The EuDML Metadata Schema: Version 1.0

Thierry Bouche, Claude Goutorbe, Jean-Paul Jorda, Michael Jost

The EuDML metadata schema, version 1.0

Thierry Bouche[1], Claude Goutorbe[1], Jean-Paul Jorda[2], and Michael Jost[3]

[1] Cellule Mathdoc (UMS 5638), Université Joseph-Fourier
(Grenoble 1), B.P. 74, 38402 Saint-Martin d'Hères, France
thierry.bouche@ujf-grenoble.fr, claude.goutorbe@ujf-grenoble.fr,
[2] EDP Sciences
17, avenue du Hoggar, BP 112, 91940 Les Ulis cedex A, France
jean-paul.jorda@edpsciences.org,
[3] FIZ Karlsruhe, Zentralblatt MATH
Franklinstr. 11, D-10587 Berlin, Germany
jo@zentralblatt-math.org

Abstract. After an extensive study of the metadata policy of each of its content partners, the EuDML project evaluated many different strategies and existing schemas that could store every detail faithfully, and yet reserve room for the enhancements foreseen in the project's work plan. The framework provided by the so-called NLM Journal Archiving and Interchange Tag Suite was selected as best readily available approximation of our needs. Some modifications of it have been endorsed by the project, defining the first version of our interchange schema for heavily math-based content.

Key words: EuDML project, metadata schema, XML, interoperability

1 Introduction

1.1 The EuDML project

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 international 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.

1.2 Why a EuDML schema?

A public well-specified EuDML schema is needed:

1. Because content providers need to know *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.
Some examples: reference citation matching, duplicate detection and records merging or 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 in parallel to the pre-existing one(s).
5. Finally, EuDML must be able to *export* its content in a predictable, reusable way, for safety backups, interoperability, and to enable content providers to retrieve the EuDML-enhanced metadata for their collection of items, in order to improve their local collections to a higher level of quality. They need to know how this new information will be encoded so that they can use it. This cannot be done with their internal format, as many added-value elements will be beyond the scope of such format.

1.3 A metadata model for EuDML

This paper presents the first specification of the EuDML metadata schema, which is already used by the current prototype of the system.

Its main goals are to:

- provide details on the structure, granularity, and encodings that should be supported by content partners (see § 2);
- provide incentive to more content providers to contribute their best metadata to the EuDML central metadata repository using adequate schemas and interoperability devices (see § 2.5 and § 3.5);
- present 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 (see § 3);
– introduce a set of best practices to ensure perfect understanding of tagging practice among EuDML partners (see § 4);
– outline directions for improvements (see § 5).

2 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.

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 The EuDML item

The central object in EuDML, used as the unit of delivery and thus as the pivot for the metadata schema, is an *item*.

EuDML defined 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".

We warn readers from the library community that this definition is incompatible with that used in FRBR [11] terminology, where it would be rather called *manifestation*.

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 mathematical community. It is the object an unambiguous citation in a scientific article would point to, thus the importance of this concept for a reference library.

Two different editions of a book would be different items, but not a new print; two digitisations of the same article would be related to a single item, but the reprinting of that article in its author's collected works would yield a new item (hopefully related to the previous one). We must be very cautious with mathematical references that, 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 them collectively, 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.

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 explicitly support 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.

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, in house 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. We won't try to record authors' contribution weights, ordered authors' list where either the first or the last name has more significance, etc.
– 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 (full text) providers.

2.4 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 formulae 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.

2.5 Quality insurance on metadata

From the above requirements analysis, we derived the following functional definitions, which help identify more objectively whether a given item's metadata is eligible to support minimal digital library operation, standard full-featured operation, or advanced operation tweaked for EuDML math-specific content.

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.

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.

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 our enhancing tool set 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 by the project.

Supplemental metadata Beyond fundamental metadata, this is additional metadata that should be stored, generated, and exploited 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.

3 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 formulae 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.

3.1 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 [4];

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) [3] ;

MODS Metadata Object Description Schema from the Library of Congress, which is pretty much an interchange format for multimedia library catalogues [7];

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 [2];

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 [10]. 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 [9].

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 work plan. 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) [5]:

> 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. [8]:

> 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 work plan (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).

In summary, here are some decisive features that highlight NLM JATS as the best available framework to host EuDML metadata:

- 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 PubMed Central, JSTOR, or Portico. Its wide deployment and large user community makes it a good reference model for outer interoperability.
- 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 enhancing workflow).

– Last but not least, it is MathML-ready (yet allowing storage of alternative representation of the mathematical content).

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/`)

Although JATS can be easily customised to fit any special need [1], we will try to adhere to its readily published DTDs 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.

3.2 The EuDML schema, initial version, based on JATS

To assess the suitability of JATS to our needs more objectively, the above 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.

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 published DTDs 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, cf. the supported item types listed in § 2.2.
   We thus decided to organise all our content in three major *containers*: journal article, single book, and multiple volume 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 § 3.3.
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 a common encoding practice and understanding among all EuDML partners and content providers. The initial version is outlined here, § 4, and available on our web site (see `http://www.eudml.eu/eudml-metadata-specification`). Further revisions will be released periodically based on feedback from other activities within the project.

3.3 EuDML metadata specification v. 1.0

The EuDML metadata schema version 1.0 as defined by deliverable D3.2 [6] 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 some 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 a decision on the formal way used to store and transport our items' metadata, pragmatically rooted in the existing JATS DTDs, and in the fact that bibliographical data cannot be structured the same way for these different items. It is not intended to restrict in any way the items' records are exposed to or navigated by end users.

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.

3.4 Conversion summary

While developing this work, 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 practices 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.

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

Collection	**EuDML metadata (Schema)**	**Notes**
Gallica-Math	2 081 (**article**)	converted from internal XML with LaTeX
CEDRAM	1 868 (**article**)	converted from internal XML with MathML
NUMDAM	43 944 (**article**)	converted from internal XML
DML-E	6 401 (**article**)	converted from SQL database
EDPS journals	200 (**article**)	slight variation of native EDP schema to obey best practices
ElibM	25 453 (**article**)	converted from internal XML
BulDML	436 (**article**)	converted from DC XML
DML-CZ	26 476 (**article**), 132 (**book**)	converted from internal XML
GDZ Mathematica	53 396 (**article**), 2 298 (**book**), 296 (**mbook**)	converted from METS XML
RusDML	16 486 (**article**)	converted from METS XML
Port. Mat.	1 347 (**article**)	converted from TEL XML
All	**180 814 records**	

In order to get the most out of the contributed metadata, all converted items meet the obligatory requirements, even when the original metadata didn't meet them. For

this, we had to heuristically split some unstructured fields into tagged bibliographic references, e.g. We also tagged all LaTeX formulae encountered with NLM superstructure with MathML alternative encoding (see § 4.3). We thus have currently 206,775 tagged formulae in our metadata. They were mostly processed from TeX encoding embedded into a text field from the provider's metadata.

3.5 A note on 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 embedded 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. A prototype implementation of this is available to the project partners in a dedicated OAI-PMH server.

4 Best practices recommendation for mapping EuDML abstract metadata to the EuDML schema

A best practices working group for representing EuDML metadata in JATS notation was formed. A set of recommendations has been developed, and has now been tested on all available EuDML items. Complete examples of EuDML XML files obeying these recommendations are available on our web site. We give some examples of the issues tackled below.

The recommendation itself is a work-in-progress, which is available to the project's partners in an internal wiki as a live HTML page they can edit. Its first version has been made accessible in an area of the `www.eudml.eu` website dedicated to developers' resources (`http://www.eudml.eu/eudml-metadata-specification`). Up-to-date documentation is in the process of being made available there for download as well: the specification, the DTDs and possible associated tools.

4.1 Special item types

Proceeding volumes were an interesting case, as they were handled in very different ways by different providers. As many EuDML partners were primarily journal digitisation projects, they had "journalised" proceedings, even when they were published as independent books: modelling a conference series as a journal, each proceedings volume as a journal issue, and each contribution as an article. However, the bibliographic metadata of a conference series publishing its yearly proceedings in a general lecture notes series, e.g., is quite different from that of a journal special issue. Our modifications of JATS standard DTDs is in this area are meant to have exactly the same details for each flavour of conference publishing. As soon as the volume holds conference proceedings, the conference details are placed in the metadata, either in <article-meta> for journal articles, or in <book-meta>. The <conference> element is the same in each case.

To set information about editors (e.g for a proceedings, when editors replace authors), <contrib> element is used together with @contrib-type attribute set to "editor".

Books, like articles, can have multiple translations of their title in JATS. For some reason, journal issue metadata is somewhat less detailed than for books. In EuDML you can have multiple <issue-title> elements, distinguished by their @xml:lang attributes. In the case of an article published in a special issue, its authors (or editors) should be distinguished from the issue editors: the <contrib> element is used together with @contrib-type attribute set to "issue-editor".

4.2 Identifiers and relations

The item-centric vision of the EuDML imposes to be very careful on identifiers and relations. Each EuDML item *must* have a unique persistent identifier. As it comes from elsewhere, it will also typically come with a number of different identifiers that we should register in order to enable further interoperability.

The item metadata record is something like a linking hub for the given item: it should hold links to various resources attached to the described item, as well as to other related items. Obviously, a relation needs some sort of identifier for the related object as well.

We identified three main classes of relations, for which we recommend to use three different JATS structures:

1. *Primary identifiers* identify an item or a container. They are assigned by the publisher (DOI, PII and specific internal identifier) or by the local DML. They are not necessarily associated with an URL.
   Such identifiers are stored in a dedicated element relative to the item's type (e.g. <article-id>). They must have a @pub-id-type attribute.
2. *Document identifiers* provide links to the different versions of the *content* pertaining to an item or a container on the provider's web site (the PDF version, the full HTML version, etc.).
   These identifiers come in the form of an activable link stored as @xlink:href attribute to the <self-uri> element. The combination of the mime-type and a controlled vocabulary for values allows to predict the nature of the resource the link points to.

3. *External identifiers* are primarily identifiers proceeding from other authorities, such as Zentralblatt MATH, Math. Reviews, CrossRef, which assign IDs to articles, authors, journals or books, or related resources.
   External identifiers must be set using <ext-link>, the value being the identifier itself. An activable link should be stored as @xlink:href attribute when applicable while the @ext-link-type attribute should keep track of the identifying mechanism and authority. Links to related items (including other EuDML items) should use <ext-link> in a similar way.

4.3 Mathematics

Although the MSC reads "Mathematical *subject* classification", and although JATS provides a <subject> element, MSC should be encoded with the <kwd> element inside a <kwd-group> with attribute @kwd-group-type set to the actual scheme: "msc" followed by the year (e.g. "msc2000").

Inline and display mathematical formulae are expressed respectively with <inline-formula> and <disp-formula> elements. Both MathML and (La)TEX version of the same formula can be wrapped up using <alternatives> element. This mechanism will be extended so that other versions (accessible, aural) can be stored in a similar fashion.

It is recommended to attach a unique ID to each formula to ease further processing.

For TEX notation, it is recommended to put compilable code into the <tex-math> element. This means that the switch to math mode should be part of the element's value, and control sequences should be standard (standard meaning: macros defined in plain and LATEX formats, possibly with AMS mathematical extensions). It is important because some environments such as `multline` change the internal grammar of their content while switching to math, and this would be lost. We currently recommend that the content of the <tex-math> element be literal TEX code with two characters (& and <) escaped using standard XML entities (`&` and `<`). Putting a full LATEX source in a CDATA section (as exemplified in JATS documentation) is explicitly disapproved.

For instance

A product of four (p,q)-sections (with $p < q$).

should be encoded the following way.

```
A product of four
 <inline-formula id="d1e4">
   <alternatives>
      <mml:math xmlns:mml="http://www.w3.org/1998/Math/MathML">
         <mml:mrow>
            <mml:mo>(</mml:mo>
            <mml:mi>p</mml:mi>
            <mml:mo>,</mml:mo>
            <mml:mi>q</mml:mi>
            <mml:mo>)</mml:mo>
         </mml:mrow>
      </mml:math>
```

```
      <tex-math>$(p,q)$</tex-math>
   </alternatives>
 </inline-formula>-sections (with
 <inline-formula id="d1e20">
   <alternatives>
      <mml:math xmlns:mml="http://www.w3.org/1998/Math/MathML">
         <mml:mrow>
            <mml:mi>p</mml:mi>
            <mml:mo>&lt;</mml:mo>
            <mml:mi>q</mml:mi>
         </mml:mrow>
      </mml:math>
      <tex-math>$p&lt;q$</tex-math>
   </alternatives>
 </inline-formula>).
```

5 Conclusion & further work

We have exposed the rationale that leaded us to base the EuDML schema for descriptive metadata on NLM Journal Archiving and Interchange Tag Suite. We provided the current specification of the schema as a diff to three existing standard JATS DTDs. We also gave some examples of the recommendations we came up with so that design choices allowed by JATS are correctly understood by all partners. This work is assessed by the fact that we could convert all EuDML partner's metadata available to us into this framework.

Now, EuDML is starting to exploit what we have generated so far. A number of tools are developed in order to improve the quantity as well as the quality of the metadata available to the project, they will evolve into various automated workflows. Using these tools, we should be able to get new metadata elements, that may superseed or just add to the items' descriptions. Duplicates, similar, and related items should also be detected. We will thus face the necessity to merge item records from a number of sources, some of them "trusted" (e.g. manual keywords or copy-edited translations), some of them much less so (computed similarity, guessed MSC, automatic translation, OCRed math formulae…). We feel that the current schema is robust enough to store all this information faithfully side by side, while retaining its origin. Indeed, we think that managing this will boil down to adding a number of rules and attribute values to our Best practices. However, we are now expecting feedback from a number of project's activities to assess and refine the work reported here.

Acknowledgement This work is partly financed by the European Union through its Competitiveness and Innovation Programme (Information and Communications Technologies Policy Support Programme, "Open access to scientific information", Grant Agreement no. 250.503).

References

1. Jeff Beck, editor. *Proceedings of the Journal Article Tag Suite Conference 2010*, Bethesda (MD), 2010. National Center for Biotechnology Information (USA). Available online at `http://www.ncbi.nlm.nih.gov/books/NBK47086/`.
2. 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.
3. DCMI Eprints Working Group. Scholarly works Dublin Core application profile. Available online at `http://dublincore.org/scholarwiki/SWAPDSP`, 2006.
4. EULER project. The EULER application profile. Available online at `http://www.emis.de/projects/EULER/metadata.html`, 2002.
5. INERA Inc. NLM DTD Resources: Introduction. Web page available online at `http://www.inera.com/nlmresources.shtml`.
6. Michael Jost, Thierry Bouche, Claude Goutorbe, Jean-Paul Jorda, et al. Delivrable D3.2: The EuDML metadata schema, initial version. Technical report, The EuDML project, 2010.
7. Library of Congress. MODS schema. Documentation available online at `http://www.loc.gov/standards/mods/mods-outline.html`, v. 3.4: 2010.
8. Mulberry Technologies Inc. JATS - The Journal Archiving and Interchange Tag Suite. Web page available online at `http://www.mulberrytech.com/JATS/index.html`.
9. 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.
10. 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`.
11. 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`.