

SCHEMAS

Contract: N° IST-1999-10100

Forum for Metadata Schema Implementers

METADATA WATCH REPORT #5

D26

Document number:

SCHEMAS-PwC-WP2-D26-Final-20010828

General Information

Title	Metadata Watch Report #5
Creator	Makx Dekkers
Contributor	Michael Day, Erik Duval, Laurie Causton
Subject-Keywords	Deliverable D26; WP2; Metadata Watch Report #5; Metadata activity reports
Description	This document comprises the top-level synthesis for D26, Metadata Watch Report #5
Publisher	PricewaterhouseCoopers
Date	28 August 2001
Type	Text Manuscript
Format	application/MSWord 2000
Identifier-	
Document Number	SCHEMAS-PwC-WP2-D26-Final-20010828
Language	English
Rights	European Commission; Internal circulation within project; External circulation via SCHEMAS Web site

Dublin Core metadata for this document

```
<?xml version="1.0"?>
<rdf:RDF xml:lang="en"
xml ns: rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xml ns: dc="http://purl.org/dc/elements/1.1/"
xml ns: smes="http://www.schemas-forum.org/registry/schemas/SCHEMAS/1.0/smes#">
<rdf:RDF rdf:resource = " ">
<dc:title> Metadata Watch Report #5 </dc:title>
<dc:creator> Max Dekkers </dc:creator>
<dc:contributor> Michael Day </dc:contributor>
<dc:contributor> Erik Duval </dc:contributor>
<dc:contributor> Laurie Causton </dc:contributor>
<dc:subject> Deliverable D26 </dc:subject>
<dc:subject> WP2 </dc:subject>
<dc:subject> Metadata Watch Report #5 </dc:subject>
<dc:subject> Metadata activity reports </dc:subject>
<dc:description> This report comprises the top-level synthesis for D26, Metadata Watch
Report #5 </dc:description>
<dc:publisher> PricewaterhouseCoopers </dc:publisher>
<dc:date> 2001-08-28 </dc:date>
<dc:type> Text </dc:type>
<dc:format> application/MSWord </dc:format>
<dc:identifier> SCHEMAS-PwC-WP2-D26-Final-20010828 </dc:identifier>
<dc:language> en </dc:language>
<dc:rights> European Commission; Internal distribution within project; External
circulation via SCHEMAS Web site </dc:rights>
</rdf:RDF>
```

Table of Contents

1. INTRODUCTION	5
2. OBJECTIVE	5
3. CO-OPERATION	5
4. MULTILINGUALITY	6
5. CONTROLLED VOCABULARIES	7
APPENDIX A: DOMAIN REPORT: CULTURAL HERITAGE SECTOR	9
APPENDIX B: DOMAIN REPORT: EDUCATIONAL SECTOR	12
APPENDIX C: DOMAIN REPORT: PUBLISHING SECTOR	16

1. Introduction

This deliverable is the fifth Metadata Watch Report from the SCHEMAS project. As specified in the project objectives, the purpose of the SCHEMAS Metadata Watch (MD Watch) is to provide a quarterly overview of world-wide progress in the metadata field, which includes work on metadata sets, schemas, frameworks, registries, and the tools needed to create and use all of these.

This fifth report describes the work that has been done by SCHEMAS partners PricewaterhouseCoopers, UKOLN and GMD, as well as by a number of correspondents in the following domains:

- Cultural Heritage sector
- Educational sector
- Publishing sector

2. Objective

This fifth Metadata Watch Report gives an overview of the developments in various domains that are relevant for the SCHEMAS audience. Two particular aspects receive special attention in some of the contributions: multilingualism and controlled vocabularies. These two aspects have also been special topics of the third SCHEMAS Workshop that was held in Budapest in May 2001.

This Report contains a top-level synthesis looking at major trends in the metadata field, based on the contributions from the correspondents that are included as appendices.

3. Co-operation

A first trend that can be observed is a trend towards co-operation between metadata activities and even concentration of efforts.

From their beginnings in the last five years, many metadata activities and standardisation efforts have been based on specific requirements with a domain or group of co-operating organisations. The specific requirements were formulated with very specific goals in mind, providing services to a specific audience.

In the last year or so, we can see a gradual blurring of boundaries. Even if projects had a fairly clear idea of their objectives when they set out, they are finding out that there are other communities with comparable objectives that may be served by their solution, originally designed for one community. For example, when solutions for rights management are developed for single-media resources (text, images, music), the same solution may be applicable to multimedia resources.

When boundaries blur, metadata solution developers are finding out that others are also busy designing metadata solutions in domains that are overlapping with theirs. And there is in fact a continuum of overlapping domains: museums, libraries, archives, book publishing, multimedia publishing, broadcasting of moving images, still images, educational resources, government information, geospatial information, statistical datasets; it is not difficult to identify types or genres of resources that are

relevant to two, three of the domains mentioned. Yet many of these domains have their own metadata approaches.

In this situation, metadata approaches start to compete for adoption within a certain community. When this happens, and metadata projects opt for either one or the other, we will see 'islands of interoperability' emerge where exchange of information between adopters of one approach is relatively easy, and exchange with others is, at best, more complex.

Such a development is obviously contrary to the concept of interoperability, and fortunately, many developers of metadata approaches have come to this realisation in recent times. They are starting to feel that it is a waste of (scarce) human resources to duplicate effort to solve the same problems and especially that it is a waste to try and win the competition and become THE metadata approach.

That having been said, it also does not make sense to strive for a single, all-encompassing metadata approach. Looking at various approaches, it can be seen that these are all firmly rooted in requirements from a certain constituency, and that no single approach would be able to meet all these diverse requirements.

In other words, we need to accept that there are and will be different metadata approaches. Some of those may be applicable to vary narrow domains or even to single organisations, others may be applicable across a scientific or business domain, while others may have relevance on a very broad level.

There are two models that we can see happening when two metadata approaches meet:

1. Formal co-operation between standards developers, where the attempt is to co-operate on the highest level, harmonising the approaches. In some cases, this can lead to the de-fact merger of metadata approaches (e.g. OeBF and EBX), to the development of a single international standard (e.g. IEEE LOM from IMS and Ariadne) or to the signing of Memoranda of Understanding (e.g. DCMI and IEEE LOM)
2. Bottom-up mapping exercises, where the attempt is to describe the mechanism of converting metadata from one standard to another. These exercises are usually done by users in a domain that is faced with a choice between two standards, or where domain-specific metadata (e.g. MPEG) needs to interoperate with cross-domain metadata (e.g. the Dublin Core)

In the first category above, a further meeting between DCMI and IEEE/IMS will take place in August in Ottawa, to discuss further practical steps on the basis of the Memorandum of Understanding that was signed in 2000.

4. Multilinguality

A second trend that can be observed is that multilingual issues are not yet on the list of highest priorities in the metadata arena.

When we look at the information and specifications produced by the leading metadata activities, we see that almost all of that is in English only. The Dublin Core Metadata Initiative attempts to provide translations of its main standard in many languages: 30 translations of the Core element set are available through the DCMI Web site.

During the third SCHEMAS Workshop (Budapest, May 2001) it was suggested that the Dublin Core set is a relatively simple case as far as translation of the specifications is concerned, because it has only 15 core elements and the semantics are only loosely defined. When it comes to more complex metadata element sets, such as the set that is being developed in the context of the Digital Object Identifier, there are more problems translating elements names and semantics that are very precise and highly culturally dependent, for example in relation to national legislation. In those cases, a “mapping” between language/culture-specific versions may be required.

It may be expected that the gradual rise in importance of non-English communities and resources on the Internet will have an effect on many metadata communities, and that at least some of them will start providing information in multiple languages.

However, it is not clear that a translation of standards specification in itself is worth the effort. In the first place, translations are bound to introduce slight differences with the original text (a phenomenon sometimes referred to as *semantic drift*). In the second place, the existence of many translations introduces a serious problem of maintenance: any change in the standard needs the involvement of many people in a very short timeframe to avoid translations to be different from the base standard.

For the implementer community, the emphasis should be on implementation guidelines and best-practice examples in local languages to explain the proper use of a standard metadata set in a specific cultural and linguistic context.

5. Controlled vocabularies

In the relatively short history of metadata for Web resources, many metadata creators have been typing in text for the metadata content without using commonly known controlled vocabularies. In the worst case, they did just type in a couple of keywords that they thought of on the spot. In more organised environments (e.g. producing resources for a specific project or in a specific domain) they may have used a list that was agreed upon within their group of collaborators.

It is becoming widely recognised that interoperability can break down completely if people enter metadata content with no underlying agreements. How can searches for a certain subject deliver relevant resources if every metadata creator assigns keywords that are not known to others? We need at least some form of consistency in order to be able to build useful discovery services.

As many people realise, controlled vocabularies can help to address this problem. Some people use words like ‘ontologies’, ‘semantic registers’, ‘concept categories’ or ‘thesauri’; whereas the exact definitions of these words or the objectives that they serve may vary depending per domain, the important point for this discussion is that they define a finite set of terms, usually maintained by some organisation and intended to bring some structure in the subjects covered in a certain domain.

Of course, the idea is not new – in various domains controlled vocabularies have been in use for a long time, sometime for centuries. We find them in the form of flat lists of terms, but they can also be structured hierarchically or be part of a networked structures that allow various types of relations to be expressed between terms. These lists help creators of metadata by giving them a list of subjects to choose from, maybe in the form of drop-down menus in their user interface. The same list can of course be

used to suggest search terms to searchers, thereby enhancing the chances that the user will find relevant results.

Some controlled vocabularies also address the multilingual issue by providing parallel lists in multiple languages. The list for a specific language can then be used by metadata creators and searchers speaking that language as above, while smart tools can either build a single language-independent index or expand searches to search parallel indexes based on the equivalence of the terms in the various languages.

Other controlled vocabularies are based on a numerical approach. These have the advantage of being language-independent and can be presented to users in the right language in the user interface. The disadvantage is that it is not obvious from looking at the raw metadata what the code stands for.

The issue of controlled vocabularies is now becoming an important discussion topic within and between metadata standardisation and implementation activities. When appropriate controlled vocabularies are publicly available and are consistently used in metadata creation, more useful search facilities can be built and long-term interoperability can be ensured.

APPENDIX A: Domain report: cultural heritage sector

Correspondent: Michael Day, UKOLN

Current state of domain

Many current developments in this domain are based on implementations of the eXtensible Markup Language (XML) and the tools being developed by the Open Archives Initiative. Despite some problems with terminology (see the paper by Hirtle), the use OAI appears to offer the cultural heritage domains (and others) a relatively 'light-weight' way to develop interoperable services based on simple metadata encoded in XML. Other topics that will be covered in this report will be an update on recordkeeping metadata, some developments originating in the museum world (including an XML DTD for the SPECTRUM standard) and the establishment in the UK of a Collection Description Focus.

Peter Hirtle, OAI and OAIS: what's in a name? *D-Lib Magazine*, 7(4), April 2001. <http://www.dlib.org/dlib/april01/04editorial.html>

Recordkeeping developments

In the general archives and recordkeeping domain, an international group of recordkeeping experts have formed an Archiving Metadata Forum in order to exchange information, to explore issues in more depth and to identify areas of potential collaboration. The group was set-up at a meeting held in the Netherlands in June 2000, and consists of recordkeeping, resource discovery and information technology experts from Australia, North America and Europe.

Archiving Metadata Forum: <http://www.archiefschool.nl/amf/>

In Australia, recordkeeping metadata initiatives have begun to build upon the Australian Recordkeeping Metadata Schema (RKMS), a general framework developed by the SPIRT Recordkeeping Metadata Project of the School of Information Management and Systems at Monash University in Melbourne. The RKMS defines a highly structured set of metadata elements that conforms to a data model based on that developed for the Resource Description Framework (RDF). The schema is designed to be extensible and can inherit metadata elements from other schemas. In June 2001, the State Records Authority of New South Wales (NSW) published the *NSW Recordkeeping Metadata Standard* (NRKMS). The standard was issued under the *NSW State Records Act 1998*, and is a mandatory standard across all NSW public offices. The Victorian Electronic Records Strategy (VERS) has also defined a metadata scheme for 'self-documenting' records, published in April 2000. This scheme has been designed to be compatible with the Recordkeeping Metadata Standard developed by the National Archives of Australia, despite being based on a different conceptual model.

Australian Recordkeeping Metadata Schema: <http://rcrg.dstc.edu.au/research/spirt/>

NSW Recordkeeping Metadata Standard:

<http://www.records.nsw.gov.au/publicsector/erk/metadata/rkmetadata.htm>

Victorian Electronic Records Strategy: <http://www.prov.vic.gov.au/vers/>

OAI-related initiatives

Another line of work centres on the technical framework being developed by the Open Archives Initiative (OAI). Version 1.1 of the OAI Protocol for Metadata Harvesting was published on 2 July 2001. The US Digital Library Federation (DLF) is now encouraging the use of the OAI. The DLF, in co-operation with the Andrew W. Mellon Foundation, are supporting the development of testbed gateways that will use OAI harvesting techniques to allow users to access distributed digital library holdings as if they were part of a single collection. A wide range of domains and organisation types has an interest in the use of the OAI protocol. In August 2001, the list of registered OAI data providers included e-print archives, subject gateways, libraries, text archives and many other kinds of organisation. One example of its proposed use is by the UK Resource Discovery Network (RDN) who have developed an experimental OAI protocol-based system for record sharing between RDN subject gateways.

Herbert Van de Sompel & Carl Lagoze, *The Open Archives Initiative Protocol for Metadata Harvesting*, v. 1.1, Open Archives Initiative, 2 July 2001. <http://www.openarchives.org/OAI/openarchivesprotocol.htm>

Digital Library Federation evaluation of the Open Archives Initiative: <http://www.diglib.org/architectures/testbed.htm>

Andy Powell, *An OAI approach to sharing subject gateway content*. Poster presented at WWW10: the Tenth International World Wide Web Conference, Hong Kong, 1-5 May 2001. <http://www10.org/cdrom/posters/1097.pdf>

The museum domain is also investigating the use the OAI technical framework. The CIMI Consortium, for example, participated as a pre-release tester of the OAI protocol. The consortium is aware that much museum data is "hidden" away from Web search engines in databases, exhibition catalogues, research papers, etc. Perkins suggests that the OAI technical framework might help provide an enabling technology to facilitate the federating of distributed information services and their discovery and use. CIMI has developed an OAI-based Metadata Harvesting project to investigate the potential of OAI for museums and to provide a testbed implementation.

CIMI Metadata Harvesting Project: <http://www.cimi.org/wg/metadata/>

John Perkins, *A new way of making cultural information resources visible on the Web: museums and the Open Archives Initiative*. Paper presented at Museums and the Web 2001, Seattle, WA., 14-17 March 2001. <http://www.archimuse.com/mw2001/papers/perkins/perkins.html>

Within the library domain, a team based at the Virginia Tech Digital Library Research Laboratory (DLRL) has recently undertaken to create an XML Schema to help support the wider distribution of MARC records in OAI contexts. The Virginia Tech team has defined an XML transport format for MARC so that MARC (presumably MARC21 or USMARC) records can be exchanged using the OAI protocol. This is not the first project to investigate the conversion of MARC structured records into XML or SGML-based formats, but it is the only one with a specific focus on the OAI. Related (but older) MARC-XML projects include the XMLMARC DTD developed by the MEDLANE project at Stanford University Medical Centre and the SGML and XML DTDs created by the Library of Congress MARC DTD project.

Virginia Tech MARCXML initiative:

<http://www.dlib.vt.edu/projects/OAi/marcxml/marcxml.html>

Stanford University MEDLANE project: <http://xmlmarc.stanford.edu/>

Library of Congress MARC DTD project: <http://lcweb.loc.gov/marc/marcsgml.html>

Museum developments

The CIMI Consortium are involved in the development by mda (formerly the Museum Documentation Association) of an XML Document Type Description (DTD) for SPECTRUM, an international standard for the description of museum objects. CIMI are setting up a SPECTRUM XML-DTD Testbed project that aims to demonstrate the DTD's potential for data migration and exchange, for the integration of systems and the reuse of content.

CIMI SPECTRUM XML-DTD Testbed: http://www.cimi.org/wg/xml_spectrum/

SPECTRUM: <http://www.mda.org.uk/spectrum.htm>

Bart Degenhart Drenth, *Building on the mda SPECTRUM-XML DTD for collections management data interchange*. Paper presented at Museums and the Web 2001, Seattle, WA., 14-17 March 2001. <http://www.archimuse.com/mw2001/papers/degenhart/degenhart.html>

Another museum-based development that may have relevance for metadata is a new European Museums' Information Institute (EMII) project that will be funded as part of the European Commission's Information Society Technologies (IST) Programme. At the present time (August 2001), there is little information available about this project, which is called the EMII distributed content framework (EMII-dcf). A preliminary description on the mda Web pages says that EMII-dcf will establish a working model for the provision of various types of content (text, images, film, video, etc.) from various sources (museums, broadcasters, archives, libraries, etc.) within EC funded research projects.

EMII-dcf: <http://www.mda.org.uk/200106c.htm>

Collection description developments

Another recent initiative is the UK's Collection Description Focus. This is a 12-month project that will build on work on collection-level description previously undertaken as part of the Research Libraries Support Programme (RSLP) and the Electronic Libraries Programme (eLib) of the Joint Information Systems Committee (JISC). Johnston and Robinson (2001) say that the Focus "will look into the development of simple tools to facilitate the creation and management of collection descriptions". These will include "tools to support transformations between different collection description schemas; enhancements to data creation interfaces; the use of a schema registry to publish and share CD schemas; and mechanisms for 'harvesting' distributed descriptions". The Collection Description Focus is jointly funded by the JISC Distributed National Electronic Resource (DNER), the RSLP and the British Library.

Collection Description Focus: <http://www.ukoln.ac.uk/cd-focus/>

Pete Johnston & Bridget Robinson, Collection Description Focus. *D-Lib Magazine*, 7(7/8), July 2001. <http://www.dlib.org/dlib/july01/07inbrief.html#ROBINSON>

APPENDIX B: Domain report: educational sector

Correspondent: Erik Duval, Univ. Leuven

1. Introduction

This report focuses on implementation activities, with a particular emphasis on multilingualism and controlled vocabularies.

2. Standardization work

2.1 IEEE LTSC LOM

Within the IEEE LTSC Learning Object Metadata (LOM) specification, *controlled vocabularies* are defined for 19 elements.

Central to support for *multilingualism* in LOM is the notion of LangString: when element values are defined as free text, lingual variants of the values can be provided in a number of human languages.

2.2 CEN/CENELEC LTWS

The CEN/CENELEC ISSS Learning Technologies Workshop includes both multilingualism and controlled vocabularies as major work items. With respect to *multilingualism*, three project teams are working:

- to ensure that the IEEE LOM, as the globally accepted solution, is capable of addressing specific European cultural requirements (such as multilingualism).
- on standardization actions to permit the identification of alternative versions of resources, in different languages, as well as the origin of the translation. This is taking place within a LOM context.
- to 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 from the Workshop's web pages. These will be replaced in due time by updated and widely accepted revised versions.

On *controlled vocabularies*, a project team is collecting and organizing a register of taxonomies and repositories relevant to European learning, via an on-line repository. 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. Focus is 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.

3. Consortia based work

The consortia mentioned below, some of which contribute to the development of standards, adapt these standards to the needs of their constituencies, a process referred to as 'profiling' in the standards world.

3.1 ARIADNE

The *ARIADNE* Foundation develops and exploits the Knowledge Pool System, a distributed database of reusable learning components, with associated metadata that describe them.

ARIADNE has developed *controlled vocabularies* of science types, disciplines and sub-disciplines: this is a hierarchical structure of 2*10*10 entries, used to indicate the semantics of a learning object. The ARIADNE tools draw heavily on the LTSC LOM vocabularies. In some cases, the ARIADNE application profile restricts the value space to a subset of the LOM value space.

The latest generation of the indexation and query tools (based on the recently balloted LOM v6.1) supports *multilingualism* in a number of ways.

- The language of the user interface can be selected from seven languages.
- Controlled vocabularies have also been translated in all these languages, so that, for the elements with controlled vocabularies, the selected or retrieved values are also translated.
- For all elements with free text values, the user can insert or search for a number of lingual variants, with an indication of the language(s) being used, as per the LangString construct in LOM.

3.2 IMS

The *IMS* consortium re-synchronized its metadata specification in May with the recently balloted LOM v6.1: the IMS information model v1.2 includes the LOM specification.

The 'Best Practice and Implementation Guide' includes suggestions of 37 schemes and *controlled vocabularies* for 13 elements. When no single vocabulary is judged to be dominant, multiple suggestions, with their origin and suggested applicability, are listed. Schemes are listed for elements such as language (RFC1766, ABS1267, ISO639, ISO3166 and Z39.53). Vocabularies are listed for elements such as classification (LOC, LCSH, DDC, UDC, CIP, SCIS, GEM, etc.)

No mention is made of *multilingualism* in the guide.

3.3 ADL & SCORM

The ADL Sharable Course Object Reference Model (SCORM) profiles LOM for raw media, content and courses.

A 'curricular taxonomy' is defined as the *controlled vocabulary* for the content structure hierarchy, mapping the U.S. Army, Navy and Marine Corps, as well as the Canadian designs to raw media, content and courses. For 8 CMI (Computer Managed Instruction) API data elements, vocabularies are defined. These are lists of 2 to 8 values.

SCORM does not address *multilingualism* explicitly.

3.4 EDNA

EDNA relies on DC with 8 additional elements (Audience, Approver, CategoryCode, Entered, Indexing, Review, Reviewer and Version).

For certain elements, schemes are specified: for DC.Subject for instance, reference is made to LCC, LCSH, APSDEP, SCIS, ASCED, DDC, UDC, MeSH and edna-KLA. For other elements, *controlled vocabularies* are listed, such as for DC.Type, where vocabularies are defined for document types, curriculum types and event types. Similar vocabularies are defined for audience, sector, user level and coverage.

With respect to *multilingualism*, the only explicit reference is to use the Alternative refinement for the Title element to include translations.

3.5 EUN

The EUN specifies two metadata element sets. For learning objects, it relies on Dublin Core, with additional elements for rights, approver, release, user level and version. For collections, the specification is based on RSLP and Renardus.

In the 'European Treasury Browser', an extensive *controlled vocabulary* has been developed that refers to individual development, learning and research, school activities, leisure activities, teaching and training and evaluation and guidance, the educational system, subject, facilities, communication, culture, political and social aspects, health, environment, society, international organisations, geopolitical areas and languages. Other vocabularies are provided for the user level, type. For pedagogy, the GEM vocabulary is used.

In order to support *multilingualism*, the EUN specifies for which elements language specification is necessary. Moreover, the subject vocabulary is available in English, French, German, Italian and Spanish. Swedish, Danish and Greek versions are under development.

3.6 Gateway to Educational Materials

The Gateway to Educational Materials (GEM) extends the Dublin Core element set, with 8 additional elements.

Controlled vocabularies are defined for format (17 values, based on Internet MIME types), grade (8 values) and resource type (29 values, incl. Course, event and tool). More elaborate vocabularies are specified for

- Audience: 18 values to indicate for whom the tool is intended and 46 for the ultimate beneficiary;
- Pedagogy: This is clearly the most elaborate GEM vocabulary: three categories of pedagogy are distinguished:
 - teaching methods: 37 values detail the type of educational activities, including for example cooperative learning and role playing;
 - grouping: 7 values, ranging from individualized to large-group instruction;
 - assessment: 10 values, including peer evaluation and portfolio evaluation.
- Relation: 20 relation types (incl. IsParentOf and isCriticalReviewOf);
- Subject: a hierarchical structure with circa 15 top-level elements and around 15 elements for each top-level element at the second level.

There is no explicit reference to *multilingualism* in the GEM specifications. Interesting to note is that the controlled vocabulary for the language elements includes 5 values (only).

4. Projects and Research Activities

Many projects include metadata tools and infrastructures in their R&D efforts. The first results of this research are beginning to appear in publications. As an example, the recent ED-MEDIA conference included about 30 papers on applications of LOM and 10 that refer to Dublin Core [Vitelli and Montgomerie, 2001]. Publications that focus exclusively on educational metadata applications (both LOM and DC-Education based) include [Greenberg, 2000] and [Duval & Robson, 2001].

Often, the conceptual remarks based on the experiences of the authors are quite relevant. Sometimes however, the comments point to 'features rather than bugs', in the sense that the original developers of the metadata specification had anticipated some of the problems. In those cases, the typical consensus in the standardization bodies was that it was better to have a 'good enough' solution now, rather than a 'perfect' solution in an all too distant future.

Many of the metadata tools seem somewhat immature: they typically start from the LOM specification, rather than from the user characteristics and tasks. Hence, the usability of many of these tools is rather poor. It can be expected that this situation will improve: whereas the current generation of tools is heavily focused on demonstrating how the technology can work, (we hope that) future tools will concentrate more on helping the user to get a job done.

5. References

[Ariadne, 2001] <http://www.ariadne-eu.org/>

[Duval, 2001] Erik Duval. *Standardized Metadata for Education: a Status Report*, Proceedings of ED-Media 2001: World Conference on Educational Multimedia, Hypermedia and Telecommunications, Tampere, Finland, pp. 458-463, June 25-30, 2001.

[Duval & Robson, 2001] E. Duval and R. Robson (eds.). Special edition on Metadata, Journal on Interactive Learning Environments, 2001.

[Greenberg, 2000] J. Greenberg (ed.). *Metadata and Organizing Educational Resources on the Internet*. Journal of Internet Cataloging, Vol. 3, No. 2/3, 2000.

[Vitelli and Montgomerie, 2001] J. Vitelli and C. Montgomerie (eds). *Proceedings of ED-Media 2001: World Conference on Educational Multimedia, Hypermedia and Telecommunications*, Tampere, Finland, June 25-30, 2001.

APPENDIX C: Domain report: publishing sector

Correspondent: Laurie Causton, Clearbay Limited

Current state of domain

The IPTC's (www.iptc.org) 2001 Spring Meeting was the first full session since the formal adoption of NewsML (version 1.0), and related metadata issues formed a large part of the discussions.

While there was no call for changes to NewsML metadata, a number of issues were seen to need further work - the handling of updates; digital rights; transport mechanisms; metadata structuring standards; standardising physical metadata; elimination of indirection; and formulating a policy on extensibility. Moreover, there was a need to ensure continuing metadata compatibility between NewsML, NITF and the IIM, and it was agreed that there should be co-operation with other organisations involved in the development and application of metadata.

January saw a number of proposals for changes and additions to the IPTC's NITF (News Industry Text Format) (www.nitf.org). Overall, the IPTC has been taking a hard look at overlaps and the need for collaboration (more on this in the next section).

CISAC (www.cisac.org) has been making progress in the implementation of CIS (Common Information System), deciding on the creation of a new authority within CISAC, the "CIS Supervisory Board" to serve as the administrative council for the CIS. Full details are being worked out, but the objective is to simplify and streamline the administration of CIS, with CISAC itself becoming more of an agency managing the technical standards.

Santa Fe is dead; long live OAI - the Santa Fe Convention of the Open Archives Initiative has been discontinued. Attention is now focussed on the Open Archives Initiative Protocol for Metadata Harvesting (www.openarchives.org/OAI/openarchivesprotocol.htm), designed to offer an application-independent interoperability framework for use by communities engaged in publishing content on the Web.

In April the PRISM Working Group (Publishing Requirements for Industry Standard Metadata, www.prismstandard.org) announced the release of version 1.0 of the PRISM metadata specification, providing a metadata vocabulary for print and online publishing.

PRISM is a good example of the collaboration and the moves to eliminate redundancies and overlap that can be seen these days – more than 20 diverse organisations were involved in the specification, and it builds on existing standards such as the Dublin Core and RDF. Moreover, the Working Group is developing style sheets to make PRISM metadata interoperate with complementary standards such as NITF and NewsML.

In November 2000 the Association of American Publishers' (www.publishers.org) Open Ebook Publishing Standards Initiative (www.doi.org/ebooks/doi-eb.html) had recommended the adoption of Digital Object Identifiers (DOIs) as the primary identification system for managing the metadata associated with the development of eBooks.

The IDF itself has been active in this area, with the first international DOI-EB (DOI for E-books) meeting held early this year, following the formation of the DOI-EB Working Group.

Metadata seems to be enjoying increased attention in the IDF, a point underlined perhaps by its recently appointed

Business Development Director, Stephen Mooney: "The persistent identification of intellectual property entities combined with interoperable metadata is the key to effective and efficient commercial rights transactions. DOI is focused on exactly this space..."

The issue of intellectual property entities for multimedia is in fact the basis of another recent initiative in the IDF, a study into a Rights Data Dictionary, which will be based on the <indecs> framework (www.indecs.org) and which will be known as <indecs>2. The DOI's namespace for metadata management will be donated to the consortium developing the RDD.

There was a good deal of discussion of metadata Application Profiles in the last Metadata Watch report, and now the IDF has come up with its own (a DOI-AP), described as "the functional specification of an application (or set of applications) of the DOI System to a class of intellectual property entities that share a common set of attributes" (<http://www.doi.org/doi-ap.html>). A DOI-AP is intended to enable implementation of an application (or set of related applications) in a particular environment, ranging from relatively simple discovery to complex e-commerce and rights management applications. It is partly defined in terms of a metadata schema that is always a superset of the DOI Kernel metadata schema. All DOI metadata schemas will be based on the DOI Namespace (DOI-NS), a project that is currently in hand to support the development of DOI-APs. This means that all DOI metadata will be fully compliant with the <indecs> analysis.

And further to eBook developments, the EBX Working Group (www.ebxwg.org) formally combined with the Open eBook Foundation (www.oebf.org) in March, with a plan to concentrate their efforts toward the efficient development and widespread adoption of electronic publishing standards. Activity now seems to be concentrated in the OeBF, judging by the respective web sites – that of the EBX WG is now static.

The OeBF has now prepared its Requirements Portal to gather participants' requirements for all aspects of e-publishing. The idea is that the requirements thus collected can be debated and considered for incorporation into the e-publishing standards to be developed by the OeBF.

Lastly, in February Final Draft Standard 15707 for the International Standard Musical Work Code (ISWC) was distributed to ISO's member bodies for voting and approval, to be completed in May.

Overlaps and gaps identification

Recent months have seen less of overlaps and gaps arising, but rather more activity in promotion of joint use and collaboration to offset or minimise overlaps.

On the NewsML front, a presentation on NewsML has been made to the SMPTE Metadata Committee, and a return presentation on the SMPTE metadata work was made during the IPTC Spring Meeting.

The PRISM specification, now at the Working Draft stage, are looking at issues involved in joint use of their specification and NewsML before finalising the specification. Although the aims of both sides are generally complementary, there has been some overlap between the standards, and a NewsML report has proposed ways of achieving compatibility.

The nature of the content carried by NewsML will influence the use and development of the standard, and this content can be in many different forms. A survey found more than four hundred standards (generally XML-based) being developed, or already in use, for specialised industries. There is already overlap between many of these initiatives, but co-operation is happening in a number of areas.

In general terms, the IPTC Spring Meeting called for co-operation with other organisations involved in metadata development and there is various other evidence, noted above, of the move towards greater co-operation or combination of effort: the “merger” of the OeBF and the EBX Working Group (there is already the OeBF’s standards co-ordination initiative, as described in the last report); the PRISM Working Group’s aim to make PRISM metadata interoperate with complementary standards; and the IDF’s collaboration with various bodies on <indecs>2.

Trends

One trend certainly seems to be greater collaboration, discussed in the previous section. Part of this might be attributable to a growing recognition of the importance (or inevitability?) of certain key initiatives - those which have become, or are becoming, a stable feature of the metadata landscape in publishing – such as, perhaps, Dublin Core and the DOI. This may arise from a *de facto* acceptance of their applicability and utility, or simply from their level of adoption by the industry.

The International DOI Foundation themselves recognise the need for collaboration and awareness, citing activities in, for example, the OeBF, the World Wide Web Consortium, the Internet Engineering Task Force (IETF) and MPEG as a motivation for their funding of a “content industry wide” Rights Data Dictionary.

This is a trend which should continue. Many metadata ventures understandably arise out of the needs of a certain domain, such as publishing and, at least initially, focus on those particular needs. But that scope can broaden and boundaries start to blur. Publishing needs to deal with audio-visual content as much as textual, but the audio-visual sector has its own metadata initiatives. There is already movement towards broader collaboration, such as <indecs>2 which is aimed at multimedia.

Another example is the moves made by the Tribune group, a US-based multimedia company covering newspapers and other publications, TV, cable news, and radio, as well as Internet news and information services. They want an economical and practical content sharing system for multimedia, and they see a solution in some form of single repository with a searchable index and common metadata. They have analysed how key areas of metadata would map between standards - difficult because standards have been developed to meet different needs as noted above, giving different metadata structures – and conclude that converging these standards would be impracticable. This raises the question of how to use the metadata assets effectively,

and as a first step they are proposing that the IPTC and SMPTE (Society of Motion Picture and Television Engineers – www.smpte.org) should get together to develop a common Media Independent Protocol (MIP), with each group defining the mapping of its metadata to the protocol. Thus far, the proposal has been favourably received by both the IPTC and the SMPTE, and work is continuing.

Perhaps another trend may be the development of more advanced approaches to metadata construction. Much of content description can be based on keywords, but these have their limitations – they can be ambiguous or over-specific. The Machine Understanding Group at the MIT Media Laboratory has a long-running “The News in the Future” (NIF) programme (nif.www.media.mit.edu); within that, there is research into ways of describing content using disambiguated concepts. A Structured Controlled Vocabulary (SCV) called BRICO controls the relationships between terms, broadly in the manner of a thesaurus, but not the terms themselves. Past and current sponsors of NIF include a number of organisations from the publishing and broadcasting sectors. As an example of the use of this idea, the existing keyword system in NITF could be complemented with suitably developed concept references.

Main issues

Activity in the area of e-books continues to increase, and the AAP, IDF and OeBF are all working hard in this area. However, while much of this work may be essential to a viable e-book market, that market may take a while to happen.

Gartner Group sees e-books as in their early stages, despite that vendors “have been trying to market these things for some time.” Device sales have been low, with Jupiter Media Metrix estimating less than 50,000 e-book hardware devices in use in the United States, and attributing the low number to lack of content and high prices; and they see an e-book audience of only just 1.9 million by the end of 2005. Jupiter’s Robert Hertzberg commented “Reading an e-book is just like reading a book ... but it’s just less fun, more expensive and heavier. That’s not much of a marketing motto.”

Lack of a single software standard or hardware platform is also seen as a contributing factor, and amongst the standards issues are those of copyright, distribution, and security, which is where metadata has a role.

Forrester Research agrees, predicting slow growth. They estimate digital delivery of custom-printed books, textbooks, and e-books to account for total revenues of \$7.8 billion by 2005, around 17.5 percent of publishing industry revenues, but only \$251 million will come from e-books and the necessary devices.

These industry observers generally all see technical and educational books as being the first which may gain acceptance, since their audience is more attuned to digital delivery. The industry appears to concur – in January, netLibrary, Inc and Houghton Mifflin announced plans to launch a digital textbook initiative, and February saw McGraw-Hill and the American Society of Mechanical Engineers International (ASME) forming alliances with technology vendors to improve digital delivery to specialised audiences, believing that education and professional services were the most immediate prospects for e-publishing.

Not all pundits agree with this gloomy forecast. IDC sees demand for e-books building slowly in 2001, then exploding in 2002, with the US market growing from \$9 million in 2000 to \$414 million in 2004. But note that, as recorded in the last Metadata Watch report, there had been estimates of \$12million sales in downloaded

books in 1999. And a recent Seybold survey showed a lukewarm attitude in North America towards reading electronic content and even less of a commitment to spending money for it. In fact, two-thirds of all respondents were "not at all likely" to purchase an e-book or dedicated device in the next year.

So, on one hand we have the publishing industry and a good number of technology vendors committing themselves heavily in this market, on the other we have an audience which is showing some evidence of indifference and little evidence of imminent growth. The standards work is necessary, and metadata initiatives will have a key role in enabling e-book production and commerce, but it does rather look like that they will not need to hurry.

Multilingualism

The publishing sector has historically seen more of a focus on trading and rights management, and therefore on developing the enabling mechanisms for business use. Accordingly, multilingual aspects have perhaps taken a back seat.

This is not to say that language is totally ignored. Clearly, in describing content, the source language of that content is a valuable descriptive element, and is commonly found in publishing metadata schemes. Indeed, various initiatives have adopted Dublin Core, and hence employ its language element. As an example, EBX – focussed on trading and distribution of e-books, employs Dublin Core for what it calls its ‘concise metadata’, while suggesting schemes such as ONIX for ‘extended’ metadata, but within its specification there is no mention of language or multilingual aspects. And in any event, employing a language descriptor is not the same as providing for multilingualism.

Moreover, with the emphasis on business processing, typically within an XML structure, the formal elements cannot admit linguistic variation; they must be consistently presented to comply with the protocol. For example, PRISM uses Dublin Core, and also describes its own elements Dublin Core-style: by way of an Identifier, a formal XML element type (a protocol element) which must be expressed as-is, such as `prism:distributor`; and a name, such as Distributor, which can be expressed in any language.

Controlled vocabularies

Controlled vocabularies occur to a limited extent in publishing metadata. Certainly, given the common adoption of Dublin Core, publishing metadata can follow any controlled vocabulary aspects of that initiative. Certain others offer vocabularies – NewsML proposes a controlled vocabulary for its Topic Types for example.

PRISM is notable here – indeed, its aim is the development of a standard XML metadata vocabulary. As the PRISM people say: “But while XML specifies how things can be encoded for exchange, it does not specify what information must be exchanged. Therefore, the publishing industry needs standard vocabularies such as PRISM to realise the potential of e-commerce in online publishing.” It defines sets of controlled vocabularies, for example for resource types and categories.

Moreover, although these are early days yet, the International DOI Foundation (IDF) is funding the feasibility study for the development of a Rights Data Dictionary (RDD) – “a common dictionary or vocabulary for intellectual property rights.”

Even so, one might expect more in this sector in the development of controlled vocabularies, since the e-business bias should encourage descriptive precision, but there is not much evidence found at present – the priorities for the moment are different perhaps.