erpatraining

Metadata in Digital Preservation

ERPANET Training Seminar, Marburg
September 3-5, 2003

FINAL REPORT

Getting what you want, knowing what you have and keeping what you need.

Metadata in digital preservation

Marburg, 3-5 September 2003

CONTENTS

Metadata in digital preservation

Marburg, 3-5 September 2003

Introduction

The seminar was opened by Dr Frank Bischoff, director of the Archivschule Marburg, who emphasised the importance of organising seminars in the area of digital preservation and the possibility to meet each other for discussion.

The briefing paper for the seminar indicated that 'preserving the right metadata [are] key to preserving digital objects', and if numbers are an indication of agreement, the about 73 people who attended the ERPANET seminar in Marburg[1] from 3-5 September 2003 agreed on the vital role metadata play in digital preservation. The participants came from 21 countries in Europe and North America to listen to the 12 speakers from Europe, Australia, New Zealand and Canada, who represented different disciplines and offered a broad range of perspectives including libraries, archives, records management, government and information technology.

The Archivschule Marburg co-hosted this seminar with ERPANET and its success owes much to their efforts. The seminar, which was reported in the local press, benefited from beautiful weather, casual dinners and a guided tour through the medieval town of Marburg . The ambience fostered relaxed conversation between the participants about metadata and digital preservation.

The seminar's goal was to provide information on the latest developments in the area of metadata and to facilitate discussion on the key issues and needs of different communities as they attempt to preserve digital objects through time. In many activities in these domains appropriate metadata sets are being developed and implemented. One of the main areas that this seminar hoped to address was to identify the commonalities between the communities and possibilities of information exchange - in short, the key issue of interoperability. Apart from that it also aimed at promoting better understanding and examining the possibilities of closer collaboration. The many distributed resources kept and managed by different

[1] 'About' because a number of unregistered students from the Archivschule joined us.

communities, organisations and institutions are increasingly accessible through the world wide web and this requires not only intelligent search facilities across domains and sources, but also advanced possibilities to exchange information.

The program consisted of four parts and discussed existing metadata initiatives and key issues, interoperability, standards and schemas, and finally implementation issues.

In between the sessions participants formed small breakout groups to discuss practical issues in more detail based on some leading questions. The sessions are an important and much appreciated part of erpaEvents as they offer a good opportunity to exchange information about practical experiences and a chance to promote interaction between participants and speakers.

Executive summary

Metadata are for documents or other information resources as water is to human beings. Lack of water will lead to dehydration and malfunctioning. If there are insufficient metadata one can't find or understand information sources any more. These documents will then wither until they are forgotten and probably deleted. This (often) unintentional loss of information and knowledge may have significant and costly consequences for an organisation.

Participants from different communities met at this seminar to discuss issues to keep information sources valuable and viable as long as they are needed through the creation and management of adequate metadata.

These discussions grow even more pressing as the possibilities and needs for exchange of information increases, and the barriers among separate, previously isolated and independent domains fall away and interdisciplinary research and work flourish. The interconnectivity offered by the World Wide Web enables new opportunities for improved and more creative communication, collaboration across domains, the sharing of experiences and knowledge among individuals from different disciplines as well as the discovery of new ways to carry out activities and provide services. However, these benefits will only be possible if organisations and communities make a considerable effort in the area of information management and more specifically metadata management. Many initiatives already exist, but they need to become better coordinated and information about them disseminated.

The main conclusions and recommendations of the seminar were:

- Senior managers have to be (made) aware of the consequences if their organisations do nothing. They have to take responsibility and ensure that metadata management is an integral part of their business and information management. The emerging e-business and e-government environment and the organisational consequences of this new situation will require that information management and the metadata issues move to the centre of an organization's strategic management initiative.
- More international and cross-domain collaboration is needed to achieve better interoperability. Metadata registries represent one approach that will improve understanding and knowledge about the various initiatives in the different

communities and countries. Furthermore registries help to document the meaning of the various metadata structures and content across time.

- The role of vendors in the management and preservation of metadata was highlighted. Vendors provide the software tools or systems to accommodate the needed metadata requirements and services. Until recently vendors have failed to become sufficiently involved in this area. In this respect XML was regarded as an important development, but one that is insufficient without the requisite software tools’ and systems’ development.
- Finally, the workshop participants called for more case studies focussed on metadata implementations. These case studies should provide us with better insight into practical experiences with the application of metadata sets, legal implications and/or impediment, the type and degree of the organisational change needed, possible new methods for implementing metadata management, and new requirements for systems development.

Perspectives and key issues

The first speaker, Wendy Duff from the University of Toronto, gave an overview of metadata initiatives and the many issues related to the topic. She identified and explained the different definitions of metadata that exist, as well as the different views that individual domains have of metadata and the various purposes and functions that metadata fulfil. The introduction provided an overview of a specialist world, which set the context for the seminar.

Metadata for preservation represent only one of many perspectives that exist; and this perspective is not exclusive. Professor Duff stressed the importance of being aware of the many perspectives on the topic and the many purposes that metadata serve as well as, the necessity of collaborating and collocating these different views into a larger picture that covers all perspectives and enables interoperability.

However, this is not a simple issue. The various views of metadata paint a complex picture and raise numerous questions. For example, how should existing sets of metadata be compared? No one schema fulfils all functions. Furthermore, creating crosswalks is a difficult, if not an impossible task because of the different contexts in which schemas are created and used. Each view on the world has its own merits and this view must be understood in order to make sense of the related metadata schema. Will standardisation be a possible solution and if so, what degree of standardisation is needed? On the other hand do or will so-called intelligent agents or search engines like Google mitigate the need to standardise? At the present time this is very unlikely because they only access part of the Web -- although many users are unaware of the ‘deep’ web that remains beyond their reach. Who will create metadata? Humans are neither the most consistent creators of metadata, nor ardent to do it. They have to understand the personal benefits that they will accrue to ensure their participation in metadata creation.

A first step to make metadata schemas viable in the long term may be the establishment of metadata registries that describe metadata schemas.[2] However, who will do this and who will maintain these registries? There are some initiatives in

[2] A schemas is a framework that specifies and describes a standard set of metadata elements and their interrelationships. Schemas provide a formal syntax (or structure) and semantics (or definitions) for the metadata elements. A metadata registry in turn describes schemas (their purpose, usability, background, version etc.) in a consistent way, so it becomes possible to compare them.

this area such as UKOLN, InterPARES and the Digital Library Federation. But the question of how best to describe the schemas remains. The answer to these questions will reflect a variety of perspectives. However, a registry is key to understanding and analysing existing metadata schemas.

Professor Duff raised many more issues for consideration:

- How much metadata do we need?
- Who will be responsible for their creation?
- What are the costs of creating metadata and what benefits justify these costs?
- To what extent can existing metadata be re-used in other contexts?
- Do metadata have intellectual property rights?

With this introduction she had set the scene for the seminar that began the process of analysing these issues in more detail.

Steve Knight of the National Library of New Zealand was the second speaker of the first session. The National Library has been working hard on preservation metadata, and he presented their work to-date on their metadata set and the schema they have developed. He pointed out that one of the driving forces behind preservation of digital objects in his country is their relevance to the identity of a nation. Therefore the library, although small, has a central role to play. Digital preservation in New Zealand also includes web-documents and publications. In 2002 and 2003 the Library published a metadata schema for preservation that built upon the Cedars and OCLC/RLG proposals.[3] The model accommodates different kinds of objects, e.g. single and complex objects as well as object groups.[4]

Extraction or collection of metadata by automated means is an important issue. An analysis of the model found that the library could automatically capture up to 90-95 % of the identified metadata elements. Subsequently they built an extraction tool, but they still face the issue of convincing vendors to incorporate this process into their products. This issue is one that must be tackled jointly in order to place enough pressure on the companies. Another one is automated creation of metadata.

[3] See http://www.natlib.govt.nz/files/4initiatives_metafw.pdf.

[4] Interestingly, the New Zealand National Library preserves only the ‘preservation master’ and not the derivates.

In a discussion of the supporting infrastructure for digital preservation Steve Knight identified some points needing further attention: the assignment of unique identifiers, the concept of trust in relation to trusted digital repositories, and the user interface. In looking at the issue of establishing trusted digital repositories a participant suggested that the reference model for the Open Archival Information System (OAIS) be used to enable or support an auditing process in connection with technological and financial sustainability.[5] An audit process based on OAIS would open up a certification of digital repositories. The combination of trust, resources and sustainability is at the core of this approach.[6]

In the following breakout session the groups discussed how to assess the amount of metadata that is sufficient, what metadata elements are needed, and what other key issues should be taken into account? Some participants took the position that more metadata is better than less. In the end there was general consensus that determining the amount of metadata that is sufficient depends upon knowing the business context. Metadata does not stand in isolation, and is dependent on many factors which include present use, and possible future uses, as well as preservation. This view is complicated with the possible trade-off between the creator and users (some of whom may be a researcher or an auditor or legal party). A distinction may be made between primary (i.e., business) use and secondary (any other) use.

Of course, it does not matter what form the metadata takes if it is not linked properly to the object it is describing. Links must be persistent. This also depends on the way metadata are stored, that is, separately or as part of the digital object. Note however, that metadata are meaningless without the object and at the same time, the digital object has little meaning without its metadata.

The importance of metadata on the one hand or the risk of information loss on the other depends on the purpose of preservation. Reasons could be the need of evidence or retrieval and reusability of valuable information. Methods to avoid loss should include:

[5] For further information on the OAIS model, see the documentation included in the ERPANET Seminar report: http://www.erpanet.org/www/products/copenhagen/ERPANET%20OAIS%20Training%20Seminar%20Report_final.pdf.

[6] See also the work being done by RLG/OCLC in this area: http://www.rlg.org/longterm/certification.html.

- automated metadata creation and extraction with involvement of professionals,
- establishing appropriate and solid procedures,
- fostering interoperability, and identifying what core set of metadata is needed,
- engagement of vendors in developing appropriate tools and standardisation, so information objects will have appropriate metadata. This includes the need to articulate the requirements properly to the vendors by the communities involved.

With respect to authenticity participants discussed numerous issues such as how to secure the metadata. Could a digital signature or another kind of seal serve this purpose? Participants agreed that security is a major issue, especially, for example, the need for securing the process of transforming one digital format or platform to another. Also needed is a description of the essence of the digital object or record. The Cedars project labelled this as the ‘significant properties’ of a digital object.

However who will decide what is significant or essential? In some countries like Sweden and the Netherlands, for instance, the judge decides in each individual case whether electronic records are acceptable as evidence. In the US the Food and Drugs Administration (FDA) has established a quasi standard for the minimum metadata requirements for evidence. Additionally an organisation itself has to identify what is significant information for its business.

Whatever the significant properties required are, metadata serve an important role in providing circumstantial evidence for authenticity. Furthermore, the purpose determines how much metadata is necessary.

An approach for safeguarding the authenticity and integrity of the digital object taken at the National Archives of Australia is to retain the original bit stream of the digital object next to the ‘normalised’ version in XML. The latter is considered to be the ‘archival master’.

The discussions that took place in the breakout session show that preservation metadata is a complicated area and that limiting the focus to one or two questions is difficult. Each question raises other questions. Many issues are interrelated which makes the discussion not only very difficult for people to participate in, but also to follow. The discussion also made clear that preservation metadata cannot be singled out as a separate entity. Each perspective on metadata is as valid as any other. The difficulty is how to compare and reconcile these different perspectives and the metadata sets they suggest or require? For example, similar metadata elements may be used for different perspectives in different sets, but we have to achieve a common understanding of their meaning: the issue of semantics. This situation also requires that the experts translate the often very theoretical issues into understandable and implementable guidelines as well as into tools or instruments that organisations and vendors in line with their business can use.

Different communities, different needs and interoperability

The goal of the second plenary session was to get an overview of the different views and purposes of metadata. Four speakers from different communities (libraries, archives, science and e-government) were invited to speak about their background, approach and experiences. The session concluded with a discussion panel that gave participants the opportunity to ask questions and provide comments on the issues presented.

The first speaker, Heike Neuroth from the Göttingen State and University Library, discussed why and how libraries approach digital preservation, including the metadata issue.
In Germany this mandate is divided between different libraries. She mentioned the role libraries play in the long-term preservation of publications, including the emerging new types of digital resources. The heterogeneity of digital formats raises problems for libraries, as it does for other custodial organisations. The Springer publisher for example delivers about 15 types of .TEX formats.[7]

As a metadata description standard Dublin Core (DC) plays an important role, while the recent METS and XML are increasingly being used as exchange formats. Interoperability and standardisation are needed to improve retrieval. On the other hand, the longevity of standards remain an issue. For instance, will DC still exist in another 5 years? Standards will likely change over time and any modification or adoption of new standards will require considerable effort. Apart from this temporal issue, the co-operation and co-ordination between producer, distributor (e.g. publisher) and provider (e.g. library) raise other challenges. We need to achieve much greater gains in this area. For example, long term preservation requires a degree of willingness on the part of the publisher to supply specific or standardised formats.

Furthermore, if producers are made aware of the importance of metadata for preservation and other benefits (e.g. better access possibilities) they may be willing to co-operate. Pre-prints in the mathematical sector have shown some success in a related area.

[7] TEX is a mark-up language for scientific documents with formulas.

An issue raised by Heike Neuroth was also the need to have shared repositories, where digital objects are stored in more than one place. This redundancy may help in safeguarding digital information. This requires the consistent use of metadata.

The next speaker, Malcolm Todd of the National Archives UK, discussed the metadata issue in the context of records management and archives, as well as in connection with e-government, as highlighted in the title of his presentation. He observed that people responsible for developing e-government do not see preservation as an issue. The main reasons for this may be that expected benefits are rather long range, substantial investments are required, and a business case for return on investment is difficult to make. Other drivers for metadata besides e-government are resource discovery, digital preservation and records management, which are not mutually exclusive.

He suggests a three-layered model to connect the different but partly overlapping domains of digital preservation, records management and resource discovery. In this model digital preservation deals with mainly technical issues and metadata at the object level enables migration or emulation. Records management deals with what other communities sometimes call ‘administrative metadata’, but what is better thought of as metadata about the activities and procedures of digital objects, as well as metadata which supports the structuring of information and describes the interrelationships between records. Finally, resource discovery metadata enable searching across domains. The Dublin Core initiative plays a dominant role here. The DC model is rather a ‘flat’ model and supports single objects, such as webpages. From the perspective of resource discovery other metadata are referred to as ‘administrative’ metadata, but this type of metadata is essential for records management as well as for digital preservation despite the dismissive term that is used for it. Another issue is the linking of object oriented resource discovery metadata and the much broader and multidimensional records management metadata. This will be necessary, because DC has been adopted world wide for webpages and publications on the web. However, attempts to include records management metadata into DC are doomed to fail, because they pursue different objectives and describe different types of objects. Nonetheless some degree of compliance with DC is still possible from a records management perspective.

Within the European Union there exist different initiatives with respect to metadata, such as the set of Model Requirements for records management (MoReq), MiReg (for information exchange in e-government), and national approaches with respect to metadata, such as the UK e-Government Metadata Set (e-GMS). They come from slightly different backgrounds and therefore there are issues with interoperability as well as problems at the sub-element level, where the differences between schemas become more apparent. There will always be different schemas for different purposes, so we need approaches that can reconcile the differences and make the various schemas interoperable.[8] Metadata capture should be automated as much as possible to avoid extra burden for users.

Finally Todd emphasised that records management metadata should be predominant to ensure the reliability and accessibility of the resources in e-Government.

Denis Minguillon from the Centre National d'Études Spatiales (CNES) in France offered quite a different perspective. The scientific data used in space research and technology organisations have to some extent different requirements. The requirements start with the recognition that space data are expensive to produce and therefore need to be preserved as well as kept understandable and accessible. Several methods and approaches are developed for these purposes within the space agencies and associated industries community (gathered in the Consultative Committee for Space Data Systems, CCSDS). They focus on data packaging in a way that enables interoperability. In order to achieve interoperability data are described both at a syntactical and a semantical level. An example of a syntactical description language is the EAST (Enhanced Ada Subset) language, which has been an ISO standard (ISO 15889) since 2000 An example of a language for describing semantics is the Data Entity Dictionary Specification Language (DEDSL) which became an ISO standard (ISO 22643) in 2003.

One of the principles of the EAST language is that structure of any type of data can be described as a tree with branches and leaves. DEDSL describes the semantic characteristics of data as values of attributes. This uses trees, leaves, and nodes. In order to enable easy use and application of the languages a tool has been developed called OASIS. All tools for applying EAST are freely available and use XML.

Although the space research and industry community has dealt with some aspects of preserving and keeping data accessible and understandable over time, the

[8] See for example the Warwick framework, http://www.ukoln.ac.uk/~lisap/BIBLINK/wf.html.

community does not discuss the process of preservation and maintenance. Is the need for information (metadata) about these processes not needed? To be able to know some key information such as whether the data are reliable and where they come from, it will be necessary to have methods that capture these metadata as well. In this respect the space and/or scientific community can learn from other disciplines such as records and archives management.

The last perspective presented in this session was of e-Government. Palle Aagard of the Danish National IT and Telecom Agency explained the approach the Danish government takes in achieving and enabling the 'free flow of information'. The key word to free flow is again interoperability, which should enable easy access to, and use of different information sources across government. Better interoperability will also minimise costs, because it facilitates the re-use of information and the use of automatic tools. In order to achieve this ambitious target existing metadata models have to be harmonised and standardised. However, practice turns out to be difficult and unruly. Only a few models exist, while at the same time international standards are not well known, let alone used. Furthermore, Danish initiatives to introduce e-government are also not well co-ordinated and lack a common framework. This results in isolated islands of information. Over the last two years efforts have been undertaken to establish better co-ordination, to improve standardisation, and to begin an XML initiative. A special working group developed XML schemas, such as XML schemas for documents and legislation, which are made available through the 'infostructure database'. Metadata models and a metadata core for exchange of information between ERM systems that use international standards, such as Dublin Core, MIReg and MoReq are also available.

One of the obstacles the working group(s) encountered is that it is easier to develop a model or schema, but it takes a long time in convincing people and achieving consensus on it. Nonetheless some interesting results have been achieved, among which an encoding scheme for authority files that will describe institutions and organisations, and a comprehensive metadata model for ERM systems. One of the lessons of this work is to make use of local expertise in establishing standards as it boosts their support.

After the presentations Seamus Ross chaired a discussion panel, during which one of the observations was that despite the different communities, there are many common concerns with respect to managing and maintaining information resources.

However, a single solution that solves the challenges in these different communities may not be the answer. The needs and purposes of each community may be too different. Nonetheless, co-operation between different parties is emphasised, where it is possible.

Dublin Core may serve as a minimal common language for discovery that can be used as a start by the 'digital tourist' searching for interesting information. However, Dublin Core was not designed to be a metadata scheme for preservation purposes. In that respect the set of preservation metadata as proposed by OCLC/RLG may serve as a core set. As it originates from the (digital) library community, again, the question may be raised to what extent this serves the needs of all communities. We may need a more European or even international point of view, in addition to the US one, as well as the input of other communities to broaden the support for a scheme in this area.[9]

Other questions made clear that there exists also much confusion about the MIReg and MoReq initiatives within the EU, including how they relate to each other as well as to other international initiatives, and finally what their objectives are. In the Netherlands MoReq together with the DoD 5015.2 standard served as a basis for developing a set of functional requirements that should accommodate the application of the ISO 15489 records management standard.

Furthermore, one participant pointed out that it is always necessary to understand first the purpose and background of a standard, before applying it. This not only prevents misuse, but also avoids wasting time and resources in modifying it if its application does not work.

More attention has to be paid in this respect to the risks and/or the benefits versus the costs of developing metadata frameworks. Little is known about these important aspects or the various projects in this area. More systematic research on costs and risks/benefits is required to help decision making. Not only do we need discussion on whether or not work should be done, but also how it should be done. The increasing interconnectivity of society and the interrelationships between communities for instance require a much more cross-domain and cross-organisational approach than in the past.

Finally, some observations on the human factor in relation to metadata were made. This relates not only to the creation of metadata, but also to their management and use. Creation of metadata is usually prone to errors, and with no instantly recognisable benefits it is hard to motivate people to take time and care. This affects the creation of metadata. Numerous recommendations for solving this problem by supporting metadata creation and capture with automated tools were made. This requires, however, a sophisticated electronic working environment where appropriate metadata frameworks are in place. Unfortunately, that is often not yet the case and therefore it is difficult to convince users of the benefits they will accrue if they create metadata. These benefits mainly relate to better retrieval, use and understanding of information resources available in an organisation and within the performance of (business) activities that people are responsible for or involved in.

[9] Apart from the RLG/OCLC initiative to develop a preservation metadata set, the DCMI also established a working group to develop a preservation metadata set based.

Standards, processes and schemas

The topic of standards and their applicability was previously noted, but was further discussed in two presentations representing different perspectives. Michael Day, from the UKOLN project of the University of Bath (UK), presented a typology of standards with respect to preservation as well as some observations about the implementation and sustainability of standards. The categorisation showed a distinction between conceptual and practical standards and Michael Day noted a gradual evolution towards more practical standards, such as VERS, METS, the New Zealand National Library preservation metadata set, MPEG-7, and so on. These provide mostly also XML schemas. These standards all use XML though they represent different perspectives. Michael Day stressed that these experiences urgently need to be shared.

He highlighted the need to learn about the practical value of these standards. Implementing standards is one method to investigate them and much work is underway in this area, for example the OCLC/RLG preservation working group is investigating issues that arise from implementation.

An important practical issue is costs that should be considered in relation to risks of data loss and the need to recover the data. People often see metadata creation and maintenance as expensive without taking into account the consequences of possible loss of information. There needs to be a proper balance between them. Minimally we need robust selection criteria and to increase people’s awareness possible ways to re-use existing metadata.

The issue of interoperability was addressed in a discussion of the possible role of registries in enabling a better understanding of existing standards including their structure and semantics. A registry may be part of the reference model of the Open Archival Information System (OAIS), as it should support different functions within that model. As such, the registry function should be part of the infrastructure for metadata.

Andrew Wilson of the National Archives of Australia presented an overview of developments from ‘Down Under’. The developments focus on resource discovery and records management, especially in standards setting. The Australian Commonwealth government has adopted the Australian Government Locator Service

(AGLS), based on Dublin Core, and have added an extension to it. This standard is mandatory for Commonwealth organisations and should enable interoperability and cross government web based information resource discovery. The New Zealand government has also adopted this standard. An observation was that some people who use DC have unrealistic expectations about it. However this remains a problem of people's expectations not a fault of DC.

In the field of records management Australia has been very active, both in doing research and setting standards.[10] The research at Monash University in Melbourne, e.g. the SPIRT project, has developed a framework for standardising records management metadata. Using this framework the Australian National Archives has established a standard metadata set for records management which includes 20 elements, and is interoperable with AGLS. Unfortunately the records management metadata standard was developed after DC has been applied in developing the AGLS,. A similar standard has been adopted at the State level by the New South Wales government. The National Archives is also active in the AtoR project (*Archives to Researcher*), which pays special attention to the audit trail as part of preservation metadata. The metadata requirements are not yet established though.

The National Library of Australia has been very active in the area of preservation metadata. The framework for this work was the OAIS reference model. However, the Library has not yet established a standard. Apart from being involved in metadata it has also developed guidelines for digital preservation for the UNESCO.

Andrew Wilson made a special recommendation for developing application profiles that will help leverage the (re-)use of existing standards, allow interoperability and formalise the adaptation of standards. These application profiles also can be shared through metadata registries.

The second practical session was dedicated to the omnipresent topic of interoperability. The questions discussed were, among others, how to co-ordinate different efforts and achieve interoperability? What methods or approaches should be taken, e.g. through standards or a meta-language (meta-level), or creating a new environment such as the semantic web? How to keep metadata meaningful over time?

[10] In Australia the term recordkeeping is used for records management.

The different groups came up with the following observations and views:

- What does interoperability really mean in this context? To what degree is interoperability necessary? Answers to these questions are important to be able to carry out adequate activities. There was, however, a warning that international definitions (e.g. of types of objects, metadata) won't work, because of the different needs of different communities.
- Standards may help in achieving interoperability, even if they are not at a mature level. Since preservation is still in an evolving phase it is now time to act. XML can be seen as a baseline in standardisation.
- Interoperability only can be achieved through co-operation and a proper understanding of each other's standards or schemas
- Maybe ERPANET could assume the role of co-ordinating these efforts

Interoperability is not simply a technological matter. As well as developing technology we need to raise awareness and training.

Implementation issues

The last session was dedicated to some examples of implementation of metadata frameworks, both at a logical and technical level.

In the domain of science publications there is a pressing need for the exchange of information among scientists. Thomas Severiens of the German Institute of Science Networking discussed the approach taken in Germany to tackle that issue including the aspects of metadata and interoperability. He depicted a landscape of 'Information Mountains' consisting of full text documents at the bottom, information about their structure in the middle and on top standardised metadata. If people want to share the documents networking is needed and for that purpose metadata play an essential role. During their life different layers of metadata are added to text documents. First the author adds metadata, followed by the institution, and then the publisher or local library and finally the preservation institution.

He provided examples of projects that use the different approaches taken in projects such as PhysNet, eDoc, Physik-Multimedial, and Vascoda.

The PhysNet developed a metadata tool for authors called Meta-Maker, that allows authors to make an abstract, keywords and provide other some information about the creator. The eDoc project of the Max Planck Society installed a central eDocument server for all connected institutes representing a variety of disciplines. To facilitate the exchange of information it established a common metadata set based on international standards. The metadata are represented in XML-schemas. It is also possible to export and import collections and metadata with the use of XSLT. The versioning and bundling of documents, however, is still problematic.

It was interesting to hear of the cases in which users create their own metadata, because the current experience is that users will not do this. However, the example given was the attribution of simple free text key words to documents from the (limited) viewpoint of the creator. In the VASCODA project the emphasis is on the semantics of metadata to enable shared web services and networking. In his conclusion Thomas Severiens stated that XML is not a solution to everything. It is a facilitating instrument that provides another dimension, but there is the need for a semantic level to fully support interoperability .

Bill Roberts (Tessella) presented another view on the topic based on his experiences in implementing systems that support archiving of information, such as the UK National Archives digital archive and the Pfizer Central Electronic Archives. His view reflected a technical perspective based on a workflow of managing digital objects, consisting of collection, import, storage, search, view and export.

He identified issues around collecting or capturing metadata. They include:

- some metadata has to be captured manually, but the user needs to be supported by tools to avoid mistakes
- it is necessary to avoid duplication in the case of hierarchies of records
- in a user environment automated tools embedded in workflow and business processes should support metadata capture. The Pfizer Electronic Archives uses a small basic set of metadata. It still needs to improved and made more accurate, though.
- information about the management processes should be captured
- and file formats should be analysed automatically as for instance is done at the UK National Archives.

XML can be used as a transfer format and it is possible to link metadata to the file during the transfer of information. A real problem may be the efficiency of XML with large transfers, because in these cases XML turns out to be very expensive to process and requires huge system capabilities.

As soon as the digital objects are imported and checked for viruses, file format, etc. these objects must be kept safely in a secure storage environment. The issue of maintaining the links between the metadata and the objects while ensuring efficient information retrieval remains. These issues can influence the way information is stored. The two main approaches are encapsulation of the objects with their relevant metadata or separately maintaining a database with retrieval information. The latter enables easier and faster access. The usual approach, certainly with huge volumes of data, is to have the metadata in a database and the files or objects on a file server. The system at the UK National Archives uses this architecture. The metadata are stored as XML documents in databases, while a subset is identified as searchable and indexed in a text-based index. Furthermore a few key elements are stored in tables that are indexed as well (e.g. on the unique identifier). The file contents are at the moment not searchable. The record and metadata file are kept separately and the links between them are managed in a database. There is an unlimited depth of hierarchy between the records, so records can contain sub-records and in turn these sub- records can contain sub-records and so on and so forth. There is a flexible relationship between records and computer files, as a record can consist of one or more computer files. Finally the design allows easy extension of metadata elements.

An alternative approach is taken by the Australian VERS project, where the digital objects are encapsulated in an XML structure with metadata. The advantages of this system are that the record is self-contained and it is easy to apply digital signatures on both the content and metadata. However the disadvantages are adding or editing metadata is difficult and more de-normalisation is needed for access. So, a dilemma between safety and ease-of-use requirements arises. The practical approach to achieve these two objectives is the use of both encapsulation and a database, but that entails the issue of consistency.

On the issue of interoperability Bill Roberts highlighted the lack of experience with interoperability and that XML may help, but only in a limited way. The point is that schemas may be similar, but will not be identical, and a schema may be implemented in a variety of ways. Semantics is the issue. What are needed in the short term are mappings between schemas. In the long term more standardisation and semantic based approaches may be the way, but this will be difficult to achieve. There will be no ‘one-size-fits-all’ schema and schemas will evolve over time. Therefore version control will be necessary. Certainly from a preservation point of view ensuring understandability, of not only of the content, but also of the metadata will remain an issue. A possible approach may be the addition of layers of metadata to the object.

In his conclusions Roberts suggested that digital preservation is still a ‘young’ discipline and that we still do not know which is the ‘best’ approach. Therefore, he recommended that we carry out more projects and experiments and learn from those experiences, as well as design systems in a flexible and modular way that will better enable and support evolution. Ultimately, the digital object has to outlive all implementations.

The final speaker provided an example from a Swedish government organisation, the Swedish National Insurance Board. Lars-Erik Hansen discussed the reference model developed for managing documents and records within this organisation. The National Insurance Board is a large organisation distributed throughout the country that deals with all social insurance for families, children, old people and in case of illness. Within the context of e-government citizens should be allowed to approach the organisation electronically and therefore a better enabling infrastructure is necessary.

They have built a metadata reference model for electronic document management (EDM) that serves three purposes: the business process (the case worker), the records management administration, and the archives. The archival part of the systems is using the ISAD/G standard which enables different levels of description. The core focus is on the digital object (one or more documents) and from the object different levels of aggregation are identified and described. This case shows a practical application of an archival standard in an early stage of the records management life cycle.

The metadata are used for controlling the work process and the related case management system as well as improving information retrieval.

In the final session discussion focused on the issues of developing a business case and the costs aspect of metadata. Questions raised were, how to make it work, how to make them effective and meaningful to people (e.g. managers), what messages do we tell the different parties involved, and how to create and manage metadata in a cost-effective and feasible way?

To get the attention of responsible managers or other people the message should not solely focus on metadata. More convincing arguments will be necessary and they need to be embedded in a broader business context that appeals to managers. Several possible incentives or triggers for attention were identified including the risk of losing valuable information more efficiently carrying out business processes, auditing, less search time for documents, avoidance of court cases, more reliable documents and even a legal mandate. To some extent they all relate to costs and benefits. Unfortunately, management is not really aware of these issues and their relevance yet. Buzz-words like knowledge management and content management lever more attention than metadata. The ‘what, if not…’ question may be helpful, but also more positively worded business cases that use/speak the language of management should be developed to support proposals for funding.

Finally, organisations have to deal with all sorts of financial, political and organisational constraints, that will influence and shape the approach or policy taken and how that will be implemented.

Conclusions

The topic of metadata is an important, broad, and yet very specialised area. Everybody agrees that without proper metadata the management of information, publications or records, digital or traditional, is impossible. The issue of metadata therefore is relevant in many communities and also includes many different perspectives. As a result metadata have to accommodate, support and meet very different needs. It starts at the level of one community, such as libraries, science, archives, cultural heritage sector, e-commerce or e-government, in each of which many initiatives related to metadata are born and developed. Within each community it is already an issue to reach a certain level of consensus of what is needed and to enable communication. On top of that the Web-environment enforces each community also to look at what is happening outside its own domain. The Web requires the exchange of information among communities and cross-domain searching facilities that can only be achieved with appropriate metadata frameworks. Access to Web based documents provides an important incentive for co-ordination and collaboration in the area of metadata. The dominant Internet metadata standard is presently the Dublin Core, which although useful reflects only one perspective and has a rather limited view on the world of information. Search engines like Google facilitate searching well, but these search tools are also not sufficient. They only cover a small part of the information sources on the Web though this is not always apparent to users. These tools in many cases will provide a sufficient retrieval rate and their limitations may not be an issue. However, it is not only about retrieval. In many areas, such as e-government, e-commerce and cultural heritage, access to and exchange of *reliable and trustworthy* information sources are essential. That requires more robustness and a broader view. Good and adequate metadata schemes are necessary and important instruments in these environments.[11]

This brings us to the key issue of interoperability and the questions that arise from it. It has at least two important dimensions: the interoperability across different information domains or communities and the interoperability through time. In other words will future generations still be able to find and understand (the context of) the current information sources? Communication between different countries or communities of people requires translation mechanisms, and the same goes for

[11] A recent discussion on the pros and cons of DC can be found in: http://www.digicult.info/pages/pubpop.php?file=http://www.digicult.info/downloads/dc_info_issue6_december_20031.pdf.

communicating and understanding information objects from different contexts or sources and from different times or episodes. The first activities for enabling interoperability are emerging, e.g. the semantic web and its community-specific ontologies or the ABC-model. It is, however, not clear whether approaches like that will be adequate and help solve the issues discussed or that they are complementary solutions. Probably more and other approaches are required.

More and broader international co-operation is needed. Similarly, projects need to be cross-domain and include communities beyond archives and libraries. Greater sharing of experiences and knowledge is necessary and in that respect the WWW is important because it facilitates the discovery and sharing of knowledge. It eliminates the need to reinvent ‘the wheel’. However, the knowledge from one domain may require translation into the language of other domains before it will be sufficiently applicable.

Metadata registries, which describe metadata standards or sets in a consistent way, will be an important instrument to enable interoperability and have a large role to play. They will disseminate knowledge about the existing standards, and improve communities’ understanding about the purposes and origins of the various schemas. Registries will be an important tool also for analysing schemas and they will possibly facilitate the reuse and sharing of the schemas across domains. They also may help in keeping track of the history of metadata schemas and as such support the enduring reliability and understandability of information objects.

There will be no overall metadata model or set for preservation, if at all, before we have built up a common understanding and agreement of the issues. Only once a broad framework that incorporates various communities’ points of view is established a basic set of elements on a generic level may be possible. The OAIS information model may perhaps serve as a basis.

Apart from interoperability other issues need to be tackled, such as cost-benefits, and the involvement of two key players, management and of vendors. The importance of management buy-in especially is important, not only for funding, but also for giving the issue of metadata the authority and relevance it needs.
We also need studies that identify and explain the role of vendors. Vendors have to provide the necessary software tools that can accommodate the needs for metadata. However, they need authoritative sources, like standards, to derive from in order to

develop those tools. That again requires a kind of consensus of communities involved. Mark-up languages, especially XML, may be at the moment the best approach available for creating metadata structures, but not the ‘miracle cure’ as somebody put it. There is still a need for further development and research.

Another issue is that more case studies on best practices or innovative approaches are needed. They will also provide better insight in costs versus benefits in the implementation of proper metadata strategies. Little is known about the implications of implementing metadata strategies, though the impression exists that metadata are (very) costly. More research into the costs and benefits may change that image or lead to the development of a more appropriate and balanced image.

Finally, the seminar brought together people from many different communities, and with varying degrees of expertise. The discussions and exchange of experiences were lively, gave new insights, opened up new worlds and ideas and even converted some sceptical people. It certainly provided seed for further thoughts, contacts between people and organisations, and maybe new collaborative projects. One result of the seminar was the organisation of a web-chat session between some speakers and participants.[12]

[12] The report of this chat session has been published separately on www.erpanet.org.