EuDML

From EuDML to WDML

Next steps

Thierry Bouche

Cellule MathDoc & institut Fourier,
Université de Grenoble

Future World Heritage Digital Mathematics Library
Washington, National academies of Sciences
June 2nd 2012

Outline

1. EuDML: The project
2. EuDML content: state-of-the-art
3. EuDML tools & services
4. What EuDML can offer to WDML
5. What next?

Conjecture (Bouche, *ca.* 2008)

$$\mathrm{U} = \frac{\mathrm{W}}{2}$$

$$\Downarrow$$

$$\text{rope} = \mathrm{W}(\text{orld})/2$$

$$\Downarrow$$

$$\text{is half the effort to build the WDML}$$

Still open!

Conjecture (Bouche, *ca.* 2008)

$$\mathrm{U} = \frac{\mathrm{W}}{2}$$

$$\Downarrow$$

$$\mathcal{E}u\mathrm{rope} = \mathrm{W}(\mathrm{orld})/2$$

$$\Downarrow$$

is half the effort to build the WDML

Still open!

Conjecture (Bouche, *ca.* 2008)

$$\mathrm{U} = \frac{\mathrm{W}}{2}$$

$$\Downarrow$$

$$\mathcal{E}u\mathrm{rope} = \mathrm{W}(\mathrm{orld})/2$$

$$\Downarrow$$

$\mathcal{E}u$DML is half the effort to build the WDML

Still open!

Conjecture (Bouche, *ca.* 2008)

$$\mathrm{U} = \frac{\mathrm{W}}{2}$$

$$\Downarrow$$

$$\mathcal{E}u\mathrm{rope} = \mathrm{W}(\mathrm{orld})/2$$

$$\Downarrow$$

$\mathcal{E}u$DML is half the effort to build the WDML

Still open!

The European Digital Mathematics Library

EuDML Vision (2008)

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

EuDML Vision (2008)

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

Consortium

IST	**Management & Technical Coordination** Instituto Superior Técnico (Lisbon, PT)
UJF/CMD	**Scientific Coordination** Université Joseph-Fourier: MathDoc (Grenoble, FR)
CNRS/CMD	Centre national de la recherche scientifique: MathDoc (Grenoble, FR)
UB	University of Birmingham: Computer Science Dpt. (UK)
FIZ	Fachinformationszentrum: Zentralblatt (Karlsruhe, DE)
MU	Masarykova univerzita: Informatique (Brno, CZ)
ICM	University of Warsaw: ICM (PL)
~~CSIC~~	~~Consejo superior de investigaciones cientificas: IEDCYT (Madrid, ES)~~
EDPS	Édition Diffusion Presse Sciences (Paris, FR)
USC	Universidade de Santiago de Compostela: Instituto de Matemáticas (ES)
IMI-BAS	Institute of Mathematics and Informatics, BAS (Sofia, BG)
IMAS	Matematicky Ustav Av Cr V.V.I. (Prague, CZ)
IU	Ionian University: Informatics Dpt. (Corfu, GR)
MML	Made Media UK (Birmingham, UK)
	~
EMS	European Mathematical Society
SUBGoe	Göttingen university library (DE)

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 (27,000 items)
Russia	RusDML (17,000 items)
Poland	DML-PL (13,000 items)
Greece	HDML (2,400 items)
Spain	DML-E (6,400 items)
Italy	BDIM (2,000 items)
Portugal	SPM/BNP (2,000 items)
Bulgaria	BulDML (450 items)

Retrodigitised BNP/SPM/IST, BDIM, 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

Selection

Process 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.

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])
physically archived at one of the partnering institutions

Summary Currently harvested: **235,000 items** in **12 collections**
(185,000 journal articles, 3,200 proceedings articles, 41,000 chapters and contributions in books, 2,500 books, 300 multiple volume works)

EuDML content

Copyright owners

Public domain few journals, most books

Public/Charity 50 Universities, Research organizations, Institutes, Academies

Foundations Compositio Mathematica, few not-for-profit bodies

Societies 20 math societies

Publishers

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

Rights status

Public access to PDF

Public domain 10%

Open access 97%

Embargoed 3%

Metadata

Public domain 76,000 items

Freely reusable All (CC0/CC-BY)

Project access to full text

Plain full text 75,000 items have text OCR or PDF extract

Processable PDF 170,000 items available for project internal processing

Servable PDF 105,000 items' re-serving after project processing

EuDML content

Metadata

Metadata babel

- Internal relational database (no XML)
- In-house proprietary DTD
- Standard DTDs (DC, Dspace, minidml, METS, TEL, NLM. . .)

We adopted the
NLM Journal Archiving and Interchange Tag Suite
for EuDML metadata storage and exchange

- Article
- Book
- Book-article (internally)

EuDML content

EuDML metadata schema

NLM Journal Archiving and Interchange Tag Suite

Pros

- Widely deployed and tested (PubMed Central, JSTOR)
- NISO standard
- Precise and flexible (structured *and* flat models)
- MathML (and *alternatives*) ready
- Covers periodical content, books, book collections
- Has room for all foreseen metadata elements, yet extensible

Cons

- Needs “application profile” (best practices)
- Geared towards item’s full text
- Had to be tweaked for some item types (chapter in edited book, multivolume works. . .)

Supports all EuDML item types so far!

EuDML tools & services

Already running

- Conversion to NLM from providers' XML (mostly on-the-fly)
- Small metadata enhancements (tagging refinement, TEX→MathML)
- EuDML reference matching, ZBmath matching (item, ref.)
- Public demo website (only journal articles currently, presentation MathML based display of formulae)
- Experimental formula search
- Experimental similarity computation
- Experimental Opensearch

EuDML tools & services

Expected soon

- Integration of all content types in the public website
- Web 2.0 features
- Service interfaces (Opensearch, LOD, OAI-PMH, REST API)
- More mathematical knowledge generated and stored in NLM records through
  - MSC and English keywords acquired from ZBMATH
  - Guessed MSC, subject categorization
  - Text+MathML extraction from born digital PDF (maxtract)
  - Text+MathML extraction from image PDF (Infty)

EuDML legacy

What EuDML can offer to WDML

- A corpus of 195,000 mathematical documents
- Almost all can be used for trying new processes
- All have metadata in a homogeneous format (NLM based)
- A number of partially evaluated tools (from basic aggregation to accessible math through math formula search)
- Routines to interconnect DML items
- Experience of the first cross-repository, transnational DML effort

EuDML legacy

Personal return on experience

- Some providers are very picky on things such a project could be willing to do with their content: You can't be successful if you don't take this into account (protecting some metadata that they could give to you for internal processing but not for public display).
- Not everything is free, you're not allowed to share what doesn't belong to you.
- On the other hand, providers are expecting good value produced from bigger aggregation: If you comply with their requests, they will be quite eager to cooperate.
- The time frame of such a project needs to leave plenty of time for testing and assessing with real users
- Technology is nothing but a toy as long as no one is convinced to give it a try
- Beware technology-only oriented partners!

What next?

Content

- It should be trivial to enlarge the content to the point where DML becomes an invaluable resource to users
- But this is not an interesting goal!
- Challenges still to be tackled:
  - Integration of collections from any digital library (typically: with wrong granularity, thus need of automated article detection & metadata generation)
  - Serve effectively the non-specialist user of mathematical results (probably a mix of formula search, competent author knowledge base, jargon mismatch recovery)
  - Heritage corpus is multilingual in essence: make it navigable seamlessly!
  - Parse the text and generate logical dependency graph...

What next?

Architecture

We should set up an open yet secure infrastructure so that

- A corpus as big as possible is integrated and made easily accessible to users
- Copyright and content owners wishes are enforced
- The eligible content is used to experiment with cutting-edge technology
- Technology partners can plug-in new processes to the system with no hassle

What next?

Some ideas

Here are some ideas that were not yet tried out to their full potential

- Crowd sourcing: seems inappropriate to mathematics where a typical wikipedian (with a high community rating) is able to screw up the mathematical meaning. But it's maybe the only way to getting proper metadata for lots of items
- Interlinking as a metadata cloud (using detailed metadata of linked items to enhance metadata of a bare item)
- "Image search" technology to provide semantics to formulae (and possibly language-neutral metadata as well)
- Tracking user path to deduce relations

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
Université Joseph-Fourier (Grenoble 1) France
MathDoc *director*
EuDML *scientific coordinator*
EMS Electronic Publishing Committee
CICM Steering Committee
IMU Committee on Electronic Information and Communication