Journal of Universal Computer Science, vol. 7, no. 7 (2001), 591-601
submitted: 1/5/01, accepted: 1/6/01, appeared: 28/7/01

Metadata Standards: What, Who & Why

Erik Duval
(Departement Computerwetenschappen
Katholieke Universiteit Leuven, Belgium
Erik.Duval@cs.kuleuven.ac.be)

Abstract: In order to be able to (re-)use digital content, interested users must be able to identify and locate relevant documents. This requires descriptive data, nowadays generally referred to as *metadata*. Technical *standards* for a scaleable deployment on a global scale are required if we want to achieve a critical mass of resources. In this paper, we present the current status of ongoing work in this area, with a particular emphasis on the IEEE LTSC Learning Object Metadata standard [IEEE, 2001] and related developments in the context of the ISSS Learning Technologies Workshop [ISSS, 2001].

Keywords: metadata, learning technology standardization
Category: H.3 - Information Storage and Retrieval

1 What are Metadata: Introduction and Background

Metadata is often defined as:

> *'data about data'*.

A somewhat more informative *definition* is [IEEE, 2001]:

> *'information about an object, be it physical or digital'*.

Thus, metadata are basically descriptive data. As such, metadata are at the heart of more general developments in the area of digital libraries [Fox et al., 2001]. Basic metadata elements indicate the title, author, year of publication and similar simple bibliographic data. Richer metadata structures also cover technical features, copyright properties, annotations and so on.

The *purpose* of metadata is 'to facilitate search, evaluation, acquisition, and use' of resources [IEEE, 2001]. Moreover, in the case of educational resources, the purpose is also 'to facilitate the sharing and exchange of learning objects, by enabling the development of catalogs and inventories while taking into account the diversity of cultural and lingual contexts in which the learning objects and their metadata will be exploited' [IEEE, 2001].

In the documents on metadata and the 'semantic web' from the web consortium, metadata is often used for descriptive data that can be processed by machines [Berners-Lee et al., 2001]. This is a more restricted interpretation than the one we adopt here. By explicitly including descriptive data that need to be interpreted by humans, we want to recognize the importance and relevance of such metadata.

We will explicitly *not* focus on metadata for geo-spatial applications, library applications (Z39.50, MARC variants) and continuous media specific standards (such as MPEG-7), though most of what we will present here is applicable to those more specific contexts as well. This paper will also not deal with the *Dublin Core* specification, that defines 15 elements for cross-domain search. The Dublin Core specification has recently been submitted for approval as an American national standard, referred to as Z39.85. There is a Memorandum of Understanding between the IEEE LTSC LOM group (see below) and the Dublin Core initiative, with the intent to investigate common mechanisms for interoperability between the two metadata schemes, potentially based on an interoperable approach that builds on the RDF framework of the World Wide Web Consortium. More details on how Dublin Core relates to the IEEE LTSC LOM specification can be found in [Duval, 2001].

2 Why Standards: Interoperability

2.1 Introduction

Generally speaking, technical standards are important because they make it possible to develop interoperable tools and services [Paepcke et al., 1998]. In this context, my favourite definition of *interoperability* is [Rust & Biede, 2000]:

> *'enabling information that originates in one context to be used in another in ways that are as highly automated as possible'*

. There are a number of noteworthy aspects in this definition:

- The central notion is that of crossing boundaries of *context*: this may involve straightforward technical boundaries (like when metadata are served from a server by a particular vendor to a client from another origin), but also more subtle boundaries, such as linguistic ones (like when metadata are to be translated), social ones (like when metadata intended for teachers need to be transformed into metadata for learners), or, more generally, cultural ones (like when metadata refer to national or regional educational contexts, such as 'bac+2' in France). It is clear that the technical boundaries are the easier ones to cross.

- The definition above mentions 'as highly automated *as possible*'. Obviously, it is to be preferred that the process of crossing context boundaries is fully automated, as in the case when documents are translated from one format (like LaTeX or Microsoft Word) into another one (like HTML or Adobe Portable Document Format). However, the definition makes it clear that this process is not always fully automatic, as for instance in the case when examples in a document need to be replaced by examples from another application domain. (For instance, a document on the concept of 'calibration', originally developed for the automobile manufacturing industry may be reused in the context of medical measuring equipment.)

Thus, the notion of interoperability is not a binary one, where systems would be either interoperable or not. Rather, there is a higher or lower degree of interoperability. This is well illustrated by the standards on paper size: in Europe, the 'DIN A4' standard, from German origins, prevails, whereas, in the United States, the 'U.S. letter' standard is more widespread. These standards have become so widely accepted that they are almost 'invisible': we assume that papers fit in binders, that binders fit in closets, that paper trays of printers have the correct size, etc. Moreover, we all take it for granted that we can buy the hardware involved from different companies and make the different components work together without any transformation. Nevertheless, many of us have the experience of printing that goes less than perfect when we download and print documents that have been formatted for the 'other' standard size.

The above illustrates the main advantage of interoperability: it prevents end users from being locked into proprietary systems. The World-Wide Web is a perfect example of how standards (in this case: URL, HTTP and HTML [Berners-Lee, & Fischetti, 1999] can be the basis of open, interoperable systems, that allow end users a choice of client and server systems alike. The Web also illustrates that interoperability is not always absolute: because of the diverging additions to the official W3C standards that Netscape and Microsoft support, some features may only be available on one platform, or the developers may be required to develop those features in non-standard ways, separately for each supported platform.

2.2 Layers of interoperability

It is useful to distinguish different levels of interoperability [see table 1]. At the most machine oriented level, there is network protocol interoperability, where the relevant standards include TCP/IP and HTTP. The HTTP standard for instance enables a Web browser and server to exchange messages, even when these software components were developed by different vendors, operating under different operating systems, on different kinds of hardware, etc.

Secondly, there is the level where data gets bound to a particular representation format or data binding. A typical web example is the representation of a document in HTML. For metadata, the most popular bindings nowadays are XML, or, more specifically, RDF [W3C, 2001].

1	Protocol	TCP/IP, HTTP
2	Data binding	HTML, XML, RDF
3	Metadata scheme	LOM, Dublin Core
4	Semantic	Ontologies, classifications, vocabularies, taxonomies

Table 1: Layers of interoperability

The level that we will focus on in the remainder of this paper is that of the conceptual data model or metadata scheme, that specifies the data elements of which a metadata instance is composed. Metadata instances based on a common metadata schema have a high degree of 'semantic interoperability' [Forte et al., 1999]. The

binding of metadata schemes in level 2 representations is typically defined in a binding specific way. As an example, an XML DTD has been developed for the IEEE LTSC LOM specification, in order to define an XML binding of LOM. Similarly, alternative mechanisms can be used to bind to the same representation format (for instance: XML Schema) or alternative representations (for instance: RDF or SQL schemas). In [Section 3], we deal with standards for layer 3 interoperability.

Finally, ontologies, classifications, vocabularies and taxonomies attempt to define common semantics. In most cases, these conceptual structures are restricted to a particular domain. Typically, they define the relevant concepts in that domain, and their interrelationships. The intent is to enable consistent interpretation of statements that make use of these concepts. An example of this layer of interoperability is the reference 'Category**:** H.3 - Information Storage and Retrieval' in the header of this paper: it refers to the ACM Computing Reviews classification widely adopted in the domain of computer science. Common classification structures are the basis of consistent descriptions that support systematic access for indexers and end users. More sophisticated such approaches are based on knowledge engineering technologies [Hill et al., 2000].

3 Who and What: an Overview of Metadata Standards

3.1 Who: an Overview

Three 'official' *accredited* standardization organisations are active in the field of educational technologies in general, which includes the more specific field of learning object metadata. These organizations are:

1. The *IEEE Learning Technologies Standardization Committee* (LTSC) was set up in 1996. Its purpose is to standardize the 'smallest, useful, doable specification that has technically feasibility, commercial viability, and widespread adoption'. Besides working groups on for instance 'Computer Managed Instruction', 'Simple Identifiers' and others, there is a group that focuses on 'Learning Object Metadata' (LOM) [IEEE, 2001].
2. The Centre Européen de Normalisation organizes a *workshop on Learning Technologies* since 1999, under the umbrella of the so-called 'Information Society Standardization System' (ISSS). The main purpose of this workshop is to 'promote the development and adoption of appropriate standards, taking into account the diversity of cultural backgrounds and languages that exists within Europe'. After an initial requirements analysis [ISSS, 2000], work has now started on LOM related work (see below), copyright, quality issues, educational modelling languages, etc. [ISSS, 2001].
3. ISO and IEC have set up a Joint Technical Committee (JTC1) that, since 1999, has a *subcommittee on Learning Technologies*. At this moment, this more formal body is initialising its operations. In the domain of metadata, it has invited the IEEE LTSC to submit its LOM standard as soon as LTSC deems appropriate.

Besides the formal standardization bodies, there are numerous *consortia* that carry out technical work in the field of educational technologies. Once this work leads to mature specifications, those specifications can be submitted to the accredited standardization organizations. Conversely, consortia often represent communities of practice that adopt standards as they are developed by accredited organizations.

The consortia with a more direct standardization impact on metadata include:

1. The *ARIADNE* Foundation regroups academic and industrial members [ARIADNE, 2001]. At the core of its infrastructure is the so-called Knowledge Pool System, a distributed repository of pedagogical documents and their associated metadata [Duval et al., 2001]. An integrated Web-Based Learning Environment supports the development of courses that reuse resources from the Knowledge Pool System.
2. The *IMS* consortium regroups vendors of Learning Management Systems, authoring tools, and related products [IMS, 2001]. IMS does not develop implementations, but focuses on the development of specifications that can then be submitted to the standardization bodies mentioned above. Work is currently ongoing in the area of content packaging, question and test interoperability, etc.
3. *ADL* was originally a U.S. Army initiative for interoperability developments in the area of learning technologies, but it has substantially increased its scope and relevance. One of its major milestones is the Sharable Content Object Reference Model [ADL, 2001]. The major aim of the SCORM model is to define an overall specification for interoperability between components of a digital learning infrastructure, based on the IEEE LTSC LOM and CMI standards.

The IEEE LTSC LOM standard is based on early work by the ARIADNE and IMS consortia, which led to a joint submission of a base document. Since then, ARIADNE, IMS and ADL have all contributed to further development of the LOM specification within the IEEE working group. At the time of writing, the LOM standard is in ballot.

ARIADNE, IMS and ADL are now developing their own so-called ‘application profiles’ of the LOM standard: these are specifications that adapt the standard to the specific needs of their communities. In practice, this can for instance involve a mandatory status for some data elements, or more restricted vocabularies than those contained in the IEEE specification (see below).

3.2 What: Learning Object Metadata

3.2.1 Learning Objects

In the context of the IEEE LTSC LOM, the term "Learning Object" should be understood in its most general sense. The definition in the standard is:

> *“a learning object is defined as any entity, digital or non digital, that may be used for learning, education or training."*

Thus, learning objects can be of any size, type, etc. In principle, they need not be digital, and can include people, rooms, equipment, etc. This concept of a generalized learning object is in contrast to that of an object as a discrete item or piece of content, often within a hierarchical content model that progresses from the level of raw media, up through content objects, learning objects and then lessons, courses, curricula, etc.

Moreover, a learning object need not be restricted to a static object or piece of content: it can also be a momentary collection or assembly of content, for instance adapted to the specific needs of a particular learner in a given situation and time.

3.2.2 Base Scheme

The IEEE LTSC LOM standard defines a so-called *base scheme*. This is basically a collection of data elements that can be used to describe a learning object. The LOM scheme regroups data elements in nine categories:

1. The *General* category groups information that describes the learning object as a whole. This category includes elements like identifier, title, language, keywords, etc.
2. The *Lifecycle* category groups the features related to the history and current state of the learning object. It also describes the individuals or organizations that have affected the learning object during its evolution. Data elements in this category include the version, status, and contributors (authors, publishers, etc.).
3. The *Meta-metadata* category groups information about the metadata, rather than about the learning object that they describe. This includes an identifier for the metadata instance, contributors to the metadata, the language used in the metadata, etc.
4. The *Technical* category groups the technical requirements and characteristics of the learning object. This category describes for instance the MIME type of the learning object, its size, location, required soft- and hardware, etc.
5. The *Educational* category groups the educational and pedagogic characteristics of the learning object. It indicates the interactivity type (active, expositive, etc.), learning resource type (exercise, simulation, questionnaire, etc.), interactivity level, semantic density, educational context (primary education, higher education, vocational training, etc.), typical age range, etc.
6. The *Rights* category groups the intellectual property rights and conditions of use for the learning object. For this category, LOM adopted a fairly simple approach, indicating whether or not any cost is involved, and whether copyright and other restrictions apply. The idea is to refer to other standards for more complex modelling of rights management metadata [INDECS, 2001].
7. The *Relation* category regroups features that define the relationship between this learning object and other ones, with an indication of the type of the relationship (‘based on’, ‘part of’, etc.).
8. The *Annotation* category provides comments on the use of the learning object and information on when and by whom the comments were created.
9. The *Classification* category describes where the learning object can be classified within a particular classification system. As any classification can be referenced, this category provides for a simple extension mechanism.

3.2.3 Data Elements

For each data element, the base scheme defines:

- *name:* the name by which the data element is referenced;
- *explanation:* the definition of the data element;
- *size:* the number of values allowed;
- *order:* whether the order of the values is significant (only applicable for data elements with multiple values);
- *value space:* the set of allowed values for the data element - typically in the form of a vocabulary (see below) or a reference to another standard (such as vCard, ISO8601 for the representation of dates, etc.);
- *data type:* a set of distinct values;
- *example*: an illustrative example.

Some data elements contain *sub-elements*. Data elements with sub-elements do not have values directly, but indirectly, through their sub-elements. As an example, the element that indicates the learning object that the described object is related to (Relation.Resource) has a value indirectly only, through one or more of its subelements (Relation.Resource.Identifier, Relation.Resource.Description or Relation.Resource.CatalogEntry).

Vocabularies are recommended lists of appropriate values, that define the value space of a data element. Other values, not present in the list, may be used as well. However, metadata that rely on the recommended values will have the highest degree of semantic interoperability, i.e. the likelihood that such metadata will be understood by other end users is highest. As an illustration, the data element Educational.LearningResourceType may have a value from the LOM vocabulary, such as for instance "Questionnaire". This option is preferred if the values in the vocabulary can adequately express the intended meaning. If the indexer wants to assign a value that is not part of the list given in the LOM document for that data element, then the indexer may designate the value as, for instance, ("http://www.vocabularies.org/ LearningResourceType", "MotivatingExample"). This option provides more flexibility to the indexer of learning objects, at the expense of semantic interoperability. User defined values will not be used consistently throughout the larger community. In the example above, a URI was used to indicate the source of the vocabulary. This approach is certainly good practice, but using a URI is not a requirement.

For each of the data elements, the specification includes the *data type* from which it derives its values, such as Date, Character string, etc. Of particular interest is the notion of ‘LangString’, used to represent a phrase in a human language. A value of this type can consist of multiple (Language, String) tuples where Language indicates the human language (according to the ISO639 standard) and String holds the actual character string (according to ISI/IEC10646-1). An example of this concept, as represented in an XML binding could be:

```
<title>
  <langstring>
    <string xml:lang="en">Draft Standard for Learning
      Object Metadata</string>
    <string xml:lang="nl">Voorstel van Standaard voor
      Metadata van Leerobjecten</string>
  </langstring>
</title>
```

In this case, two titles are defined for the learning object: one in English, and one in Dutch.

A LOM metadata instance may contain *extension* data elements. Such elements cannot replace data elements in the LOM structure.

3.3 European efforts on LOM

The IEEE LTSC Learning Object Metadata schema is explicitly recognized by the ISSS Learning Technologies Workshop as the commonly accepted global standards solution for describing learning objects through metadata. The Learning Technologies Workshop is complementing this global activity with a number of projects that address Europe's specific requirements [ISSS 2001]:

- A first project will ensure that the IEEE LTSC LOM, as the globally accepted solution, is capable of addressing specific European cultural requirements (such as multilinguality). The outcome of this project may be a proposed addendum for LOM.
- A second project is investigating standardization actions to permit the identification of alternative versions of resources, in different languages, as well as the origin of the translation, all within a LOM context. The outcome may be an application profile of LOM to deal with this specific requirement.
- A third project will ensure that LOM is localized and translated in the languages of the EU and EFTA countries. Translations of earlier versions of LOM already are available. These will be replaced in due time by updated and widely accepted revised versions.
- On the semantic level of interoperability, the workshop will collect and organize a registry of taxonomies and repositories relevant to a European learning society, via an on-line repository. The registry will indicate the applicability of taxonomies and vocabularies, their interrelationships, as well as mappings and translations between different structures. This will benefit interoperability between European learning technology systems and services as metadata implementations will be able to rely on standardized taxonomies and vocabularies. It is expected that many will be developed and implemented at national level. Actions will focus on the identification of existing taxonomies, their applicability and interrelationships. Where possible, mappings or translations will be made between various taxonomies and vocabularies used in multilingual and multicultural learning domains.

All in all, it looks likely that LOM will be widely adopted in Europe, as this standard is well suited to deal with the multilingual European context. This is not surprising, as much of the original research and development took place under the European ARIADNE umbrella.

4 Open issues and problems

We believe that the LOM standard provides a sound basis for educational metadata. Even if we take this for granted, a series of important issues and problems require further attention.

Awareness about the relevance of metadata for knowledge management in general, and for educational purposes in particular, has increased sharply these last years. In itself, this is obviously a positive evolution. However, it also raises issues about *information dissemination* towards the community of end users and developers, and about *expectation management*. Even though most of the relevant organisations have a very open approach, where basically anyone can participate and contribute, the technical nature of the work and the somewhat obscure formalisms involved ('acronym soup') may make the field somewhat intimidating to newcomers or those who just want to assess the impact on their own work. Moreover, an analysis is required of how different communities adopt and support educational metadata for their constituencies [Duval, 2001].

Authoring of data and metadata is (too) *hard and time consuming*: automatic generation of obvious metadata is useful and possible, but especially semantic metadata will in most cases need to be provided through human intervention. Metadata templates can help to make this process of indexation easier, especially when similar documents need to be described regularly. Moreover, the development of interesting educational resources, that *really* add value when compared with their paper counterparts (books, slides, etc.) is extremely time consuming and quite complex, the more so as it often requires a multidisciplinary team of context experts, graphical designers, technical experts, pedagogical experts, etc.

We basically argue that standardized metadata help end users to identify and locate relevant educational material. However, that doesn't mean that, once such material has been identified, no *further barriers to (re-)use* remain: the user interface or look-and-feel of the resource may need to be adapted to the overall context it will fit in, there may be technical, organisational or legal reasons that prevent the material from being made available to new users, the pedagogical style of the resource may not be appropriate for the intended new context, etc. This issue raises the question of adaptation of resources, either through human intervention (which requires interoperable authoring environments) or automatically (which is beyond the current abilities of so-called adaptive systems).

In the more general sense, there is the open question on *what exactly needs to be standardized*, and in what order. This is to be considered in the context of political, legal and other sensitivities. The question of appropriate priorities is even more important when one realizes that the standardization process takes a long time: between the first stable ARIADNE metadata specification and the LOM standard ballot, five years were spent on consensus building, evaluations and testing!

Finally, there is the issue of interoperability in a wider sense: as the goal is to realize an infrastructure for interoperable tools and services, the question arises what the appropriate building blocks or components for such an infrastructure are. Should there be document and metadata servers? Should these be the same? Can their data modelling and management requirements be met by traditional database technologies? Should we rather opt for a peer-to-peer approach? How will management of educational resources tie in with general knowledge management? What about security? Etc. Etc.

5 Conclusion

In this paper, we have argued that standardized metadata are a prerequisite for large-scale deployment and (re-)use of educational resources. The standardization process in this area is maturing rapidly, with the first stable specification (IEEE LTSC LOM) now under ballot. This seems to suggest that the first requirement for a worldwide pool with a critical mass of reusable pedagogical documents can be met. This situation creates exciting opportunities for further research and development in this area (design for reuse, semantic interoperability, metadata and document authoring).

References

[ADL, 2001] http://www.adlnet.org/.

[ARIADNE, 2001] http://www.ariadne-eu.org/.

[Berners-Lee, & Fischetti, 1999] T. Berners-Lee & M. Fischetti. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor.* Harper, 1999.

[Berners-Lee et al., 2001] T. Berners-Lee, J. Hendler and O. Lassila. *The Semantic Web*. Scientific American, May 2001 <http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html>.

[Duval et al., 2001] E. Duval, E. Forte, K. Cardinaels, B. Verhoeven, R. Van Durm, K. Hendrikx, M. Wentland Forte, N. Ebel, M. Macowicz, K. Warkentyne, F. Haenni, *The ARIADNE Knowledge Pool System: a Distributed Digital Library for Education*, Communications of the ACM, May 2001.

[Duval, 2001] E. Duval. *Standardized Metadata for Education: a Status Report*, EdMedia2000, 26 June-1 July 2001, Tampere, Finland.

[Forte et al., 1999] E. Forte, F. Haenni, K. Warkentyne, E. Duval, K. Cardinaels, E. Vervaet, K. Hendrikx, M. Wentland-Forte, and F. Simillion, *Semantic and Pedagogic Interoperability Mechanisms in the Ariadne Educational Repository,* ACM SIGMOD Record Vol. 28, no. 1, 20-25, March 1999.

[Fox et al., 2001] E. Fox & G. Marchionini. *Special issue on Digital Libraries.* Communications of the ACM, May 2001.

[Hill et al., 2001] L. Hill and T. Koch (eds.), *Networked Knowledge Organisation Systems: introduction to a special issue*, Journal of Digital Information, Volume 1, Issue 8 <http://jodi.ecs.soton.ac.uk/Articles/v01/i08/editorial>.

[IEEE, 2001] *Draft Standard for Learning Object Metadata,* draft 6.1, April 2001 <http://ltsc.ieee.org/wg12/index.html>.

[IMS, 2001] http://www.imsproject.org/

[INDECS, 2001] http://www.indecs.org/.

[ISSS, 2000] *A Standardization Work Programme for "Learning and Training Technologies & Educational Multimedia Software".* CEN Workshop Agreement 14040, 14 July 2000.

[ISSS, 2001] <http://www.cenorm.be/isss/Workshop/lt/>.

[Paepcke et al., 1998] A. Paepcke, C. K. Chang, H. Garcia-Molina & T. Winograd. *Interoperability for Digital Libraries Worldwide.* Communications of the ACM, Volume 41, No. 4, pp. 33-43, April 1998.

[Rust & Biede, 2000] G. Rust & M. Bide. *The <indecs> metadata framework. Principles, model and data dictionary.* June 2000 <http://www.indecs.org/pdf/framework.pdf>.

[W3C, 2001] *Metadata and Resource Description*, April 2001 <http://www.w3.org/Metadata/>.