

Metadata the Gateway to the Web? BERA 2001, Leeds University: 13 - 16 September 2001

Paul Shabajee
*Graduate School of Education &
Institute for Learning and Research
Technology, University of Bristol*
paul.shabajee@bristol.ac.uk

Andy Dingley
*HP Laboratories,
Bristol*
andy_dingley@hpl.hp.com

Abstract

Genuinely valuable educational resources are becoming widely available via the Internet and our pedagogic understanding of how to make use of these valuable resources is increasing rapidly. However locating information and other resources is often a serious barrier to genuinely effective use of the Internet.

This paper explores and demonstrates how the systematic use of metadata can make publishing, locating and using Internet based resources more effective. The evolving national and international educational metadata standards will be critically reviewed noting a number of concerns and issues regarding their development and use.

Key words: Internet, ICT, educational resources, metadata.

1. Background

This paper is based on research activities focused on the development of ARKive (Wildscreen Trust 2001) a large multimedia project working to digitise and make available, on the Web, a very large database of video, still images, audio and supporting information and resources about the world's wildlife and habitats.

The project has funding from the UK Heritage Lottery Fund (£1.6M) to produce the 'British Chapter' i.e. a dataset about UK biodiversity and £0.5M. from the New Opportunities Fund to begin the production of a dataset based on the worlds most endangered species and habitats.

In parallel Hewlett Packard Labs, is funding a research programme (\$2M) to help develop the technological infrastructure for the project. HP Labs is also funding research project (ARKive-ERA) based at the University of Bristol investigating how ARKive type systems can be developed to be as useful as possible to the 'education' communities around the world.

ARKive and HP Labs wish to develop an infrastructure that will enable maximum use of the use of ARKive resources, by all users. Re-purposing individual resources (i.e. the re-use in different contexts of a given resource) is fundamental to such an endeavour. Metadata and the related technologies are the building blocks of the ability to re-purpose resources effectively.

2. Introduction to Metadata

Metadata

"... there are no results."

Oxford English Dictionary On-line 12/08/01

The definition of the term Metadata is not yet stable e.g. 'information about information' (EdNA 2001), 'an item of metadata is a relationship that someone claims to exist between to entities' (Rust & Bride 2000) and Dublin Core (DCMI 2001) state that "the simplest definition of metadata is " structured data about data."

A working defination used by the ARKive-ERA project is:

"Structured data about data which facilitates the management, discovery and retrieval of resources"

In the case of electronic resources these may or may not be machine readable form i.e. in a form that a computer program can extract and utilise e.g. as context when searching a database.

Clearly 'metadata' is not new or specific to electronic media. Library catalogues and many other cataloguing and indexing systems are based on the same principles.

3. What is it for, why do we need it?

Hodgson (1998) has developed a list of 'metadata functions' which lists what appears to be a comprehensive list of current (c1998) uses of metadata.

- **Summary** - to summarise the meaning of the data (ie what is the data about)
- **Finding** - to allow users to search for the data
- **Advisement** - to allow users to determine if the data is what they want
- **Selection** - to help decide which instance of the data should be retrieved (if multiple formats are provided).
- **Retrieval** - to retrieve and use a copy of the data (i.e. where do I go to get the data)
- **Restriction** - to prevent some users from accessing data. Content ratings describe attributes of a resource within a rating scheme (e.g., PICS - the Platform for Internet Content Selection standard)
- **Interpretation** - to instruct on how to interpret the data (e.g., format, encoding, encryption).
- **Specifications** - to give information that affects the use of data (e.g., legal conditions on use, its size, or age); terms and conditions for use of an object...
- **History** - to describe the history or provenance of data, such its original source and any subsequent transformations (filtering, decimation, etc.)
- **Data administration** - to give specifications for the management of an object within a server or repository (date of last modification, date of creation, and the administrator's identity)
- **Data linkages or relationships** - to give specifications about the relationship between objects (between a set of articles and a containing journal, ...)
- **Data structure** - to list the logical components of complex or compound objects and how to access those components (table of contents; the list of components of a software suite)

This list is the most complete that the authors have identified in our review of the literature. However it might be useful to re-label and re-organise these e.g. "data linkages or relationships" could be replaced with the now widely used term 'interoperability' (Paul Miller 2001b) i.e. the ability of systems to 'inter-operate' or more simply to be compatible.

'Re-purposing' is not explicit in this list which reflects the fact that this was not a widely identified purpose c1998. However effective metadata 'tagging' is functionally essential if 'resources' are to be re-purposed – i.e. one needs to know what the object is and what it can be used for, before it can be reused meaningfully.

The 'users' of metadata can be divided up in many ways, one that the authors find helpful is:

Publishing: by those who wish to publish information who utilise metadata to help in management, re-purposing and publishing of the data, including publishing the relevant metadata to help consumers or other third parties access their data.

Consuming: by end users who may explicitly utilise the metadata using indices or implicitly via for example a Web search engine which may utilise metadata.

Manipulating: third parties may utilise the metadata to provide other 'information services' e.g. Web search engine and Web portal providers.

3.1. Schemas and Serialisations

Metadata standards must define it at two levels; schema and serialisation.

Schema

A Schema consists of a list of conceptual elements, together with any further constraints such as qualifiers to increase the specific nature of elements, or lists of allowed values.

One element from the Dublin Core's schema (of fifteen elements in total) is Creator. There is no need for an "author" element, because this is represented as a Creator with a further Creator Type qualifier of "Author".

Educational schemas may define a relatively simple list of elements, but have great complexity on their vocabulary of allowed values (the UK National Curriculum is an extreme example). Allowing effective machine processing that is more powerful than previous free-text searches depends as much on these constrained and meaningful values as it does on the structure.

The most popular schema is that of Dublin Core. This began as a non-web, and indeed non-IT, format and has weathered many changes. Its resolutely general purpose nature and refusal to be defined by any one single serialisation has meant that it is still useful and free of technical obsolescence. Despite being often derided as a lowest-common denominator format, it is extremely capable when used correctly, particular with its qualifier mechanism. Although Dublin Core is indeed a lowest common denominator, it still represents a considerable advance in capability over current typical practice.

Schemas are often described in terms of free-text descriptions. This is a useful technique in that it is easily understood, but it is also risky as it permits vague definitions and misunderstandings. As the processing of metadata becomes more automatic, the need for accuracy will increase. Web crawling metadata consumers are very literal-minded and do not have a human's ability to infer correct meanings – and the major audience for this metadata will always be machines, not humans.

Metadata published without reference to a schema is effectively useless. The web is already full of HTML <meta> elements that were created because they were easy for the author, not because they were useful to a consumer. Authors should guard very carefully against this risk; there is simply no point in publishing metadata unless it has a future use, and it is also described by some published method so that the consumer can identify it as being useful to them. An author's own invented schema conveys absolutely no information to a consumer who does not possess or understand it.

Serialisation

A serialisation is a description of how the conceptual data model is written into a document. A few well-known formats are already popular; HTML <meta> elements, MARC records and RDF (W3C 1999). This is an area of considerable development at present, as tools for manipulating the far more sophisticated RDF make access to it as easy as previous <meta> elements.

In HTML, the long established <meta> element might look like this:

```
<meta name="DC.Identifier"
content="http://www.antiguanracer.org/" />

<meta name="DC.Title" content="The Antigua
Racer Conservation Project" />

<meta name="DC.Subject.keywords"
content="snake, endangered, rare, Antigua,
ARkive, Wildscreen, Caribbean " />
```

```
<meta name="DC.Creator" content="The
Wildscreen Trust, Fauna & Flora
International" />
```

In RDF, a small part of this same content may be represented as:

```
<dc:creator>
<rdf:Bag>

<rdf:li rdf:parsetype="Resource" >
<rdf:value>The Wildscreen Trust</rdf:value>
<dc:relation dcq:relationType="website"
dcq:relationScheme="URI"
>http://www.wildscreen.org</dc:relation>
</rdf:li>

<rdf:li>Fauna & Flora International
</rdf:li>

</rdf:Bag>
</dc:creator>
```

RDF is a complex topic, and an example of this size can convey only a tiny fraction. Note that although this example is bulkier than the simple <meta>, it also contains far more, and more structured, content about the creator. Where the structure is simple (as for FFI here), then it can reduce to be no more complex than the simple case.

As for schemas, the only useful serialisations are those to widely understood standards. Accuracy is also important, and even more so than for the schema. An inaccurate schema may cause a slight misunderstanding, but an invalid serialisation is likely to be discarded entirely.

3.2. Educational Metadata

By definition, broadly speaking, 'educational metadata' is metadata which provides 'educational' data about data. What constitutes 'educational' in this context is not yet clear as the number and nature of the evolving educational standards demonstrates (see below)

However the need for a 'standard' for educational metadata is clearly articulated by the members of communities working in this area e.g. in the UK the MEG Concord (MEG 2001) and internationally the 'Memorandum of Understanding between the Dublin Core Metadata Initiative and the IEEE Learning Technology Standards Committee' (DCMI and IEEE LTSC 2000). The reasoning reflecting the wider value of metadata (see above).

Examples of some of the major international initiatives include: IEEE LOM (USA/Global), EdNA (Australia), IMS (USA/Global), ARIADNE (EU) and DCMI Education Working Group (Global) – for details see references. All of these groups have now signed the Memorandum of Understanding above either as full signatories or as ‘concurring projects’. The IEEE LOM standards is becoming the most widely referenced standard.

Internationally these initiatives have been under development for some time the IEEE LTSC, Learning Objects Metadata Working Group for example had it first meeting in Dec. 1997 (IEEE LTSC 2001b). Only in the last few years has there been significant convergence.

In the UK there are a number of initiatives working to develop metadata standards focused on particular aspects or resources for education in the UK e.g. National Curriculum Metadata Standard (QCA 2001), Ufi/LearnDirect (Ufi 2000) and BECTa/Virtual Teachers Centre (BECTa 2001).

Each of these developing standards focused on it’s particular needs e.g. national curriculum levels and key words. Each has it’s own additions or amendments to the ‘core standards’ e.g. Dublin Core or IEEE LOM.

The MEG concord (MEG 2001) (see above) in broad terms indicates members agreement that they understand and commit to, the need for a common and collaborative development of educational metadata standards and its use.

It is worthy of note that the membership of the International and local communities include diverse range of organisations and interests, including public bodies, commercial companies and academic institutions e.g. the IMS international participant list (IMS 2001) includes ADL CoLab (Department of Defense), Apple Computer, Artesia Technologies, Blackboard, The Boeing Company, ..., California State University, Campus Pipeline, ..., Cisco Systems, Click2Learn, Inc.,...

MEG members include (MEG 2001): Aberdeen College , Archaeology Data Service (ADS)..., British Broadcasting Corporation (BBC) , British Educational Communications & Technology Agency (BECTa) , Book Industry Communication (BIC) ..., MicroCompass Systems Ltd..., Scottish Cultural Resources Access Network (SCRAN)..., and University for Industry (Ufi).

Example: It is beyond the scope of this paper to describe the detail of the various metadata standards. However as an example the IEEE LOM (IEEE, 2001) contains the following sections:

- | | |
|----|---|
| a) | The <i>General</i> category groups the general information that describes the learning object as a whole. |
| b) | The <i>Lifecycle</i> category groups the features related to the history and current state of this learning object and those who have affected this learning object during its evolution. |
| c) | The <i>Meta-metadata</i> category groups information about this metadata record itself (rather than the learning object that this record describes) . |
| d) | The <i>Technical</i> category groups the technical requirements and characteristics of the learning object. |
| e) | The <i>Educational</i> category groups the educational and pedagogic characteristics of the learning object. |
| f) | The <i>Rights</i> category groups the intellectual property rights and conditions of use for the learning object. |
| g) | The <i>Relation</i> category groups features that define the relationship between this learning object and other targeted learning objects. |
| h) | The <i>Annotation</i> category provides comments on the educational use of the learning object and information on when and by whom the comments were created. |
| i) | The <i>Classification</i> category describes where this learning object falls within a particular classification system. |

The details of the element educational element set, e) above is:

Interactivity Type	e.g. Active , Expositive , Mixed , Undefined
Learning Resource Type	e.g. Exercise , Simulation , Questionnaire , Diagram , , Graph, Self Assessment
Interactivity Level	e.g. very low, low, medium, high, very high
Semantic Density	e.g. very low, low, medium, high, very high
Intended End User Role	e.g. Teacher, Author, Learner, Manager
Context	e.g. Primary Education, Higher Education, University First Cycle, Continuous Formation, Vocational Training
Typical Age Range	e.g. 7-9, 0-5, 15, 18-,
Difficulty	e.g. very easy, easy, medium, difficult, very difficult
Typical Learning Time	-
Description	-
Language	-

Each standard, and localised amendments to core standards, is utilised either by those who developed the standard e.g. the UK National Curriculum Metadata Standard is used to provide indexing for sites that have the status of ‘trusted partner’. Thus providing the benefits of metadata ‘tagging’ described above.

The diversity of standards is seriously problematic for those like ARKive who are wish to provide access to their resources via many different 'search engines' or 'portals' in many different countries targeting many different user groups. Remember too, that 'educational users' are but one, others include be 'media researchers' and 'scientific researchers'. Conforming to the needs of multiple standards in this environment is likely to become a greater maintenance and overhead than the actual data management itself.

The MOU between the international organisations (see above) signifies that a core set of specifications and accredited standards is emerging, it is hoped that these will support an interoperable infrastructure for world-wide e-learning technology.

However this may be a long way off. The current situation is summed up well by Paul Miller “For those creating content, it is often unclear how best to create metadata which effectively describes what they create. For aggregators and portal managers, it is extremely difficult to provide comparable descriptions of resources from different sources. For the poor learner, it may be almost impossible to effectively gauge the relevance, value, or quality of many resources. In short, the current situation probably benefits no one.” (Miller 2001)

4. Solutions and Good Practice

This paper isn’t intended as a tutorial, but it can provide a few pointers and useful references.

4.1. Authoring Practice

Authoring metadata requires simple text-editing tools to embed it into web pages, good documentation and examples from the standards creators, and a little effort in learning techniques.

There are many additional tools available to automatically format it, often on the basis of filling-in web forms. These can offer a saving in the early stages, or a source of examples with the author’s own data, but they are not a good long-term solution because they do not address maintenance. Some tools (e.g. MKDoc) maintain a parallel database about the content, and can insert the necessary information when the pages are published. This is a better solution, as it allows updating, but it still distances metadata handling from the authoring process.

With a sophisticated content production process, metadata can be generated at the same time as the content, and preserved with it. In some systems, such as that proposed by the authors for ARKive, the metadata is generated as the results of a user query, then used to retrieve items of content from a large content store before assembling them as a document for publication.

4.2. Multiple Standards

The web is an inherently international medium, and this includes metadata. Educational authors should be aware of the different major standards, and support each of them. Unfortunately, at present this involves manually duplicating information, often with only minor formatting changes. This is an area of current research interest, and matters are improving.

4.3. Translation between Schemas

Most schemas are similar, and many share a common heritage in Dublin Core. Translation between them is relatively simple in principle, although tiresome. Automatic translation can rapidly become unworkable for several reasons; vaguely defined serialisation formats make tools such as XSLT unwieldy, simple n^2 growth as the number of schemas increases the translations required, and most awkwardly, the need to translate between vocabularies rather than elements.

Vocabulary translation is difficult because it is a semantic translation, not just a structural transform. This affects human translation too. It is often difficult to translate important educational concepts simply because there isn’t an obvious mapping between such items as calendar age and grade-based educational categories. These mappings are often dependent on context, and this may be a context which an external translator does not have access to. For this reason, it is often better for the authors themselves to publish to these multiple schemas.

4.4. Application Profiles

A solution to the chore of translation between schemas is that of Application Profiles (Heery & Patel 2000). Each working group will always continue to express its own requirements, but now instead of defining them as a separate schema (with much duplication) the Application Profile approach is to define them as a collection of elements taken from pre-existing schemas. Only if an element has a new function, or a new vocabulary of values, must a new one be defined. As most elements are still the basic Dublin Core set of title, authors and publishing dates, then this is a large saving, especially when many international formats must be supported.

Application Profiles cannot merge elements between different serialisations. Achieving real flexibility here requires progress in how metadata is formatted for publishing, moves towards greater use of shared formats such as RDF and smarter user agents that are more accepting in the serialisations they understand.

5. Push and Pull

An interesting area in current web development uses RSS newsfeeds (Oasis 2001). These began as a means of exchanging “headlines” between web sites, with an increasingly broad definition of their scope. Early newsfeeds either summarised the site, or offered time-based news wire services. Now, with RSS version 1.0, they’ve expanded into many new areas of content; tailored directories of available content, “animal of the day” services, lists of deep-links between sites that are encouraged as a “federation” of sites.

RSS uses the “push” model of publishing rather than “pull”. In the pull model, browsers visit a site and select items of content. In the push model, sites offer a package of related items. Any clients connecting to them receive the whole package. This is a broader view of “web push” than was once used (delivery is still by pulling), but it’s distinct because it is the supplier of content that is making the selection of items, not the recipient. Making this work outside the scope of simple news headlines relies on good metadata handling. Clients must be able to request content within appropriate limits

An exciting part of the RSS architecture is that of aggregation. Third party sites do not create content or deliver it to viewers, but they do select items from several input feeds and offer them to different outputs. In the ARKive project, a feed about ARKive and Wildscreen themselves might be combined with a feed on general environmental news, then filtered down to just those about animals and biodiversity. At present, most intelligent aggregators are operated by human editors. As attached metadata improves, more and more can be processed automatically, yet accurately and with relevance.

6. Concerns: Thinking in the Box and Disconnections

6.1. Thinking in the box

Whilst researching the literature on educational metadata, the majority of which is focused on implementation and the positive potential of metadata (see references). In the time that we had available we have been unable to identify papers which offer a detailed critical review of possible problems or negative implications for the wide spread use of metadata standards in education.

We are concerned that it is hard to find such works. This is because, it may be that systems being developed to enhance machine readability and interoperability are not fundamentally reflective of the principles of good pedagogy.

It is also the case that they are being developed to meet both the pedagogical and commercial needs of traditional School and University based ‘educationalists’ and commercially focused training bodies. Developing a single coherent system that will be effective in meeting the needs of all those needing to use the standards will, we believe, necessitate many compromises on all sides.

These are serious concerns, standards once developed are of by definition and necessity constraining. There is a risk that it becomes impossible to ‘think outside the box’ and still be interoperable or compatible and thus ‘approved’ or accessible via search engines or portals that require ‘standards conformance’.

It seems to the authors that there is a danger that educational resource developers be they school teachers, lecturers or commercial developers may be at risk of letting metadata structures dictate the structure of their publications. Authoring resources with the pigeon holes of metadata categories in mind.

6.2. Disconnection

The standards are of necessity developing very quickly, the authors have noted that there is a near total disconnection between the end user community and small scale publishers (e.g. university lecturers) on the one hand and the standards development community on the other.

Given that the technological tools to make use of the metadata are not yet available and thus the standards are not in widespread use, it is not possible for widespread involvement of the majority of users or indeed developers at this stage in development.

The majority of those involved in the development of the standards are either focused at Higher/Further Education, vocational training or are commercial companies (e.g. see IMS 2001 and MEG 2001). The involvement of school focused educationalists and practitioners at this stage is very limited.

This is not unusual in technological standards development, however given the nature and potential impact of the standards involved it may be that such involvement needs to be encouraged.

This disconnection has two risks. One is obvious, that standards developed without close contact with users are often less relevant than they ought to be. The second is less obvious, but more immediately threatening. As noted previously, there is a great tendency to publish metadata in approximate formats that are still human-readable, but no longer machine processable without great effort.

Standards bodies must place as much emphasis on simple, useful examples and relevant (if minor) tools as they currently do on formal specifications. The user community that most needs to be educated about metadata creation is not that from a IT background, it's the community of teachers and content authors who are primarily interested in the documents, not the tools. Past experience in software teaches that good, plentiful and relevant examples communicate this knowledge more readily and accurately than any number of specifications.

7. Summary and Conclusions

Broadly speaking the effective use of metadata provides resource developers, publishers and end users with a very effective means to produce and consume information.

Educational metadata standards are still evolving. There is a realisation that to enable effective interoperability between systems, generic and widely published and utilised standards for the production and use of educational metadata is necessary.

Good practice is still evolving, however the significant and urgent interest in this area to promote e-learning worldwide is driving the very fast development of these standards.

However there are concerns about a number of issues including, the nature and impact of the evolving standards on the development of educational resource, the competing needs of stakeholders and the apparent disconnection between end users and standards development especially with respect to school age education.

8. References

- Apache (2001) Cocoon, available online at <http://xml.apache.org/cocoon/>
- ARIADNE (1999) Educational Metadata Recommendation, version 3.0, (December 1999), available online at <http://ariadne.unil.ch/Metadata/>
- ARKive (2001) ARKive project homepage, available online at <http://www.arkive.org.uk>
- BECTa (2001a) BECTa Metadata index page , available online at <http://vtc.ngfl.gov.uk/metadata/>
- DCMI (Dublin Core Metadata Initiative) and the IEEE Learning Technology Standards Committee (LTSC) (2000) Memorandum of Understanding, available on-line <http://dublincore.org/documents/2000/12/06/dcmi-ieee-mou/>
- EdNA (2001) Metadata Homepage, available online at <http://standards.edna.edu.au/metadata/index.html>
- Hodgson, Katrina. (1998) Metadata: Foundations, Potential and Applications, School of Library and Information Studies, University of Alberta Edmonton, Alberta, Canada, available online at <http://www.slis.ualberta.ca/538/khodgson/metadata.htm> [viewed 6/4/01]
- IEEE (2001) Draft Standard for Learning Object Metadata (LOM) v6.1 (18-4-2001), available online at (<http://ltsc.ieee.org/wg12/index.html>)
- IEEE (2001b) LTSC, Learning Objects Metadata WG website <http://ltsc.ieee.org/wg12/>
- ILRT (2001) Institute for Learning and Research Technology homepage, available online at <http://www.ilrt.bris.ac.uk>
- IMS (2001) IMS Participants List, available online at <http://www.imsproject.org/members.html>
- MEG (The UK's Metadata for Education Group) (2001) , index web page, available online at <http://www.ukoln.ac.uk/metadata/education/>
- Rachel Heery and Manjula Patel (2000) "Application profiles: mixing and matching metadata schemas", Ariadne, Issue 25 available online at: <http://www.ariadne.ac.uk/issue25/app-profiles/>
- Paul Miller (2001) "Towards consensus on educational metadata", Ariadne, Issue 27 available online at: <http://www.ariadne.ac.uk/issue27/meg/intro.html>
- Paul Miller (2001b) Interoperability Focus homepage available online at: <http://www.ukoln.ac.uk/interop-focus/>
- QCA (2001) The National Curriculum Metadata Standard - index page available online at <http://www.nc.uk.net/metadata/index.html>
- Rust, Godfrey and Bride, Mark (2000) The <indec> metadata framework: principles, model and data dictionary, <indec> available online at <http://www.indec.org/pdf/schema.pdf>
- Ufi (2001) Ufi Qualified Suppliers homepage, available online at <http://www.ufilt.co.uk/>
- W3C (1999) Resource Description Framework (RDF), available online at <http://www.w3.org/RDF/>
- Wildscreen Trust (2001) ARKive, available online at <http://www.wildscreen.org.uk/arkive/>
- Oasis project (2001), introduction to RSS (Rich Site Summary), <http://www.oasis-open.org/cover/rss.html>
- MKDoc (2001), web site content management system for publishing metadata , <http://mkdoc.com/>