

---

# Creating metadata practices for MIT's OpenCourseWare Project

---

*Rebecca L. Lubas*  
*Robert H.W. Wolfe and*  
*Maximilian Fleischman*

---

## The authors

Rebecca L. Lubas is the Special Formats Cataloging Librarian, Robert H.W. Wolfe is Metadata Specialist and Head, Metadata Unit and Maximilian Fleischman is Metadata Production Assistant, all at Massachusetts Institute of Technology Libraries, Cambridge, Massachusetts, USA.

---

## Keywords

Online cataloguing, Libraries, USA

---

## Abstract

The MIT libraries were called upon to recommend a metadata scheme for the resources contained in MIT's OpenCourseWare (OCW) project. The resources in OCW needed descriptive, structural, and technical metadata. The SCORM standard, which uses IEEE Learning Object Metadata for its descriptive standard, was selected for its focus on educational objects. However, it was clear that the Libraries would need to recommend how the standard would be applied and adapted to accommodate needs that were not addressed in the standard's specifications. The newly formed MIT Libraries Metadata Unit adapted established practices from AACR2 and MARC traditions when facing situations in which there were no precedents to follow.

---

## Electronic access

The Emerald Research Register for this journal is available at [www.emeraldinsight.com/researchregister](http://www.emeraldinsight.com/researchregister)

The current issue and full text archive of this journal is available at [www.emeraldinsight.com/0737-8831.htm](http://www.emeraldinsight.com/0737-8831.htm)

Until 2002, MIT Libraries' Bibliographic Access Services only dabbled in non-MARC metadata. The reliable MARC format met most cataloging needs for decades. In the last few years, encounters with other metadata schemes began to occur with increasing regularity. Frequently, digital objects carried metadata that could be used in the online catalog if harvested and converted to MARC. The libraries converted a sampling of Dublin Core (DC) and Federal Geographic Data Committee (FGDC) metadata records into MARC to make them compatible with MIT's local integrated library system. The focus of these experiments was to make the data as MARC-like as possible rather than to exploit the features of the alternative standards.

With the advent of DSpace, MIT's digital repository, the libraries' staff gradually realized that not all metadata would ultimately be converted to MARC and that in some cases MARC would not be the most desirable standard. DSpace contains items in a variety of formats. Some of the items are born digital and others are converted from print resources. Dublin Core was chosen as the DSpace metadata standard because it is well-developed and flexible enough to address the needs of a wide array of formats.

DC's similarity to MARC also made it a logical choice for MIT's first major venture into non-MARC metadata. Its advantage over MARC is in being tailored specifically for describing and providing access to electronic resources. The DC records in DSpace function in the same way that the MARC records do in an ILS. They are primarily surrogates to aid discovery. The libraries' approach to creating metadata practices for DSpace was heavily influenced by MARC traditions.

These early metadata experiences demonstrated that much needed to be learned to appreciate and utilize the capabilities of other metadata standards. The libraries formed a Metadata Advisory Group in 2001 for the express purpose of creating inhouse expertise about metadata beyond MARC.

In the spring of 2002, representatives from MIT's OpenCourseWare Project (OCW)[1] approached the libraries. OCW plans to make much of the course materials from 2,000 of MIT's course offerings available on the Web, free of charge, to any user anywhere in the world. They plan to do this by 2007, with 25 per cent off the courses available in Fall 2003. OCW's planners recognized the need for metadata to make the courses with their associated learning objects

---

Received August 2003

Revised September 2003

Accepted November 2003



searchable, retrievable, and readily preserved. They also recognized that the MIT Libraries would be the place to begin looking for metadata experts.

## The OCW metadata proposal

The MIT Libraries were engaged to propose a metadata scheme in the summer of 2002 for OCW. OCW needed a metadata scheme to begin using immediately for courses going into production the following fall. The libraries agreed to recommend an accepted metadata standard and to suggest best practices for the standard's application in OCW. The proposal also included staffing suggestions for the creation of a metadata unit to provide metadata application services to OCW. The unit would be organized under Bibliographic Access Services, the libraries' monograph cataloging department.

The OCW project presented another opportunity to take a step away from MARC in the application of metadata to electronic resources. The libraries recommended the SCORM/IEEE Learning Object Metadata (LOM) standard for the foundation of OCW metadata. Sharable Content Object Reference Model (SCORM) treats educational digital objects as active, dynamic items – an aspect which traditional library organization usually neglects. SCORM includes a more robust architecture for describing the digital objects' structural and operative relationships.

IEEE LOM has nine basic element areas:

- (1) General;
- (2) Lifecycle;
- (3) Meta-Metadata (information about who created the metadata and how it was created);
- (4) Technical;
- (5) Educational;
- (6) Rights;
- (7) Relationship;
- (8) Annotation; and
- (9) Classification.

The General, Lifecycle, Meta-Metadata, Technical, Annotation and Classification areas provide information much like a familiar MARC record with local information fields added. The Rights area addresses the complex aspect of rights management in the digital world. OCW chose to handle rights information in a separately-staffed Intellectual Property division with its own metadata record. The educational area, which describes the use of the object in a learning context, is an aspect of SCORM/IEEE LOM that needed the most work in establishing best practices for the libraries staff to fully exploit the

standard's capabilities. The Relationship area provides a way to describe how resources interact with one another.

After examining the 20 pilot courses, libraries staff identified a need for three types of metadata records for each OCW course. The initial proposal labeled these record types:

- Course Level;
- Course Item; and
- Content Item.

The Course Level record was an overall record, describing the entire course site and its contents. The Course Item record would describe items created for the specific course, such as syllabi, lecture notes, and exams. The Content Item record would describe items that could stand by themselves out of the context of the course, such as journal articles. The Course Item records and the Content Item records would use the IEEE LOM Relationship area to link the object back to the course of origin. This aspect of the initial proposal came out of the common cataloging and library collection management practice of treating published items differently from more ephemeral material.

## Project planning and implementation

After the initial proposal was accepted, the libraries then participated in OCW's workflow and production planning. During this phase, the original proposal was adjusted as more decisions were finalized about the actual course content, production timelines, and available labor. It became increasingly clear that the distinction between Content Item and Course Item was not as important as it had been with more traditional formats in a library collection. As planning for the metadata service developed, the practices for Course Items and Content Items appeared nearly identical. By the time full production began, the only discernable difference between a Course Item and a Content Item was that a Content Item might have additional contributors listed. OCW project personnel use the term "resources" to refer to all items, and the libraries staff ultimately adopted this term for simplicity.

In practice, a fairly rigid content hierarchy of the metadata records developed. Each course was organized on three levels:

- (1) Course Level;
- (2) Section Level; and
- (3) Resource Level.

The course in general is represented by a single HTML document, and the Course Level record became the most robust of the metadata records in

OCW's content management system. It is only at this level that name authority work and subject analysis are provided. In the suggested practice of the original proposal, every resource within a course received detailed subject analysis and name authority work. This requirement was dropped in the libraries' final agreement with OCW because of the labor-intensive nature of such work. With an average of 35 resources per course, there would have been 70,000 resources to analyze. OCW's tight deadlines did not grant enough time to do subject analysis of every resource with the amount of staff hours that were funded.

The next level in the hierarchy, the Section Level, developed into aggregations of resources around functional activities such as exercises, exams, lecture notes, syllabi, etc. These section pages receive the least amount of metadata. Principally they record the organization and operation of the complex digital objects. The Section Level records inherit metadata from the Course Level, and a Relationship field is generated that refers every section back to the course in which it resides.

Resources, the last level of the hierarchy, are non-HTML single bitstreams. Adding to the complexity of the course is the fact that resources refer to, employ and relate to each other. Digital objects, and educational digital objects especially, present a challenge for MARC in their recombinable nature. LOM addresses this aspect in a more logical way than MARC does.

Unlike a catalog record, the metadata records created for the OCW project are more than search and recovery surrogates. The content management system relies upon the metadata records in constructing the navigation environment of the course. Course and Section HTML pages are not the source of global and local navigation. Rather, the CMS includes a "frame" or "skin" which contains navigation menus for each course. These menus are built from the structural metadata in the records the Libraries help create.

The workflow for the OCW metadata is based on traditional library technical services models, with the added feature of working with author- and auto-generated metadata. The process requires the OCW faculty liaison, working with the course's creator, to submit a preliminary set of metadata following guidelines developed by the libraries and OCW staff. Some of the metadata is supplied by the system, drawing on pieces of information from the course framework, such as the course's number, title, and the semester in which the course was originally taught. Other metadata elements are filled in by the faculty liaison. The records then come to the libraries'

Metadata Unit[2], where a metadata production assistant and a metadata specialist revise them. The division of labor in the libraries' unit mirrors cataloging departments. The metadata production assistant, the equivalent of a library technical cataloging assistant, revises the existing descriptive metadata much like a copy cataloger would revise a MARC record provided by another library. The metadata specialist, acting in the manner of a traditional professional cataloger, performs the specialty duties of classification, subject work, and name authority work. In practice, during the first few months of production, the metadata specialist did much resource metadata revision and creation in order to see a representative number of resources as well as meet OCW's course publication deadlines. This activity guided the establishment of best practices.

Using SCORM's implementation of IEEE's LOM allows the cataloging entity to employ the major concepts of AACR2 and MARC in its areas, including description, subject access, and classification. An attractive feature of LOM is that it offers considerable flexibility by allowing multiple thesauri to be plugged in the Classification area as long as the thesaurus in question is identified. The proposed OCW implementation of IEEE LOM employed methods from traditional cataloging such as standardization of contributors' names via authority work in the Contribute area and the use of the Library of Congress Subject Headings in the Classification area. Additionally, OCW requested that the Libraries include the National Center for Education Statistics' Classification of Instruction Programs. The Libraries proposed that during the course creation process, the faculty be allowed to use any other classification scheme they might be familiar with, and that such suggestions be included in the metadata. In practice, this capability has not yet been used to its full potential. For the first 500 courses, the OCW faculty liaisons provided keyword lists in a free form manner. The liaisons create a list of keywords derived from the course descriptions and resource titles. When the metadata are enhanced in the libraries, the keywords are examined for redundancy. Some keywords may be added. For example, if a Library of Congress Subject Heading is added to the Classification area in the Course Level metadata, and that heading has cross-references, the cross references may be used as keywords. Both extremes of electronic search systems, uncontrolled keywords and controlled vocabularies, are available in OCW. OCW metadata are developing into a hybrid of time-tested cataloging practices and Web searching methods.

Name authority work is performed for the course authors and contributors that appear in the Contribute area. First, the metadata creator searches the OCLC authority file. If the name is not found there, the authority file in Barton, MIT's catalog, is searched. If a match is still not found, the MIT roles database of current students and faculty members is examined. When an authoritative form of the name is found, the metadata creator opens a record in a FileMaker Pro database, and records the name, any cross-references, and the course number in which the name appears. In the event of conflicts, AACR2 methods for distinguishing names are employed. For example, if there are two James Smiths, distinctions such as a middle initial or birth date are sought. The authoritative form is then entered in the Course Level metadata, and this information is replicated in all levels of metadata for the course. At this time, the database of OCW author names stands alone, but the Metadata Unit is planning ways in which the information may be utilized more interactively.

Much labor is saved via inheritance. All of the resources within a course inherit the Course Level metadata, which includes contributors, basic technical information, and classification terms. The name of the course populates the Relationship area in each resource. If a resource is used in more than one course, it has multiple relationship entries.

It was originally anticipated that OCW learning objects would take many forms such as video, audio, text, and program code. IEEE LOM includes a technical requirements element set, and the libraries' suggested implementation is an expansion of the System Requirements note in MARC. The courses encountered in the first months of the Metadata Unit's work tended to be heavily reliant on text documents in Portable Document Files (PDFs). The focus on text resources allowed the metadata creators to borrow heavily from the text-centric MARC standard for best practices. For example, the best practices for the Unit use spacing and punctuation rules from AACR2/MARC to guide the metadata in the General Description area of IEEE LOM.

The original proposal was made and much of the workflow planning happened before a content management system with a metadata interface was chosen. The first 50 pilot courses were published in a temporary system. The permanent content management system became live shortly before the production process for the 500 Fall 2003 courses began, so much of the development of best practices needed to happen on the job.

## Creating resource item record best practices

The arrival of the Content Management System (CMS), an out-of-the-box Microsoft solution, brought the real challenge to preservation of rich MARC/AACR2 practices. Where the SCORM/IEEE LOM standard is extensible and allows for that most sacred of AACR2 traditions, cataloger's best judgment, the CMS is a rigid tool that forces the metadata creator to be dogmatic about metadata application, especially in applying metadata to resource level objects.

The need to balance the impulse toward traditional practices against the flexibility required by the OCW resources and the limitations of some of the available tools is best explored through examination of the Metadata Unit's efforts towards best practices in the application of two of the SCORM element sets to content item or resource level objects.

## Technical requirements

After choosing a standard that allowed the most flexibility in describing technical requirements, the difficulty in applying these metadata elements focused upon interpretation of the SCORM instruction for this element to provide "technology required to use this learning object". The Metadata Unit identified two ways in which one could understand the intended use of the objects, the original educational use versus the OCW use.

To provide technology required to satisfy the original educational use, one would follow the point of view of the disseminator of the course objects, providing all technology that a student would require to employ the objects in an actual learning situation. A second definition for OCW would require one to follow the point of view of the end-users of the OCW Web site, providing just that technology that would be required to access the material. Note that this expectation of use is closer to that which MARC records hold for library catalog users.

While the libraries' staff recognized and planned the metadata application for the intrinsic, complex operational nature of the OCW objects, in implementation that operation is restricted to something closer to the surrogate nature of MARC records. The Metadata Unit chose to follow the second interpretation and not provide all the information a student would require. This decision was based upon OCW's statement that it does not intend these materials to constitute credit for a course or in any way substitute for MIT courses of instruction. Both the object and the

metadata are a snapshot in time. The material intended for public consumption, while in format instructive, in function is illustrative. This is not to say that the Metadata Unit does not take advantage of the richness of the SCORM Standard[3]. The unit's rules of application are no technical requirement needed unless:

- A Web browser will not display file and text editor displays file as gibberish.
- For resources that require software of which there are multiple vendors, add a general technical requirement. Do not recommend one vendor over another.
- Analyze the contents of Zip files for other possible technical requirements.
- Many programs use generic file extensions. Examine the context in which the file is included in the section materials to determine the software that might be needed.

### Best practices for learning resource types

A second element that highlights the Metadata Unit's attempts at good cataloging practice for electronic objects in a non-MARC standard is the "Learning Resource Type" (LRT). IEEE LOM defines the value space for this element as the following list:

- exercise;
- simulation;
- questionnaire;
- diagram;
- figure;
- graph;
- index;
- slide;
- table;
- narrative text;
- exam;
- experiment;
- problem statement;
- self assessment; and
- lecture.

Not having the resources or time to provide the level of effort required to properly create a custom taxonomy for all of MIT's learning objects, the Metadata Unit elected to adopt this list. The justification for adoption focused on the benefits of a having an immediately employable controlled vocabulary, focusing the unit's effort at identification and improving the claim to interoperability. Interoperability is a prime concern at the end of the lifecycle of these objects, when they will have to be crosswalked to other metadata schemes in order to be permanently archived in a repository such as DSpace.

SCORM's implementation of IEEE LOM describes the Learning Resource Type element as specifying, "the kind of learning object". For those used to AACR2 and MARC, the LRT functions much as General Material Designations and Specific Material Designations. Unlike AACR2, the LRT terms of SCORM do not have glossary definitions. IEEE LOM refers implementers to the OED and "communities of practice" for definitions of its terms. The Metadata Unit found that the OED provided an adequate means of interpreting these terms as kinds of learning objects. Kind is understood to define a "class of objects distinguished by attributes possessed in common". The unit discovered that the three attributes these terms describe are format, function, and association. Lacking an equivalent of the Library of Congress Rule Interpretations for IEEE LOM, the Metadata Unit created rules of application after seeing a critical mass of roughly 300 courses with associated resources.

Here is the list of the Best Practice guidelines that the unit devised for applying LRTs:

- (1) When the attributes suggest more than one value for the learning resource type, choose the dominant kind in this fashion: format dominates function dominates association.
- (2) Per SCORM specification that, "the most dominant kind shall be first", enter the LRT that is most dominant. As the content management system does not allow for multiple instances of this element, place any other applicable LRTs in the description field.
- (3) Take a narrow interpretation of narrative text. If the dominant format is determined to be textual but not narrative, then apply the rule: Function dominates association dominates format. If neither a function nor an association is identifiable for the object, only then resort to narrative text.
- (4) For images that are not readily identifiable by their format as a diagram, figure or graph, look to Function and then Association. If neither of these is identifiable, then best practices recommend the use of "Figure" with an appropriate description.
- (5) When applying an LRT for binary code use Simulation as format and LRT. Code that displays text in a browser or text editor should be treated as text, apply rule 2.

The way in which images were captured in the LRT best practice vocabulary was not entirely satisfactory. The generators of the list did not accurately consider the instructive power of images that were not communicating some quantifiable data. "Figure" became a catch-all for images of this sort, while the metadata creators relied upon the educational description element to provide better information.

## Conclusion

In selecting metadata formats and creating best practices, MIT Libraries' Metadata Unit attempted to preserve AACR2 and MARC cataloging traditions developed from generations of library experience. The adoption of other standards is an attempt to accommodate the diffuse natures of the libraries' growing collection of electronic resources within the standards and practices of traditional cataloging. Sometimes it is the tools that make things difficult, when dealing with a CMS that is not as extensible as the standard. Sometimes it is the youth of the area of endeavor, when controlled vocabularies still have gaps.

The Metadata Unit has found that even with all the forward thinking and cutting edge technologies used in the OCW metadata effort, it

is the traditional cataloger's sensibilities regarding good description and access – as derived from the AACR2/MARC heritage – that is most valuable in discovering access to the library's new class of electronic objects.

## Notes

- 1 OpenCourseWare, available at: <http://ocw.mit.edu/OcwWeb>
- 2 MIT Libraries Metadata Unit, available at: <http://libraries.mit.edu/guides/subjects/metadata/index.html>
- 3 SCORM Standard, available at: [www.adlnet.org/index.cfm?fuseaction=scormabt](http://www.adlnet.org/index.cfm?fuseaction=scormabt)