EuDML
The EUROPEAN DIGITAL MATHEMATICS LIBRARY

ICT PSP
ICT POLICY SUPPORT PROGRAMME
part of the Competitiveness and Innovation Framework Programme CIP

DELIVERABLE

Project Acronym:	**EuDML**
Grant Agreement number:	**250503**
Project Title:	**The European Digital Mathematics Library**

D3.5: Final report on external imported metadata

Revision: 1.1 as of 30th January 2013

Authors:

Nicolas Houillon	UJF/CMD
Thierry Bouche	UJF/CMD

Contributors:

Claude Goutorbe	UJF/CMD
Jean-Paul Jorda	EDPS
Petr Sojka	MU
Romeo Anghelache	FIZ
Vlastimil Krejčíř	MU
Vittorio Coti Zelati	SIMAI/UMI

Project co-funded by the European Commission within the ICT Policy Support Programme

	Dissemination Level	
P	**Public**	
C	**Confidential, only for members of the consortium and the Commission Services**	✓

Revision History

Revision	Date	Author	Organisation	Description
0.1	20th Dec, 2012	Nicolas Houillon	UJF/CMD	First draft. Document structure and data available from CMD
0.2	10th Jan, 2013	Nicolas Houillon	UJF/CMD	Additional data for individual providers
0.3	11th Jan, 2013	Jean-Paul Jorda	EDPS	Details on EDPS
0.4	11th Jan, 2013	Romeo Anghelache	FIZ	Details on ELibM
0.5	16th Jan, 2013	Nicolas Houillon	UJF/CMD	First part
0.6	18th Jan, 2013	Vlastimil Krejčíř	MU	Details on DML-CZ
0.7	18th Jan, 2013	Nicolas Houillon	UJF/CMD	Future part
0.8	22nd Jan, 2013	Thierry Bouche	UJF/CMD	LaTeX enhancements + editorial revamping
0.9	23rd Jan, 2013	Petr Sojka	MU	DML-CZ update + some copyediting
0.10	23rd Jan, 2013	Nicolas Houillon	UJF/CMD	General description of joined diagram
0.11	23rd Jan, 2013	Thierry Bouche	UJF/CMD	Diagram integrated, presentation enhanced
0.12	24th Jan, 2013	Nicolas Houillon	UJF/CMD	Examples for process description
1.0	24th Jan, 2013	Thierry Bouche	UJF/CMD	Final candidate
1.1	28th Jan, 2013	Nicolas Houillon	UJF/CMD	Review from Jiří + adjusted numbers

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Contents

Short Summary

This deliverable summarizes the results of harvesting metadata from content providers, including detailed statistics. In total, almost 225 thousand articles and books were collected, which is estimated to be roughly about 7% of the whole corpora of mathematics literature ever published (probably next to 10% of the mathematical literature in digital form), and fulfils more than the goals of Milestone MS033.

The internal formats of the content providers were so heterogeneous that non-trivial efforts had to be done to transform them into a uniform collection encoded according to the EuDML metadata schema. This is expected to be necessary for eventual new data providers, given that they do not yet have data already in JATS metadata compatible format.

1 Introduction–Executive summary

This is the final report on metadata aggregation from EuDML partners. It is an update of Deliverable D3.4 [1] that gave a similar overview halfway through the project.

We report here significant progresses in the area of metadata aggregation, where the project has reached its objectives—it went somewhat beyond its objectives, indeed, as an associated content partner joined successfully, and full text harvesting has been set-up, which was not considered in the beginning.

The total number of harvested records is 225,163, the total number of ingested items is 227,520 (plus 45,520 book chapters that are identified but not really supported as items in the EuDML system: they are not searchable individually, e.g.). This is to be compared with the figure of 170,000 documents expected from the DoW. The total number of harvested full texts currently used for indexing is 135,989.

To achieve this, each content provider had to look carefully at each problem reported in D3.4 and take action. UJF/CMD and IST provided support in order to fine-tune metadata shaping, export, harvest, and transformations.

The area where the picture is still not perfect, but is considerably better than when D3.4 was produced, is automated updates. A number of content providers that did not have an OAI-PMH server now have one, and all of those who have one now export what we called EuDML-ready metadata (not necessarily EuDML 2.0 format, but some format that is automatically converted to EuDML 2.0 without trouble).

1.1 Summary of harvested and validated records

This section lists the number of records (XML documents) available to EuDML after the harvest and validation process, in each collection.

Provider/Collection	EuDML metadata (Schema)
BNF/JMPA	2,081 (article)
BNP/PM	1,347 (article)
CMD/CEDRAM	2,084 (article)
CMD/NUMDAM	50,240 (article) 426 (book)
CSIC+USC/DML-E	6,358 (article)
EDPS Math. journals	2,879 (article)
FIZ/ElibM	36,835 (article)
ICM/PL-DML	14,704 (article) 67 (book)
IMI-BAS/BulDML	715 (article)
IU/HDML	3,272 (article) 6 (book)

Provider/Collection	EuDML metadata (Schema)
MU+IMAS/DML-CZ	28,705 (article) 173 (book)
SIMAI/BDIM	2,138 (article)
SUBGoe/Mathematica	55,520 (article) 2,266 (book)
SUBGoe/RusDML	15,347 (article)
Total	222,225 (article) 2,938 (book)
	225,163 records

1.2 Summary of item types

This section lists the number of items present in the records above. The numbers differ because a record my contain multiple items, like a book and all its chapters, or one items may be present in multiple records, like multiple volume works.

Item type	Number
Journal article	221,293 items
Proceedings contribution	2,962 items
Book chapter	42,520 items
Book: monograph	1,724 items
Book: conference	66 items
Book: volume	1,179 items
Multiple volume work	296 items
Total	270,040 items

1.3 Summary of interoperability and updating

This section presents whether updates of data from the providers are automated or not, first if the transfer of data is done through an OAI-PMH server, and second if the data exposed by the providers is automatically updated when their primary data is.

Provider/Collection	OAI-PMH server	Automatically updated
CMD/CEDRAM	yes	no[a]
CMD/NUMDAM	yes	no[a]
EDPS Math. journals	yes	no[a]
FIZ/ElibM	no	no
ICM/PL-DML	yes	yes

Provider/Collection	OAI-PMH server	Automatically updated
IMI-BAS/BulDML	yes	yes
IU/HDML	yes	no[a]
MU+IMAS/DML-CZ	yes	yes
SIMAI/BDIM	yes	yes
SUBGoe/Mathematica	yes	no

[a] CMD, EDPS and HDML process is partly manual at the time of writing, but they are working on making it automatic.

2 Publication and harvest: An overview

2.1 Transformations

Overview

The material that the various providers wish to publish in EuDML is usually stored in a format specific to the provider, sometimes exposed in a generic format (such as oai_dc), but in no case it is exactly what the EuDML systems expects (in the case of EDPS it is very close), as described in Deliverable D3.6 [3].

The EuDML schema was designed to be very close to the NISO JATS Journal Archiving and Interchange Tag Set [5] to minimize the difficulty of writing such a transformation for a provider that already generates its data in this format, and we even provide a transformation (written by UJF/CMD) that will make the basic conversion, but as the base format is very permissive it cannot guarantee that the result will be valid according to the EuDML best practices.

This means that future EuDML providers will probably all have to transform their data even if they already provide it in the NISO JATS format, but in this case the transformation should be relatively simple (as it was for EDPS).

The data of all current providers had to be transformed to the expected format, either by the provider itself, or by the EuDML harvesting system, and in some specific cases data transformed by the provider is transformed again by the harvesting system but for a different purpose, explained further below.

The providers that write and execute these transformations for their own collections are the following[1]:

- CMD,
- EDPS,
- FIZ,
- MU+IMAS,
- SIMAI/UMI[2].

[1] We refer to deliverable D3.1 [2] and its annex for the naming scheme of content providers, collections, and the description of the metadata formats.

[2] SIMAI/UMI are the Italian mathematical societies that joined as associated partner during the project and contributed the BDIM collection (2 journals).

Most of them serve the resulting data in EuDML format through a standard OAI-PMH server, except for FIZ that was not able to do so yet.

Some providers were not able to do this transformation themselves, but serve their data through their respective OAI-PMH server in a different format. In these cases the data is harvested in the provided format and transformed to the EuDML format by the central harvesting system.

The providers whose collections are transformed in this manner are the following (with the harvested format given in parenthesis):

- BNP (oai_dc),
- ICM (bwmeta),
- IMI-BAS (oai_dc),
- IU (oai_dc).

As the specifications of the oai_dc format allows many interpretations, each of the providers using it did so in a different manner, especially, the most important information is usually stored as a bibliographic citation that is constructed differently by each provider.

This means that a transformations had to be written for each provider even though a priori they served the same format.

In addition, for the providers that serve multiple collections (ICM and IU), slight differences in the usage of the original format and the different EuDML result format from one collection to the next also meant that one transformation had to be written for each.

All these transformations were written by UJF/CMD.

Finally, UJF/CMD acts as the surrogate provider for a few collections that neither can be served nor transformed by their original provider, by writing the required transformations and by serving them in EuDML format as if they were its own. These are the collections from:

- BNF/Gallica,
- CSIC+USC,
- SUB Goe.

Range of transformations

In addition to the transformations meant strictly to change the format of the data, mentioned above, a variety of other transformations have been written to enhance the data or modify its structure.

These transformations are usually generic by nature and not tied to a collection, each do a small specific task, and most can be run on any given collection in any order.

Here are some that the central harvesting system and UJF/CMD use:

- TeX conversion to MathML.
  The TeX formulae in titles, keywords and abstracts are converted to MathML and presented in a JATS formula element that contains both versions (or more) of the formulae, using TeX2NLM [6].
- Splitting keywords.
  In a number of cases, keywords come as a comma separated list. However, in many cases the keywords contain themselves commas and formulae. This transformation

makes sure that the keywords are separated without breaking the formulae (it needs to be run after the TeX formulae in the keywords are properly tagged by the one mentioned just above).

- Adding links to full text.
  IST has been running the OCR program Infty on multiple collections to provide indexable full text that would not be otherwise available. UJF/CMD and MU have been doing the same for NUMDAM and DML-CZ, respectively. This transformation adds links to the full text results in the data.
- Merge book chapters.
  Some providers serve their book data at the chapter level, which is not what is expected by the EuDML harvesting system. This transformation merges the multiple chapter records into one complete book record.

These transformations were written by UJF/CMD, though the one that adds full text links to IST result depends on a service at IST to provide the actual links.

Transformation framework

To make running all these transformations a smooth process, UJF/CMD developed a framework, called eudml-transform, that allows to chain together an arbitrary number of transformations and apply them to a collection of documents. It can either be used on the command line or launched as a service that is directed with REST requests.

Along with the framework, UJF/CMD provides the actual transformations that exploit it and are used during harvest for collections that need any transformation on the EuDML side, including the transformations to eudml format and the ones described above.

An instance of this framework runs on the central harvesting system and is used to harvest all collections that need any transformation on the EuDML side. It is also used at UJF/CMD for its own collections.

2.2 Process

General description

In Figure 1 on the facing page, HTTP responses are omitted for clarity unless it is where a transfer of data happens. In this case they are named with the suffix **r**.

The process starts with REPOX sending a request **1** or **1′**. This happens periodically for each collection.

If the provider publishes data in EuDML format and as one set per collection:

1′ : REPOX sends an OAI-PMH request to the server of the provider, requesting a set.

1′r : the response from the server contains the records from the requested set, each correspond to an item from the corresponding collection, and is stored in REPOX.

If not :

1 : REPOX sends a REST request to eudml-transform that describes where data should be taken from and how, lists the transformations that should be applied to it, and where it should be stored when it is done.

5 : REPOX starts sending periodically another REST request to eudml-transform, asking whether it is done with the previous request, to know when the data is available on the file system.

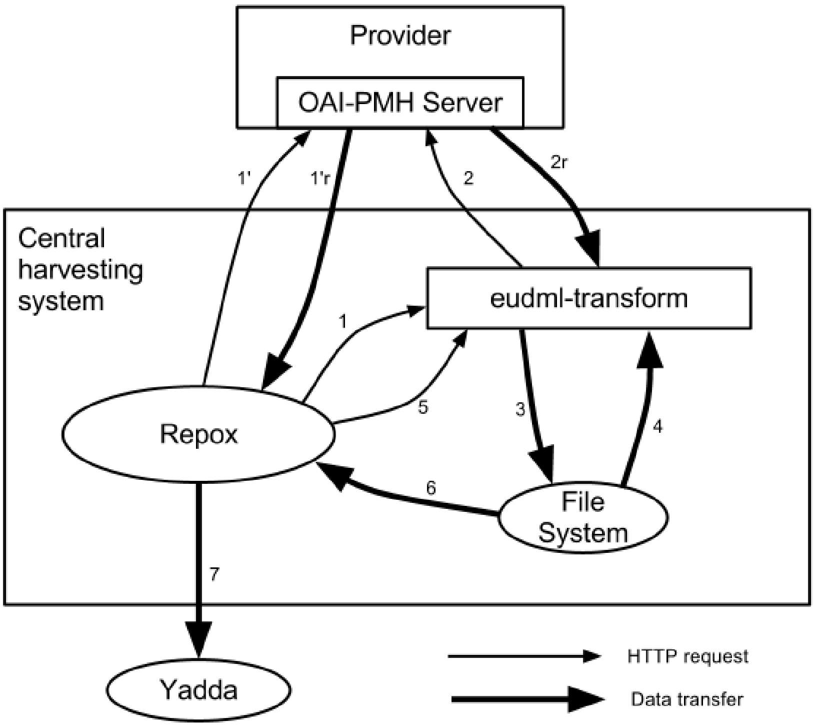

Figure 1: Ingestion and transformation process workflow

At this point eudml-transform can get the data either from an OAI-PMH server or from the local file system, depending on what was requested by REPOX in the request **1**.

From an OAI-PMH server :

2 : eudml-transform sends an OAI-PMH request to the server of the provider, requesting one set or sometimes the whole content of a server that doesn't support sets.

2r : the response from the server contain the requested records, they may correspond to an item or part of one.

From the local file system:

4 : eudml-transform can read the content of a directory and load each file as a record, or list the content of a directory to be all opened in a single transformation (used for merging items).

Once the records are loaded with requests **2** or **4**, eudml-transform applies sequentially the transformations that REPOX requested it to in the requests **1**.

3 : When the transformations are done, eudml-transform saves the resulting records to the file system.

The next time REPOX sends a request **5** it is notified that the result is ready for further processing or harvesting.

If further processing is required, the cycle starts again from request **1**.

6 : When REPOX has finished directing all the necessary processing, it stores the result by reading the file system.

At this point, REPOX performs a validation (described below) on the stored items, marks the items that don't pass it as invalid for the system, and sends a report to the provider with any encountered problem.

7 : The items that passed validation are transferred to Yadda storage, the central data repository.

Examples

The final part of the process, after the data is stored by REPOX following a request **1′r** or **6** data transfer, happens in all cases and will be omitted from the following examples.

CEDRAM This is an example for the most simple harvest process, where REPOX harvests the provider's OAI-PMH server directly as the data for the collection is already in eudml-article 2.0 format and in one set.

1′ : REPOX sends an OAI-PMH request to the provider's server, requesting the records from one set.

1′r : The provider's server responds by sending the records from the requested set to REPOX, which stores them.

BDIM This is also one of the simplest cases, the data is already in eudml-article 2.0 format on the provider's OAI-PMH server, but it is split in two sets. As REPOX uses sets to define collections, eudml-transform is used to group the two provided sets' records.

1 : REPOX sends a REST request to eudml-transform describing the following process, and starts sending periodically requests **5**.

2 : eudml-transform sends an OAI-PMH request to the provider's server, requesting the records from the first set.

2r : The provider's server responds by sending the records from the requested set to eudml-transform.

2 : eudml-transform sends an OAI-PMH request to the provider's server, requesting the records from the second set.

2r : The provider's server responds by sending the records from the requested set to eudml-transform.

3 : eudml-transform saves all the received records in a single location on the file system.

When this process is completed,

5 : REPOX is notified that the data is ready.

6 : REPOX stores the data by reading the file system.

PM In this case the data is stored on the provider's OAI-PMH server in oai_dc format, in one set. Some titles contain formulae in TeX, and IST provides links to Infty OCR results.

1 : REPOX sends a REST request to eudml-transform describing the following process, and starts sending periodically requests **5**.

2 : eudml-transform sends an OAI-PMH request to the provider's server, requesting the records from one set.

2r : The provider's server responds by sending the records from the requested set to eudml-transform.

- eudml-transform applies a series of transformations:
  - Transformation from oai_dc to eudml-article 2.0
  - Add links to Infty OCR results
  - Convert TeX formulae to MathML

3 : eudml-transform saves the transformed records on the file system.

When this process is completed,

5 : REPOX is notified that the data is ready.

6 : REPOX stores the data by reading the file system.

DML-CZ_proceedings The data for this collection is in eudml-book 2.0 format but is represented at a different level than expected, the records each describe one presentation (conference article) when it is expected that a record describes the whole conference proceedings and contains all presentations. The data is also spread among a few sets, and the provider gave a rule on set's names to determine what sets from their OAI-PMH server belong to this collection.

This is a two step process, first the records are gathered (and if any transformation to the actual data was needed it would happen in this first step), then they are merged.

1 : REPOX sends a first REST request to eudml-transform describing the following process, and starts sending periodically requests **5**.

2 : eudml-transform sends an OAI-PMH request to the provider's server, requesting the names of all the available sets.

- eudml-transform follows the rules given by the provider to determine which sets belong to the collection. The following two steps are repeated for each selected set.

2 : eudml-transform sends an OAI-PMH request to the provider's server, requesting the records from one set.

2r : The provider's server responds by sending the records from the requested set to eudml-transform.

3 : eudml-transform saves all the records in a single location on the file system.

When this first process is completed,

5 : REPOX is notified that the data is ready.

1 : REPOX sends a second REST request to eudml-transform describing the following process, and starts sending periodically requests **5**.

4 : eudml-transform makes a list of all the records available in the directory where all the records were saved previously.

- eudml-transform loads all the listed records for a transformation that merges the records belonging to the same conference, creating one record per conference.

3 : eudml-transform saves all the resulting records on the file system.

When this second process is completed,

5 : REPOX is notified that the data is ready.

6 : REPOX stores the data by reading the file system.

2.3 Validation

The validation of harvested items happens after all the required transformations are applied, it is a two-steps process:

1. First the items are validated against their corresponding EuDML XML schema. The result of this validation is binary decision—an item either passes or is rejected.
2. The second validation is made with the Schematron language[3]. It is designed to make sure that the items follow the rules defined in the EuDML best practices. The result is more informative than an XSD validation and is more fine grained: items can be rejected if an error is detected, but it can also give warnings when some recommended data is missing or when a common error, that could also be legitimate data, is detected.

The Schematron rules implementing EuDML best practices (as published in D3.6) have been primarily written by EDPS.

A library to run these validations on a collection of items and sort the results in a meaningful way has been developed by UJF/CMD, using the official EuDML schemas and the Schematron rules from EDPS. It is used by the central harvesting system to validate all incoming data and give feedback to the providers about the status of their data. The validation service is run on-the-fly at ingestion time as a service in the REPOX central installation. There is also an interactive web service where providers can submit data and look at the validation results, which helps them check and fine tune metadata export to EuDML. A demonstration of this service is available at `http://eudml.mathdoc.fr/eudml-validation-demo/`.

3 Full text harvesting

We reported in D3.4 questions discussed within the consortium regarding full texts harvesting which were open at that time. After few rounds of error-and-trial, we ended up with the following set-up.

1. Content providers declare whether they agree that EuDML harvests some flavor of their full texts by adding the reference to that full text in an item's record. We extended the EuDML schema's Best practices specification so that it can hold this information together with the rights licensed to EuDML (harvest, index, enhance, re-serve). By "flavors of full texts" we mean various possible formats such as PDF, plain text in UTF-8, simple indexable XML (text with MathML formulae), TeX or XHTML file created by OCR (Infty), etc.
2. When the record is harvested, a service is called to download the full text(s). As these are generally not served through OAI, we agreed on the following rationale: if an item's record is updated (in the sense of OAI record datestamp, even if it didn't change), then its full text should be harvested anew.
3. When the full text is provided as PDF, a service is run to extract a textual version suitable for indexing [6]. Depending on rights granted and eligible technology, other versions of the full text are generated (accessible formats to be re-served, e.g.).

[3] See `http://www.schematron.com/`.

This workflow has been run over all incoming records and has generated so far 135,989 full texts currently indexed in the EuDML system.

4 Future-proofing: Report on interoperability and updating

Some of the collections are fixed and may not be updated in the future:

- BNF/JMPA,
- BNP/PM,
- CSIC+USC/DML-E,
- SUBGoe/RusDML.

The others are expected to be updated, and this poses the problem of reflecting these updates in EuDML.

As most of these collection are served by OAI-PMH the transfer of updates from the provider's server to EuDML is not really a problem, but what might be problematic is for the provider to keep his OAI-PMH server up to date.

In some cases the served data is automatically updated when the original data is (the served data may actually be the original), but in others a manual intervention is required for the data exposed to EuDML be brought in line with the original data.

5 Detailed report per provider

5.1 BNF/CMD

Bibliothèque nationale de France and Cellule Mathdoc

Data sets

- **Gallica-Math** (*articles* from one journal)

Pre-publication The original data is stored in an internal *ad hoc* XML format.

2,081 items are available and eligible for export to EuDML.

The data is transformed to the eudml-article 2.0 format using XSLT by UJF/CMD.

Published 2,081 records are published, all of them in the eudml-article 2.0 format, through a standard OAI-PMH server hosted at CMD.

Harvest The harvest is a simple OAI-PMH transfer.

Validation There a some minor problems with recommended information that is unknown or just does not exist (missing authors for anonymous works, missing publisher), but nothing that would make an item ineligible.

Final count of imported items 2,081 articles.

Note The Gallica-Math collection also has collected works from various mathematicians. The metadata is not readily exploitable for EuDML: These collected works are in fact collections of volumes, some of these volumes should be dealt with as volumes in multiple-volume (edited) works while others are just monographs. The existing metadata doesn't allow to produce the correct EuDML item type in each case. It is thus necessary to edit manually the metadata in order to make it eligible in EuDML. This had to be postponed.

5.2 BNP/IST

Data sets

- **PM** (*articles* from one journal)

Published The National Library of Portugal publishes a large number of records in a variety of XML formats through a standard OAI-PMH server. Out of these, EuDML harvests 1,347 records from one set in the oai_dc format.

Harvest After a simple OAI-PMH transfer, the data is transformed to the eudml-article 2.0 format using XSLT, enriched with links to full text from Infty results by IST, and TeX formulae are converted to MathML with TeX2NLM. These transformations are provided by the eudml-transform framework and are executed by the central harvesting system.

Validation There are some minor problems with recommended information that is unknown or just does not exist (missing language), but nothing that would make an item ineligible.

Final count of imported items 1,347 articles.

5.3 CMD

Data sets

- **CEDRAM** (recent *articles* from 10 serials)
- NUMDAM
  - **NUMDAM** (*articles* from 61 serials)
  - **NUMDAM_book** (*books* (memoirs) from two series)

Pre-publication For CEDRAM and both NUMDAM collections the original data is stored primarily in their respective internal CEDRAM or NUMDAM XML format, but some specific data is spread in a variety of other sources (database, other XML files). CEDRAM items have formulae both in TeX and MathML. NUMDAM items have formulae in TeX.

There are, at the time of writing, 2,242 items in CEDRAM and 53,611 in NUMDAM. All CEDRAM items are also present in NUMDAM, but the data quality being higher in CEDRAM only the CEDRAM version is exported. These numbers also include very recent items, that are not exported due to the fact that they are not yet publicly posted.

Out of the previous numbers, 2,084 CEDRAM items and 50,666 NUMDAM items (of which 50,240 are articles and 426 are books) are eligible for export to EuDML.

The data is transformed internally to the corresponding EuDML format. The transformation has several steps, to convert the XML to EuDML format using XSLT, merge the data from the different sources (also using XSLT) and convert NUMDAM TeX formulae to MathML with TeX2NLM.

Published 2,084 CEDRAM and 50,240 NUMDAM records are published in eudml-article 2.0 format, 426 NUMDAM_book records are published in eudml-book 2.0 format, through a standard OAI-PMH server.

Harvest The harvest is a simple OAI-PMH transfer.

Validation There are some minor problems with recommended information that is unknown or just does not exist (missing authors for anonymous works, missing ISSN for older journals that don't have one, missing ISBN for books), but nothing that would make an item ineligible.

Final count of imported items CEDRAM: 2,084 articles, NUMDAM: 50,240 articles, NUMDAM_book: 426 books.

5.4 CSIC+USC

Data sets

- **DML-E** (*articles* from 22 journals)

Pre-publication The original data is stored in an SQL database that was provided to CMD by USC at the beginning of the project. 6,401 items are eligible for export to EuDML.

The data is transformed to the eudml-article 2.0 format using an *ad hoc* program by CMD.

Published 6,401 records are published, all of them in the eudml-article 2.0 format, through a standard OAI-PMH server hosted at CMD.

Harvest The harvest is a simple OAI-PMH transfer.

Validation There are some missing page numbers in 43 items that make them ineligible. Other than that, some minor problems with recommended information that is unknown or just does not exist (missing authors, missing publishers, missing language), but these do not make items ineligible.

Final count of imported items 6,358 articles.

5.5 EDPS

Data sets

- **EDP Sciences math. Journals** (*articles* from 7 journals)

Pre-publication The collection is a selected subset of the EDP Sciences Journals articles. For all these articles, an XML document containing the metadata and the bibliography is generated by the typesetter from the LaTeX file. All the documents are validated against the *EDP Publishing DTD*, a slightly modified version of the *NLM Publishing DTD version 3.0*. These documents include formulae both in LaTeX and MathML. The MathML version is not used at EDP for displaying, and thus is still considered as "experimental".

The collection include about 2,800 articles from seven EDP Sciences Journals.

Articles in eudml-article 2.0 schema are output from a two steps XSLT transformation. The first one cleans the specific EDPS information and structures, and generates the EuDML specific information. The second one adds the EuDML namespace. Some semi-automatic checking and cleaning has been necessary for a small group of articles, mainly to correct some identifiers' duplication, and to deal with inconsistencies between the DTD and the XML Schema version of the MathML 2.0 format.

Published 2,879 records are published, all of them in the eudml-article 2.0 format, through a standard OAI-PMH server, distributed among 7 sets each representing a journal.

Harvest The harvest is a simple OAI-PMH transfer, but it is executed through eudml-transform to merge the seven original sets into one collection.

Validation All items validate.

Final count of imported items 2,879 articles.

5.6 FIZ

Data sets

- **ELibM** (*articles* from 101 journals)

Pre-publication Collection available as semi-structured text metadata records containing the results of a web harvesting and article/URL parsing process combined with Zentralblatt-MATH data.

Some scripting, XSLT sheets and some manual fixes were needed to bring these records into the eudml-article 2.0 format.

Published 36,835 records are published, all of them in the eudml-article 2.0 format. They are provided as a compressed archive containing one XML file per item, and made available as an HTTP transfer by giving the direct link via email. Metadata provided is minimally compliant with the eudml-article 2.0 schema, it does not provide OCRed or full text items, but provides instead links to the user readable full text and EuDML indexable articles available in the collection.

Harvest The archive mentioned above is manually downloaded and expanded. The resulting files are then ingested in the central harvesting system.

Validation There a some minor problems with recommended information that is unknown or just does not exist (missing publisher, missing ISSN), but nothing that would make an item ineligible.

Final count of imported items 36,835 articles.

5.7 ICM

Data sets

- PL-DML
  - **PL-DML** (*articles* from 11 journals)
  - **PL-DML_book** (*books* from 3 book series)

Published 14,771 records are published, all of them in the bwmeta format, through a standard OAI-PMH server, without any set. Out of these, 14,704 belong to PL-DML and 67 to PL-DML_book.

Harvest After a simple OAI-PMH transfer, records are sorted into their respective collection based on their data. Records belonging to PL-DML are then transformed to the eudml-article 2.0 format, those belonging to PL-DML_book to the eudml-book 2.0 format, both using XSLT. They are then enriched with links to full text from Infty results by IST. These transformations are provided by the eudml-transform framework and are executed by the central harvesting system.

Validation There a some minor problems with recommended information that is unknown or just does not exist (missing authors, missing ISSN, missing ISBN for books), but nothing that would make an item ineligible.

Final count of imported items PL-DML: 14,704 articles, PL-DML_book: 67 books.

5.8 IMI-BAS

Data sets

- **BulDML** (*articles* from 7 serials)

Published BulDML publishes 6,850 records, all of them in the oai_dc format, through a standard OAI-PMH server. Out of these, EuDML harvests 715 relevant records from 7 sets.

Harvest After a simple OAI-PMH transfer, the data is transformed to the eudml-article 2.0 format using XSLT and enriched with links to full text from IMI and from Infty results by IST. These transformations are provided by the eudml-transform framework and are executed by the central harvesting system.

Validation There is one minor problems with recommended information that is unknown or just does not exist (missing author), but it doesn't make the item ineligible.

Final count of imported items 715 articles.

5.9 IU

Data sets

- HDML
  - **HDML_journals** (*articles* from 5 journals)
  - **HDML_conferences** (*articles* from one conference series)
  - **HDML_books** (*books*)

Published 3,630 records are published, all of them in the oai_dc format, through a standard OAI-PMH server, in a single set. Out of these, 2,340 belong to HDML_journals, 929 to HDML_conferences, and 361 to HDML_books.

Each HDML_conferences record represents a presentation item, and each HDML_books record represents a chapter item.

Harvest After a simple OAI-PMH transfer, records are sorted into their respective collection based on their data, then the record is transformed to the eudml-article 2.0 format for journals and conferences, or to the eudml-book 2.0 format for books, using XSLT. Even though they are then enriched with links to full text from Infty results by IST, the full text results are unusable because of very poor quality (resolution of 72 DPI).

The book records, originally at chapter level, are also merged into book level records. The 361 chapters are merged as 6 books.

These transformations are provided by the eudml-transform framework and are executed by the central harvesting system.

Validation There are some minor problems with recommended information that is unknown or just does not exist (missing authors, missing publishers and ISSN for journals, missing ISBN for books), but nothing that would make an item ineligible.

Final count of imported items HDML_journals 2,340 articles, HDML_conferences 932 articles, HDML_books 6 books.

5.10 MU+IMAS

Data sets

- DML-CZ
  - **DML-CZ_serial** (*articles* from 13 serials)
  - **DML-CZ_monograph** (*books* from 4 collections)
  - **DML-CZ_proceedings** (*books* from 6 conference series)

There are three main types of documents/publications in DML-CZ: *serials*, *proceedings* and *monographs*. In these collections are several sets (in the terms of OAI-PMH) that expose metadata for the DML-CZ content.

To make the handling of the items of different types easier, they were grouped into three separate collections, one for each type of items.

- 13 *serials* with 28,705 articles, out of which 21,588 have Infty files (with full text math).
- 6 *proceedings* series (65 'books') with 2,030 papers, out of which 1,642 have full text with math (by Infty).
- 4 *monograph* collections (108 'books'), with 1,860 chapters (1,074 of them having rich full text with math).

Pre-publication In addition to sets described in previous subsection, there is also published a special set which collects the work of the famous Czech mathematician Otakar Borůvka. However, this set is not finished yet and also has many 'virtual' items so the set has not been exposed to the EuDML yet. This and other items will be provided to EuDML when completed/published.

Most of the full texts are publicly available, and if there are not rich full texts with MathML by Infty, plain full texts—extracted from PDFs by DSpace (PDFBox)—are published. The only exception are four journals that have a moving wall (up to 24 months) policy for the most actual issues. Nevertheless, for the purposes of EuDML enhancing (similarity and indexing) even these full texts are also provided securely (IP protected).

Original metadata for further processing is stored in the Metadata Editor (DML-CZ internal tool for managing the contents) in an internal XML format in several XML files. There are separate metadata records for the whole journal, volume, issue, article etc.

DML-CZ internal formats use their own schemas to describe the contents.

For public viewing, the DSpace system is used so the metadata are also exposed via OAI-PMH server that is a part of the DSpace system. [4]

The DML-CZ exporting workflow can be depicted as:

> internal metadata files → bash scripts → XSLT → pre-final EuDML JATS → upload to DSpace (OAI-PMH) → EuDML harvest.

The internal XML metadata files are stored in a special file system directory structure. Several bash scripts prepare the appropriate XML files for XSL Transformation made by Saxon. For example, to produce an article metadata, four XML metadata files are needed: Journal metadata, volume metadata, issue metadata, and article metadata. The bash scripts seeks data related to given article and passes them as parameters to Saxon.

Saxon uses XSL templates to produce a pre-final JATS file. Pre-final means that there is not all information yet—there are no final links to full texts (the '`<ext-link>`' elements), for example. These are added in the following step.

XSL templates are written for the DML-CZ internal format transformations. Some functions from the eudml-transform framework are used. After the transformation is finished the output is stored back in the directory structure at the same place where the DML-CZ internal metadata is stored, i.e. at the article/chapter level.

The pre-final JATS XML file is then stored to the DSpace. DML-CZ uses its own importing applications written using DSpace JAVA API. DSpace provides OAI-PMH functionality—default settings provide only oai_dc format. However, DSpace can be configured to also expose different formats in many ways via the so-called crosswalk mechanism.

DML-CZ uses its own customized crosswalk written in Java implementing DSpace Crosswalk interface. This crosswalk exposes JATS file as eudml-article 2.0 (for articles) and eudml-book 2.0 (for books) formats in terms of OAI-PMH. Crosswalk also on-the-fly modifies pre-final JATS by adding necessary `<ext-link>` elements according to the moving wall settings for the given article.

Published 28,705 DML-CZ_serial records are published in eudml-article 2.0 format, 1,860 DML-CZ_monograph records and 2,030 DML-CZ_proceedings records are published in eudml-book 2.0 format, through a standard OAI-PMH server.

Each DML-CZ_proceedings record represents a presentation item, and each DML-CZ_monograph record represents a chapter item.

The records are sorted into OAI-PMH sets representing journals, monographs or conferences, the name of a set can be used to determine which type it is following rules given by the provider.

Harvest After a OAI-PMH transfer, records are sorted into their respective collection based on the OAI-PMH set they belong to. The monographs records, originally at chapter level, are merged into monograph level records (1,860 chapters become 108 monographs), and the proceedings records, originally at presentation level, are merged into conference level records (2,030 papers become 65 conference volumes).

These transformations are provided by the eudml-transform framework and are executed by the central harvesting system.

Validation There are some minor problems with recommended information (provided in the validation report) that is unknown or just does not exist (missing authors, missing ISSN, missing ISBN for books), but nothing that would make items ineligible.

Final count of imported items DML-CZ_serial 28,705 articles, DML-CZ_monograph 108 books, DML-CZ_proceedings 65 books.

5.11 SIMAI/UMI

Data sets

- **BDIM** (*articles* from 2 journals)

Pre-publication The original data is stored in an internal XML format. The records already contain formula both in TeX and MathML (produced using LaTeXML).

Some material available internally (Collected Works) has not been exposed since it is not yet complete.

2,138 items are eligible for export to EuDML.

The data is transformed internally to the eudml-article 2.0 format using XSLT, adding links to the full text files produced by IST with Infty (stored on the local server).

Published 2,138 records are published in eudml-article 2.0 format, through a standard OAI-PMH server, in 2 sets each representing a journal.

Harvest The harvest is a simple OAI-PMH transfer, but it is executed through eudml-transform to merge the two original sets into one collection.

Validation All items validate.

Final count of imported items 2,138 articles.

5.12 SUB Goe

Data sets

- GDZ
  - **GDZ_Mathematica** (*articles* from 47 journals)
  - **GDZ_RusDML** (*articles* from 10 journals)
  - **GDZ_Band** (*books*)
  - **GDZ_Monographs** (*books*)

In the tables outside this section, GDZ_Mathematica represents a grouping of GDZ_Band, GDZ_Mathematica and GDZ_Monographs.

Pre-publication The original data is stored in the METS XML format. It was provided to CMD by SUB Goe.

57,944 GDZ_Mathematica records, 16,748 GDZ_RusDML records, 747 GDZ_Band records, and 1,549 GDZ_Monographs records are eligible for export to EuDML.

Items from GDZ_Mathematica and GDZ_Monographs are transformed respectively to eudml-article 2.0 and eudml-book 2.0 formats using XSLT.

Records from GDZ_Band are also transformed to the eudml-book 2.0 format using XSLT, but it is a process in multiple steps, as muliple-volume data, originally separate, is integrated.

Records from GDZ_RusDML are transformed to the eudml-article 2.0 format using an ad hoc program, that relies on XSLT for XML format changes, but also detects Cyrillic strings and transliterate them.

All these transformations are done by UJF/CMD.

Published GDZ_Mathematica and GDZ_RusDML records are published in eudml-article 2.0 format.

GDZ_Band and GDZ_Monographs records are published in eudml-book 2.0 format.

All records are published through a standard OAI-PMH server hosted at CMD.

Harvest The harvest is a simple OAI-PMH transfer.

Validation There are missing article titles in 2,424 items of GDZ_Mathematica, that make them ineligible.

There are missing publication year in 1,401 items of GDZ_RusDML, 21 items of GDZ_Band and 9 items of GDZ_Monographs, that make them ineligible.

Other than that, there a some minor problems in all collections with recommended information that is unknown or just does not exist (missing authors, missing ISSN, missing publisher, missing language, missing ISBN for books), but these do not make items ineligible.

Final count of imported items GDZ_Mathematica 55,520 articles, GDZ_RusDML 15,347 articles, GDZ_Band 726 books, GDZ_Monographs 1,540 books.

6 Conclusion

We have managed to harvest and collect almost a quarter million items of mathematical literature and the library is still growing. As most cooperating content providers will update their holdings and continue to export them via OAI-PMH, EuDML will grow automatically. Given the expertize reached during transforming data from current content providers, and tools and transforming frameworks developed, it may be rather straightforward to add new content providers in the near future. Negotiations are ongoing.

References

[1] Thierry Bouche, Claude Goutorbe, and Nicolas Houillon. Report on partners imported metadata, August 2011. Deliverable D3.4 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library.

[2] Thierry Bouche and Hugo Manguinhas. Report on available collections and metadata, November 2010. Deliverable D3.1 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library.

[3] Claude Goutorbe, Thierry Bouche, and Jean-Paul Jorda. EuDML metadata schema - final version, January 2013. Deliverable D3.6 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library.

[4] Vlastimil Krejčíř. Building Czech Digital Mathematics Library upon DSpace System. In Petr Sojka, editor, *Proceedings of DML 2008*, pages 117–126, Birmingham, UK, July 2008. Masaryk University. `http://www.fi.muni.cz/~sojka/dml-2008-program.xhtml`.

[5] National Information Standards Organization. JATS: Journal Article Tag Suite, ANSI/NISO Z39.96-2012, August 2012. See `http://jats.niso.org/`.

[6] Petr Sojka, Krzysztof Wojciechowski, Nicolas Houillon, Michal Růžička, and Radim Hatlapatka. Toolset for Image and Text Processing and Metadata Enhancements – Value Release, March 2012. Deliverable D7.3 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, `https://project.eudml.org/sites/default/files/D7.3.pdf`.