

INFORMATION SOCIETY TECHNOLOGIES
(IST)
PROGRAMME

Project IST-2001-37057 MKM-NET

Report D5.2
Metadata for Mathematical Knowledge Management

A. Asperti (University of Bologna)
G.Gogvadze (DFKI — Saarbrücken)

Project Acronym: MKM-NET
Proposal/Contract no.: IST-2001-37057 MKM-NET

Contents

1 Introduction	3
1.1 Data and Metadata	3
2 Markup for Metadata	3
2.1 General Metadata Principles	3
2.2 Resource Description Framework RDF	4
2.3 DAML+OIL	4
3 Existing Metadata Standards	5
3.1 Dublin Core Metadata Element Set	5
3.1.1 IEEE Learning Object Metadata and Extensions	6
3.1.2 IMS content packaging	6
4 Mathematical Metadata	7
4.1 Mathematical Classification Systems	7
4.1.1 Mathematical Relations and Dependencies	7
4.1.2 Searching for Mathematical Expressions	8
5 Summary	9

1 Introduction

1.1 Data and Metadata

The word Metadata should be literally understood as “data about data”, that is data describing “the content, quality, condition, and other characteristics of data”¹. This definition is widely accepted; nevertheless, it is often misleading, and a frequent source of misunderstanding. As a matter of fact, the notion of data, that is the informative content conveyed by a piece of information, is largely dependent by our particular point of view, and the intended processing of the information. As we change the interpretation of data, the relevant notion of metadata is typically modified accordingly. For this reason, the list of functionalities embodied in metadata is normally expanded to include (at least as the term is utilized in the Web context) support for “...any data used to aid the description and location of networked electronic resources”², and also the management of information resources (including rights management) and their long-term preservation³.

Of course, this is not much more informative, and the list of functionalities supported by “metadata” in the future is probably going to grow, and in ways that are hard to predict at present. Perhaps it is for this reason, that the World Wide Web Consortium defines metadata as “machine understandable information for the web”⁴. Unfortunately, this definition, inspired by the so-called “Semantic Web”, is even more confusing: was not providing a machine understandable, but platform and language independent representation of the information, the very aim of XML? In consequence, it is not uncommon to encounter the claim that *any* markup – being an annotation over a data – is a form of metadata. While this may be a reasoned position, it somewhat distant from the original meaning of the word.

The meaning of the word “metadata” is thus so ill-defined and so dependent on the particular context and perspective one is interested in, that any attempt of developing a general model, even in restricted domain such as Mathematical Knowledge, is an impossible (and probably even useless) task. It seems by now accepted that “no single type of metadata can suit every application, every type of resource, and every community of users”⁵.

This document aims to provide a short survey on the topic, listing the main existing metadata systems, research approaches and state-of-the-art technologies in this field.

2 Markup for Metadata

Independent of the precise definition of the word, metadata are still data, and must be expressed according to some given format; moreover, since they are supposed to help in locating, managing and processing other information in a completely automatic way, the requirements that the metadata specification should satisfy are quite demanding.

In this section, we first briefly discuss these requirements and then look at the main “standard” markup languages for expressing metadata that are currently available.

2.1 General Metadata Principles

According to [6], metadata should satisfy the following principles: modularity, extensibility, refinement, and multilingualism (internationalization).

Modularity permits a clear distinction between metadata coming from different sources and provides a way to use existing standards instead of redefining them in the concrete application. The use of name spaces plays an essential role here.

¹<http://www.fgdc.gov/metadata/metadata.html>

²<http://www.fgdc.gov/metadata/metadata.html>

³<http://www.ukoln.ac.uk/metadata/>

⁴<http://www.w3.org/Metadata/>

⁵<http://www.schemas-forum.org/project-info/>

The need for extensibility and refinement is clear: every particular application should be able to extend the metadata set according to its needs as well as refine the existing structures (e.g. sorting with respect to some additional properties or defining particular schemas for value sets).

Some metadata elements in multi-cultural environments can be interpreted differently depending on the country. For example, the date formats differ in North America and Europe, so that in order to store the date information correctly one might consider defining separate fields for day, month, and year rather than representing the full date as a string, where the order of tokens plays a role. Another example is the differences in education systems that makes it difficult to annotate some pedagogical levels of learning material.

2.2 Resource Description Framework RDF

The W3C Resource Description Framework (RDF) [4, 15, 14] provides a general, domain-neutral model for representing metadata as well as a syntax for encoding and exchanging these metadata over the Web. It supports interoperability of independently developed Web-servers and clients, and more generally, between applications that exchange machine-understandable information on the Web. Documents described by RDF-metadata can potentially be indexed by search engines in a more effective way.

The basic construct in RDF is a URI. RDF introduces the notion of a resource that is anything that has a URI (Uniform Resource Identifier) and way of describing a resource using so-called statements that are triples of the form (*resource, property, value*) where the elements are also referred to as the 'subject', 'predicate' and 'object' of a statement. Schematically, one could imagine an RDF document as a labelled directed graph, consisting of nodes and arcs. The nodes of the graph correspond to resources or the values of the properties assigned to this resources and the arcs represent properties themselves. By definition, a node can be represented by a URI, as a blank node or a string (so-called "literal") and an arc is always labelled by URIs.

There is a straightforward translation of RDF statements into XML that constitutes the intermediate format for exchange between RDF applications. In this format, the nodes and arcs of the RDF-graph are turned into XML-elements, attributes, element content and attribute values, while the URI labels for properties and object nodes are written in XML using name spaces.

The basic syntax of RDF is defined by the RDF-schema. Using this schema one can create documents or define one's own RDF languages by specifying a new schema using the syntax of RDF.

Like XML-schema, RDF-schema have a mechanism to describe constraints on the elements. The main difference from XML is that RDF focuses on the communication of its *classes* and *properties* using nested elements to simulate the ordered graph of an RDF statement, while the difference of nature between elements and attributes is ignored as they are freely converted one to another, whereas for XML, maintaining this distinction is fundamental. This means that an RDF class is similar to a class in object-oriented programming in that it can have super-classes, subclasses, instances and properties. The RDF-schema provide a mechanism for specifying constraints on the use of properties and classes in RDF-documents. These constraints are specified by declaring domains and ranges for properties that are the instances of particular classes.

Compared to the XML-schema, an RDF-schema is weaker in respect of the constraints it may express on the XML-structure of a document. Furthermore, the semantics of the data types defined as abstract classes in RDF depend on the application. RDF does not specify whether or how an application must process the constraint information, so that different applications might use these constraints in different ways.

2.3 DAML+OIL

DAML+OIL is a is an RDF-based semantic markup language for encoding Web-resources. It extends the basic RDF-schema with modelling primitives from Description Logic. DAML+OIL provides a semantic interpretation for the parts of an RDF graph that instantiate the DAML+OIL schema.

DAML+OIL is a development of the DARPA Agent Markup Language (DAML)⁶, which aimed to express more sophisticated RDF class definitions than permitted by RDFS (RDF schema), and Ontology Infer-

⁶see <http://www.daml.org>

ence Layer (OIL)⁷, an alternative approach building on constructs from frame-based AI.

The additional classes and properties defined in DAML+OIL schema can express far more sophisticated classifications and properties of resources than RDFS and, therefore, provide a powerful representation format for ontologies of machine processable knowledge. For example, one can express boolean combinations of classes or specify that two classes are disjoint. In particular, the `Property` class of RDF has some important refinements. It can be enriched with some qualifiers such as *inverseOf* that provide information about the relation of one property to another or *TransitiveRelation* providing meta-information on the structure of relations. Another important facility that DAML+OIL provides is property restriction, that provides a way to restrict classes to a set of resources satisfying particular properties. The cardinality or the values of these properties can also be specified.

Another key to the expressiveness of DAML+OIL is that apart from the RDF-mechanism for defining types it also allows the use of XML-schema data types simply by including their URIs within the DAML+OIL ontology.

3 Existing Metadata Standards

The rise of the World-Wide Web has created an urgent need to define standard methods and vocabularies for describing its contents in a consistent and orderly manner. Since 1995, a number of related initiatives have arisen in what has been called a Metadata Movement. “Metadata” is a broad term that covers many types of “structured data about data” – from traditional resources such as library catalogues, subject indexes, book reviews and abstracts, to new forms of technical and descriptive data for Web resources ranging from digital signatures and digitised map co-ordinates to on-line mail-order catalogues. In its most familiar form, metadata may be used to list the Author, Title and Subject of resources in a collection. Other types of metadata may list the price or quality rating of those resources, specify the format of their computer files, name the administrators responsible for their preparation, or clarify the terms and conditions under which they may be viewed or copied. Some of these metadata types are meant to be read by humans, while others are designed to be processed directly by computers. No single type of metadata can suit every application, every type of resource, and every community of users. Rather, the broad diversity of potential metadata needs can best be met by a multiplicity of separate but functionally focused metadata packages, or schemas.

3.1 Dublin Core Metadata Element Set

The Dublin Core Metadata Initiative⁸ proposes a minimal metadata element set comprising the basic metadata needed by most Web applications, and in particular, this covers administrative information concerning the document.

According to version 1.1 of the Dublin Core Metadata Element Set (DCMES), there are fifteen metadata elements:

Title, Creator, Subject, Description, Publisher, Contributor,
Date, Type, Format, Identifier, Source, Language, Relation,
Coverage, Rights

Each Dublin Core (DC) element is defined using a set of ten attributes from the ISO/IEC 11179 [10] standard for the description of data elements. Every metadata element has a *Name*, *Identifier*, *Version*, *Language*, *Definition* etc.. This meta-metadata specifies some technical properties of the metadata element itself and suggests the usage.

Note that the DCMES is not sufficient for describing even its own elements, that is, some of the ten attributes used to describe a DC-element are not DC-elements themselves (such as *Maximum Occurrence* indicating a limit on the number of instances of the data element). This, however, does not speak against the

⁷see <http://www.ontoknowledge.org/oil/>

⁸see <http://www.dublincore.org>

aim of the Dublin Core Metadata Initiative, that is to provide a minimal set of elements used by an application in order to administrate the data.

Various applications implement DCMES in their data representation formats with different levels of refinement, attributing some of the elements by sub-properties such as `role` of Contributor with values “`edt`” for editor, “`trl`” for translator etc..

3.1.1 IEEE Learning Object Metadata and Extensions

IEEE Learning Object Metadata (LOM) together with IMS Learning Resource Metadata propose a metadata element set for use in annotating learning objects. According to the LOM specification,⁹ the purpose of this standard is to facilitate search, evaluation, acquisition, and use of learning objects by learners or teachers. Another goal is to facilitate the sharing and exchange of learning objects.

The elements of the Base Scheme of LOM are grouped in nine categories:

General, Lifecycle, Meta-metadata, Technical, Educational,
Rights, Relation, Annotation, Classification

Each of these categories groups data elements and every element has an *explanation* that defines this element:

- *size* indicates the number of values
- *order* of these values
- *value space*
- *data type*
- *example* providing a suitable illustration.

The Dublin Core element set is represented as a subset of LOM, and some refinements of the DC elements are provided. For instance, the element `relation` is refined into separate category, including important properties of relation such as *kind* specifying the nature of relation, *keywords*, or other classifying metadata.

In addition, LOM defines learning situation specific metadata. For example, one can specify the type of learning resource (`exercise`, `table`, `self assessment` etc.), intended role of the user (`learner`, `teacher`, `author` etc.), learning context (`secondary education`, `university first cycle`, etc.), semantic density and technical difficulty of the content.

Finally, LOM also provides metadata for specifying the degree of interactivity. There is the interactivity type that is defined according to criteria such as the balance of the information flow between the learning object and the user, and the interactivity level for measuring the quantity of communication.

Some extensions of IMS such as EML (Educational Modelling Language)¹⁰ are designed to serve not only as static descriptions of properties of learning objects, but also try to annotate the dynamic process of learning and teaching.

3.1.2 IMS content packaging

The IMS (Instructional Management Systems) Global Learning Consortium¹¹ has issued a proposal attempting to standardize the organization and distribution of learning materials in the form of so-called content packages. The objective here is to define standardized structures in order to enable the exchange of content.

According to the IMS Content Packaging information model, a *package* represents a unit of reusable content together with information about the organization responsible for it. A package may be a part of another package, but it must contain all information needed in order to use it stand-alone.

The package consists of a so-called *manifest* that describes the package itself and of the actual content files. The manifest contains metadata about the package, an organization element describing the organization of the content within manifest, a resources element consisting of resources that are records of metadata,

⁹see <http://ltsc.ieee.org/doc/wg12/LOM-WD3.html>

¹⁰<http://eml.ou.nl>

¹¹<http://www.imsglobal.org/>

dependencies and identifiers of the physical resource items (files), and possibly one or more (sub)manifests. Any of the IMS metadata elements can be used to express this metadata.

4 Mathematical Metadata

Dublin Core, LOM and IMS Content Packaging are all examples of “wide range” metadata models. Focusing on a particular domain of knowledge, such as Mathematics, for instance, it may become necessary to extend or refine the model in order to meet new requirements for particular or enhanced functionalities.

As an example, the Euler project¹² has recently defined an extension of the Dublin Core explicitly tailored to mathematical documents, and in particular supporting classification and transmission of such documents across the Web. In this case, however, the refinement has been driven more by expected new functionalities (web management and transmission), than by the actual content of the information (i.e. mathematics). This latter is merely reflected in the use of standard Mathematical Classification Systems for describing the subject of the document.

4.1 Mathematical Classification Systems

Currently, there are three Mathematical Classification Systems in use: the Mathematics Subject Classification¹³, the Dewey Decimal Classification¹⁴, and the Universal Decimal Classification System¹⁵. The latter two are general schemas of which mathematics forms a part.

The Dewey Decimal Classification (DDC) was conceived by Melvil Dewey in 1873 and first published in 1876. The DDC is published by Forest Press, a division of OCLC Online Computer Library Center. The DDC system is one of the most widely used classification systems in the world. It is familiar to librarians but almost unknown amongst mathematicians. At the broadest level, the DDC is divided into ten main classes meant to cover all areas of knowledge. Each main class is further divided into ten divisions and each division into ten sections. Not surprisingly, for a given area of knowledge, such as mathematics, the final classification is not as fine-grained as that of MSC 2000.

The Universal Decimal Classification system (UDC) is a similar “universal” classification schema which was originally inspired by DDC and then independently evolved into an international standard. UDC is currently widely spread in Russia but less used than DDC in western countries.

The common standard in mathematics is MSC 2000 which currently is supervised by Jane Kister (editor-in-chief of Mathematical Reviews) and Bernd Wegner (editor-in-chief of Zentralblatt MATH). The MSC is used to categorize items covered by the two reviewing databases, Mathematical Reviews (MR) and Zentralblatt MATH (Zbl). The MSC is broken down into over 5,000 two-, three-, and five-digit classifications, each corresponding to a discipline of mathematics (for example, 11 = number theory; 11B = sequences and sets; 11B05 = density, gaps, topology). A recognized problem in practice with MSCS is that different reviewers may annotate a document with different classification keywords.

It is to be expected that following the development of this area of research, the organizations maintaining the classifications will gradually gain interest in standardizing symbol sets which will be used by authors and will subsequently be sought in databases.

4.1.1 Mathematical Relations and Dependencies

In learning environments such as ACTIVEMATH¹⁶ [13] additional metadata for educational purposes are particularly important. These metadata may be essentially classified as follows:

¹²<http://www.emis.de/projects/EULER/index.html>. European Libraries and Electronic Resources in Mathematical Sciences, European Project, Telematics for Libraries sector.

¹³MSC 2000, <http://www.ams.org/msc/>

¹⁴DDC <http://www.oclc.org/dewey/>

¹⁵UDC, see for example <http://www.lib.demokritos.gr/udceng.htm>

¹⁶<http://www.activemath.org/>

- description of administrative/legal characteristics of a document or mathematical item;
- information needed to update the user model when the user has worked on the material;
- information needed for choosing automatically the most suitable material in the current learning situation.

While all of them are important, only the third category is strictly related to the mathematical content of the information (for the first category a Dublin-Core like metadata model may suffice, while the representation of feedback on user's actions is still an active and somewhat obscure subject of research).

In order to choose material depending on the learning situation one has to introduce metadata modelling the learning situation and refine the structure of relations of the mathematical items with respect to the representation of learning situation. This work requires first of all a content description of the mathematical document at a sufficient level of granularity (at least at the level of individual statements, examples, etc.), and then a classification of the relevant dependencies between these items, which could be of interest in the learning environment (see also [5]). For instance, given a definition the user may be interested to:

- see an example;
- check other notions the definition is relying on;
- view where the definition is actually used;
- be warned if the definition is a particular case of a more general notion;
- ...

All these complex operations requires a sophisticated model of relations between mathematical items. In respect of this, the MOWGLI project¹⁷ has recently developed and is currently validating a model (see [1, 7]) based on the following *kinds* of mathematical relations:

- `requires` indicates a reference to the required knowledge (converse: `is_required_by`).
- `is_instance_of` means that the current item is an instance of the concept it relates to (converse: `has_instance`)
- `is_generalization_of` means that the current item is a generalization of the item it relates to (converse: `has_generalization`)
- `for` is used when the current item serves the item referred to e.g., `proof for` a theorem, `definition for` a concept (converse: `has`).
- `example_for` indicates the item is an example for some concept (converse: `has_example_for`).
- `counterexample_for` is used if the given item is a counterexample for some concept (converse: `has_counterexample`).
- `lemma_for`, `corollary_for` and `assumption_for` indicate that the current item is a lemma, corollary or assumption respectively for the item referred to, that is, an assertion can be a lemma with respect to another assertion (counter values: `has_lemma`, `has_corollary` and `has_assumption`).
- `citation` is a reference to a bibliographical entry (converse: `is_cited_by`).

These are further extended by a set of relations pertaining to OMDoc[11]:

- `alternative` states that one mathematical item is an alternative to another
- `entailed_by`, `entails` and `equivalent_by` provide references to other mathematical resources proving that either the two given items are mathematically equivalent or one entails (or is entailed by) the other.

4.1.2 Searching for Mathematical Expressions

The requirements for effective search techniques for mathematical notions based on a content-driven encoding of the information may be highly demanding. In particular, typical queries may require complex elaborations that cannot be trivially compiled into standard database query languages. Typical examples are:

1. matching “equivalent” notions, such as say n^{-1} and $1/n$. Supporting this feature requires the implementation of some form of “reduction” eventually comprising the unfolding of definitions.
2. taking into account “isomorphic” shapes. When looking for a formula of the shape $A \wedge B$, we would like to match $B \wedge A$ as well.

¹⁷<http://www.mowgli.cs.unibo.it> European Project IST-2001-33562 MOWGLI

3. supporting unification, as opposed to “pattern matching”.

An essential prerequisite for this kind of elaboration is the low-level markup of the *content* of mathematical expressions, such as that provided by, *e.g.* MATHML-content. Thus, mathematical expressions (and sub-expressions) becomes structured data, possibly deserving of their own metadata.

The exploitation of metadata for mathematical expressions has been pursued for the first time inside the HELM Project¹⁸ [2] with promising results. In this case, metadata provides an approximate description of the mathematical formula, comprising a list of the identifiers occurring in it, with some additional information about the position of these occurrences with respect to the tree-like structure of the formula [9]. These metadata allow a fast trimming of the search space, which can then possibly be refined by other techniques. In other words, because of the complexity of queries and the dimension of the data base it may be interesting to divide the search into two phases, where we first select a restricted number of “candidate” documents, and then secondly interrogate this subset in a more precise way, to get the final answer [8]. The first “selection” step, could be based on metadata automatically generated from the source document in the spirit of the HELM approach. This approach is really modular, since by choosing metadata in different ways we may enhance different searching functionalities.

5 Summary

We have provided a brief motivation and background to the origins and use of metadata, up to and including DAML+OIL. We have summarized the relevant aspects of some of the key existing metadata standards – Dublin Core for generic information and two education oriented metadata schema. In conclusion we explore three existing taxonomies for mathematics, two being general library classifications and the third developed by the American Mathematical Society: the Mathematics Subject Classification (2000). Considering the problem of querying mathematical content, it is observed that conventional regular or structured (data base) query languages are likely to be inadequate and that (mathematical) domain-specific knowledge and even significant computational effort, using mathematical knowledge, may be necessary for even apparently quite straightforward queries.

References

- [1] Asperti, A., Goguadze, G., Melis, E., “Structure and Meta-Structure of Mathematical Documents”, Deliverable n.D1b of Project IST-2001-33562 MOWGLI, <http://www.mowgli.cs.unibo.it/> (last accessed January 2004).
- [2] Asperti, A., Padovani, L., Sacerdoti Coen, C., Schena, I., “HELM and the semantic Math-Web”. Proceedings of the 14th International Conference on Theorem Proving in Higher Order Logics (TPHOLS 2001), 3-6 September 2001, Edinburgh, Scotland. Published in Springer Lecture Notes in Computer Science (LNCS) volume 2152, pp59–76.
- [3] Berners-Lee, T., “Universal Resource Identifiers in WWW”, RFC 1630, CERN, June 1994. Available from <http://www.w3.org/Addressing/rfc1630.txt> (last accessed January 2004).
- [4] Bray, T., “What is RDF”, published by O’Reilly [xml.com](http://www.xml.com/pub/a/2001/01/24/rdf.html), January, 2001. <http://www.xml.com/pub/a/2001/01/24/rdf.html> (last accessed January 2004).
- [5] Dahn, I., Schwabe, G., “Personalizing Textbooks with Slicing Technologies – Concept, Tools, Architecture, Collaborative Use”, Proceedings of the 35th Hawaii International Conference on System Sciences – 2002. In the addendum to the proceedings, p.3ba. ISBN 0-7695-1435-9
- [6] Duval, E., Hodgins, W., Sutton, S., Weibel, S.L., “Metadata Principles and Practicalities”, D-Lib Magazine, April 2002, Volume 8, Number 4, ISSN 1082-9873 (electronic publication). Available from

¹⁸<http://www.helm.cs.unibo.it>

- <http://www.dlib.org/dlib/april02/weibel/04weibel.html> (last accessed January 2004).
- [7] Gogvadze, G., “Metadata for Mathematical Libraries”, Deliverable n.D3a of Project IST-2001-33562 MoWGLI, <http://www.mowgli.cs.unibo.it/> (last accessed January 2004).
- [8] Guidi, F., Sacerdoti Coen, C., “Querying Distributed Digital Libraries of Mathematics”, in Proceedings of Calculemus 2003, Aracne Editrice S.R.L., Thérèse Hardin and Renaud Rioboo editors, ISBN 88-7999-545-6, pp. 43–57.
- [9] Guidi, F., Schena, I., “A Query Language for a Metadata Framework about Mathematical Resources”, in Proceedings of the second International Conference on Mathematical Knowledge Management, Bertinoro, Italy, February 2003. LNCS 2594, pp. 105–118.
- [10] International Standards Organization. ISO/IEC 11179:1–6 Information technology – Specification and standardization of data elements – Parts 1–6. Available from www.iso.ch (last accessed January 2004).
- [11] Kohlhase, M., “OMDoc: Towards an OPENMATH Representation of Mathematical Documents”, Seki Report, FR Informatik, Universität des Saarlandes, 2000.
- [12] Mathematical Markup Language (MathML) 2.0 W3C Recommendation, 21 February 2001. <http://www.w3.org/TR/MathML2/> (last accessed January 2004).
- [13] Melis, E., Büdenbender, J., Andres, E., Frischauf, A., Gogvadze, G., Libbrecht, P., Pollet, M. and Ullrich, C., “ActiveMath: A Generic and Adaptive Web-Based Learning Environment”, Artificial Intelligence and Education, Volume 12, Number 4, 2001, pp385–407.
- [14] Resource Description Framework (RDF) Schema Specification 1.0, W3C Candidate Recommendation 27 March 2000. <http://www.w3.org/TR/rdf-schema/> (last accessed January 2004).
- [15] Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation 22 February 1999. <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/> (last accessed January 2004).