

# **SCHEMAS**

**Contract: N° IST-1999-10100**

**Forum for Metadata Schema Implementers**

**METADATA WATCH REPORT #1**

**D22**

**Document number:  
SCHEMAS-PwC-WP2-D22- Final-20000602**

## General Information

<b>Title</b>	Metadata Watch Report #1
<b>Creator</b>	PwC
<b>Subject-Keywords</b>	Deliverable D22; WP2; Metadata Watch Report #1
<b>Description</b>	This document comprises the introduction and top-level synthesis for D22 Metadata Watch Report #1 plus the domain reports
<b>Publisher</b>	PricewaterhouseCoopers
<b>Contributor</b>	Project team
<b>Date</b>	06 June 2000
<b>Type</b>	Text Manuscript
<b>Format</b>	application/msword
<b>Identifier-URL</b>	www.schemas-forum.org/folder/filename
<b>Identifier-Document Number</b>	SCHEMAS-PwC-WP2-D22-Final-22000602
<b>Language</b>	English
<b>Rights</b>	European Commission; Internal circulation within project; External circulation via SCHEMAS Web site

```
<META NAME="DC.Title" CONTENT="Metadata Watch Report #1">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#title">

<META NAME="DC.Creator" CONTENT="PricewaterhouseCoopers ">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#creator">

<META NAME="DC.Creator.Address" CONTENT="PwC MCS Luxembourg">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#creator">

<META NAME="DC.Subject" CONTENT="Deliverable D22">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#subject">

<META NAME="DC.Subject" CONTENT="WP2">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#subject">

<META NAME="DC.Subject" CONTENT="Metadata Watch Report #1">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#subject">

<META NAME="DC.Subject" CONTENT="Metadata Watch Report #1">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#subject">

<META NAME="DC.Description" CONTENT="This document comprises the introduction and top-
level synthesis for D22 Metadata Watch Report #1 plus the domain reports">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#description">

<META NAME="DC.Publisher" CONTENT="PricewaterhouseCoopers">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#publisher">

<META NAME="DC.Date" CONTENT="(SCHEME=ISO8601) 2000-06-02">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#date">

<META NAME="DC.Type" CONTENT="Text.Manuscript">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#type">

<META NAME="DC.Format" CONTENT="(SCHEME=IMT) application/msword">
```

```
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#format">
<LINK REL=SCHEMA.imt HREF="http://sunsite.auc.dk/RFC/rfc/rfc2046.html">

<META NAME="DC.Identifier" CONTENT="http://www.schemas-forum.org/folder/filename">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#identifier">

<META NAME="DC.Identifier" CONTENT="(SCHEME=URN) SCHEMAS-PwC-WP2-D22-Final-
2000602">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#identifier">

<META NAME="DC.Language" CONTENT="(SCHEME=ISO639-1) en">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#language">

<META NAME="DC.Rights" CONTENT="European Commission; Internal circulation within project;
External circulation via SCHEMAS Web site">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#rights">

<META NAME="DC.Date.X-MetadataLastModified" CONTENT="(SCHEME=ISO8601) 2000-06-
02">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#date">
```

# Table of Contents

<b>1. INTRODUCTION</b> .....	<b>5</b>
<b>2. TOP-LEVEL SYNTHESIS</b> .....	<b>6</b>
2.1 COMMON THEMES .....	6
2.2 OVERLAP .....	6
2.3 GAPS .....	7
2.4 CONTENT IS CONTENT – OR IS IT?.....	7
2.5 CONVERGENCE AND CROSS-FUNCTIONALITY .....	8
2.6 CO-OPERATION.....	8
2.7 GEOGRAPHY AND LANGUAGE.....	9
2.8 “BRIGHT LIGHTS” AND THE RELATIVE LEVEL OF DEVELOPMENT OF THE SECTORS.....	9
2.9 WEB VS. NON-WEB .....	9
2.10 AUTOMATION.....	10
<b>3. DOMAIN REPORTS</b> .....	<b>11</b>
3.1 INDUSTRIAL SECTOR .....	11
3.2 PUBLISHING SECTOR .....	12
3.3 AUDIO-VISUAL SECTOR .....	17
3.4 EDUCATIONAL SECTOR.....	21
3.5 ACADEMIC SECTOR .....	24
3.6 RESEARCH SECTOR .....	26
3.7 GEOGRAPHICAL INFORMATION SECTOR.....	29
3.8 OTHER SECTORS .....	31

# 1. Introduction

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 things. The added value that the MD Watch provides consists of giving readers (a) the ability to get the information they need from one easy-to-use source, (b) expert opinion and (c) a multi-tiered format that allows readers to get information at three levels of granularity.

The three levels consist of the following:

- Top level: An overview of key issues in the metadata field;
- Middle level: Individual reports, by sector, of work in the metadata field; and
- Bottom level: Reports of individual activities in the metadata field.

The middle level is comprised of the following sectors:

- Academia
- Audio-Visual
- Cultural Heritage
- Education and training
- Geographic information
- Industry
- Publishing
- Research, i.e. non-commercial laboratories, corporate research, and professional societies
- Other, which includes, but is not limited to, environmental work, government (including military), health care, mail and delivery, and transport and logistics.

SCHEMAS examined but did not make full MD Watch activity reports for activities that simply announce that they use metadata. Rather, we focused on projects and services that make a point of publishing their schemas, participate in standards-making activities, or otherwise promote metadata, whether with software tools or with working groups. These projects could be of any type - commercial, not-for-profit, government, subsidised, military, or a combination of the above. SCHEMAS also tried to separate the hype from the reality, i.e. announcements of a particular activity or product that were not accompanied by a usable product, e.g. standard, registry, or tool, were not included in the individual MD Watch activity reports. However, they were recorded and will be monitored for inclusion in each succeeding quarterly MD Watch Report.

## 2. Top-Level Synthesis

This top-level synthesis of the SCHEMAS sectoral Metadata Watch (MD Watch) reports provides an overview of the findings of the first quarterly MD Watch report. It highlights the common threads running through the world-wide metadata scene as well as key differences among the sectors.

The sectors covered in the sectoral MD Watch reports are as follows:

- Academia
- Audio-Visual
- Cultural Heritage
- Education and training
- Geographic information
- Industry
- Publishing
- Research, i.e. non-commercial laboratories, corporate research, and professional societies
- Other, which includes, but is not limited to, environmental work, government (including military), health care, mail and delivery, and transport and logistics.

### 2.1 Common Themes

The themes common to the metadata-related activities in the above sectors are similar to those encountered throughout the field of information and communication technology, and include:

- Interoperability and standardisation;
- Human- and machine-readability;
- Security;
- Provisions for intellectual property issues;
- Simplicity yet broad applicability of core metadata sets;
- The development of unique identifiers to link content with related documentation through the process of creation, delivery, use and re-use;
- The semi- and fully automatic creation of embedded (wrapped) metadata;
- Automated indexing of content and meaningful subsets of content;
- Version management;
- The development of methods of extraction, searching, evaluation and validation to assist content queries; and
- Multilinguality.

### 2.2 Overlap

Clearly, from reading the list of sectors (above), there is significant scope for overlap. For example, if a commercial firm creates an education and training package that specifically makes use of metadata and schemas in a novel way, does that fall into the

Industry category or the Education and Training category? If an audio-visual product is comprised of cultural content, does that fall into Audio-Visual or Cultural Heritage? To solve this problem, the SCHEMAS correspondents allocated in advance the various activities they were to cover. In those cases in which it was unclear in which category an activity should be placed, a judgement was made based on a close examination of the activity.

### **2.3 Gaps**

A number of sub-sectors, such as transportation and logistics, did not appear to be sources of activity in the metadata field. That, in itself, is information. Taking the case of transportation and logistics, it could be that (a) there *are* metadata-related activities taking place, but they are not easily locatable, and may appear in the next quarterly MD Watch report, or (b) players in the transportation and logistics field are participating in metadata-related activities that fall into other categories.

### **2.4 Content is Content – or is it?**

Content is content, and nobody denies that content needs metadata to be truly useable, especially with the current proliferation of digital content. However, different sectors deal with different types and amounts of content, and this is one reason for the divergence – both qualitative and quantitative – in the activities across the various sectors.

For example, the Industry sector is populated by activities focusing on business information systems for commerce, especially business-to-business commerce on the Web, i.e. the supply chain. Schemas for purchasing, bills of lading, and invoicing abound in this sector. Also quite common are schemas related to the internal business processes that any firm undertakes – human resources, employee records, and salary administration, for example. The above can all be represented perfectly adequately in text (alphanumeric) format and, as one would expect, activities in the Industry sector focus primarily on the description of textual information.

On the other hand, the Audio-Visual sector is quite different. Here we are dealing with staggeringly large amounts of information, only a portion of which is textual. In addition, the means by which this content is distributed varies greatly, e.g. terrestrial broadcast, Internet, and CD- or DVD-ROM. Furthermore, while the technical infrastructure to search, sort, and tag alphanumeric content has been around for decades, the techniques for searching *directly* on “multimedia” content, e.g. sound, still images and moving images, are non-existent or nascent at best.

So the two sectors have quite different problems to solve based on the type of content and the media through which that content is disseminated. Other differences exist as well, as the issues of rights and security can illustrate. The question of who owns the intellectual property rights to an invoice used in business-to-business commerce is not a very important one, but security is extremely important. When considering the online *payment* of that invoice, security becomes even more important. In the Audio-Visual sector, however, the issue of rights is probably of the highest importance – for example, the potential damage caused from a single broadcast programme being

“lost” or illegally intercepted pales in comparison to the potential damage caused from a single business communication being lost or illegally intercepted.

## ***2.5 Convergence and Cross-Functionality***

The emerging problem (or opportunity) for all metadata specifications of all types, especially those in the Publishing, Industry, and Audio-Visual sectors, is that they tend to end up covering more or less the same ground. The definitions of product and market are becoming hazy in the world of “physical” product. Book publishers release audio and video materials. DVDs include audio, text, visuals, audiovisual. Serials, magazines, news all come now in all media types. The conventional divisions neatly represented by physical content types and their identifiers do not apply to metadata schemes which must increasingly embrace all forms of creations. The fact that one sector is more biased towards “text” or “visual”, one more towards “audio” and another more “audiovisual” stuff is not much of a useful distinction when it comes to designing metadata systems in which all must be well described irrespective of their predominance or otherwise, and pretending a recording is a kind of book, or vice versa, as corporate and library systems once did, is no longer adequate.

In addition, metadata is becoming multifunctional. For example, all of the major record companies are currently engaged in establishing their own corporate international databases. It is a reasonable assumption that these systems will be designed to support in due course the requirements of marketing, “label copy” for product packaging, Web metadata, rights and royalty management, and sales, as well as incorporating business rules, and the data in them will be derived in part directly from production workflow systems.

## ***2.6 Co-operation***

The various sectors differ not only in the level of activity going on within them but also in the level of co-operation shown by those activities. The Geographic Information sector shows relatively low levels of co-operation across activities, though it does show a high level of organisation in the sense of division of labour, i.e. the various activities are cleanly divided along traditional GI lines – geospatial data, hydrographic data, geological data, etc.

By comparison, the Research sector shows a high degree of co-operation among activities, possibly due to the well-established tradition of international co-operation that has characterised scientific endeavour this century, with activities often building on the results of other activities and ensuring that various concurrently-developed projects remain compatible.

Co-operation also depends on the sometimes financially-motivated politics that characterise a particular sector. Where commercial gain can be affected by the outcome of standards activities, such as are described in the individual MD Watch Activity Reports, openness and co-operation is often the first victim. In the Industry and Audio-Visual sectors, for example, players must reconcile conflicting motivations

– co-operation can benefit all players and can create commercial opportunities where none had previously existed, but it also runs counter to industrial notions of product differentiation and secrecy. In the Academia, Research, Cultural Heritage and Geographic Information sectors, on the other hand, commercial gain is less of a factor, and therefore less of a barrier to co-operation.

## **2.7 Geography and Language**

Despite talk of globalisation, including the “globalising” effects of the Internet, geography and language still matter. In fact, geography and language are growing in importance as barriers as content originating from increasingly diverse locations comes into contact with content consumers from increasingly diverse locations. Metadata sets, for example, are still only translated into a handful of languages, if at all. Registry sites are often only in English. And the vast majority of metadata-related activities take place in the same two dozen or so countries, i.e. the countries comprising Europe and North America, Australia, Japan, and a handful of others. But co-operation in metadata-related activities, when it exists, is strong among these countries, and the need for multilinguality is widely recognised, if not widely yet acted upon.

The information industry is by now accustomed to US domination in a variety of areas such as Internet services, personal computers, software and operating systems, and servers, to name a few, and American activities in some sectors do tend to focus on American players and their needs. However, this is often in the context of US government or military activities, which, in those cases, is to be expected.

## **2.8 “Bright Lights” and the Relative Level of Development of the Sectors**

A few metadata-related activities stand out. Resource Description Framework (RDF) and Dublin Core (DC) are two of the most notable. They are widely accepted from a philosophical standpoint, enjoy the support of a number of activities, whether implicit or explicit, and co-operate with each other. The Publishing and Audio-Visual sectors have their own short lists of well-organised initiatives that, taken together, cover large sections of their respective fields, and the Education domain has four dominant metadata initiatives. The other domains have fewer, if any, “bright lights” leading the way, culminating with the Industry sector which is characterised by inward-looking activities focusing on a small sub-sector of Industry and even, sometimes, competing activities that seem to take no notice of each other.

## **2.9 Web vs. Non-Web**

Metadata-related activities can be divided into two classes: “old school”, or pre-Web, and Web-oriented. The majority of the working, tested and viable software tools fall into the former category. There, the terms metadata and schema are generally applied to SQL databases, especially groups of SQL databases which conform to different schemas. What is needed here are tools that combine the quality of the old school with the Web-awareness of the “new school” into products that can work transparently with content in relational databases as well as Web content at the same time.

Although a number of sectors are heavily Web-focused, it would be a mistake to consider SCHEMAS to be a Web-oriented effort. Although Internet and Web protocols can be transmitted using almost any physical medium, e.g. terrestrial broadcast, various types of telephonic transmission, and amateur (“ham”) radio, much of the content upon which various metadata-related activities are focused is not Web-based, nor will it be in the near future. The Audio-Visual sector addresses much, of not most, of this type of content.

## **2.10 Automation**

Because we believe that processes should, where possible, be automated and because of the ever-increasing amount of content with which we will have to deal, we want machines to do things for us wherever they can. In the domain of metadata, that means we want machines to do things like understand schemas, transform data between schemas (i.e. understand data conforming to one (source) schema as data conforming to another (destination) schema), and find and retrieve schemas for us. This is related to what Tim Berners-Lee calls the "semantic Web" or "the Web for machines".

This idea of machine-readability for metadata sets and schemas, of automated tools to tag content, and, ultimately, of a machine-readable universe of metadata associated with content of all types, is lacking in the activities described so far in the MD Watch. In some cases, machine-readability is probably assumed; in others, it is probably too early in the process to consider machine-readability. However, other activities run the risk of failing to take machine readability into account and having to do some significant catch-up work later. In the case of one registry in particular, the schemas are even difficult for *humans* to read. Machine readability, while not impossible for a clever programmer on the user side to implement after the fact, was apparently not taken into account.

### 3. Domain reports

#### 3.1 Industrial sector

##### **CURRENT STATE OF DOMAIN, MAIN ISSUES, TRENDS, AND OVERLAPS AND GAPS**

Activities are generally carried out by:

- 1) Not-for-profit consortia, mostly comprised of for-profit companies
- 2) Individual for-profit companies

Outputs generally fall into the following categories:

- 1) Applications of XML
- 2) Tools
- 3) Frameworks, schemas, rules and metadata sets
- 4) Registries

The subcategories in the industry domain are shown below. An asterisk indicates that a number of bona fide SCHEMAS-worthy activities were found in each respective subcategory:

- General\*
- Advertising\*
- Automotive\*
- Banking, finance and insurance, a.k.a. financial services
- Computer hardware manufacturers\*
- Computer software manufacturers
- Computer software integration and services
- Electric power distribution
- Manufacturing
- Retail and B-to-B commerce and communications\*
- Search engines
- Telecoms
- Transportation and logistics

The majority of the work in the Industry domain is targeted to b-to-b Web commerce and in the creation and management of internal business information systems.

Applications of XML are fairly numerous and include such things as FinXML, FpML, FIXML, adXML, and Acord XML and are nearly always defined by consortia. Applications of XML are generally targeted at a particular industry or industry subsector, such automotive, advertising, or financial services. One would think that these applications might have something in common and could therefore develop a core application from which industry-specific applications could be more easily developed, but this does not seem to be happening.

XML has been a real boon and is leveraged by a wide variety of initiatives.

The idea here seems to be to get a large enough group of important enough players to lend momentum to any particular XML application, especially in industries such as financial services where there are competing applications (an example of an overlap).

The main gap here is that most industries simply don't have standard applications of XML they can call their own.

Tools are generally created by individual for-profit companies and fall into two classes: "old school", or pre-Web, and Web-oriented. The vast majority of the working, tested and viable tools fall into the former category. There, the terms metadata and schema are generally applied to SQL databases, especially groups of SQL databases which conform to different schemas and which use different metadata sets to describe themselves.

What is needed here are tools that combine the quality of the old school with the Web-awareness of the new school into products that can work with content in relational databases and Web content at the same time with ease.

Frameworks don't have much of a place in this domain, i.e. industry, although there is the BizTalk framework, which, in principle, covers every industry. The BizTalk Registry of schemas is notable in part because of its size - 86 organisations have submitted a total of 401 schemas which are used by 652 parties.

Web-oriented tools are much more rare, and some of those that exist do so only in the form of vaporware or low-quality (poorly tested, feature-poor) packages.

Frameworks, schemas, rules and metadata sets are less advanced than in the other Schemas domains (publishing, broadcasting, etc.). I would expect to see frameworks, schemas, rules and metadata sets in the industry domain to leverage those created in other domains, ones that are farther ahead in the process.

The only working registry found so far is the BizTalk Registry of schemas. Unfortunately, it is not designed to be machine-readable, although a clever programmer probably could create a machine to read it, and is more difficult than necessary for humans to read.

Research into the subcategories of (a) transportation and logistics and (b) electrical power distribution revealed little or no activity. The automotive industry, well known for being one of the early pioneers of b-to-b communications and commerce (as well as b-to-c) showed much less activity than expected.

### **3.2 Publishing sector**

#### **PUBLISHING METADATA: BABEL BACKWARDS**

For an accurate model of the current state of metadata development, the publishing business can look to one of its earliest examples: the story of the Tower of Babel. In this digital version, though, the plot is reversed: the participants have started out with a multitude of languages, only coming later to the dawning realization that they are building the same tower to a multimedia heaven (or hell – that part is unclear).

Strong initiatives are coming out of most of the traditional content sectors: books (EPICS, ONIX), recordings (RIAA, IFPI), audiovisual (MPEG, SMPTE), copyright (CIS), news (NITF, NEWSML), magazines (PRISM), academic journals (CROSSREF); and from the emerging e-books business (EBX, EBOOKS). (Although

audiovisual is dealt with in another Metadata Watch report, it is impossible to ignore it entirely in this review).

At the same time, metadata activity is being stirred up around the main members of the “ISXX” group of creation identifiers under ISO Technical Committee 46: ISBN (books), ISRC (recordings), ISSN (serials), ISAN (audiovisuals), ISWC (musical works) and the fledgling ISTC (textual works). These developments are normally related to but not necessarily integrated with the initiatives already mentioned.

The International DOI Foundation (IDF), adopting (like EPICS) the INDECS metadata framework, is approaching metadata from a multimedia perspective at the outset.

### ***Convergence: format wars?***

The emerging problem (or opportunity) for all of these specifications is that they all finish up covering more or less the same ground. The definitions of product and market are becoming hazy in the world of “physical” product. Book publishers release audio and video materials. DVDs include audio, text, visuals, audiovisual. Serials, magazines, news all come now in all media types. The conventional divisions neatly represented by physical content types and their identifiers do not apply to metadata schemes which must increasingly embrace all forms of creations. The fact that one sector is more biased towards “text” or “visual”, one more towards “audio” and another more “audiovisual” stuff is not much of a useful distinction when it comes to designing metadata systems in which all must be well described irrespective of their predominance or otherwise, and pretending a recording is a kind of book, or vice versa, as corporate and library systems once did, is no longer adequate.

A key role is currently being played by e-tailers seeking coherence in the way in which metadata is delivered, and they are increasingly multimedia in their product ranges: Amazon.com promotes “books”, “music” and “DVD and video” without priority. Amazon played a key role in shaping the international solution for ONIX, the most significant recent development from the book sector, and the discussion has now moved quickly on to the possibility of extending ONIX to music and audiovisual products.

But the underlying driver is of course digital delivery. The explosion in the use of unlicensed MP3 files and the record industry’s collective response through the Secure Digital Music Initiative (SDMI) has most publicly breached the dam; but all sectors have been wrestling with early methods of securely and profitably providing content on the Web.

The functional specifications for these tend towards an obvious commonality: in metadata at least, the sectoral Chinese walls are collapsing. SDMI, though nominally for “music”, must recognise data files of any type. MPEG, nominally for “Motion Pictures”, does the same. The “book” industry dictionary EPICS (the “parent” of ONIX) is structured to accommodate any media type to any hierarchical level of content. From different start points in traditional sectors, all major initiatives are being forced to address the same multimedia range of content.

## **MULTI-NATIONAL, MULTI-LINGUAL, MULTI-FUNCTIONAL**

To stretch matters further, retailers like Amazon, HMV and Barnes & Noble, and “publishers” like Bertelsmann, Warner and Sony are not just *multimedia*, they are *multinational* and *multilingual*, and so their metadata requirements are inevitably moving in that way too. Language and territorial dimensions are becoming the norm in data specifications.

Even more importantly, metadata is becoming *multifunctional*. For example, all of the major record companies are currently engaged in establishing their own corporate international databases. It is a reasonable assumption that these systems will be designed to support in due course the requirements of marketing, “label copy” for product packaging, web metadata, rights and royalty management, sales and more besides, and the data in them will be derived in part directly from production workflow systems.

It is also a reasonable assumption, without betraying any corporate secrets, that similar changes are going on in “publishing” business in all sectors. The need for definitive product and rights data, sourced once and ultimately digitally protected, is gradually dawning as a corporate imperative: it is a sign of the rapidly-changing times that the International Federation of Phonographic Industries, custodians of the metadata-less ISRC, has recently appointed a full time “Metadata Executive”.

## **CONSEQUENCES OF MULTI-FUNCTIONAL METADATA**

As publishers plan and build systems where metadata will be called upon to do more and more, the specifications for industry standards are expanding rapidly in their complexity. The EPICS data dictionary has grown far beyond the scope originally envisaged when it began as the more modest “Title Information Project” in 1998. The SMPTE (audiovisual) Data Dictionary has expanded similarly, and the editors of both of these will tell you that there are still major areas, especially in the description of rights management, which are at present dealt with in the most cursory manner, and that many of the multimedia issues are no more than doors or hooks left open for future expansion.

The EPICS/ONIX pairing illustrates another characteristic we will see a lot more of as a result of multifunctional metadata: *families* of specifications. In this case, the EPICS Data Dictionary is the wider resource of which ONIX is one subset, expressed in a specific (XML) markup format.

Is this simply “scope-creep”, lack of definition or over-ambitious design? Is it not possible to “keep metadata simple”? It seems not. Just as the MARC cataloguing format has grown into a catalogue in its own right, so now commercial industries are addressing even wider description requirements and finding there are no adequate quick metadata fixes.

The SDMI initiative’s experience illustrates the point. The original draft functional specification included the identification of general metadata requirements, for both descriptive and rights purposes. It rapidly became a monster, requiring SDMI-compliant technology to interpret descriptions of any kind of content subject to any permutation of business rules, in an environment where creators and rights owners do not yet even have established unique identities. Fairly rapidly, all but the most basic

metadata requirement has since been pushed out of the SDMI specification, relying on embedded identifiers to provide the links to metadata in other systems.

While the consequences of this complexity for a specific sector are daunting, the impact of the convergence of competing metadata schemes is even more of a concern. While organisations like Amazon.com are making it known that they are intent on having a single metadata delivery format, the sectors that supply them are preparing a whole set of not necessarily compatible schemas.

It might be argued that some activities in the “publishing” sector fall outside this convergence. News media and academic journals, for example, surely belong in relatively integral domains, and might proceed to develop their own vocabularies and interchange specifications with relative freedom?

Analysis of the functions and content of these, and their overlap with other content types, suggests this would be a brave assumption. The news media, for example, benefits (or suffers) from the fact that it is critically dependent on rapid and accurate data interchange through increasingly complex supply chains, and nowadays in all forms of media. This has led to the early and widespread use of standard mark-up formats throughout the industry, and a tradition of industrial standard-setting. New XML versions of these have appeared (NITF, NEWSML) the latter grappling with the fuller implications of multimedia content.

#### **THE IMPACT OF RIGHTS**

In any case the question of rights metadata makes the suggestion of “safe havens” for content metadata irrelevant. It provides is the functional requirement to end all such, and it is waiting in the wings, likely to make its main entrance and in all probability dominate the commercial metadata stage in the coming few years.

“Digital Rights Management” (DRM) has become a buzz-term, but it is somewhat misleading. DRMs at present are generally concerned with content protection and delivery, and pay little or no attention as yet to the underlying complex rights transactions which attend digital dissemination and manipulation. Early flirtations with generic formats in initiatives such as INDECS and the recently announced XrML from Microsoft/Xerox are only the stage-setters. MPEG7’s IPMP work has, rightly, pushed the issue out of scope and deferred to an as-yet non-existent framework of rights metadata. The next year will see some serious and heavyweight activity towards directly implementable rights metadata systems.

The consequences for descriptive metadata are potentially serious, because most important descriptive terms (contributors, links, formats, events of creation and publication) are loaded with legal implications in the right context – they are, in fact, essential parts of rights statements and agreements. In the CIS community, for example, Author and Publisher are implicitly rights owners. Indeed, it is actually irrelevant in rights management whether someone really was the author, provided that the “real” author agrees they were (or is sufficiently unwilling to sue, or is dead). Lennon and McCartney are each credited, by agreement, with “co-writing” many of each other’s songs to which they in reality contributed nothing. This is just one of the more famous examples among (literally) millions where commercial interests have framed bibliographic reality.

The draft MPEG7 IPMP metadata specification makes it clear that no terms used as MPEG descriptors may be deemed to have any legal implications: the first of many such “disclaimers” with which metadata schemas will have to deal. The ability to recognize “bibliographic” relationships alongside parallel “legal” relationships will add a general further level of complexity in due course.

### **THE ROLE OF IDENTIFIERS**

Identifiers play a central role in most developing publisher schemes. The more recent identifiers come with their own mandatory metadata, and it is a measure of the importance of creator/publisher allocated “ISXX” identifiers that the ISO TC46 subcommittee responsible for them (sc9) intends to clear its decks of all issues not related to identifiers.

ISBN (books) has no formal metadata, but the *Books In Print* and similar trade bibliographic services provide a practical metadata context for ISBN which has been the backbone of the book industry supply chain since the 1970s.

In stark contrast with books, the recording industry, though centre-stage in digital delivery, has no tradition at all of collective or standardized metadata. The ISRC has functioned for over a decade with no related metadata or registration database, which means a lot of them are issued but nobody else has any idea what they are (your PC can read the embedded ISRCs from your CD, but it will be none the wiser if it does). As the recording industry’s focus shifts from the bar-coded album to the digital track, ISRC will become a hugely important identifier, but not until it is associated with metadata which enables its discovery and verification.

ISSN (serials) has metadata, though exactly what a serial is requires some clarification in the digital age. ISAN (audiovisuals) and ISWC (musical works), the new kids on the block, have mandatory core metadata, and if ISTC (textual works) makes it to the start line it will follow suit, with metadata drawn almost certainly from the EPICS dictionary.

### **POSSIBLE MULTIMEDIA SOLUTIONS**

Several developments are worth watching as they may provide some ways to avoid the brewing multimedia confusion.

The MPEG21 initiative is the most ambitious “umbrella” framework proposal for bringing all technical and metadata standards together in an integrated program. MPEG’s track record and strong support in technology sectors.

The International Digital Object Identifier (DOI) Foundation (IDF) is now deploying an approach to content description and management which implements several potentially valuable principles: the notion of interoperable and overlapping content “genres”, the “declaration” of kernel metadata independent of , and “actionable identifiers”. The DOI structure allows it to act as a “meta-identifier working in parallel with other more limited Its first implementations include CROSSREF.

The INDECS framework (of which EDItEUR, IDF and IFPI are members among others) provides a high level generic model and vocabulary which acts as a framework for the development of interoperable schemas. It is especially focussed on the

integration of rights and descriptive metadata, recognizing “events” as the underlying common denominator.

Finally, the ONIX International initiative under the EDItEUR umbrella has emerged since April as perhaps the most likely candidate to provide a true multimedia content interchange standard (or set of specifications). It is a measure of the fragility and rapidity with which events are moving throughout this sector that it looked unlikely as recently as March that there would even be a single agreed EPICS/ONIX specification for the book industry, let alone now the possibility of its extension to its audio and audiovisual product relatives. The agreed spec represents a healthy marriage of American market-driven urgency and European analytic thoroughness without ultimately seriously compromising either, so hopes are high for its widespread adoption and success within “the industry formerly known as the book industry” at least. A number of intermediary organizations, include MUZE Inc who are pioneering the use of RDF in this sector, are committed to ONIX compliance.

#### **SCHEMA MAPPING/REGISTRIES**

As elsewhere this is in its infancy in the publishing sector. The most public activity is the MPEG7 medmet initiative, which has begun with the relationship between the SMPTE and Dublin Core sets. At this stage it would be best to describe this as experimental.

### **3.3 Audio-visual sector**

#### **CURRENT STATE OF THE DOMAIN**

A description of the domain could be: "Production, distribution and archiving of digital audiovisual material". The application fields are: radio and television broadcasting; audio and video (post)production; audiovisual archives; multimedia library systems; image banks; news agencies; WEB TV. Due to the convergence of television, computer and communication technologies, the separate area's within the digital media domain are approaching one another in many respects. They share a number of common focus points and interests :

- The increase of the number of digital documents in the organization.
- The need for standards for storage, exchange, cataloguing and indexing of digitized material.
- The necessity to fill multiple new channels and services.
- The need for migration scenarios from analogue to digital collections, procedures and systems.
- The need to integrate media-technology and business systems for efficiency savings, cost control and support for commercial transactions.
- The need of fast and solid delivery to increasing groups of (new) customers within and outside of the organization.
- The pressure on the preservation of cultural heritage via digitization.
- The need to guard, protect and track copyrights.
- The need to improve research possibilities for digital audiovisual material.

At the moment numerous differently scaled and shaped activities can be observed to go on in this area. The work is done on the one hand internationally, but on the other hand also locally and the developments are rather different in scope and effect. To name a few : the design of national infrastructures for digital broadcasting; the

distribution of radio and television programmes via the Internet; the work of the international standardization committees; the setting up of various kinds of online multimedia-catalogues; the development of educational and commercial Video on Demand Services. These activities are being executed in several project formats, as part of regular business developments, as government programmes, as European funded initiatives and so on. The variety in players is great. Involved are national and commercial broadcasters, publishers, producers, universities, national and private audiovisual archives, industrial partners, IT, broadcast engineering, governmental departments, national and international bodies and professional federations.

Important part of these activities and projects is the development of metadata standards. Metadata development might be the core objective of an activity (e.g. the standardization committees), in other cases it is (inevitable) part of a smaller or larger digital (private or public) media project. The dynamics and objectives of the developments show numerous differences and the level of synchronicity is low. Some activities concern all forms of access to and (re)use of all digital multimedia on the web as well as for professional use, others again are dedicated to one single area. At the moment, in many cases, the lack of common standards forces the industry, the broadcast companies and the distributors of WEB TV alike, to develop proprietary methods and pragmatic solutions to be able to actually run their implementations and systems.

It can be observed that 'audiovisual' metadata still are subject to different interpretations. On the one hand IT would consider the concept as: data relevant to an information system including data models for technical appliances. In this sense metadata are really a type of information depending on form, characteristics, classification, storage and structuring; version management, integrity and performance are also parts of this concepts. On the other hand 'broadcast IT' interpret metadata primarily as media systems, i.e. information that focuses on media content identification and analysis. Here, audiovisual production metadata are logical data and the manner of their classification, purely related to programmes on radio and television.

Broadcasters have therefore cast this appropriate definition : content = essence + metadata (essence, i.e. the audiovisual content itself). Information analystst in this area will often even think of an additional layer, metadata for them refers to information regarding the structure of metadata proper, that again refer to the document content.

## **MAIN ISSUES**

Metadata in the digital media domain are rather more complex than metadata related to digitised textual information. Though the principles are identical, image and sound productions, more than text, contain much implied information that again constitutes separate groups of metadata. Finally, the digitised production process generates a quantity of technical metadata used to facilitate transfer, internal and external distribution or storage. It is clear that the increase of digital documents in this areas causes many problems, but it could also offer a possible solution: some of the identifying information can be embedded in the video or audio stream itself (the electronic metadata). To be able to maximize the benefits from this, the need for standardization is high.

The various types of 'audiovisual' metadata can be categorised as: signal related metadata, programme documented related metadata and pointers. The

standardization of the various groups and classes of metadata is linked to the global user requirements for digital media systems.

The main issues are:

- Interoperability between mediaformats and systems with automatic exchange of metadata.
- Standardization of data structures and data models.
- The development of unique identifiers to link the stored audiovisual material with the related documentation through the process of creation, delivery, use and re-use.
- The semi automatic and full automatic creation of embedded (wrapped) metadata within the content streams and files.
- Automated (semantic) indexing of important sequences or parts in video, sound recording or stills.
- Standards for automated indexing to be able to use image analysis, teletext information, speech recognition and sound analysis.
- The migration of the embedded metadata into unwrapped metadata, stored in databases, that can be managed and controlled over networks.
- Versionmanagement: the incorporation of different versions of the same document in different stages of the production process, as well as descriptions that cover different copies of the same document
- The development of methods of extraction , searching, evaluation an validation to assist content queries.
- Multilingual searching.

## **TRENDS**

The work of the standardization committees and working groups may be summarized as follows:

- Regarding the definition of a central registration structure that implements the mapping of different metadata schemes.
- Regarding the development of tools with generic functionalities, to develop and use metadata repositories.
- Regarding metadata taxonomies, so as to structure the metadata discourse.
- Regarding ontologies related to the characteristics of metadata and to the definitions of data elements such as designation of fields, types, classification and semantics.

Amongst users and committees alike there is a growing assumption that metadata and metadata standardization will eventually solve all problems connected to the issues of interoperability. The questions surrounding the protection and guarding of the integrity of the metadata itself, seem somewhat neglected here.

One of the problems standardizing committees face, is that current technical equipment has become so inexpensive that the program makers and other users can generate near-broadcast quality themselves and thus not wait for "the next standard". Some committees operate closely with manufacturers. Therefore, industrial politics is involved, and this might be an obstacle. Other groups are very large and their list of required data is often so extensive that difficulties are foreseen. The work of a fair number of groups and committees is concentrating on practically every subject attached to the audiovisual domain. In these cases each workpackage could be a project of its own.

At the moment the broadcast world (production, broadcasting and archiving) is specifically looking at the work of SMPTE in this area (Metadictionary and Engineering Guidelines) and SMPTE-UMID. The UMID (Unique Material Identifier) has only very recently become a standard (SMPTE 330). The SMPTE Metadata Dictionary is out for publication in two or three weeks from now. It is expected that the Dictionary will be a standard in a few months. The broadcast world is also highly interested in the BBC Media Asset Management Data Model SMEF (Standard Media Exchange Framework) which only this week has been made widely available via free licence. The European Broadcasting Union Projectgroup P/Meta is aiming at developing common media exchange formats for broadcasters, publishers and archives and will incorporate the work of SMPTE, MPEG-7 and SMEF.

As for multimedia content in the professional environment, the activities of the MPEG-7 group are being monitored. The development of this standard is still at a very early stage. The MPEG-7 Draft International Standard is not to be published until medio 2001. Before that time the many different sections within MPEG-7 will have to be integrated and connected. Within the EBU community a initiative has recently started to establish a cooperation between the MPEG-7 Description Schema Work and the SMPTE Metadata Dictionary Taskforce. This initiative is aimed at providing interoperability between the two standards and to link the broadcast environment to the objectives of MPEG-7.

#### **OVERLAPS AND GAPS**

In general there is still a proliferation of activities that does unfortunately carry the danger of producing incompatible standards, even if those had not been -as they are, in different stages of development. At the moment there are many different contributions to standards developments for various areas and at different stages of growth, that stem from the various professional qualifications and commercial interest of the participants. For as long as this is found wanting, it has to make with ad-hoc solutions that cannot fail to contribute to the general confusion. As a consequence it is observed that while some standards have already a framework that only need to be completed, other- even though of comparable nature - are quite a distance from any completion. It might be concluded that-within the area of 'audiovisual' metadata - too many organizations are currently developing far too many 'standards', methods and procedures. This situation inhabits the danger of both overlaps and gaps.

MPEG-7 is aiming at solutions for the entire audiovisual content. The SMPTE work is dedicated to the production and distribution of digital television and radio material. Audiovisual material however, contains textual information including catalogues entries, production information, scripts etc. Offering access to this textual information is not within the objectives of MPEG-7 or SMPTE. Therefore they will have to tackle this issue by incorporating descriptive metadata as a seperate node within the Metadata Dictionary. This particular node still needs to be developed, i.a. with contributions from other initiatives that specifically focus on descriptive metadata.

The membership of most commissions and workinggroups in the digital media domain is highly technically. Despite the fact that the convergence of documentation fields and technology is extremely manifest here, the technical expertise generally exceeds the archival and 'documentational' know-how. On the other hand it can be observed that the level of understanding among members of some committees, of the issues and techniques associated with media technology, is not very even. On top of

this, the available resources are usually highly constrained. In addition, development of thinking in the field is proceeding fast while some clear priorities are emerging.

Audiovisual archives is not evenly represented in most of the workpackages even though subjects as migration scenario's, user requirements and datamodeling are highly documentation and archive fields. Matters documentalists and archivists should more address are: storing metadata (embedded or separate); unique identifiers: how will these work on a day-to-day basis, significance and practical implications etc. Additionally, these professional groups could offer input as to matters covering quality control and information. In the end most committees will have to come up with both an archive format and a common data format. It is expected of broadcast archives to play a big role in developing technical formats, even though these archives have up till now not really done any serious work within this area.

For all committees and workingsgroups it will be important to stay aware of the work of on another in order to guarantee compatibility, mappability and interoperability. The general question remains whether all these local, national and international efforts will in the end provide the audiovisual world with a solid, workable standard. A better option perhaps, would be for all the involved organisations and commissions to specifically focus on the metadata part they are experts in. An overall metadata infrastructure could then be designed, to facilitate for all media organizations to interoperate. For the broadcast environment the Advanced Authoring Format(AAF)-initiative moves in this direction.

### **3.4 Educational sector**

#### **CURRENT STATE OF DOMAIN**

Basically, there are four important metadata initiatives that dominate the education domain:

- a) IEEE LTSC LOM: The IEEE Learning Technology Standardization Committee includes a number of working groups, that deal with issues ranging from a Learning Technology Architecture to work on a Learner Model. Metadata activities are regrouped in the Learning Object Metadata group. This three year old initiative has developed an elaborate metadata scheme with a hierarchical structure. Data elements are regrouped under categories (general, lifecycle, metametadata, technical, educational, rights, relation, annotation and classification). Especially relevant in this domain is the educational category, that includes elements such as:
  - interactivity type (active versus expositive)
  - learning resource type (exercise, simulation, questionnaire, etc.)
  - interactivity level (from very low to very high)
  - semantic density (idem)
  - intended end user role (teacher, author, learner, manager)
  - context (primary education to vocational training)
  - typical age range
  - difficulty (from very low to very high)
  - typical learning time
  - description
  - language of the typical intended user

Besides these data elements that are specifically geared towards the domain of education, LOM also includes a rich set of data elements in the other categories. Consequently, many organisations are starting to deploy LOM in a more general context that just education and training.

- b) DC-Education: Since August 1999, the Dublin Core community is investigating the use of Dublin Core for description of educational materials. Since November 1999, a new group has been formed to investigate the use of the Dublin Core metadata set in education and training. That group had a first meeting at DC-7, with representatives from the IEEE LTSC LOM group, and had a first real meeting in February 2000.

The present recommendation is to add a number of data elements to the DC set, explicitly borrowing from the IEEE LTSC LOM work. Whereas the use of namespaces to achieve semantic and syntactic interoperability between DC-Education and IEEE LTSC LOM is mentioned, it remains somewhat unclear how this would be organized in practice. Moreover, if such an approach would be operational, then there would be little value to DC adopting LOM elements, as any combination of these elements would be acceptable anyway.

- c) CEN/CENELEC ISSS LTWS: Since March 1999, CEN/CENELEC ISSS has set up a new workshop, on Learning Technologies. In a first phase, two projects have been started. A workplan for future European work in this area is being finalized. That workplan lists a number of concrete recommendations. Metadata related such recommendations include:

- promotion of standards
- taxonomies and vocabularies
- profiling of LOM for specific communities
- bindings of LOM to RDF and XML
- internationalization of LOM

In the second project, a French and German version of IEEE LTSC LOM have been developed. Catalan, Dutch, Greek, Italian and Spanish versions are under development.

The CEN/CENELEC ISSS LTWS is complimentary to the Prometheus initiative, in that the former focuses more on formal standardization work, whereas the latter is more concerned with consensus building.

- d) Recently, a new subcommittee has been set up under the umbrella of the ISO/IEC JTC1 Joint Technical Committee. As far as formal standardization goes, this venue represents the top in the hierarchy, with a global scope and world-wide recognition. The IEEE and CEN both have formal relationships with the ISO/IEC JTC1. In fact, both IEEE LTSC and CEN/CENELEC ISSS LTWS have liaison status with ISO/IEC JTC1.

The IEEE LTSC decided not to actively propose any of its work items into the ISO/IEC JTC1 SC36 at this moment, as it felt that it had to finalize the work before forwarding it to ISO, so as to make sure that no parallel divergent standards development paths arise. However, the IEEE LTSC has also stated that it seeks active collaboration in any field that the SC36 would want to tackle and that is covered by the activities of the LTSC. As the German delegation to SC36 has indeed proposed a work item on metadata, it seems like that some activity in this area will be initiated soon.

## **MAIN ISSUES**

It is interesting to note that, whereas the IEEE LTSC LOM was originally developed specifically for the domain of education and training, and is becoming gradually more and more deployed for metadata of general resources, it is rather the other way around with DC-Education, as the Dublin Core metadata element set was originally developed for general resources, and is now being adapted for the specific field of education and training.

All in all, the most pressing issue for these two organizations is to convince the field of education and training that IEEE LTSC LOM and DC-Education will not "fight for world domination", as this will prevent a large portion of the potential users of metadata in this field from starting any activity whatsoever. As those potential actors will prefer to wait "until the battle is over", the net result will be a hindrance to the uptake of metadata in this field, something that would be detrimental to both DC-Education and IEEE LTSC LOM. It seems that at least some of the more important actors in both communities are aware of these ramifications.

Another important issue at this moment is the finalization of LOM: there seems to be an agreement that the core of the specification is not evolving dramatically anymore, but that most time now seems to be spent on developing the exact standards wording. As this is a consensus based initiative, the time involved is somewhat difficult to estimate and impossible to control by the LOM group.

Work in the CEN/CENELEC ISSS LTWS seems to be moving faster, as there is a report being finalized after no more than 12 months of operation. On the other hand, it now remains to be seen what the member state representatives will decide. It seems quite likely that funding will be provided for the implementation of the recommendations. In the mean time, limited funding still remains to continue work on the localization of LOM.

How soon the ISO/IEC JTC1 SC36 will start working in earnest, specifically on educational metadata, is somewhat unclear. The next meeting is in September 2000, and it doesn't look like any serious work will happen beforehand.

## **TRENDS**

The main trend is that awareness in the sector of education and training on the issue of learning technology standardization in general and on metadata specifically, is definitely growing fast.

On the other hand, with awareness of the importance of these issues also seems to grow the confusion and misunderstandings. For many interested parties, the difference in status between for instance consortia (that can build their own internal specifications) and accredited standards organizations (that work mostly under an open consensus-based model) is not at all clear.

And, as has happened so often before with educational technologies, there is a definite danger that expectations are being raised unreasonably high, either by conscious action from those involved in the development of these technologies, or because of the genuine enthusiasm of all those involved.

## **OVERLAPS AND GAPS**

To a certain extent, the standardization bodies mentioned above all have overlapping activities. On the other hand, the CEN/CENELEC ISSS LTWS has explicitly stated that it would not duplicate any effort going on elsewhere, unless that work would not take into account genuine European needs and requirements. The current localization work of the IEEE LTSC LOM, taking place in the LTWS, proves that this approach can indeed work in practice.

There is also considerable overlap in the work of the different consortia (ADL, AICC, ARIADNE, IMS, etc.) that work on actual implementations of the metadata related and other standards. This does not seem to be a problem, the less so as all these consortia contribute to the development of the standards, and then adapt them to the needs of their constituencies, a process referred to as 'profiling' in the standards world. In fact, having several independent implementations of the metadata standards is a good thing, as it increases the probability that problems or shortcomings will be identified early on. It would be most useful though to have some interoperability development taking place now that the specifications and implementations are maturing, so as to prove that the standard does indeed serve its ultimate goal.

### **3.5 Academic sector**

#### **CURRENT STATE OF DOMAIN**

The academic domain - as interpreted by this report - is quite diverse in nature. The majority of the metadata-based initiatives described in this report are funded as part of relatively short-term projects. Others are linked to longer-term initiatives (e.g. the Recordkeeping Metadata Standard for Commonwealth Agencies), part of standards development processes (e.g. the Reference Model for an Open Archival Information System) or part of wider discovery-type services (e.g. the Resource Discovery Network).

#### **MAIN ISSUES**

In the metadata initiatives of the academic domain that have been reviewed here, the main issues relate to topics like: metadata for the Web, Internet information gateways and metadata for recordkeeping and digital preservation. Each one of these different strands raises different issues.

Web metadata developments are, in part, based on the work of standards bodies like the World Wide Web Consortium (W3C). For example, since the publication of the HTML 4.0 specification, it has been possible to embed metadata into the headers of Web pages. W3C have also developed the Platform for Internet Content Selection (PICS) that enables labels (metadata) to be associated with Internet content. Most current metadata developments coming from the W3C are centered on the development of the Resource Description Framework (RDF)- which combines other W3C work on PICS, digital signatures (DSig) and Platform for Privacy Preferences (P3P).

The first Internet information gateways - often then known as subject gateways or subject-based information gateways - were created in the mid-1990s. Many of these services have been funded as part of short-term projects (e.g. as part of the Electronic Libraries (eLib) Programme) and many continue with this form of business model.

Increasingly, however, Internet information gateways are becoming part of more formal collections of services like the UK's Resource Discovery Network (RDN).

Software projects like ROADS were funded to support the development of Internet information gateways. ROADS, for example, developed a software toolkit for such gateways together with documentation and support. Similar work has been carried out by European Commission-funded projects like DESIRE - e.g. by publishing a 'DESIRE Information Gateways handbook' - and Renardus. Co-operation between gateways has also been encouraged, especially through multi-partner projects like DESIRE and Renardus and through the IMesh collaboration. The development of software tools remains important. A project known as IMesh Toolkit is attempting to develop a configurable, reusable and extensible toolkit for subject gateway providers.

Recordkeeping metadata has been an active topic of research for the archives and records management communities since the mid-1990s, when two prominent North American-based projects were concerned with looking at the long-term preservation of electronic records. The first was funded by the US National Historic Publications and Records Commission (NHPRC) and was entitled 'Functional Requirements for Evidence in Recordkeeping'. The project was carried out by the School of Information Sciences at the University of Pittsburgh and is usually referred to in the literature as the Pittsburgh Project. The second project was entitled 'The Preservation of the Integrity of Electronic Records' and was undertaken by a team of researchers based in the School of Library, Archival & Information Studies at the University of British Columbia (UBC) and was funded by the Social Sciences and Humanities Research Council of Canada. Neither project was purely concerned with metadata but the Pittsburgh Project published a metadata specification for evidence based on a model known as the Reference Model for Business Acceptable Communications (BAC). The UBC project created a set of eight templates (i.e. metadata) that were intended to identify the necessary components of records in a variety of recordkeeping environments. The work of the UBC-based project has been carried forward in a wide-ranging international project known as InterPARES.

Despite all of these developments, there has not been widespread implementation of recordkeeping metadata by the archives and records management community. This may be due - in part - to the complexity of the systems that would result from an implementation of something like the Pittsburgh Project's BAC-based scheme. Despite this, the Australian archives and records management communities - in particular - have begun to develop metadata element sets for recordkeeping and associated documentation. For example, the National Archives of Australia have produced a Recordkeeping Metadata Standard for Commonwealth Agencies (May 1999). Also, researchers based at the School of Information Management and Systems at Monash University have produced a metadata element set for recordkeeping as part of the SPIRT Recordkeeping Metadata Project.

A fourth strand of metadata developments in the academic sector relates to the subject generally known as the long-term preservation of digital information (digital preservation). Some of the initiatives relate to the preservation (and other) metadata that needs to be collected as part of digitisation projects. Examples of these would be the collection of metadata elements used by the Making of America II Testbed Project and the sixteen elements defined by a Research Libraries Group (RLG) commissioned Working Group on the Preservation Issues of Metadata (May 1998). Other work is

more general - based on the concept of being able to describe any digital object for long-term preservation in association with an identified preservation strategy. The National Library of Australia (NLA) has been very active in this area in its proof-of-concept project called PANDORA (Preserving and Accessing Networked DOcumentary Resources of Australia) and in its draft document 'Preservation metadata for digital collections'. Other - typically library-based - projects like the UK Cedars (CURL Exemplars in Digital Archives) project and the European Commission-funded NEDLIB project have based their metadata schema development upon a model defined in an draft ISO standard known as the Reference Model for an Open Archival Information System (OAIS).

## **TRENDS**

The main trends differ according to which strand is being considered.

The metadata work of the W3C is still mainly centered upon the development of RDF. This standard has not yet been implemented widely, although this is expected to change in the near future.

The main trends that relate to Internet information gateways are the increased international co-operation between gateways as evidenced by broker system for various European gateways that is envisaged by the Renardus project and by wider IMesh activity. Another trend that can be identified - in some contexts at least - is the development of gateways from short-term project funding to unified services like the RDN.

Metadata standards being developed for recordkeeping and for digital preservation have not - as yet - been widely implemented. Those implementations that exist tend to be part of limited-scale pilot projects. In the digital preservation area at least, much future development will be based upon the OAIS model. The RLG and the Online Computer Library Center (OCLC) have also recently (March 2000) announced some co-operative work on digital preservation, including the production of a document entitled 'Preservation Metadata for Long-Term Retention'.

## **OVERLAPS AND GAPS**

The activities described in this report are quite diverse so it is difficult to adequately identify overlaps and gaps. We can see, however, that there are functional similarities between metadata schemas being developed for electronic recordkeeping and for digital preservation. Internet information gateway initiatives also overlap with library-based initiatives like OCLC's Cooperative Online Resource Catalog (CORC). In some sectors (e.g. the RDN), Internet information gateways are increasingly becoming known as Internet resource catalogues.

### **3.6 Research sector**

#### **STATE OF THE DOMAIN (TRENDS, ISSUES)**

The scope of this first watch report was research institutes and professional societies. We spent some time examining the Web sites of major private research institutes, such as Rand, PARC, SRI, and Mitre. Trying to locate metadata activities through the Web sites of corporate laboratories seemed not to be very cost-efficient -- apart from the occasional white paper on corporate extranets (General Electric) or data

warehousing (Rand), our efforts turned up little of interest. We also looked at the titles of papers in several digital library and Web conferences for possible leads.

We examined but did not make full Watch records for projects and Web services that simply announce that they use metadata. Rather, we focused on projects and services that make a point of publishing their schemas, participate in standards-making activities, or otherwise promote metadata, whether with software tools or with working groups.

In Europe, research laboratories such as Fraunhofer and the members of ERCIM (European Research Consortium for Informatics and Mathematics, of which GMD is a member) typically participate in metadata activities as minor partners in national or European projects, which seemed out of scope for our information-gathering.

One important project among European research laboratories is the DELOS Network of Excellence, a Fifth Framework project (2000-2003) organized by ERCIM, whose Standardisation Forum provides a context for ad-hoc working groups to address issues related to the deployment of digital library standards, particularly in the area of metadata.

Nordic countries have long been active in promoting metadata. The Nordic Metadata Project, a cooperative venture of the Nordic countries and an organizer of the Fifth Dublin Core Workshop in 1997, pioneered the notion of a metadata-based distributed index across national boundaries. In particular, the NetLab at Lund University has been a center of technical research and development of metadata tools and harvesting robots. NetLab has been active in the European DESIRE project since 1996 and maintains SAFARI, a search engine that harvests Dublin-Core-based metadata embedded in documents located on the Web servers of participating organisations, which include universities and public research organisations in Sweden. The SAFARI Project provides particularly good documentation of its schema, with crosswalks to related schemas.

Among the professional societies, the German Mathematical Society (DMV) has taken a leadership role internationally in defining and using metadata to create a Web portal of mathematics resources. In the context of an Information and Communication (IuK) initiative among German professional societies, the MathNet Project has created a search engine that indexes materials held at a wide range of institutions -- primarily in Germany, but now including France, Austria, Italy, Sweden, and USA. The indexes are structured by type of material -- eg, MPRESS for preprints and PERSONA MATHEMATICA for home pages -- with a general index, SIGMA, that redundantly captures everything held in the more specific indexes. MathNet encourages the use of (Dublin-Core-based) metadata by distributing a free metadata editor and displaying search hits that have metadata in a separate result set, making clear the difference in quality. The project is governed by a steering committee and related committees of the German Mathematical Union (DMV) and the German Information and Communication (IuK) Initiative.

MathNet was a major focus of a workshop in Berkeley in December 1999 on the "Future of Mathematical Communication". At that workshop, the Committee on Electronic Information and Communication (CEIC) of the International Mathematical Union (IMU) decided to internationalise the MathNet initiative. CEIC foresees the

development of an international Web service based on the German MathNet model. In addition to the countries mentioned above, this committee includes representatives of Austria, Canada, Brazil, Russia, China, the UK, India, and Australia.

The EULER Project of the Telematics for Libraries Programme aims at providing a "one-stop shopping site" for users interested in mathematics. It differs from efforts such as MathNet inasmuch it includes "for pay" resources from commercial publishers and document delivery services. This diversity of resources is integrated via descriptions based on Dublin Core.

Physicists are undertaking a similar project, but the role of metadata here is less clear from the materials I have seen. The Institute of Physics (successor of a Physical Society founded in 1874) has created an extensive portal of information on physics, including PhysicsNet, a buyer's guide for physics-related products and services; PhysicsWeb, a subscription newsletter; an online bookstore; and several electronic journals (for a fee). This site would have been represented by a Watch record were it not that the technical documentation, which includes a reference to Dublin Core, seems to date from 1996.

The PhysNet project -- a network of research centers and university departments for physics centered to some degree on the European Physical Society -- seems closer to the goals of MathNet. Its server is located at the University of Oldenburg, likewise a partner in the German IuK initiative. Like MathNet, PhysNet is creating a portal for preprints and other information in the physics field in cooperation with national societies and local institutions in Germany, Australia, Ireland, Denmark, Australia, Korea, Hungary, India, and Russia. From the materials I have seen, however, the role of metadata in this project is less clear.

The project Dissertations Online, funded by the German Science Foundation (DFG), is creating national standards for cataloging and providing access to dissertations produced at German universities. Project members include German regional library networks, university libraries, and Die Deutsche Bibliothek. Since 1998, dissertations in digital form are deposited and archived at Die Deutsche Bibliothek.

The metadata schema for dissertations, METADISS, is available on the Web (<http://deposit.ddb.de/metadiss.htm>). It was developed in consultation with German scientific professional societies and is based on Dublin Core. Distinctions are made between bibliographic and discipline-specific metadata versus intellectual property rights metadata and between metadata prepared by the author and metadata prepared by professional catalogers.

A sub-project at the University of Duisburg is continuing the development of metadata for dissertations. For example, metadata packages are needed for multimedia components such as video sequences. Expected products of the project include: standard DTDs for dissertations in several fields; agreements on standard document formats for digital dissertations; workflow models for the authentication and archiving of dissertations; and strategies for the non-commercial and commercial use of dissertations on the Web or CD-ROMs.

In Germany, the activities of professional societies in the area of metadata are coordinated by a the umbrella initiative Information and Communication (IuK). Relevant activities include MathNet, PhysNet, the DFG-funded project Dissertations Online, and the German Educational Server. Representatives of these and other

projects meet in an IuK Working Group on Metadata and Classification. More recently, IuK has become a member of W3C. Founded with a vision of distributed repositories for academic information, the IuK initiative emphasizes simple standards and cost-effective methods.

IuK operates within the context of the broader funding framework Global Info. Global Info, widely known as The German Digital Library Project, is the result of a 1996 program of the German Ministry for Education, Science, Research, and Technology (BMBF) called "Information as the Raw Material for Innovation". This program was based on the notion that knowledge was the key to future prosperity in Germany, and that the key to knowledge lay in delivering richly structured information to the desktop of every scientist and technician. From the outset, Global Info has emphasized the necessity of cooperation between scientists, libraries, information centers, publishers, and end users.

Also worthy of mention is the software development effort that has created Protégé, an editing environment for generating knowledge-based systems from formal ontologies. A product of digital library projects at Stanford University, it has been adopted by many projects as a basis for further development, as listed on their Web page. This seems at present to be one of the leading editors for RDF structures.

### **3.7 Geographical information sector**

#### **CURRENT STATE OF DOMAIN**

First, we can distinguish two main geographical zones: the United-States and Europe. These two groups do not seem to cooperate a lot together.

The American initiatives I found, were really developed and a little bit ahead of the European because they began earlier. The FGDC (Federal Geographic Data Committee) has developed an impressive number of tools, training material, awareness activities, guidelines, etc. on metadata intended to the GI community.

Nevertheless, the European projects are still rather numerous. They are regrouped under common banners such as the European Commission projects or EUROGI. Under these banners, they have more strength and impact. Besides, they share their metadata experience and exchange ideas and concepts.

The GI community did not seem to agree yet between its own members on a metadata definition.

I found many different ways of working with metadata, and therefore of seeing metadata. Metadata is still often seen, as in the old school, as a means to facilitate the search of data in a database. The concepts of: "metadata for search on a database", "metadata for catalogues and queries", "clearing house", "catalogue and metadata level information", are really present in the projects I discovered. Yet, I believe that it is important to also take these projects into consideration, otherwise the SCHEMAS project will have to put aside a quite impressive number of initiatives.

I actually found very few projects dealing purely with metadata for the Web, as we do in SCHEMAS.

The GI community does not seem to work towards a common goal regarding this subject. Even when I attended GI workshops or meetings, I felt that a long way is still to be done in order to arrive to a common understanding between people working in that field.

Actually, a lot of work has been done and is still done, yet every initiative works independently and on its own.

If the metadata definition is not clear yet, the standards to adopt are even more blur. Too many various standards are developed by people working on projects but without any joint work with other initiatives or projects. To a certain extent, each project follow its own definition and standards.

The main standards for the moment are:

- The work of ISO TC 211 is really important in the domain and should be taken into account, even if they will not be reporting out the final version of the global GI metadata standard until Sept. 2001 (according to the current chart). Yet, when it will be finally delivered, many large organisations, including national agencies such as defense mapping agencies, will be adopting it.
- The CEN TC 287 standard for metadata, which was completed and released as a voluntary standard (ENV 12657), which many organisations, such as MEGRIN's GDDD and others, have adopted in one form or another.

Finally, I do not want to appear too critical and negative.

Actually it is a domain where metadata is present, which is already very good.

The GI community realises the need for metadata and seems to work hard on finding definitions and standards, which can suit their needs.

Very good project, such as ETeMII, tries to bring the GI community together in order to reach a common agreement on metadata and the standards to adopt.

#### **MAIN ISSUES**

To try to work more together and in the same direction

- To establish a common metadata definition and standards
- To focus more on metadata for the Web
- To open the GI community to other working fields in order to share experience and ideas on the domain
- To develop more material: tools, training material and guidelines

#### **TRENDS**

It seems that the GI community really wants and tries to work towards an agreed definition and standards of metadata.

They also try to open their field to other ones dealing also with metadata in order to have a broader horizon on what is happening on that front and try to reach a common agreement, and not working together alone in its own domain.

## **OVERLAPS AND GAPS**

I did not notice any important overlap regarding the subjects involved. Various sub-domains and fields in the geographic information field are treated. The project are really varied, always dealing with a precise geographic field like: geospatial data, hydrographical data, digital spatial data, geological data, etc...

Yet, we can maybe identify overlaps and gaps in the way metadata are seen. Too many projects are dealing with metadata catalogues and queries. Maybe, these projects could be regrouped in order to work on a common multidisciplinary database, instead of developing too many various databases.

Furthermore, strong gaps subsist in the development of metadata for the Web.

Finally, too many projects develop metadata for their own use and we notice gaps in developing joint metadata standards and definitions. Projects such as ETeMII are essential in trying to bring people to the same table and to discuss about the decisions to take towards metadata. Yet, few projects as this one exist.

### **3.8 Other sectors**

#### **CURRENT STATE OF DOMAIN, MAIN ISSUES, TRENDS, AND OVERLAPS AND GAPS**

Activities in this domain are generally carried out by not-for-profit consortia or by government agencies.

Outputs generally fall into the category of frameworks, schemas, rules and metadata sets. This contrasts sharply with the Industry domain and its outputs which fall into the following categories:

- 1) Applications of XML
- 2) Tools
- 3) Frameworks, schemas, rules and metadata sets
- 4) Registries

The subcategories in the industry domain are shown below. An asterisk indicates that a number of bona fide SCHEMAS-worthy activities were found in each respective subcategory:

- Electrical power distribution
- Environment\*
- Government (including military)
  - General\*
  - Environment\*
  - Military
- Health care\*
- Mail and delivery
- Transportation and Logistics

Of the above, most activity falls within the Government category. The primary goals of the government projects are to facilitate metadata-enabled content creation, delivery, use and reuse:

- among government agencies within a particular country
- within a particular government agency within a particular country
- between business and government
- between scientific research and government

In one case, the primary goal was to facilitate metadata-enabled content creation, delivery, use and reuse between agencies of different governments (in this case, the US and Canada).

As is the case with many of the activities in the Industry domain, many activities in the “Other” domain consisted of proposals and descriptions of an activity, but no tangible, usable programme or output.

Also like the Industry category, the trend here seems to be vertical, i.e. activities are oriented towards one particular government agency or the link between one particular agency and its non-governmental associates. The same is true of the field of health care which is served by the HL7 (Health Level 7) initiative.

There seems to be little interaction between military and non-military initiatives in government. The gap in the government domain becomes apparent when one surmises that perhaps one method of facilitating metadata-enabled content creation, delivery, use and reuse might be extensible to an entire government or even a consortium of governments, instead of having different methods (or no method) for each government agency (and its non-government associates, if necessary).