DELIVERABLE

Project Acronym: EuDML

Grant Agreement number: 250503

Project Title: The European Digital Mathematics Library

D3.4: Report on partners imported metadata

Revision: 1.1 as of 24th August 2011

Authors:

Thierry Bouche	UJF/CMD
Claude Goutorbe	UJF/CMD
Nicolas Houillon	UJF/CMD

Contributors:

Ioannis Karydis	IU
Radoslav Pavlov	IMI-BAS

Project co-funded by the European Comission 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	2011/06/30	Thierry Bouche	UJF/CMD	First draft, from CG and NH input
1.0	2011/07/20	Thierry Bouche	UJF/CMD	version ready for review, with input from IK and RP
1.1	2011/08/24	Petr Sojka, Georgi Simeonov	MU, IMI-BAS	version with input from internal reviewers

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This deliverable contains original unpublished work except where clearly indicated otherwise.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

1 Introduction

This is a report on the state of metadata aggregation from EuDML partners. It builds on the methodology, strategic choices, and results of a number of EuDML project's work packages.

- Deliverable D3.1—"Report on available collections and metadata" [7] provided a detailed picture of the content available from EuDML partners, metadata quantity and quality describing it, as well as interoperability devices to interchange it, at project start.
- Deliverable D3.2—"The EuDML metadata schema, initial version" [6] defined the NLM Journal Archiving and Interchange Tag Set as the general framework adopted for encoding bibliographic metadata, and specified three privileged item types' description (multiple volume work, book, article) as the first instance of the EuDML schema.
- Deliverable D4.1—"EuDML Global System Functional Specification" [4] depicted a preliminary vision of supported use cases and overall system architecture.
- Reviews of the state of the art and partners' provided technology for implementing the system and its user interfaces, augmenting metadata, annotating and associating items were conducted in work packages 5–8 [10, 8, 2, 1] as well as working demos of such software components [5, 11, 9, 3].

Metadata harvest and aggregation is an iterative process

Deliverable D3.1 helped identify all metadata sources and harvesting methods available from EuDML partners when the project started. A number of harvests were performed during the first year of the project in order to assess D3.1 findings, and inspect the various encodings of available metadata. While D3.2 was developed and discussed within our consortium, a number of further partial harvests were done, followed by a number of conversions to other formats, especially NLM.

After completion of D3.2, a full harvest of all available metadata sources providing item-level obligatory metadata (as defined in D3.2) was completed. Out of these, the team at UJF/CMD converted to the possible extent the harvested metadata to EuDML v 1.0 XML schemas. This work (known as milestone M3.2—first metadata harvest) was reported at the Madrid meeting in February 2011. At that time, a total of 174,391 items from 8 sources were converted to EuDML schema and served through OAI-PMH to all project partners. A subset of this was imported into the central architecture at ICM and used to feed the first public EuDML demo.

From lessons learned at this stage, we studied an *ad hoc* ingesting path for all items that had been left out in the first run. We also identified a number of problems with our converted metadata (some of them implied to contact the metadata owner in order to fix the problem in the original source).

Probably the most interesting problem is that of tagged formulae. As many partners had TeX mathematical formula representation embedded into metadata elements declared as text strings (typically: title, abstract, keyword. . .), these were just considered as text in the conversion process, although they should have been tagged as formulae if we wanted them to be actually managed as such (for retrieval and display, e.g.). WP7 had developed tools to generate MathML (needed for math searching, accessibility, etc.) from LaTeX source or PDF, but adding a MathML version to a LaTeX formula needs proper identification of TeX formula in the first place. We used the tool TeX2NLM from D7.2 in order to generate these tags. We thus got MathML and LaTeX representation for every supported formula in the harvested metadata, making at last the mathematical knowledge of our content more accessible to further processing. Overall, we processed 40,937 input strings from all harvested items, generating 206,775 formulae. Only 3,219 errors were encountered. We hope that math-aware mining will now have enough correct data to start producing amazing results!

A summary of harvested metadata

We will give details on each source in the next section, and discuss the issues that remain to tackle in the next iterations of this work.

Provider/Collection	EuDML metadata (Schema)	Notes
CMD/CEDRAM	1 868 (**article**)	converted from internal XML with MathML
CMD/NUMDAM	44 121 (**article**)	converted from internal XML
CSIC/DML-E	6 401 (**article**)	converted from SQL database
EDPS journals	455 (**article**)	slightly tweaked to obey best practices
FIZ/ElibM	25 678 (**article**)	converted from internal XML
IMAS/DML-CZ	26 476 (**article**), 132 (**book**)	converted from internal XML
IMI-BAS/BulDML	436 (**article**)	converted from OAI-DC XML
IU/HDML	2 340 (**article**), 28 (**book**)	converted from OAI-DC XML
SUBGoe/Mathematica	57 357 (**article**), 2 297 (**book**), 296 (**mbook**)	converted from METS XML
SUBGoe/RusDML	16 486 (**article**)	converted from METS XML

Provider/Collection	EuDML metadata (Schema)	Notes
BNF/JMPA	2 081 (**article**)	converted from XML
BNP/PM	1 347 (**article**)	converted from TEL XML
All	**187 799 records**	

The above records are available to project partners from two sources, and consolidated in a central REPOX installation at ICM:

- EDP Sciences records are available from http://oai.edpsciences.org/ (one set per journal: cocv, ita, m2an, mmnp, proc, ps, ro; metadata format: eudml (article)).
- All other records were made available from http://math-thar.ujf-grenoble.fr/repox/OAIHandler (one set per collection: BulDML, CEDRAM, DML_CZ_Serial, DMLE, ELibM, Gallica, GDZ_Mathematica, GDZ_RusDML, HDML_journals, NUMDAM, PM: NLM-AI metadata format (article); DML_CZ_Proceeding, DML_CZ_Monograph, GDZ_Monographs, GDZ_Band, HDML_Books, HDML_conferences: NLM-Book metadata format (book); GDZ_MBook: NLM-MBook metadata format (mbook)). This server is IP-protected during the testing phase of the project.

ICM handles directly the items from the DML-PL collection, it should contribute 111 books and 11 journals, for a total of 13 494 articles encoded in BWmeta 2.0 schema.

Summary of item types

According to item types as defined in D3.2, here is a summary of the harvested metadata:

Item type	Number
Journal article	185 046 items
Proceedings contribution	3 211 items
Book chapter	41 145 items
Book: monograph	1 590 items
Book: conference	119 items
Book: volume	748 items
Multiple volume work	296 items
Total	232 155 items

Interoperability and updating

For sustainability of the EuDML service, it is crucial that we reach a state where sharing metadata, including new metadata as partner's collections grow, requires the minimum effort from all involved parties. This is an area of our activity that didn't take up as smoothly as it should, where there is still a lot of room for improvements. We update here the table from § 2.2.2 of D3.1.

Provider/Collection	OAI-PMH	Quality	Up-to-date
CMD/CEDRAM	Yes	EuDML v1.0	○
CMD/NUMDAM	Yes	EuDML v1.0	○
CSIC/DML-E	No		
EDPS journals	Yes	EuDML v1.0	●
FIZ/ElibM	No		
ICM/DML-PL	No		
IMAS/DML-CZ	Yes	Partial	
IMI-BAS/BulDML	Yes	EuDML-ready	●
IU/HDML	Yes	Partial	
SUBGoe/Mathematica	Yes	EuDML-ready	
SUBGoe/RusDML	Yes	EuDML-ready	
BNF/JMPA	Yes	EuDML v1.0	○
BNP/PM	Yes	EuDML-ready	●

Notes The first column tells whether a given collection has an up-to-date OAI-PMH server attached to it. "Yes" might hide the fact that it sometimes has *two*: one for normal operation (serving subbasic OAI-DC metadata), and one set up especially for EuDML with EuDML-ready metadata. This is the case for CMD collections, e.g., and it is not the best choice for future smooth operation and easy maintenance of the system.

The second column gives an estimation whether the OAI-PMH server delivers metadata that is exploitable for EuDML service (possibly after some conversion or straightforward restructuration of the served records).

"Partial" means that the metadata is sufficient for enabling minimal EuDML operation. However, it might be that the collection holds, and is willing to provide, more metadata than is delivered through OAI-PMH, so that another path would be preferred for ingestion in EuDML system for advanced operation.

"EuDML-ready" means that the OAI-PMH server serves metadata that is readily exploitable for EuDML service, and as complete as the provider is able and willing to provide to EuDML.

"EuDML v1.0" means that the OAI-PMH server serves metadata that is encoded according to EuDML schema with best practices recommendations, following D3.2 recommendation.

The last column proposes to highlight the usability and sustainability of the current interoperability devices.

- ● A full bullet means that the situation is as good as possible: the OAI-PMH server delivers an up-to-date version of its best metadata contributed to EuDML, in exploitable form.
- ○ An empty bullet means the same, but some additional process has to be performed, which is not automatic when updates to the source service are made.

The other cases still need work from partners. Even when a full EuDML-ready OAI-PMH is out of reach, some cases could be automated using various existing interoperability devices.

2 Detailed report per provider

2.1 CMD: CEDRAM and NUMDAM

There is an overlap between CEDRAM and NUMDAM collections, as 7 series were digitised by NUMDAM before being produced by CEDRAM as born digital, as well as retro-born-digital in some cases. Moreover, the new production from these series is transfered to NUMDAM some time after their initial publication by CEDRAM. The idea is to promote good practice of sustainable publishing where the content is transfered to a third party independent library that curates this content (full text included) over the long term. For those items produced by CEDRAM, we had much more accurate and advanced metadata (XML generated from LaTeX source of the articles themselves, and LaTeX/MathML alternatives for all formulae, e.g.). It was thus decided that items produced by NUMDAM would be part of the NUMDAM collection, while items produced by CEDRAM would be part of the CEDRAM collection, without creating artificial duplicates.

This is why “only” 45 988 items are contributed to EuDML from these two collections, although cedram.org holds 6 852 items, and numdam.org holds 45 902 items.

For both collections, the metadata was directly exported as EuDML v1.0 obeying best practices from D3.2. The TeX strings in NUMDAM titles, abstracts and keywords were converted to NLM tagged formulae with TeX and MathML alternatives.

2.2 CSIC: DML-E

The “harvesting” of this collection resulted in the transfer of a MySQL database dump. An *ad hoc* program was developed by UJF/CMD for the generation of EuDML metadata from the database content.

The original data encodes formulas in HTML using the Symbol font. Symbol font code points have been mapped to Unicode during the conversion.

There is currently no procedure to retrieve new or modified items.

2.3 EDPS: Mathematical journals

NLM journal publishing DTD being the native format for EDP Sciences' new production system, all 2011 papers' metadata was already served on EDPS' OAI-PMH server as NLM (identified as pmc metadata format). After the Madrid meeting, EDPS modified the configuration of their OAI-PMH server so that it also serves the same metadata tweaked to comply with EuDML v1.0 article DTD, and associated best practices recommendations.

EuDML export of metadata is thus part of the normal publishing infrastructure at EDPS, which is the most robust solution for continued service and up-to-date content.

The older years from EDPS math journals were contributed through NUMDAM, as pre-1997 articles have been digitised by this project, and an ingestion procedure had been set up in order to archive newer material as well. While some older born digital production was ingested in NUMDAM from LaTeX sources, using tools from CEDRAM, year 2010 was ingested from internal NLM.

2.4 FIZ: ELibM

The harvesting of ELibM metadata resulted in file transfer in their own XML format, describing 29 159 items.

The original metadata has unstructured bibliographic references and has been enhanced by using a calls to the Zentralblatt database. The formulae were encoded in TeX, and converted to the extent possible to MathML using TeX2NLM tool.

The structure transformation, as well as the needed metadata enhancements, were produced using the combination of an *ad hoc* program and XSLT style sheet. During this process, a large number of items were found to lack an associated full text (PDF file). They have thus been discarded (3469 items).

Given the awkward ingestion procedure, there is currently no workflow to retrieve new or modified items.

2.5 IMAS/MU: DML-CZ

The DML-CZ collections were available in two ways to EuDML: the standard OAI-PMH server that serves an advanced version of the metadata inside an OAI-DC superstructure, and the internal production format which represents the hierarchical structure of supported publications in a file system with internal XML representation for each relevant level. As this second format was much more complete and coherent, it was decided to ingest DML-CZ metadata from this source.

Of course, this leaves us with no plans for further maintenance and update, but DML-CZ team is currently working on this issue.

Serial articles

26 558 articles were described in the contributed archive.

The XSLT transformation produces EuDML article XML files. Some restructuration (separating keywords, e.g.) is performed, while TEX formulae (in title, abstract, keyword) are converted to NLM structure with MathML using the TeX2NLM tool.

Proceedings and monographs

Proceedings volumes have been mapped to EuDML book metadata objects, with individual articles becoming book-parts. The transformation used XSLT style sheets.

2.6 IMI-BAS: BulDML

BulDML currently holds full digitization of printed articles and recent born digital production from 3 journals and one book series. Full digitization in BulDML is considered non-use of automated OCR software generating images combined with hidden layer of erroneous text. Therefore all content is manually retyped in text file format for electronic documents. All articles in BulDML are manually submitted and described by Qualified Dublin Core metadata schema. For all articles related to EuDML, metadata has been manually selected and enriched with Citation information (volume, number, number of start/end pages) and strictly keeping Dublin Core schema.

BulDML metadata is stored in Unicode (UTF-8) and correctly handle and display UTF-8 characters providing multilingual content in English, Bulgarian, Russian, etc.

According to long term preservation policies for digital repositories and institutional data preservation and archiving: BulDML have implemented the handle net system which includes an open set of protocols, a namespace, and a reference implementation of the protocols. The protocols enable a distributed computer system to store identifiers, known as handles, of arbitrary resources and resolve those handles into the information necessary to locate, access, contact, authenticate, or otherwise make use of the resources. This information can be changed as needed to reflect the current state of the identified resource without changing its identifier, thus allowing the name of the item to persist over changes of location and other related state information.

The associated OAI-PMH server provides up-to-date metadata. It needed to be parsed so that the minimal structure of a journal issue could be reconstructed. This parsing is done on-the-fly using an XSLT style sheet so that no manual work is required to have an up-to-date EuDML version of the metadata from the REPOX installation.

2.7 IU: HDML

HDML currently holds 5 journals, 6 books, and one series of conference proceedings. The associated OAI-PMH server provides up-to-date minimal metadata. It needed to be parsed so that, on one hand, the minimal structure of a journal issue could be reconstructed, and, on the other hand, parts of the same books delivered separately could be reunified.

The books and conferences were converted to EuDML book DTD, with each contribution or chapter stored as a book-part. Unfortunately, there is no book-level identifier provided in the metadata.

Journal articles can be converted on-the-fly using an XSLT style sheet so that no manual work is required to have an up-to-date EuDML version of the metadata from the REPOX installation. As other item types would be represented as book-parts within a book in EuDML, a similar on-the-fly conversion at item level is currently not possible.

Technical issues concerning the hdml.gr

All data was provided by the Hellenic Mathematic Society. The data was provided in 3 different formats. The first included almost intelligible images of the documents and all metadata were in an sqlite database. The second set included the original tiff uncompressed scanned images of the documents and no metadata whatsoever. The third set included medium quality jpg images and xml supporting metadata. The metadata of the third set required extensive check as these contained numerous errors (e.g. sometimes the author names were in the "keyword" tag!).

The technical team developed a series of programs that enabled a team of content editors to check manually each and every document now included in the database of the HDML in a distributed manner.

The website was developed with simplicity of the interface in mind: minimum graphics and 2 main choices, the browse mode and the search mode. The site uses

1. User- and search-friendly URLs
(e.g. http://karydis.ionio.gr/hdml.gr/en/browse/Journals/Ευκλει΄δης/05 instead of
http://karydis.ionio.gr/hdml.gr/browse.php?lang=en&Journals=Ευκλει΄δης&issue=05)
2. A custom search engine supported by Google that displays search results only from the HDML website in an overlay, thus offering all the extra powerful capabilities of Google while not requiring either extra pages or displaying results in differentiated environments
3. Two modes of showing an item: one based on the actual retro-scanned images using a lightbox overlay for fast page-per-page online reading and a second based on a pdf format that includes the item's images embedded and can be saved to the user's system for asynchronous reading

In addition, HDML offers an OAI-PMH server in the address http://karydis.ionio.gr:8080/repox/ based on the Repox software (http://repox.ist.utl.pt/). The schema used therein is based on dml_dc after the "Recommended Best Practice for Unqualified Dublin Core Metadata Records" following the directions of http://projecteuclid.org/documents/metadata/dml_dc/. The OAI-PMH server supports the efficient dissemination of the contents of HDML to the EuDML service for the initial harvest phase but also for any future update as well.

HDML is hosted in brand-new virtualisation servers of the Ionian University directly interconnected on the GRNet network offering almost 100% uptime, numerous fail-over alternatives, increased processing capability, high degree of customisation and high bandwidth interconnection.

2.8 SUBGoe: Mathematica

The Mathematica collection is the output of a digitisation project of physical *volumes*, so this is how the metadata is organised, stored and exchanged. The volumes can be single volumes in a multiple volume work, single volume books, or journal issues.

As SUBGoe's OAI-PMH server doesn't support massive download, it was not possible to initiate metadata harvesting in this way. However, an alternative path was indicated by SUBGoe so that, starting with a number of PPNs (SUBGoe's stable identifiers), the METS metadata could be retrieved over the network, using the standard interface provided by GDZ (`mets_download`).

The transformations to NLM were performed via dedicated XSLT style sheets. 57 353 articles in 35 journals were generated out of the 2 298 original volume files encoded in METS. As several journals from this collection have been imported into other projects (namely DML-CZ), it was decided to convert only metadata from journals that were not duplicated. The reason is that the differences between two sources would only have been improvements, as the newer project started from SUBGoe data, adding more details and corrections to it.

The 296 multiple volume works were reconstructed with their 748 individual volumes.

There is currently no procedure to retrieve new or modified items. There are some indications that the initial list of PPNs was incomplete, or a snapshot of an ongoing work. For instance, the multiple volume works refer to 779 individual volumes although it was only possible to get hold on the METS metadata for 748 among them. However, an automated updating procedure could be set up using the working interoperability tools freely available at GDZ. Once the main obstruction of harvesting the huge amount of already digitised works has been overcome: the standard OAI-PMH could be used to retrieve the relatively small number of PPNs since first harvest, e.g.

2.9 SUBGoe: RusDML

RusDML is a digitised collection of journals published in Russia. The metadata actual content has many idiosyncrasies, but the general interoperability infrastructure relies on GDZ standard system. It is thus very similar to SUB Goe's Mathematica collection.

Among the 11 digitised journals, one was left out as the metadata didn't allow to produce any meaningful insight in the content, which is probably better browsed page by page online.

The transformations to NLM were performed via dedicated XSLT style sheets and a specific program handling the recognition of Cyrillic test (e.g. to separate English keywords from Cyrillic keywords). 16 486 articles in 10 journals were generated out of the 584 original volume files encoded in METS.

As individual articles have no explicit identifier, *ad hoc* identifiers have been generated using the enclosing volume's identifier, and the ordinal position of the article within the volume. It is not clear if this method is robust enough.

Some formulae encoded in TEX have been converted to MathML.

There is currently no procedure to retrieve new or modified items, but the previous remark applies and, anyway, there are no known plans to enlarge this collection.

2.10 BNF/CMD: JMPA

The early years of *Journal de mathématiques pures et appliquées* were digitised by Gallica project at BNF. Cellule MathDoc created an article-level catalogue for this corpus, as well as an access frontend (see http://math-doc.ujf-grenoble.fr/JMPA/) as part of a more general Gallica-Math project, which did the same for a number of mathematician's Collected works.

The JMPA metadata uses a specific, rather minimal format, which was transformed via an XSLT style sheet. TEX formulae (in title) were converted to NLM structure with MathML using the TeX2NLM tool.

2.11 BNP/IST: PM

The years 1937–1993 of *Portugaliæ Mathematica* were digitised by BNP with support from IST and Sociedade Portuguesa de Matemática.

BNP's OAI-PMH server provides the full extent of available metadata in OAI-DC, which needs some parsing that can be done on-the-fly using an XSLT style sheet so that no manual work is required to have an up-to-date EuDML version of the metadata from the REPOX installation.

The collection is frozen.

3 Conversion and postprocessing

We provide here a list of the various processes that have been used at UJF/CMD to produce the master EuDML version of all ingested metadata.

For each of BulDML, HDML, and PM (those whose source format was OAI-DC, where untagged bibliographic citation is captured in a single element), an *ad hoc* parser was developed in order to get EuDML structure out of a character string.

For every ingested collection, some normalisation of the data was performed:

- Language code conversion to ISO 639-1, as required by EuDML v1.0.
- Transformation with TeX2NLM of TEX strings in text elements (titles, abstracts, keywords) into NLM formula tagging with <tex-math> (TEX) and <math> (MathML) alternatives.
- In a number of cases, keywords come as a comma separated list. However, in many cases the keywords contain themselves commas (think of examples as (a, b)-*modules*). A tool was developed to separate individual keywords without breaking math formulae.

4 Further work and Open issues

In this section, we try to keep track of the work related to EuDML aggregation, which is either ongoing within the project, or should be scheduled.

4.1 Interoperability

Towards sustainable service?

DML-CZ is working on a EuDML-ready OAI-PMH server.

As SUBGoe initial import has been satisfactorily completed, the fact that its OAI-PMH server doesn't support massive downloads might not be an impeding factor anymore. One could get identifiers of new or modified content through a simple ListIdentifiers request to the OAI-PMH server restricted to recent records. Then one could use each identifier to collect the full metadata from another source. As long as such hacks are well documented, and corresponding services are maintained, it is a manageable way of insuring automated updates.

This scheme could be used for other content providers, and for other content (such as full texts).

Relaxing EuDML schema?

The notion of items privileged for interchange service that emerged from EuDML schema specification in D3.2 seems to pose problems, notably as it prevents on-the-fly conversions for items that “cannot” be transported by themselves in EuDML schema. Two typical examples:

1. Some projects are volume-oriented, and might encode article-level information inside a volume description file in a way similar to a book’s table of contents.
2. On the opposite, some projects are based on the smallest granularity available, and would rather transport book chapters individually.

The EuDML v 1.0 specification requires that journal issues be split into individual articles (with issue metadata inefficiently repeated) while book chapters (including contributions to an edited book) be collected into a single file. More flexibility could be pragmatically allowed when interchanging these objects, as long as the ability to reconstruct whatever organisation fits better at both ends is not impacted.

4.2 Structure transformation vs. Enhancements

It is not always obvious to decide which processing should be done *before* harvesting (preprocessing by a partner of its metadata contributed to the project), *while* harvesting (on-the-fly reorganisation of the incoming metadata), or *after* harvesting.

From our experience reported here, we conclude that it is best that each content provider takes care of making its metadata EuDML-ready. It is usually much easier for the content provider, who holds the best shape of his metadata—usually in a database or as a collection of files—to generate a properly tagged interchange format with best granularity.

The rationale for this is that each content provider knows best what metadata details it holds, which elements it wants to contribute to EuDML, and how it is encoded. D3.2 documents what is expected from EuDML. Examples of such processing that should be much easier for the producer of the data to be converted are: author names splitting (into first name, last name), LaTeX formula detection (tagging *as* formula) and conversion (to MathML), generating structured bibliographic data, etc.

When such preprocessing has been done, on-the-fly conversion to any meaningful format relevant to EuDML service is easily achieved. It is also much straightforward to prepare a harvest mechanism.

Most trivial tasks such as structure or encoding conversions can (and should) be done on-the-fly, e.g. using an XSLT inside REPOX so that the harvester can serve immediately restructured metadata.

Enhancements using EuDML own tools should probably be done centrally after a full harvest has been done, and all available metadata is homogeneously stored.

4.3 Deduplication and metadata merging

The work packages boundaries are not always obvious either: does WP3 stop when each content partner has set up an automated export procedure to EuDML central system and leave it to WP5 to actually harvest and store the metadata? Does it run the full harvest and pass the resulting archive to WP5? Does it go as far as removing duplicates or merging various sources of information for the same item when applicable? Some of the last task was performed in a rather aggressive way when we discarded one source for some duplicated items, which was based on the analysis that some sources' metadata superseded another one's.

As the notion of EuDML item was modelled on that of a Zentralblatt item, we feel that the Zentralblatt database should be the central device to manage EuDML items and their relations.

We would call duplicates two provider's items describing the same EuDML item, thus contributing metadata for the same object, but possibly different copies of its full text (such as: born digital from publisher's platform, digitised from a digitisation project). In such a case, we would need a strategy for creating a reference EuDML item, whose metadata record would be created with best elements coming from all contributed metadata for this item, including Zentralblatt's when possible.

This activity has not yet started, although one could consider that it is at the core of building a reference database for the European mathematical literature. A strategy for doing it is to match all EuDML contributed items against Zentralblatt database so as to create a database modelled on Zentralblatt's containing all Zentralblatt items as well as all EuDML items only once. The matching tools developed in D8.2 should allow this easily.

4.4 Full texts as metadata

For full text search, as well as for a number of processes foreseen in work packages 7–9, we need a mechanism allowing partners to contribute various versions of their full texts to the central system. The number of possible scenarios is high, as for instance a content provider could want to provide its born digital PDFs so that the MathML Extractor (cf. D7.2) can make MathML versions of the formulae, or a similar scheme with scanned PDFs and math OCR, but one would like also to boostrap the search engine with what searchable full text are available right now, which most often means some sort of raw text.

This part of metadata aggregation didn't go as fast as expected. We ended up with the following recommendation, and we are now urging partners to follow it:

1. Full NLM encoded item. The royal road to contributing such "metadata" full texts to the EuDML central system should be to embed it into the <body> element with NLM encoding. An EuDML-ready OAI-PMH server could serve three formats (which can be mechanically derived from the richest one): eudml+ (EuDML with

NLM-encoded full text in <body>), eudml (same with <body> omitted: just front and back matter), and the mandatory oai_dc).
We notice however that there is currently no NLM full text produced by or partners willing to contribute full texts.

2. Other cases (OCR, TXT extracted from PDF, flat XML. . .). When the full text is not available in NLM encoding, it should not be embedded into the <body> element of the NLM record. On the contrary, it should be made available to the system along the general frame of "document" identifiers with links to the relevant versions pertaining to an item. In order to remove this information from a user interface, we agreed on special values so that the consuming system easily distinguishes user formats of the full text from machine-oriented ones.
To set the value of such links in EuDML schema, the <self-uri> element must be used in <mbook-meta>, <book-meta>, <book-part-meta> or <article-meta>.

5 The EuDML collections

Provider	Collection	Website	OAI-PMH server
CMD	CEDRAM	http://www.cedram.org/	http://math-thar.ujf-grenoble.fr/repox/OAIHandler
	NUMDAM	http://www.numdam.org/	http://math-thar.ujf-grenoble.fr/repox/OAIHandler
CSIC	DML-E	http://dmle.cindoc.csic.es/	
EDPS	Math J.	http://publications.edpsciences.org/	http://oai.edpsciences.org/
FIZ	ElibM	http://www.emis.de/elibm/	
ICM	DML-PL	http://pldml.icm.edu.pl/	
IMAS/MU	DML-CZ	http://dml.cz/	http://oai.dml.cz/request
IMI-BAS	BulDML	http://sci-gems.math.bas.bg	http://sci-gems.math.bas.bg:8080/oai/
IU	HDML	http://www.hdml.gr	http://karydis.ionio.gr:8080/repox/
SUBGoe	Mathematica	http://gdz.sub.uni-goettingen.de/	http://gdz.sub.uni-goettingen.de/oai2/
	RusDML	http://www.rusdml.de/	http://gdz.sub.uni-goettingen.de/oai2/
BNF/CMD	JMPA	http://portail.mathdoc.fr/JMPA/	http://math-thar.ujf-grenoble.fr/repox/OAIHandler
BNP/IST	PM	http://purl.pt/index/pmath/PT/index.html	http://oai.bn.pt/servlet/OAIHandle

References

[1] Josef Baker, Łukasz Bolikowski, Wojtek Hury, Mark Lee, Petr Sojka, and Volker Sorge. Association Analyzer Implementation: State of the Art, November 2010. Deliverable D8.1 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, http://eudml.eu/.

[2] Josef Baker, Alan Sexton, Petr Sojka, and Volker Sorge. A State of the Art Report on Augmenting Metadata Techniques and Technology, December 2010. Deliverable D7.1 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, http://eudml.eu/.

[3] Łukasz Bolikowski, Wojtek Hury, Mark Lee, Radim Řehůřek, Petr Sojka, and Volker Sorge. Toolset for Entity and Semantic Associations – Initial Release, May 2011. Deliverable D8.2 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, http://eudml.eu/.

[4] José Borbinha, Aleksander Nowiński, Gilberto Pedrosa, and Wojtek Sylwestrzak. EuDML Global System Functional Specification, November 2010. Deliverable D4.1 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, http://eudml.eu/.

[5] José Borbinha and Gilberto Pedrosa. The EuDML Metadata Registry and Repository, February 2011. Deliverable D5.1 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, http://eudml.eu/.

[6] Thierry Bouche, Claude Goutorbe, Jean-Paul Jorda, and Michael Jost. The EuDML metadata schema—Initial version, November 2010. Deliverable D3.2 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, http://eudml.eu/.

[7] 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, http://eudml.eu/.

[8] Vanessa Gorman and Mark James. User Interface Design, December 2010. Deliverable D6.2 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, http://eudml.eu/.

[9] Radim Hatlapatka and Petr Sojka. Toolset for Image and Text Processing and Metadata Editing – Initial Release, February 2011. Deliverable D7.2 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, http://eudml.eu/.

[10] Tim Kitchen. Usability Study, September 2010. Deliverable D6.1 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, http://eudml.eu/.

[11] Aleksander Nowiński, Michał Politowski, and Tomasz Rosiek. The EuDML Search and Browsing Service, February 2011. Deliverable D5.2 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, http://eudml.eu/.