The EuDML metadata schema

Michael Jost, Thierry Bouche, Claude Goutorbe, Jean-Paul Jorda, Miroslav Bartošek, Peter Stanchev, Michal Politowski

EuDML | The EUROPEAN DIGITAL MATHEMATICS LIBRARY

ICTPSP

DELIVERABLE

Project Acronym: EuDML

Grant Agreement number: 250503

Project Title: The European Digital Mathematics Library

D3.2: The EuDML metadata schema

Revision: 1.6 as of 15th December 2010

Authors:

Michael Jost (FIZ)
Thierry Bouche (UJF/CMD)
Claude Goutorbe (UJF/CMD)
Jean-Paul Jorda (EDPS)

Main contributors:

Miroslav Bartošek (MU)
Peter Stanchev (IMI-BAS)
Michał Politowski (ICM)

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	2010/07/23	Michael Jost	FIZ	Structure creation with elements from Paris meeting. Sections 1-2
0.2	2010/09/28	Claude Goutorbe	UJF/CMD	Section 3 added
0.3	2010/10/25	Thierry Bouche	UJF/CMD	First attempt toward a self-contained document reflecting Prague decisions
0.4	2010/11/08	Thierry Bouche	UJF/CMD	EuDML schema v. 1 specification, examples
1.0	2010/11/19	Claude Goutorbe, Thierry Bouche	UJF/CMD	Full featured document for partners review
1.1	2010/11/23	Thierry Bouche	UJF/CMD	Included Best practices edited by Jean-Paul Jorda. Tuned examples accordingly
1.2	2010/11/24	Thierry Bouche	UJF/CMD	Light restructuration, small typos fixed, new tables presentation in § 4 contributed by Michał Politowski
1.3	2010/11/25	Thierry Bouche	UJF/CMD	Added executive summary. Expanded some preliminary and final sections.
1.4	2010/11/30	Alan Sexton	UB	Corrected minor typos and English problems as part of internal review
1.5	2010/11/30	Thierry Bouche	UJF/CMD	Last check taking into account feedback from all reviewers
1.6	2010/12/15	Thierry Bouche	UJF/CMD	Small edits taking late feedback into account (mostly layout)

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This deliverable contains original unpublished work except where clearly indicated otherwise.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

1.1 The EuDML schema: What for?

A public well-specified EuDML schema is needed:

1. Because we need to explain to content providers *which metadata* they should expose to EuDML harvesters, which details and granularity is *required* (obligatory metadata), *appreciated* (fundamental metadata), and which further enhancements are expected to *provide added value* to their cooperation with the EuDML project (supplemental metadata).
   Thanks to the specification, they can see what information is wanted by EuDML. They can expose their holdings directly encoded in this way, or they can expose their "richest" format which contains all the relevant information to the extent possible.
2. Because the search engine has to know *where to look* for e.g. an author when a user searches for an author. (Search for "Hilbert" as author must produce quite different results than searching for "Hilbert" as free text or within the whole record of a given item, think of "Hilbert space", "Hilbert transform", which can appear as key words, in titles, cited titles, etc.). The schema serves as a pivot norm for various provider formats and schemas.
3. Because the search engine has to know *what to display* when showing search result lists (Author, Title, bibliographic source, link to full text...) as well as *how to display* complex structures (multilingual information, reference lists, mathematical formulae...). It has thus to know *how they are encoded* in order to present them in the best shape for a given user in a given environment.
4. Because metadata enhancers toolsets have to work on a *defined basis* so that they know what they start from and where they store their results.

1.2 The Journal Archiving and Interchange Tag Suite

After an extensive study of the existing metadata schemas and their actual use by its content providers (deliverable D3.1), the EuDML project evaluated many different strategies and existing schemas that could store it faithfully, and yet reserve room for the enhancements foreseen in the project's work plan. It was then decided to investigate further the framework provided by the Journal Archiving and Interchange Tag Suite (JATS), which has been created by The National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM) in the United States of America. The tag suite defines a set of XML schema modules for storing and exchanging scholarly publications' content. It is a *de facto* standard adopted by a very wide array of institutions worldwide, many of them very similar to EuDML (PubMed Central, JSTOR, Portico...). It provides readily usable schemas for journal articles and books which cover over 97% item types in EuDML. Moreover, some decisive features made it much easier to adapt to our needs:

- It has been adopted as the internal format of one of our partners (EDP Sciences), and is already vastly deployed as an interchange format by many scientific publishers because of their interoperability with one of the above mentioned services.

Its wide deployment and large user community makes it a good reference model for outer interoperability as required in our first point above.
- It is highly customisable and meant for customisation (nevertheless, we decided to keep minimal any deviation from the standard schemas in order to maximise wider interoperability).
- It has room to store any kind of scholarly content up to the full text itself, and to store parallel versions of the same content encoded differently (which is crucial for our points 3 and 4 above).
- Last but not least, it is MathML-ready (yet allowing storage of alternative representation of the mathematical content).

To assess it more objectively this analysis was completed by an attempt to transform large samples of available EuDML metadata as contributed by their providers to one of the JATS DTDs (this was conducted in parallel with Task 3.1—analysis of the collections).

1.3 The EuDML schema, initial version, based on JATS

From this experience, we concluded that JATS needed more work to suit our needs, in two opposite directions:

1. The item types currently supported by JATS are: journal article, book, and book collection (which is defined as "a series of books related in some manner"). While the EuDML "first class citizens" are more diversified: journal article; article or contribution in an edited book such as collective book, conference proceedings...; monograph, doctoral dissertation or similar memoir; multi-volume work such as large books published in separate volumes, collected works, complete œuvres, etc. We thus decided to organise all our content in three major containers: journal article, single book, and multiple books. The two first item types required very minor extensions to existing JATS schemas for article and book (such as allowing a conference description in a book metadata for conference proceedings that are not published in a journal). As the last one doesn't fit perfectly the JATS collection model, we created a new one, called mbook, which has the metadata of a book, but whose content is a list of separate books as in JATS collection.
   These slight deviations from the three standard JATS schemas form the initial version of our EuDML schema specification: see § 5.
2. A drawback of JATS versatility is that it doesn't impose strict constraints on metadata encoding, and often allows for different ways to encode the same information. For efficiency of metadata interchange and exploitation in EuDML, we felt that we needed guidelines so as to have common encoding practice and understanding among all EuDML partners and content providers. The initial version is presented here, § 6, and detailed as Annex A. Further revisions will be released periodically based on feedback from other activities within the project.

2 Introduction: A metadata model for EuDML

The EuDML project aims to design and build a collaborative digital library service that will collate the mathematical content brought by 11 of its partners and make it accessible from a single platform, tightly integrated with relevant infrastructures such as Zentralblatt MATH. As such, it is the first attempt toward a large-scale implementation of a Digital Mathematics Library (DML), and is expected to pave the way towards a truly inclusive and global DML. In this direction, we will try to accommodate new associated partners and to interoperate with relevant infrastructures in the fields of scientific information. Interoperability needs published and documented standards, which is one of the tasks undertaken by EuDML's third work package.

This document contains the first specification of the EuDML metadata schema, which will be used, tested and informed by other work package's tasks, and refined throughout the project duration. It is the first milestone of an ongoing work that will be completed at the project's end, where a definitive specification will be delivered (D3.6, which is due in January 2013).

Its main goals are to:

- identify the essential metadata entities that are relevant to the project;
- provide details on the structure, granularity, and encodings that should be supported by content providers;
- allow content providers to contribute their best metadata to the EuDML central metadata repository based on existing schemas or interchange devices;
- specify the NLM Journal Archiving and Interchange Tag Suite (JATS) as the general frame adopted to encode and exchange the EuDML metadata and list the changes needed in order to support all content types contributed to EuDML;
- develop a set of best practices to ensure perfect understanding of tagging practice among EuDML partners;
- identify the EuDML items' metadata elements that are still not properly handled by the system developed in this work, and suggest directions for improvements.

2.1 Scope of this work

Metadata is usually defined as "data about the data", so in order to target our work on metadata, it is important to make explicit what is the central data we expect to describe with our metadata.

EuDML being the digital metaphor of a mathematics oriented professional library, important concepts that will necessarily be handled in the system, and thus need some internal metadata schema, are: publication (publication containers such as journals or books, as well as individual contributions aka items), person (contributors, and users aka patrons), legal person (person's affiliations, publishers, etc.), user community, user annotation.

However, given the nature of the EuDML central repository, which will be assembled by aggregating content from a number of partner's catalogues, we are lead to single out the individual mathematical *works* in the library as our main relevant data,

and hence to focus on a metadata schema designed to bear all relevant information that can be gathered, consolidated or generated for each integrated full text.

We thus consider publication containers, persons and their affiliations, and publishers, as peripheral information attached to some full text (yet supporting the ability to link to an authority list of such). We also discard all registered users information as well as their possible annotations in this iteration of our work, for these are considered private concepts to the EuDML system, thus inappropriate in a static, exportable representation of the library's content metadata.

2.2 Conceptual Requirements

A firm conceptual metadata model is needed for EuDML.

EuDML harvest

- EuDML needs to explain to data providers which metadata they should expose to EuDML harvesters. This explanation can come in two ways: EuDML can provide them with one or several "EuDML recommended (standard) schema(s)" as an example so that they can transform and expose their stuff in exactly this way (if not already available). Or they can see from the EuDML conceptual schema what information is wanted by EuDML and expose their "richest" format which contains all this information (however, encoded differently). Typing metadata as e.g. "obligatory", "fundamental", "supplemental" will give additional guidance and serves the purpose of maintaining a defined level of quality.
- EuDML content providers must have the possibility to retrieve the EuDML-enhanced metadata for their collection of items, in order to improve their local collections to a higher level of quality.

EuDML operation

- The EuDML search engine has to know where to search for and on which criteria to filter when given a user query or when guiding users through a browsing process, or when providing alerts (e.g. RSS feeds).
- The EuDML system has to know which details to present to users when constructing search result lists or browsing features, in order that the user can identify and select the publications of interest (on bibliographic or contents-related criteria), and obtain the full text thereof.
- The EuDML system has to know which version of a relevant metadata field has to be chosen and how it has to be transformed in order to be properly displayed at user or search engine request in various formats (examples: a formula in a title which might be stored as text, LaTeX, or MathML needs typically to be converted to regular XHTML on-the-fly when the search result is displayed in a Web page, this conversion requires knowing what formats are available for this particular formula in the EuDML metadata registry, choosing the best available format for the specific display context, and translating accordingly; another example is the text-math encoding conversion involved to generate BibTeX as well as Endnote XML, etc.).

- The EuDML search engine has to know where mathematical formulae can occur, how they might be encoded, how they can be searched for, including wildcard search, and how they should be presented to users (depending also on browser capabilities).
- Multilingual information and information present in various character sets or transliterations (e.g. Latin, Greek, Cyrillic and their transliterations to other alphabets) must be adequately dealt with in most of the above scenarios.
- A defined basis for "metadata enhancers" (WP7-10) is needed, in order to specify and implement the respective functionalities. Some examples: reference analysis, association analysis, duplicate detection, metadata enrichment from various sources. More specifically to our corpus: a metadata enhancer should be able to scan an existing metadata record to find, for instance, a reference to a formula, generate a new format for this formula (e.g., by OCR or translation to MathML from LaTeX code), and to store the resulting object as an alternative format to the pre-existing one(s).

EuDML interoperability

- EuDML needs to explain to aggregators (such as Europeana) what metadata they can expect from EuDML, and in what format they can harvest it.
- Enhancement and internal processing, as well as feedback to providers, will require certain administrative information attached to the items of the EuDML collection.
- Regular "EuDML content dumps" are necessary for security and sustainability reasons, as well as replication or mirroring. Schemas, formats and encodings need to be clearly defined.

2.3 Out of scope functionality

The following do not constitute requirements on EuDML services and are thus not in the scope of a EuDML metadata schema:

- Handle material that is not considered as having been persistently and formally published (e.g. preprints, personal web pages...).
- Special provisions for papers not generally accessible online (e.g. on paper only, inhouse access only, library catalogue...).
- Version control for documents, as EuDML only considers works in published final form.
- Complicated author/contributor structures for documents, as this is of no significance in math publishing.
- Description of access embargo periods (moving wall) and other licensing, access barriers, digital rights management issues, since EuDML follows an eventual open access policy and leaves those issues under full control of the respective content (fulltext) providers.

2.4 Outline of the deliverable

The next section, 3, introduces the basic concepts and definitions needed throughout this document. A thorough analysis of the EuDML schema requirements follows in section 4.

Based on this analysis, the EuDML schema specification is detailed in section 5. The work on best practices is introduced in section 6, while open questions are left to the concluding section 7.

At the time of delivering this document, a initial set of recommended best practices has been developed, which is included as Annex A, while some full XML examples of items encoded according to the EuDML specification and recommended best practices are included as Annex B.

Following this work, three XML DTDs have been written, based on JATS DTDs. These, as well as a sample of carefully prepared XML full examples will be made available from a dedicated area of the project's Web site, in its "Resources" area.

Massive transformation of partner's metadata is ongoing based on the included specification. The results will be available to all partners through OAI-PMH and hopefully help the next step of integration in EuDML project, that is Task 3.4—Metadata export.

3 Methodology & Definitions

This section describes the principles, methods and notions that are used to define the EuDML metadata schema in the next sections. As this document, and these specifications, will be revised during the EuDML project, we also describe how this further development might be carried out.

The central object in EuDML, used as the unit of delivery and thus as the pivot for the metadata schema, is an *item*, which was the main perspective on analysing the contributed content in D3.1. Loosely put, an item is the kind of mathematical content that would be reviewed in *Zentralblatt MATH* or *Mathematical Reviews*, so the relevance of this concept is quite consensual in the scientific community.

3.1 Definitions

3.1.1 The EuDML item

Deliverable D3.1 ("Report on available collections and metadata") defines relevant logical units that can be delivered in the context of EuDML in the following way:

"An item is a self-contained mathematical text which has been scientifically validated and formally published".

According to FRBR [9] terminology, an item is a single *manifestation* of a mathematical *work*. The concept of expression (different incarnation of the same "idea") does not make sense for the mathematical corpus as, although the same "ideal work" can be manifested through various channels such as a conference, an abstract, a full paper, or a monograph, it is not possible in subsequent works to refer uniquely to any of them, as the actual details contained in each manifestation could differ enough to make the reference ambiguous. Even a solid abstract reference such as "the Hahn-Banach theorem" might be stated with quite varying hypothesis and conclusions depending on the context where it is manifested. FRBR introduces a concept of item which is a *copy* of a given manifestation (e.g. two copies of the same edition of a given book). This has no relevance either to EuDML, as EuDML describes the distinguished reference copy of a manifestation that has been published (either on paper and later digitised, or digitally).

As the main focus of the EuDML project is to ease discovery, access, use and exchange of mathematical items, *the EuDML item* is thus the primary entity type described by the schema.

The identification of an item, as a formally published text, essentially requires bibliographic data which describes *where* and *by whom* it was published and depends on the *type* of publication (journal article, book, etc.).

For this version of the schema, we specifically define the following publication types, which are logical subclasses of the generic "item" class:

- a multivolume work;
- a book, namely
  - a single volume from a multivolume work,

  - a monograph (which might be a doctoral dissertation, a memoir...),
  - an edited book (a book that contains chapters or articles that have been written by different authors and collated by scientific editors, which might be a conference proceedings volume);
- a part of a book such as a chapter, or a contribution in a proceedings volume;
- a journal article.

3.1.2 Obligatory metadata

We define obligatory metadata as the bare minimum of metadata information that is requested from EuDML data providers. This is not exactly a functional category but rather a policy requirement.

Obligatory metadata is the required minimum of metadata in order to unambiguously identify and handle a relevant mathematical publication in the scope of EuDML: Item type, authors, original title, bibliographic reference for this publication with enough structure so as to enable collection's browsing, unique identifier, URL of full text.

3.1.3 Fundamental metadata

Fundamental metadata is what satisfies the functional requirements for browsing, searching and reference matching over the collections at item level. It enables basic digital library interaction with the EuDML corpus: it is equivalent to the "basic" metadata described in D3.1.

The term fundamental was chosen so that it is clear which information is needed to provide the fundamental functionality expected by typical users. It is a qualitative superset of obligatory metadata.

If this information is relevant to the item described, then it must be present in the metadata. If it is absent from provider's original metadata, then WP7 must provide a solution in order to enable this publication in EuDML.

It contains obligatory metadata (see above) as well as standard optional information (abstract, key words, main language) that should be there, or generated in WP7.

3.1.4 Supplemental metadata

Beyond fundamental metadata, this is additional metadata that should be stored (WP3-5), generated (WP7-10), and exploited (WP5-6) within EuDML.

Supplemental metadata is whatever goes beyond fundamental metadata (e.g. relations to subject ontologies, authority lists, MR/ZM IDs, multilingual, multiscript, bibliographies/references, interlinking, math handling...), yet has relevance to the EuDML's corpus specificities and EuDML system functionalities as depicted in the Description of Work of the project.

3.2 Detailed analysis of the EuDML metadata requirements

Metadata exists to support the functionalities expected from the system. In this section we describe the functional aspects of a Digital Mathematics Library (DML) that we intend to provide:

- Uniquely identify an item not only within EuDML, but across the whole mathematical literature.
- Discovery of published items by
  - fielded search on various attributes such as author names, titles, publication year, subject, abstracts, journal title, key words,
  - browsing collections by selecting a starting point such as a given journal name, mathematical classification code, author name,
  - sorting and filtering search or browse results,
  - automated reference matching to help external resources turn their citations into links to EuDML items.
- Retrieving a specific item through a known identifier such as a DOI, URI or other unique identifier.
- Assert the relevance to the user of a given item through the display of attributes such as subject, abstract, language, and citations to and from that item.
- Display and indexing of attributes in multiple languages or transliteration systems.
- Interlinking as a powerful access tool to mathematical resources. Examples of this consist of links to reviews in the major reviewing databases (Jahrbuch, Zentralblatt MATH, MathSciNet), and links to and from citations from subsequent works.
- Linking to other material such as user provided annotations, author identification services.
- Display of mathematical formulas in various formats based on the user's choice or capabilities (e.g. MathML, TeX, graphics, speech synthesis).

Besides the end user oriented functionalities, the schema should also serve as an exchange model.

3.3 Methodology

Following the above analysis of EuDML metadata requirements, and in parallel with Task 3.1—analysis of the collections, we developed an abstract model for EuDML metadata elements which is detailed in the next section.

As part of our study, we converted large sample metadata sets from a number of partners to plain NLM DTD and inspected where conversion was difficult to achieve, when doubts or choices had to be made, when the target structure did not accommodate the source structure, etc. When we faced the necessity to choose between different structures offered by NLM DTD, we took note and started an open discussion within our working group which ended up in a number of best practice recommendations. When we found metadata that could not be faithfully stored in the existing NLM DTDs, we took note of this for further processing. Finally, we took the design decision to adhere as closely as possible to the existing NLM DTDs, implemented the small modifications that were required to faithfully store all encountered item types, and left aside some more

modifications, waiting for more feedback from the actual implementation of the project to be realised in the forthcoming months. These three sets of actions (expand JATS to support all EuDML items, document encoding preferred decisions, record questions left unanswered at this stage of the work) yield the three last sections of this document.

3.4 Conversion summary

The following table summarises the number of item types from various EuDML collections which were converted in order to evaluate our results presented here.

Provider/Collection	EuDML metadata (Schema)	Notes
CMD/CEDRAM	1 242 (**article**)	converted from internal XML with MathML
CMD/NUMDAM	40 478 (**article**)	converted from internal XML
CSIC/DML-E	6 401 (**article**)	converted from SQL database
EDPS journals	2 723 (**article**)	need slight tweaking to obey best practices
FIZ/ElibM	25 497 (**article**)	converted from internal XML
IMAS/DML-CZ	26 476 (**article**), 132 (**book**)	converted from internal XML
SUBGoe/Mathematica	53 396 (**article**), 2 298 (**book**), 296 (**mbook**)	converted from METS XML
SUBGoe/RusDML	16 486 (**article**)	converted from METS XML
BNP/Port. Mat.	1 347 (**article**)	converted from TEL XML
All	**176 772 records**	

EDP Sciences records are available from `http://oai.edpsciences.org/` (one set per journal: cocv, ita, m2an, mmnp, proc, ps, ro; metadata format: pmc).

All other records are available from `http://math-thar.ujf-grenoble.fr/repox/OAIHandler` (one set per collection: CEDRAM, DML_CZ_Serial, DMLE, ELibM, GDZ_Mathematica, GDZ_RusDML, NUMDAM, PM: NLM-AI metadata format; DML_CZ_Proceeding, DML_CZ_Monograph, GDZ_Monographs, GDZ_Band: NLM-Book metadata format; GDZ_MBook: NLM-MBook metadata format). This server is IP-protected during the testing phase of the project.

4 The EuDML metadata elements

The following sections contains an abstract description the metadata elements that are defined and required in the EuDML metadata schema, version 1.

The EuDML item is a hierarchy of metadata elements, including composite metadata elements and simple metadata elements. Composite elements consist of an aggregation of subelements and do not have values by themselves. Simple elements are the leaf nodes of the hierarchy and carry actual data values.

The functional requirements described in section 3.2 suggest three broad categories of metadata elements:

- elements that together identify a given publication;
- elements that describe the intellectual content of the publication;
- elements that allow a given publication to be "put into context" through relations with other publications.

Although there is an amount of overlap between these categories, e.g. an article's title is used for identification and hopefully also conveys some information about its intellectual content, we feel that this classification makes for a clearer description.

Each element description consists of tabular information with the following structure:

- **Name:** a name for the metadata element.
- **Description:** a short narrative description of the element, including its intended usage with regards to the functional requirements.
- **Attributes:** metadata elements carry content and information about this content, which is captured through attributes (e.g. the language in which a title is written). These are abstract attributes and does not imply the existence of a corresponding attribute in concrete XML notation.
- **Occurs:** cardinality in the form "minimum number of occurrences:maximum number of occurrences".
- **Value:** For simple elements, the schema also describes the value type which may be one of the following:
  - enumerated: the set of allowed values for the metadata element,
  - string: a plain character string,
  - text: a character string possibly including tagged subparts (also known as *rich text*),
  - math text: text that may include mathematical formulas.

  Text content is not described in detail in this section. In particular 1) the description of valid internal markup is left to the detailed description of the concrete XML notations that may be used to implement this abstract specification and 2) the same is true of the mechanism used to encode alternative versions of a given math formula.

In addition to this tabular information, a **Notes** subsection explains the intended usage and functionality supported by the metadata elements.

4.1 Identifying an item

The most basic requirement defined in section 3.2, is the ability to uniquely identify and retrieve a given item among the whole mathematical literature. The set of metadata elements needed for this identification consists of standard bibliographic information, including author names, title and other elements that depend on the *item's type*, as defined in section 3.1.1.

We first introduce bibliographic metadata elements that are common to all these document types, then describe elements that are specific to a given document type. *Identifiers* are also described in this section.

4.1.1 Common bibliographic elements

Item type, identifiers

Name	Description	Attributes	Occurs	Value
item type	the type of item described		1 : 1	enumerated list[1]
eudml identifier	the unique identifier of this item in the context of EuDML		1 : 1	string
item's url	the URL where this item can be found[2]		1 : 1	string
url of full text	the URL where the full text of this item can be found	file format	0 : ∞	string
oai identifier	the identifier used by the OAI exchange protocol	OAI-PMH server ID	1 : 1	string
provider identifier	the identifier of this item in provider's information system (needed to allow for bidirectional exchange of item descriptions)	provider ID	1 : 1	string

Contributor

Name	Description	Attributes	Occurs	Value
contributor	Contains information about a single contributor to an item	Attributes: type of contributor (author, editor, translator, other)	0 : ∞	structured

1. Journal article, conference article, book part, monograph, Ph. D., proceedings volume, edited book, multivolume work.
2. Usually an HTML page, exposing bibliographic information and links to the public full text.

Name	Description	Attributes	Occurs	Value
name	the name of a contributor. This may be in the form of simple string or a structured name, in which case the following three subelements may be used.		1 : 1	string or structured name
last name	family name of a person		1 : 1	string
first names	all given names of a person		0 : 1	string
suffix	a qualifier that follows a person's name such as "Jr", "III",		0 : 1	string
name transliteration	the name of a contributor, transliterated from a non-Latin writing system	transliteration system	0 : 1	string or structured name
affiliation	the name of the contributor's institution		0 : ∞	text
uri	a URI for this contributor (see notes below)		0 : ∞	string
address	the address of the contributor as it appears in the given item		0 : 1	text
email	the email of the contributor as it appears in the given item		0 : 1	text

Notes

Basic functionality requires only the *name* subelement. *Affiliation*, *address*, *email* are additional non essential information that may help a user identify or contact an author. An *URI* may be used in the case where the actual contributor has been identified (e.g. through the Zentralblatt-Math author identification service). There is no provision for institutions identifiers, as we do not expect EuDML to be able to deal with these.

Title

Name	Description	Attributes	Occurs	Value
title	Original title of the item	language	1 : 1	math text
translated title	a title translated from the original title	language	0 : ∞	math text
alternative title	a "different" version of the title, usually created so that it can be processed or displayed in a special way	language, transliteration system	0 : ∞	math text

Notes

Basic functionality requires only the original title. The alternative title will mainly be used to capture the Latin transliteration of Greek and Cyrillic titles.

Item type dependent metadata

The following sections are arranged by item type. Depending on the item type, metadata includes a description of the parent publication ("container", which may or may not exist as an EuDML item) if applicable.

4.1.2 Journal articles

Name	Description	Attributes	Occurs	Value
contributor	The contributors of the article. See contributor section 4.1.1			
title, translated title, alternative title	Title information. See titles section 4.1.1			
issue	a container for this article. See description below		1 : 1	structured
first page	the page number on which an article starts		0 : 1	string
last page	the page number on which an article ends		0 : 1	string
page count	the number of pages in the article		0 : 1	string
article sequence number	Some journals do not use a continuous paging for articles. This element is used to record the article sequence number within the issue		0 : 1	string
article ID	Some electronic journals do not use a continuous paging for differentiating articles. This element is used to record the article unique identifier within the issue or whole journal		0 : 1	string

Journal issue

Name	Description	Attributes	Occurs	Value
journal title	the title of the journal in which the article appeared	language	1 : 1	math text
journal subtitle	the subtitle of the journal, if it has one	language	0 : 1	math text
abbreviated journal title			0 : 1	text
alternative journal title	a "different" version of the title, usually created so that it can be processed or displayed in a special way	language, transliteration system	0 : ∞	math text
journal identifier	any identifier a content provider is able to provide for this journal	type of identifier (issn, uri, Zbl identifier, provider identifier...)	0 : ∞	string
journal description	A full text description of the journal contents as a whole, e.g.: "The journal of advanced arithmetics publishes advanced level articles in the field of arithmetics"		0 : 1	math text
volume number			0 : 1	string
volume series			0 : 1	string
issue number			0 : 1	string
issue date			0 : 1	string
issue identifiers	any identifier a content provider is able to provide for this issue	type of identifier (provider identifier, url...)	0 : ∞	string
issue title	The title of a special or thematic issue		0 : 1	math text
alternative issue title	a "different" version of the title, usually created so that it can be processed or displayed in a special way	language, transliteration system	0 : ∞	math text
issue description	The description of a special or thematic issue		0 : 1	math text
issue editors	The editors of a special or thematic issue		0 : ∞	
publisher name			1 : 1	text
publisher location			0 : 1	string

4.1.3 Books

Monographs and edited books

Name	Description	Attributes	Occurs	Value
contributor	The contributors of a monograph, or the editors of an edited book. See contributor section 4.1.1			
title, translated title, alternative title	Title information. See titles section 4.1.1			
series	the series in which a book was published. See description below			
edition statement	such as “2nd ed.”		0 : 1	text
publication date			0 : 1	string
isbn	the ISBN identifier for the book		0 : 1	string
series number	the number of the book in its series		0 : 1	string
publisher name			1 : 1	text
publisher location			0 : 1	string

Proceedings volume

A proceedings volume is essentially described by the same metadata elements as a book, with the addition of elements that describe the conference.

Name	Description	Attributes	Occurs	Value
conference title	The formal name of a conference		1 : 1	math text
alternative conference title		language, transliteration system	0 : ∞	text
conference subtitle			0 : 1	math text
conference acronym			0 : 1	string
conference location			0 : 1	string
conference date			0 : 1	string
conference subject			0 : 1	math text
conference number			0 : 1	string
conference organiser			0 : ∞	string

Name	Description	Attributes	Occurs	Value
proceedings volume number	The proceedings of a conference may be published in several distinct volumes.		0 : 1	string
publication date			0 : 1	string

Book series

Name	Description	Attributes	Occurs	Value
series title			0 : 1	text
alternative series title	a "different" version of the title, usually created so that it can be processed or displayed in a special way	language, transliteration system	0 : ∞	text
series subtitle			0 : 1	text
series description		language	0 : ∞	text
series identifier	any identifier assigned by the content provider		0 : ∞	string
publisher name			1 : 1	text
publisher location			0 : 1	string

4.1.4 Book parts

Name	Description	Attributes	Occurs	Value
book	The description of the parent book, described in section 4.1.3, books		1 : 1	
part type	chapter, conference...		1 : 1	string
contributor	The contributors of the article. See contributor section 4.1.1			
title	Title information. See titles section 4.1.1			
alternative title	a "different" version of the title, usually created so that it can be processed or displayed in a special way	language, transliteration system	0 : ∞	text
first page	the page number on which the part starts		1 : 1	string
last page	the page number on which the part ends		1 : 1	string
page count	the number of pages of the part		0 : 1	string

4.2 Describing the intellectual content

The following metadata elements are used to describe the subject of the work. They allow a user to

- decide whether the item is relevant to his needs.
- search or browse by subject (in a broad sense)

Name	Description	Attributes	Occurs	Value
abstract	A summary of the content of the given work	language	$0:\infty$	math text
translated abstract	A translated summary of the content of the given work	language	$0:\infty$	math text
alternative abstract	A different version of the abstract. In the context of EuDML, this element will be used to hold Latin transliterations of Greek and Cyrillic abstracts	transliteration system	$0:\infty$	math text
msc	codes from the mathematical subject classification that have been assigned to the item	msc revision	$0:\infty$	string
keyword	keyword or keyphrase assigned to the item	language	$0:\infty$	math text

4.3 Putting an item into context: interlinking

A given mathematical work exists in a broader context. Metadata should enable the discovery and access of logically related works, by providing navigation links.

4.3.1 Reviewing databases

Links to the three major reviewing databases in mathematics give access not only to the item's review itself, but also to related works that exist outside of EuDML, and are a fundamental feature of the interlinking infrastructure. The following three metadata elements are used for this purpose.

Name	Description	Attributes	Occurs	Value
zblid	identifier of the given item in the Zentralblatt database		$0:1$	string
mrid	identifier of the given item in the MathSciNet database		$0:1$	string
jfmid	identifier of the given item in the Jahrbuch database		$0:1$	string

4.3.2 Bibliographies

Bibliographies are also a fundamental tool for the discovery and access of related works. The EuDML metadata for bibliographies should not only consist of the references themselves, but also of any identifying system that allows direct or indirect access to the cited item.

Reference metadata

The bibliography is a list of references consisting of the following elements.

Name	Description	Attributes	Occurs	Value
reference	a bibliographic reference usually found in a "reference" or "bibliography" section at the end of the item's content		$1:\infty$	math text
zblid	identifier of the referenced work in the Zentralblatt database, if any		$0:1$	string
mrid	identifier of the referenced work in the MathSciNet database, if any		$0:1$	string
jfmid	identifier of the referenced work in the Jahrbuch database, if any		$0:1$	string
eudmlid	identifier of the referenced work in the EuDML database, if any		$0:1$	string

Notes

The EuDML system must be able to handle two kinds of reference, namely unstructured and structured. An unstructured reference is simply a string, possibly with identified (tagged) substrings such as author name, title, etc. The detailed list of subparts that may be identified (through tagging) is not specified here but must be specified by a concrete XML notation.

Structured references are composite elements aggregating the standard fields that may appear in a reference. These fields are not specified here but must be specified by a concrete XML notation.

4.3.3 Related items

A related item is an EuDML item that bears some semantic relation to this one. This element is not used to capture structural relationships such as "is part of". Examples include: corrigendum, addendum, next/previous article in a series, similar work.

Name	Description	Attributes	Occurs	Value
related item	an EuDML item that is somehow related to this one	type of related item. A list of allowed values should be specified by a concrete notation	$0:\infty$	string (eudml identifier)

4.4 Full text

In mathematics, as there is no single universal primary format from which one could derive all formats needed for various usages of a mathematical text, we are bound to collate many versions of the full text of an item, one of them being considered "data" (typically a PDF) while others are considered metadata (this can range from flat Unicode OCRed text to semi-structured versions such as HTML with math expressions stored as images, or highly structured XML with MathML formulae).

Name	Description	Attributes	Occurs	Value
body	some searchable version of the full text	encoding type	$0:1$	rich text math

Notes

Rather than having multiple full texts with different encodings for one item in order to serve different exploitation scenarios, we favour one full text that may have as many parallel versions for certain constructs as needed (such as: a formula in flat OCRed text, LaTeX, MathML, image, sound, SVG...).

5 The EuDML metadata schema

This section is about how EuDML metadata will be encoded and physically appear or be transported in certain given scenarios (such as during metadata harvest from EuDML data providers, or exposition of EuDML metadata to aggregators, e.g. Europeana, or for a "snapshot" or "dump" of EuDML contents).

As we do not want to reinvent the wheel, a quick survey of existing XML encodings was conducted, paying special attention to the following requirements:

- mathematical formulas should be supported in a variety of formats, including MathML;
- rich text should be allowed where applicable, in other words the encoding used must account for a number of basic formatting elements such as typographical attributes;
- the description of reference lists (bibliographies) should be taken into account, as they are an essential tool for researchers;
- using a recognised and widely deployed standard would be a bonus. However, as we do not expect an existing XML document type definition or schema to be able to describe our data "out of the box", it should be easily customisable.

5.1 Review of evaluated metadata encodings

We evaluated the following schemas which all provide some partial solution to our query:

EULER Euler FP5 project metadata, which was developed for cataloguing (non-digital) resources existing in various European libraries [3];

SWAP Scholarly Works Application Profile in qualified Dublin Core, which essentially provides granularity to describe (with raw text metadata) any digital scholarly work (detailed bibliographic description, eprint versioning, validation status) [2] ;

MODS Metadata Object Description Schema from the Library of Congress, which is pretty much an interchange format for multimedia library catalogues [5];

DML-DC Euclid/NUMDAM/GDZ recommendation on presenting DML metadata in simple Dublin Core, which was an attempt to qualify simple Dublin Core for making metadata interchange more useful between DMLs by URI-like prefixing repeated elements, as well as some best practices recommendations for mathematical expression encoding in titles and abstract [1];

MLAP Mathematical Literature Application Profile for Dublin Core by David Ruddy, which is a relatively strict yet very generic schema for interchanging precise bibliographic records of scholarly works [8]. Besides the fact that mathematicians are eager to exchange this kind of information in order to build larger DMLs and further the interlinking of existing DMLs, this proposal has nothing specific to mathematical content;

JATS NCBI/NLM Journal Archiving Tag Suite, which was created with the primary intent of providing a common format in which publishers and archives could exchange journal content [7].

While Dublin Core metadata is nowadays a central device for wide interoperability, especially for enhancing visibility of heterogeneous collections, it was felt that DC based formats would be useful for exporting EuDML metadata but not for storing the consolidated master with all information and additions foreseen in the project's DoW. In fact, DC is so generic that, among its 15 elements, few are relevant to a digital library project such as EuDML, and a lot of structure has to be added to qualify and organise information we would expect from each of the principal elements. This is what application profiles such as EULER, SWAP, DML-DC and MLAP are aimed at, each of these developed with a specific aspect of literature interchange in mind. MODS is a more constrained framework that can be used, together with METS, in order to describe a precise bibliographic record of a catalogued object, as well as its physical description—no room exists, apart from using relations to external objects conforming to some other format, for encoding parts of an item's textual content like bibliographies. However, none of these provide support for mathematical knowledge encoded as such: the mathematically oriented standards just favour TEX notation as it can be embedded into any XML file as text modulo some escaping.

Inera Inc. provides some introduction to the NLM Journal Archiving Tag Suite (JATS in the following) [4]:

> The NLM Journal Archiving and Interchange DTD Suite, co-authored by Inera Inc., Mulberry Technologies, and NCBI, is the de facto standard full-text DTD for scholarly publishing.
> Since the DTD was first released in April 2003, it has been (for the scholarly publishing world) rapidly adopted. Whereas ISO 12083 never achieved broad acceptance, the NLM DTD has already been adopted by hundreds of journals (probably north of 500) worldwide. Many small and medium-sized publishers have adopted the NLM DTD, and a number of larger publishers are preparing to deliver content according to the NLM DTD when asked. Most of the major journal publishing compositors and service suppliers are up to speed on the DTD and happy to deliver content tagged with it.
> The NLM DTD has also proven popular with aggregators. It is the "house" DTD of Atypon Systems and the recommended DTD for full-text content at Ingenta and Highwire Press. And, of course, NLM uses it for PubMed Central.
> The NLM DTD has been no less popular with libraries. In a joint press release, the British Library and the Library of Congress announced that they would support the NLM DTD as their archiving standard for electronic content. It has also been adopted by Portico (a major Mellon-funded archive effort).

Complemented by Mulberry Technologies, Inc. [6]:

> The Journal Archiving and Interchange Tag Suite (also called the NLM DTD although it is available in DTD, XSD, and RNG forms) provides a common XML format for preserving the intellectual content of journal articles, independent of the form in which that content was originally delivered. The Tag Suite consists of Tag Sets for Archiving, Publishing, and Authoring journal article content and a Tag Set for Books and book material. The Tag Sets have been widely adopted by archives, libraries, and publishers and are supported by many data conversion vendors and XML tools.

NISO (the National Information Standards Organization) is now working to make the JATS into a NISO standard.

As JATS was already used internally by one of our partners (EDP Sciences), and proved to have room to store faithfully all of the metadata encountered while reviewing the EuDML content to be integrated, and moreover provided standard structures for most of the new elements foreseen in the DoW (full text encoding, native support for MathML and alternative versions of formulae, notably), it was an easy task to select it as best candidate for our purpose. It is a trivial task to derive most DC based metadata from carefully organised JATS files (while the converse would require a JATS application profile in Dublin Core).

Moreover, as NLM DTD is being used by PubMed Central, one can expect that most generalist STM publishers will have a workflow to generate this format, so that being interoperable with EuDML would not put too much technical burden on their side. This is why, although JATS can be easily customised to fit any special need, we will try to adhere to it to the largest extent possible, specifying best practices recommendations in order to attain maximum compatibility among EuDML partners, and reliability of exchanged metadata with third parties.

JATS provides three DTDs that we will adapt for describing our three main content types:

- **The Journal Archiving and Interchange Tag Set** implements article3.dtd for journal articles (cf. `http://dtd.nlm.nih.gov/archiving/`)
- **The NCBI Book Tag Set** implements book3.dtd for books and bookcollection3.dtd for collections of books (cf. `http://dtd.nlm.nih.gov/book/`)

We will now detail how our conceptual metadata elements can be expressed in XML using JATS.

5.2 EuDML metadata specification v. 1.0

The EuDML metadata schema version 1.0 as defined by this document is implemented in three DTDs providing the 3 root elements holding XML metadata for three major types of items, namely journal articles, books, and multivolume works. A consequence of this choice is that book parts (typically individual articles in a proceedings volumes), while being "first class citizens" in our abstract model, are described and exchanged within the whole book they belong to (this is further discussed in section 7).

Journal articles are described with a minimal extension of the Journal Archiving and Interchange Tag Set version 3.0 with root element <**article**>:

- the xml:lang attribute is allowed for the <**issue-title**> element.

Books are described with a minor extension of the Book Tag Set version 3.0 with root element <**book**>:

- a child <**conference**> element (as in <**article-meta**>) is allowed in <**book-meta**>; this element is needed to describe conference proceedings volumes;
- a child <**book-part-id**> with attribute pub-id-type is allowed in <**book-part-meta**>; this element is used to preserve item-level identifiers, when parts of a book are EuDML items;

- the pub-id-type attribute to **<book-id>** and **<book-part-id>** can have values beyond a restricted list; it is used in particular to identify the authority who assigned the identifier.

Multivolume works are described by a new root element **<mbook>**. Multivolume works' metadata is identical to **<book>** metadata, with the addition of references to individual constituents (volumes). The element **<book-meta>** is replaced by **<mbook-meta>** with same structure, except:

- a child **<mbook-list>** element is required in **<mbook-meta>**. It is a container for individual volumes, as in JATS collection DTD;
- each component volume reference is captured by an **<mbook-volume>** element (child of **<mbook-list>**), with the following children:
  - **<title>**: the title of the volume,
  - any number of **<book-id>** and **<ext-link>** elements.

  While the EuDML internal machinery only needs **<book-id>**s in order to implement the multivolume work/individual book relationship, the **<title>** and **<ext-link>** elements should be useful to external applications for display and access purposes. Each individual volume in a multivolume work is encoded with the Book DTD.

5.3 EuDML interoperability

When acquiring metadata from different partners, it was observed that any reasonably structured format is rather easily converted to JATS format. The big drawback with many OAI-PMH servers is that they only serve the mandatory OAI-DC format in such a way that many different metadata elements are stored in the same, repeated Dublin Core element. As a consequence, only heuristics based on order of appearance, or pattern matching on an element's value allows disambiguating the metadata thus contributed. For instance, **<identifier>** can be used to transport an ISSN, a textual bibliographic reference, a URL, etc.

Qualified versions of Dublin Core that are modelled on the metadata schema with finest grain available to the content provider allow faithful interchange of metadata. Qualification can be imbedded into the value of simple Dublin Core elements as in the DML-DC recommendation or similar qualification using URN-like prefixes, or it can use qualified elements and a documented application profile such as SWAP or MLAP.

As of writing this report, the best scenario for returning EuDML metadata to providers is to use the EuDML schema over OAI-PMH communication channels.

For interoperability and visibility beyond EuDML partners or associated partners, a simple transformation has been developed to represent a subset of EuDML metadata in OAI-DC (compliant with DML-DC) so that general harvesters can manage our metadata. This is available in the OAI-PMH server `http://math-thar.ujf-grenoble.fr/repox/OAIHandler`.

6 Best practice recommendation for mapping EuDML abstract metadata to the EuDML schema

A best practices working group for representing EuDML metadata in JATS notation was formed in Prague: Jean-Paul Jorda (chair), Thierry Bouche, Claude Goutorbe, and Michal Politowski.

A set of recommendations has been derived, and is being tested on large amounts of EuDML items. Complete examples of EuDML XML files obeying these recommendations are included as Annex B to this document.

The recommendation itself is a work-in-progress, which is available to the project's partners[3] as a live HTML page. When mature enough, it will be placed in an area of the `www.eudml.eu` website dedicated to developers' resources. Up-to-date documentation will be available there for download as well: the specification, the DTDs and possible associated tools.

The current content of these recommendations is included as Annex A to this document.

3. `https://wiki.eudml.eu/eudml/D3.2-Work-in-progress_area/Best_Practices`.

7 Current limitations and open questions

7.1 Limitations

While preparing this specification, and conducting test transformations on large parts of the EuDML corpus (circa 170,000 items from 9 collections analysed), we noticed some limitations in the JATS design that were not overcome by our small number of changes.

- JATS has no issue-level metadata (special issue title, description, editors...). Any such metadata must currently be placed in <**article-meta**>, for instance with <**ext-link**> to point to some resource describing the special issue.
  Some collective articles have scientific editors in the place of authors, cf. `http://www.numdam.org/numdam-bin/fitem?id=RSMUP_2005__113__129_0`. A way to distinguish special issue editors and article editors is to use a dedicated contrib-type. This decision is left to the best practices working group (§ 6), as well as the conclusion whether the lack of issue-level metadata is critical or harmless to EuDML.
- There is no provision for a complete description of the series (or collection) a book belongs to (if any). Only the collection name can be given currently.
- The relationship between a given volume and the work it is part of (if any) is only described at the collection level, e.g. a multi-volume work is said to contain volumes (books). While this is not expected to be a problem for the internal EuDML machinery, some consumers of the EuDML metadata may need to harvest a completely self-contained book description, including a description of the enclosing work.

7.2 Open questions

- It is not possible in NLM DTD to store a structured and an unstructured version of a given citation side-by-side. We could decide that a single <**ref**> element only contains *one* reference, but can contain one or both of <**element-citation**> and <**mixed-citation**>. This would have the same benefits as having <**string-name**> together with <**name**> elements. However, this "best practice" would clearly break compatibility with recommended JATS practice where each <***-citation**> element in a <**ref**> encodes a distinct reference. That would however fit better our understanding and reserve room for expected enhancer module's functionalities, as one could imagine that a matching module could find the DOI, or Zentralblatt ID from the string citation, thus be able to add a structured version of the same citation together with more <**ext-link**> elements.
- It is still not obvious to what extent our item-centric model will support all EuDML metadata needs. For instance, it is obvious that some knowledge about serials, authors and institutions would help in navigating the collection and making internal relations. Alternative spellings of authors, journal nick names, etc., would enhance retrieval. For these, we could either rely on external specialised databases providing identifiers and specific supplementary metadata, or we could try to reconstruct this knowledge from the harvested metadata and store it in the EuDML schema (this can be done using the available machinery for alternative titles, person names, etc.).

References

[1] Thierry Bouche, Thomas Fischer, Claude Goutorbe, and David Ruddy. A recommended best practice for unqualified Dublin Core metadata records. Available online at http://projecteuclid.org/collection/euclid/documents/metadata/dml_dc.html, 2009.

[2] DCMI Eprints Working Group. Scholarly works Dublin Core application profile. Available online at http://dublincore.org/scholarwiki/SWAPDSP, 2006.

[3] EULER project. The EULER application profile. Available online at http://www.emis.de/projects/EULER/metadata.html, 2002.

[4] INERA Inc. NLM DTD Resources: Introduction. Web page available online at http://www.inera.com/nlmresources.shtml.

[5] Library of Congress. MODS schema. Documentation available online at http://www.loc.gov/standards/mods/mods-outline.html, v. 3.4: 2010.

[6] Mulberry Technologies Inc. JATS - The Journal Archiving and Interchange Tag Suite. Web page available online at http://www.mulberrytech.com/JATS/index.html.

[7] National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM). Journal archiving and interchange tag suite. Documentation available online at http://dtd.nlm.nih.gov/, v. 3.0: 2008.

[8] David Ruddy. Developing a metadata exchange format for mathematical literature. In Petr Sojka, editor, *Towards a Digital Mathematics Library*, pages 27–36, Brno, Czech Republic, 2010. Paris, France, July 7-8th 2010, Masaryk University Press. Paper available online at http://www.dml.cz/bitstream/handle/10338.dmlcz/702570/DML_003-2010-1_4.pdf, XML application profile available at http://projecteuclid.org/documents/metadata/mlap/mlap_dsp.xml.

[9] The International Federation of Library Associations and Institutions (IFLA). *Functional requirements for bibliographic records*, volume 19 of *UBCIM publications; new series*. K.G. Saur, München, 1998. Current version available online at http://archive.ifla.org/VII/s13/frbr/frbr_current_toc.htm.