
TEL	1996 Ph. D. Th.
Gallica	800 books
IMU	ICM proceedings

EuDML content

Copyright owners

Public domain few journals, most books

Not-for-profit Universities, Research organizations, Institutes, Academies, Foundations, Math. societies...

Commercial

Birkhäuser	5 journals (GDZ)
EDPS	7 journals (5 updated in NUMDAM)
Elsevier	5 journals, 1 updated (NUMDAM)
de Gruyter	2 journals (GDZ)
Heldermann	6 journals (5 updated in ELibM)
Hindawi	12 journals (up-to-date in ELibM)
Noordhoff	1 journal (NUMDAM)
AK Peters	1 journal (ELibM)
Springer	2 periodicals (NUMDAM, 1 journal updated up to 2007)
	9 journals (GDZ)

EuDML content

Selection

Process Cascading: the project selects the partnering institutions, each institution selects contributed collections.

Criteria **Published** texts holding mathematical knowledge that has been **validated** through a scientific editorial process, so that they can serve for further **reference** in future works based on this mathematical knowledge.

Items A EuDML *item* is the relevant logical unit to be ultimately delivered to our users.
A monograph, a journal article, each individual contribution in a proceedings volume or an edited book, as well as the whole book
Concretely, it is a pair
(**digital full text** [PDF], **metadata** [XML])
archived at one of the partnering institutions

Summary **235,000 items** (1000 books, 300 Theses) in **13 collections**

EuDML content

Current content overview

Collections **235,000 items, 2,600,000 pages**

Germany	ERAM/JFM, GDZ, ELibM (120,000 items)
France	Gallica-Math, NUMDAM, CEDRAM, TEL (50,000 items)
Czech Rep.	DML-CZ (25,000 items)
Russia	RusDML (17,000 items)
Poland	DML-PL (13,000 items)
Grèce	HDML (8,000 items)
Spain	DML-E (6,500 items)
Portugal	SPM/BNP (2,000 items)
Bulgaria	BulDML (270 items)

Retrodigitised BNP/SPM/IST, DML-CZ, DML-E, DML-PL, Gallica, GDZ, HDML, NUMDAM, RusDML

Born digital BulDML, CEDRAM, DML-CZ, DML-E, DML-PL, EDPS, ELibM, NUMDAM

EuDML content

Metadata schema babel

Metadata notations are of highly varying granularity and shape among EuDML partners and contributing publishers:

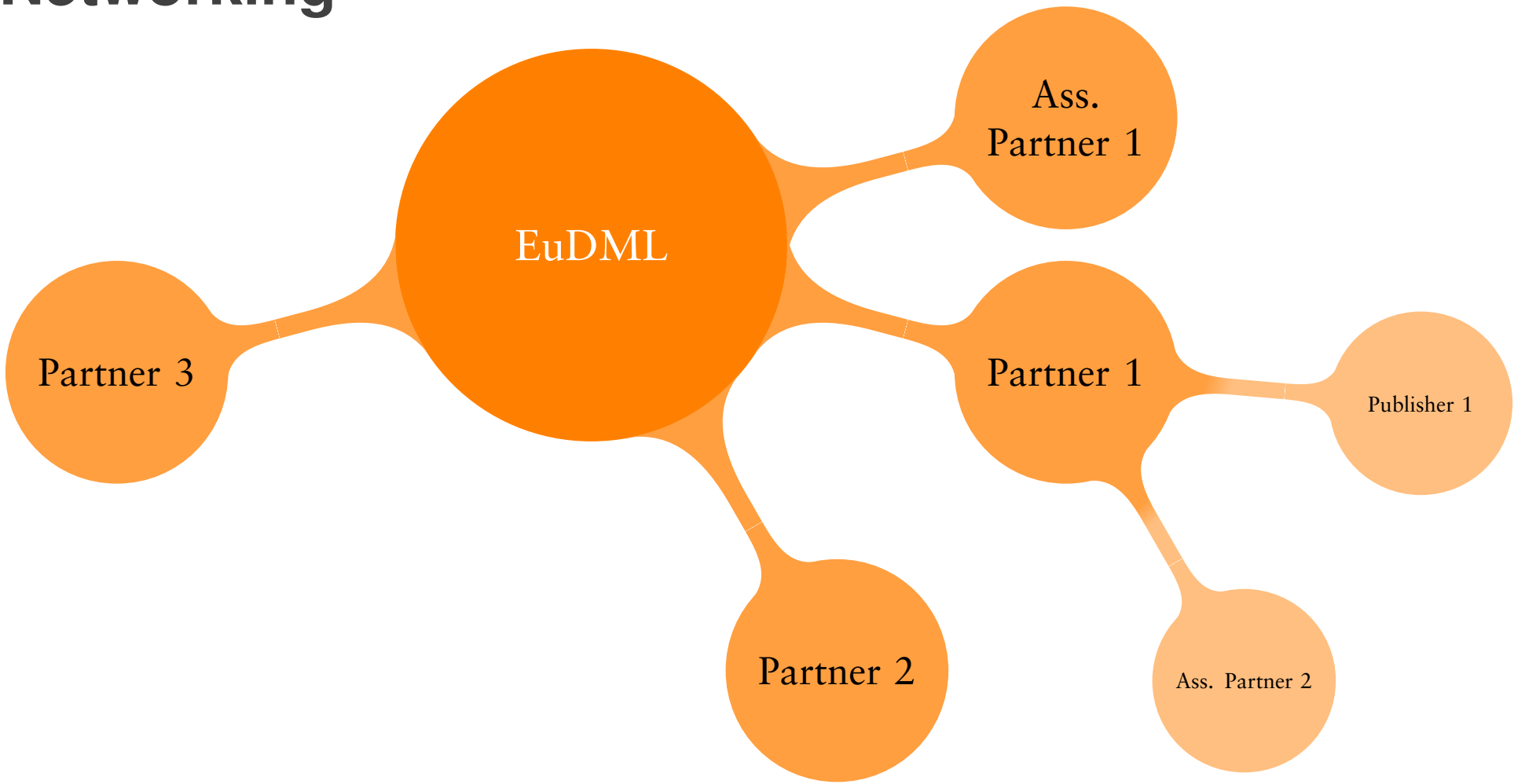
- Internal relational database
- In-house *ad hoc* DTDs
- Proprietary but publicly well documented DTDs
- Standard DTDs from library of publishing industry

We have defined the

NLM Journal Archiving and Interchange Tag Suite

as the standard basis for EuDML metadata storage and exchange (cf. PubMed Central, JSTOR, Portico, NISO...)

Networking


Ass.
Partner 1
EuDML
Partner 3
Partner 1
Publisher 1
Partner 2
Ass. Partner 2

Technical networking

Aggregation tactics

- The preferred method to ingest partner's metadata is OAI-PMH with some partner-dependant EuDML-ready schemas
- **Non customized OAI-PMH servers tend to deliver unusable metadata**

Before project start

- **9 collections from 7 providers** were served through OAI-PMH (5 with enough metadata for basic operation)
- Only **3 collections from 2 providers** served EuDML-ready metadata

Today

- **10 collections from 8 providers** are serving EuDML-ready metadata through OAI-PMH
- **14 collections from 7 providers** are served with EuDML v. 1.0 format

Tomorrow

- **3 collections** are setting up an export procedure
- The big challenge is to set up a smooth process for continuous updates

Networking policy

Scientific institution

(Proposed) Institution charter

- Be aligned with the project's goals
- Keep commited over the long term
- Develop a preservation policy for the full texts
- Acquire new items (retrodigitisation or direct from publishers)
- Sort out rights and licences
- Quality insurance for metadata
- Manage communication with central registry

Networking policy

Publisher

(Proposed) Publisher charter

- Support project's goals
- Select one partnering institution as entry point to EuDML
- Set up transfer and update mechanism for new items
- Licence local DML to store transfered files for ever
- Determine metadata granularity policy (exploited, exposed, reserved)
- Determine moving wall duration

Conclusion

- We need more discussions between research organisations, publishers, math societies, funding bodies, librarians and researchers
- The mathematical corpus is too important and fragile to be handled by for-profit organisations for long periods of time
- Future mathematicians' and scientists' ability to work will depend on the decisions that are taken now
- Some areas of mathematical literature are already endangered by the bad decisions that have been taken during the last decade...
- The EuDML operating model is scalable and could provide a reliable back-end for the future mathematical literature

$\Longrightarrow$ WDML

Conclusion

- We need more discussions between research organisations, publishers, math societies, funding bodies, librarians and researchers
- The mathematical corpus is too important and fragile to be handled by for-profit organisations for long periods of time
- Future mathematicians' and scientists' ability to work will depend on the decisions that are taken now
- Some areas of mathematical literature are already endangered by the bad decisions that have been taken during the last decade...
- The EuDML operating model is scalable and could provide a reliable back-end for the future mathematical literature

$\Longrightarrow$ **WDML**

This being partially funded by the CIP
COMPETITIVENESS AND INNOVATION FRAMEWORK PROGRAMME
(ICT PSP Digital Libraries, project # 250.503)

We will *deliver* a truly open, sustainable and *innovative* framework for *access* and exploitation of Europe's rich heritage of *mathematics*.

Thierry BOUCHE
Institut Fourier & Cellule MathDoc, Grenoble
http://www-fourier.ujf-grenoble.fr/~bouche/

MathDoc *director*
EuDML *scientific coordinator*
EMS Electronic Publishing Committee
IMU Committee on Electronic Information and Communication