



ERPANET



*erpa*training

ERPANET OASIS Training Seminar Report

Project acronym:	ERPANET	Contract nr.	IST-2001-32706
Title of Document:	ERPANET OASIS Training Seminar Report		

Date Prepared:	13 th January 2003		
File Name:	OASISReport.pdf		Version draft <input checked="" type="checkbox"/> final <input type="checkbox"/>
Date of issue:	13 th January 2003		
Partner (1-4):	1: UofGlasgow	Responsible Director:	Seamus Ross
Primary Author: (add phone, fax & email)	Monica Greenan		T: +44 (0)141 330 3549 F: +44 (0)141 330 3788 E: british.editor@erpanet.org
Distribution:	Internal, until final approval		
Document Status:	Draft, awaiting final approval		
Action Required:			
Publication:	Available in restricted section of website		
Classification Code:			

Revision History

Date	Version	Revision	Author
13.01.2003	1.0	First draft	Monica Greenan
15.01.2003	1.1	Revisions by second editor	Georg Büchler

Table of Contents

Introduction	3
Aims and Objectives of the Seminar	3
Presentations and Discussions	4
Interpretation of the OAIS Reference Model	5
Key Concepts of the OAIS Reference Model.....	5
OAIS Functionality.....	7
Interpretation of the OAIS Model.....	9
Implementation of the OAIS Model	15
Current Applications of the OAIS model – Scientific Data	17
OAIS at Edinburgh University Library	18
OAIS, Boundaries, and Issues	21
Conclusions	24
Appendix One: Speakers at the OAIS Training Seminar	25
Appendix Two: Participants at the OAIS Training Seminar	27

Introduction

Developed by the Consultative Committee for Space Data Systems (CCSDS)¹, the OAIS reference model presents a framework for the long-term preservation of information. In 2002 the model was formally approved as an ISO (International Standards Organisation) Standard (ISO 14721:2002), and is currently awaiting publication. This model has generated a significant amount of discussion and action among a wide range of information creators and preservers, and has been taken as a basis for the development of several high-profile preservation initiatives. The reference model also provides a basis for further standardisation in the area of preservation, establishing common terms, concepts, and significant relationships, and aiming to improve vendor awareness of preservation functions. While the model does not consist of specific requirements, it identifies the recommended processes and metadata necessary to preserve and make available digital information.

There are numerous projects and initiatives designed to investigate the major problems inherent in digital preservation, and the main challenge lies in identifying the necessary data and information management framework and processes. The OAIS model provides clarification and structure to data and information management processes, and an understanding of the model will enable implementers to decompose structures and relationships into meaningful strategies and actions. These are necessary to meet the challenges posed by preserving increasingly complex, distributed and irreplaceable digital resources.

Aims and Objectives of the Seminar

The training seminar provided accessible and transferable training to participants on the OAIS model itself, its functionalities, and its wider potential applications. Participants gained familiarity with the main features of the model, and examined some of the processes involved in its interpretation and application. In particular, the seminar focused on how to move from model to implementable preservation solution. Lecturers shared the experience gained by their organisations and institutions in employing the model in the development of their own preservation initiatives. These ranged from commercial organisations and public sector institutions, digital libraries, archives and scientific data centres. Discussions and analyses during breakout sessions were essential to the better understanding of the model's range of potential applications, as well raising awareness of its limitations.

With over 60 participants from 13 countries, the ERPANET OAIS Training Seminar was an opportunity to discover, analyse, dissect, and discuss the OAIS Reference Model with others with similar concerns. Participants came from Austria, Denmark, Estonia, France, Germany, Iceland, the Netherlands, Norway, Spain, Sweden, Switzerland, United Kingdom, and the USA.

Presentations and Discussions

Numerous experts with a range of institutional backgrounds and preservation perspectives presented focus papers, exploring the model's functionality, application, and limitations. The seminar presented an **Overview of the OAIS Reference model**, explored **OAIS functionality, Interpretation, Metadata, Current applications of the OAIS model**, and **potential limitations**.

These topics allowed participants to investigate some of the OAIS Reference Model's key challenges, and discussions focused on some of the following questions:

- What are the principle concepts and preservation principles of the OAIS model?
- How can this high-level model be applied in practice?
- Why should I build an OAIS compliant system?
- How can the functionality of the OAIS model be built into a system?
- Can this model apply to all types of data and information?
- How does this model and its development relate to other digital preservation systems and initiatives?
- What can be learned from existing implementations of the model?
- Where will the standards and specifications needed to implement the model come from?
- Will there be a certification process for OAIS compliant preservation systems?
- Does this model provide a useful mechanism to preserve digital information in the long-term?

¹ <http://www.ccsds.org/>

Interpretation of the OAIS Reference Model

The first day of the seminar focused on the interpretation of the OAIS Reference Model and was an opportunity for participants to understand the functions and processes and acquire an appreciation of how to move from the model to an implementable preservation solution. The main concepts, preservation principles, and functionality of the OAIS model were presented and analysed.

Key Concepts of the OAIS Reference Model

The OAIS Reference Model was developed in response to the need to preserve scientific data. The document was a response to threats to the long-term storage of data and assumptions made about the future. None of the hardware in existence today will be commonplace in the future, hardware that might exist is not likely to function, no software in existence today will be commonplace, and there will be a loss of knowledge about data.

Assumptions made about the future include:

- the concept of a sequence of bytes will still exist;
- the ASCII (American Standard Code for Information Interchange) character set will still be understood (even if it is not used);
- we will still use a representation of alpha-numeric characters;
- we will not understand such formats as Microsoft Word, GIF (Graphics Interchange Format), or Excel;
- we will not be able to read any of today's media; and
- storage densities will increase.

The OAIS document itself was designed for large organisations such as NASA, that handle very large quantities of data, but smaller institutions will still find it valuable. The model contains very useful information about information and how to handle it.

Understanding the principle that data is abstract is key to the model. OAIS is a generic model because it specifies style, not detail. It will not, for example, specify what bit streams should look like. A conforming OAIS archives will conform in architecture, operational procedures, but not on a hardware level.

Central to the OAIS Reference Model are the concepts of **Representation Information (RI)**, which outlines how the intellectual content is represented and how to extract meaning from a stream of bytes; **Preservation Description Information (PDI)** which allows the understanding of the content over time; the **Designated Community** which outlines the users of the preserved data, and which as a concept itself will change over time; as well as concepts such as **ingest**, **information object**, **Archival Information Package**

(AIP), metadata, and archival storage. All of these principle concepts would be returned to frequently during the two days of presentations and discussion.

There are several responsibilities that an OAIS must fulfil as outlined in the reference model document. The archive has the responsibility to negotiate for, and accept appropriate information from the information producers before submission of information; there must be sufficient control obtained over the information to ensure long-term preservation; there should be a determination of what communities should become the Designated Community; and the information preserved should be made available and independently understandable to the Designated Community. Documented policies and procedures must also be followed to ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated copies of the original, or as traceable to the original.

Participants discussed the concept of the information object (a data object together with its Representation Information) in relation to the model's ability to handle all kinds of data and information. It was pointed out that a data object could be as small as one word, or as large as a database. There will be difficulties in trying to determine the level of granularity of any OAIS system, and in trying to determine how many objects there actually are, which could become problematic for certain types of data sets, such as open-ended meteorological data or financial data. It was pointed out, however, that by considering the unit of management at the level of the information object such concepts and terminology as records and documents could be made redundant, along with the confusion that surrounds them. Many of the concepts will need to be specifically tailored to any institution that wishes to implement an OAIS system. How concepts are interpreted and understood in the context of the information they will preserve must be done on an individual basis of arbitrary and policy decisions.

OAIS Functionality

OAIS functionality relates to the various characteristics and relationships inherent in the model and in the management and processing of data and information. The functional entities can be seen from this diagram of the OAIS model.

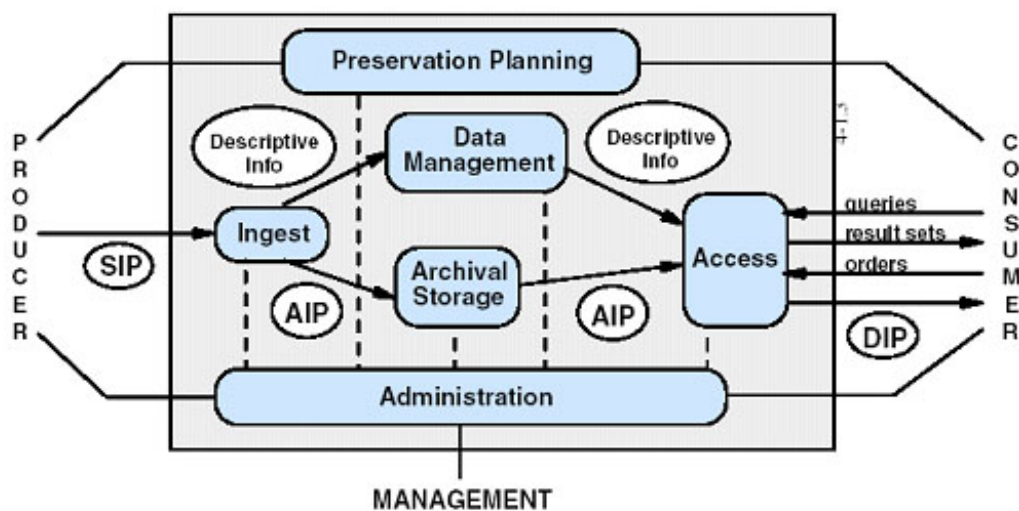


Figure 4-1: OAIS Functional Entities, CCSDS 650.0-B-1 Page 4-1, January 2002

The main functional entities of the OAIS model show the Ingest, Archival Storage, Data Management, Administration, Preservation Planning, and Access Functions.

Several important tasks take place within the **Ingest** function, and quality assurance is vital in this area. Here, Submission Information Packages (SIPs) are received and accepted, and quality assurance is performed to highlight any errors. Archival information packages (AIPs) are created and descriptive information is extracted for the Data Management function enabling the search and retrieval of AIPs. The ingest function also coordinates updates to Archival Storage and Data Management, transferring AIPs and metadata to the appropriate locations.

Archival storage receives the AIPs from the Ingest function and adds them to permanent storage. It manages the storage hierarchy, refreshes and replaces media where necessary, performs error checking, provides disaster recovery capabilities such as duplication of content for off-site storage, and provides AIPs for Access.

Data Management administers the archive database functions, receives and performs database updates (for example when new descriptive information or administration data is provided), performs queries, and produces reports such as summaries of holdings and usage reports. Other activities can be included here such as full text searching capabilities employed by the National Library of the Netherlands.

Administration solicits and negotiates submission agreements, audits the SIPs to ensure they meet the necessary standards, manages the system configuration (hardware and software), monitors the systems operations, establishes and maintains archival standards and policies, provides customer support, and activates requests.

Preservation Planning evaluates the contents of the archive and periodically recommends updates to migrate the current holdings; develops recommendations for archive standards and policies; monitors changes in the technology environment, user's service requests, and knowledge base; designs information package templates; and develops detailed migration plans, software prototypes, and test plans to assist Administration. The preservation planning function was added in one of the later drafts of the Reference Model to provide more of a top-down view of these activities, bringing them together from other modules.

The **Access** function receives requests from consumers (the OAIS term for users), applies controls to limit access as needed, coordinates requests and the execution of requests, generates Dissemination Information Packages (DIPs), and delivers DIPs to consumers.

The numerous diagrams in the OAIS Reference Model also contain other responsibilities, but the ones outlined above are the principle ones. The functions are linked together by a constant report system that is in place. OAIS functionality is already part of what many institutions and organisations undertake in their efforts to manage information, and it overlaps the general functionality of many digital asset management systems.

It is important to remember that the model is a conceptual reference model and *not* a system design model. Essential to remember too is the fact that the functions set forth in the OAIS reference model may not necessarily correspond with the functional modules of a system that would implement the model. In other words, there are a range of potential implementations of the functions of the OAIS model, and indeed actual functionality may be more limited than that in the OAIS model.

Participants agreed that it is essential to understand your own business processes before trying to interpret the functions of the model. In addition it was felt that different types of technology would be needed to implement the different functional areas of the model, which would need at the interpretation stage systems analysis from the information technology community. The CEDARS (Curl Exemplars in Digital Archiving) project focuses on the 'significant properties' of digital objects that need to be preserved (properties which affect the quality, usability, rendering, and behaviour of digital objects) in order to retain the functionality

of those objects². Identification of these significant properties is not a straightforward task, but the answer can be found by examining the designated community and its needs.

Interpretation of the OAIS Model

To interpret the OAIS Reference Model it is essential to understand who and what will likely be involved, establishing what data objects there are to preserve, and to begin to use the OAIS vocabulary to ensure that everyone is referring to the same thing. The OAIS model provides a specific vocabulary for digital preservation practitioners, specific advice on how to sub-divide a complex task, and a logic and structure to allow digital holdings to be visualised and processed. Participants agreed that one of the most important features of the model is the terminology, which allows everyone involved in preservation activities to speak of the same concepts and develop a common understanding. As a high level and complex model, it is encouraging for all preservers to note that when CEDARS project staff first encountered the model it took them several months to grasp it. Much of the model does not need to be understood by the majority of people working in digital preservation; and only some of the detail is required to implement a solution.

It is important to understand how the OAIS model handles data and information. There are three types of information packages: the producer and the OAIS communicate to produce the Submission Information Package (SIP); the OAIS and the consumer communicate to produce the Dissemination Information Package (DIP); and the OAIS preserves Archival Information Packages (AIPs). The AIPs contain both Content Information and Preservation Description Information. Content information is the digital object that is to be preserved, and the Preservation Description information is a description and additional information to explain what the content actually is. Content information itself contains the actual data object and the representation information that makes that object meaningful; together they can be considered the intellectual content. Representative information is needed that will keep the Content data understandable in the long-term.

To preserve and make available digital objects such as e-journals, e-theses, CD ROMs, it became clear to the University of Leeds Library that these objects had to be stored in a convenient form that allowed a library user to download the object which would be a duplication of the storage copy. After some investigation, the OAIS model was found to describe what had to be done in order to preserve and make available the digital objects under their responsibility.

² For more information, see <http://www.leeds.ac.uk/cedars/>

One of the initial stages in interpreting the OAIS model was identifying the producers, consumers, and management, the three 'persons' who form the central OAIS relationships. The producers of the digital objects at Leeds were numerous, including University departments, and e-journal publishers. The consumers were identified as university students, university staff, and researchers. The potential for new consumer groups in the future also has to be considered. In examining the OAIS model, it is also necessary to identify the roles of management. These were identified as: long term equipment planning; review of OAIS performance; ratification of pricing policy; relationship development (between the producer, the OAIS, and the consumer); and the promotion of OAIS uptake within the appropriate spheres of funding. Some conclusions were reached about the roles of management, including the fact that some of the roles were very close to the current roles of the library management, and that a new management group should be formed that was made up of some existing library management and other senior university strategy managers.

In addition to identifying the persons involved, there needed to be an identification of the OAIS system itself. As the library was intending to preserve the digital objects themselves, it would provide the role of the OAIS, including both the archival store and the administration. The archival holdings themselves consisted of both present and future holdings, where present holdings consist of the e-theses, CD-ROM book supplements, and e-journal subscriptions, and where future holdings might include more internal publications and more e-journals.

The Archival Information Package (AIP) also had to be interpreted in terms relevant to the University Library. The Preservation Description Information, and Content Information had to be understood in terms of their own institution. It was concluded that, upon analysis, they did not have all of the components needed for an AIP. They felt confident of having the content data object (part of the Content Information) for all electronic holdings, but only a small amount of Preservation Description Information for the E-theses.

Several lessons were learned from the CEDARS project that were of use in the interpretation of the OAIS model. The main lesson was the identification of *Significant Properties* for the digital objects to be preserved, a concept from the project which relates to those attributes of an object that constitute the complete intellectual content of that object. The identification of these should be done as early as possible. As an example, the significant properties of an e-thesis would include the complete text with divisions into chapters and sections, the layout and style, and diagrams.

It is essential to understand and interpret the functions and processes of the model in relation to actual practice.

- **Ingest** was interpreted as establishing agreements with the Producers, taking the digital data from them (Submission Information Packages), and processing the SIPs into AIPs, recording any current software dependencies for using the Content Data Object.
- **Archival Storage** places the AIPs received from the ingest function into archival storage. The data management database must be updated to keep track of the OAIS holdings. This archival storage function also included the storage, maintenance, and retrieval of AIPs.
- The **Data Management** function was understood as keeping track of the AIPs currently in archival storage and producing discovery information to pass onto the Consumer to allow them to choose suitable AIPs.
- **Access** provides support for consumers and delivering DIPs in an appropriate form for the particular consumer.
- **Administration** was considered to include the overall control of the OAIS working to record and make submission agreements with producers. It also records and implements archiving standards and policies.
- **Preservation Planning** monitors the environment of the OAIS, ensuring that AIPs remain accessible and understandable to current consumers, and develops templates for SIPs and DIPs and other assistance for working with Producers and Consumers.

The organisational view of the reference model was then presented stressing the importance of establishing the designated community. This was defined as the users the organisation services by preserving information for them. In order to preserve this information for them, however, the knowledge base of this Designated Community must be determined and this must be monitored over time for change. Concerns were raised over the nature of the Designated Community, and it was agreed that in many cases that this would be a challenging issue with important repercussions on the preserved material.

All actions must be considered with a long-term perspective, and preservation activity must take into account changing technology and the changing user community. The organisation must decide whether digital objects need to be transformed (e.g. migrated), and if they are, ensure that nothing significant is lost to future consumers. It is useful also to consider alternatives to transformation, and to examine preserving the source code for original software or emulation.

The OAIS Reference Model explores the possibilities of archive interoperability, and this is something that the organisation should consider for the benefit of consumers, producers, and the management. There are four basic models for interoperability provided by the Reference Model: independent (i.e. no interoperability); co-operation (common producers and dissemination standards); federated (most interoperated); and sharing resources (reduce

costs by sharing equipment). A federated archives might involve a central site with distributed finding aids and access aids, but as some participants pointed out, this would raise issues such as unique and duplicate AIPs and the level of autonomy of management.

To prepare a digital resource for lasting preservation it must be given a unique name, assigned metadata, and have identified significant properties. For obsolete data formats, the original byte stream must be kept. The idea of accessioning proprietary formats was discussed, and a participant provided a reference to a format registration project. The preserved Representation Information could result in software capable of rendering the information. The archive management must look out for digital objects that are dependent on rendering software that is about to become obsolete, and should use software preservation techniques to preserve rendering software.

Awareness that a byte stream can be stored forever is considered central to interpreting the model. Complex data streams must be mapped into byte streams and then mapped back again for use. The representation information preserves access to the intellectual content and makes emulation possible in the future. The ends of representation nets (i.e. the basic level of understanding and interpretation that may be defined externally such as plain ASCII text or HTML) must be monitored for obsolescence, having the potential to render much of the representation information useless, and the data object unreadable. The current best archival method of mapping the digital resource into a byte stream must be employed. Metadata must include rendering instructions, format descriptions, representation information, technical metadata, evolving technologies, and representation networks.

Concerns were raised by some participants about the discrepancies between the model and archival practices, the deeper into the model one reaches. The different approaches to the management of different types of materials were discussed, questioning the model's ability to manage all types of information. This was a problem that was returned to in the second day.

OAIS and Metadata

The OAIS model does not provide a metadata set, but a data model, on which metadata sets have been proposed. The OAIS model is both a functional model and a data model. With data objects there has been a new emphasis on non-descriptive metadata as a result of at least two problems: authentication and preservation. The authentication problem has arisen as a result of the ease of transformation and sharing of digital objects, the preservation problem because of the increase in technical mediation which in the analogue world was handled by only one device, but in the digital encompasses many more including software and hardware. Metadata sets proposed include the National Library of Australia

Metadata Standard³ in 1999, CEDARS Metadata Specification⁴ in 2000, NEDLIB (Networked European Depository Library) Metadata⁵ in 2000, and the OCLC/RLG (Online Computer Library Center/Research Libraries Group) Metadata Framework⁶ in 2002.

Julien Masanès presented two different preservation scenarios. The first examined scientific data archiving, highlighting the challenges of working with missions that generate huge amounts of data. Each mission will have its own format and specific structure unique to each mission. The second looked at web archiving and the general harvesting of sites, which can result in vast amounts of different formats under little formal control. These are two very different types of preservation activity, both of which require metadata to enable them to be preserved, managed, and accessed over time.

Representation Information, containing structure and semantic information, is an important concept, specific to the digital era, introduced by the OAIS model. The Structure Information contained in the Representation Information can be considered as a layer analysis which helps keep track of all transformations, decomposing into a physical layer, binary layer, structure layer, object layer, and application layer. The NEDLIB Representation Information Metadata set can be examined in this manner. All of these layers present information in an easily understandable manner. Semantic information is another important part of representation information, but is considered mostly for use by scientific domains. Measures used, such as unit or conditions of measure, are described here to enhance the meaning of the scientific data.

Preservation description information specifically focuses on describing the past and present states of the Content information ensuring it is uniquely identifiable, and ensuring it has not been unknowingly altered. This is made up of reference information, context information, provenance information, and fixity information. Reference information contains information about the archival system identification, global identification, and resource description. The context information is important in cases where an object is in relation with another object because of its content or if there are other manifestations. This information, which contains reasons for creation and relationships (manifestation and intellectual content), can be important in cases of migration for example. Provenance information relates to the life cycle of the document, containing information on origin, pre-ingest, ingest, archival retention, and rights management. For each of these, the event must be documented with information on the designation, procedure, date, responsible agency, outcome and notes. Provenance information can identify and authenticate content. Fixity Information must include the type, procedure (pointer to documentation or tool), the date, and the result.

³ <http://www.nla.gov.au/metadata.html>

⁴ <http://www.leeds.ac.uk/cedars/colman/metadata/metadataspec.html>

⁵ <http://www.kb.nl/coop/nedlib/results/D4.2/D4.2.htm>

Metadata proved to be one of the areas of greatest concern among the seminar participants. Many were concerned that metadata is not static and that there should be adequate allowance in the OAIS for metadata to mature over a period of time. The two main sources of information for preservation planning come from both the inside and the outside. The inside information relates to the metadata with archived collections (OAIS compliant), and the outside relates to the platform and format environment evolution. This environment information needs to be shared, for example by the development of format registries. Urgency is essential: the information that exists must be documented, and formats and technologies must be monitored.

After listening to presentations on the interpretation of the OAIS Reference Model, participants were keen to develop on some of the issues that had been raised. The cost and effort involved in designing a preservation system of this kind was discussed at some length, as the commitment to such a system by the entire organisation is substantial. Risk analysis can be a very important tool, providing both an initial business case as well as limiting expenditure in the long-term. As the OAIS is a generic model, there should only be a one-time cost, and another architecture may be more limited which would result in an organisation having to spend additional money in order to make it fit their information needs.

Participants were keen to examine how closely the OAIS model is related to other digital preservation systems and initiatives. Three other developments mentioned by participants included the Model Requirements for the Management of Electronic Records (MoReq)⁷, developed for the European Union and a useful checklist for active records systems; the MIT (Massachusetts Institute of Technology) developed D-Space⁸ initiative, an archive repository designed to capture, distribute and preserve the intellectual output of MIT staff and researchers, which has been designed to be compliant to OAIS on many different levels; and the LOCKSS (Lots of Copies Keep Stuff Safe)⁹ project, a peer-to-peer polling strategy for electronic journal content. In making decisions about what system is best to consider, the needs of the Designated Community must be paramount.

⁶ <http://www.oclc.org/research/pmwg/>

⁷ <http://www.digitaleduurzaamheid.nl/bibliotheek/docs/moreq.pdf>

⁸ <http://dspace.org/index.html>

Implementation of the OAIS Model

The second day of the seminar concentrated on a variety of applications of the OAIS Reference Model. The aim was to provide participants with actual examples from a range of institutions, highlighting decisions made and steps taken to develop an OAIS preservation system. Participants agreed that there is not sufficient information in the reference model itself to be able to implement an OAIS archive. Best practices are articulated and easily communicated about system design as a result of this document, and this fills in a previous gap where articulation about needs had been difficult. By presenting a framework, the model forces you to think about the components you need to include in any system design. OAIS makes explicit what had been done intuitively up until that point. As one participant pointed out, if nothing else, it is a very useful checklist of components.

The OAIS experience at the British Library

The British Library receives digital materials by voluntary deposit, purchase, and digitisation, resulting in a wide range of digital holdings. The Library required a method or system for the long-term storage, preservation, and access of this digital material. They embarked on developing such a system in 2000. The initial work involved developing a detailed functional specification of a system aligned with OAIS model concepts. The British Library looked to the OAIS model as it provided a good match for the system that they were looking for, and it was used to present requirements to their vendor.

There were a variety of difficulties in putting the OAIS model into practice, as there were no rules about how it was supposed to be done. There is little current expertise available, and no 'off-the-shelf' systems to purchase. However, the OAIS model is well developed and considered to be the guidance for best practice. It also provides an excellent high level framework and convincing back-up argument for political justification for the development of such a system. It also provided a standard terminology for communication and a good match for almost the entire system they were planning to build.

Even though the OAIS sounds like one system, it is not necessarily, or likely to be, one single entity. For example, there were parts already in existence in the Library. It was essential at an early stage to define the boundaries of the OAIS system. There was no formal method of implementation used, but a regular computer systems development analysis was employed. The business processes were analysed and matched to the OAIS functions. Therefore, it was important to establish what else was needed external to the OAIS functions,

⁹ <http://lockss.stanford.edu/>

and what parts of OAIS were not needed, as OAIS does not determine the boundaries for any particular system.

It takes a lot of work to understand the model and terminology, and so a glossary was created to help align terminology. Specific difficulties arose over defining an object, naming preservation users, the concept of packaging information (how and where the bits are stored), content and representation information (how to interpret the bits into data), and preservation description information (reference information, context information, provenance information, fixity information, and how to interpret the data into information). The concept of significant properties, in addition, was deemed vital, but proved difficult to define.

The British Library are in the process of publishing the metadata element set and description on their website. They do not expect that every element will be used, and there is no consensus yet on how each field should be filled in. The Library was aware of other activity in relation to metadata, and some of the metadata groups developed were inspired by the work done by the National Library of Australia. It was felt that the Dublin Core Metadata¹⁰ set was much too limited for their purposes. The Library's main concern is to develop a 'future proof' metadata set, which is not necessarily what they have right now. It is important to record metadata in a way that meets current needs and is useable into the future, which proved also not to be a simple task. The metadata set adopted has 12 core elements and about 30 in total, however not all are fully defined yet.

There were several important issues which the OAIS model did not address, but that the British Library's own system had to face. Decisions had to be made regarding which materials to store in the system, whether descriptive information should be stored internally, whether object relationships should be stored internally, whether a retrieval manager component should be included, and whether an exit strategy should be built in from the outset. The Library did not want to be vendor specific, as there were different expectations and different ideas over how things should be done. Questions also arose over possible changes to metadata, and whether changes should be allowed without delivery and re-ingest as a new item. Questions of object deletion were also not addressed by the OAIS document, and issues of whether to remove the content or the access to the content were discussed by participants. A holding area was proposed as a potential solution, but this also posed problems because of linkages. The nature of the unique identifier is another problem yet to be fully solved, with questions remaining of where it should be generated and what structure it should have.

The biggest problem the British Library had with implementing the OAIS model was the definition of the scope of the system. Despite the challenges that it poses, the Library felt that

the model provides useful terminology and concepts, and is a valuable tool that can form the basis of a strategy for the long-term preservation of digital information. It provides a checklist that allows an institution to make sure that all the necessary components are in place, and also to remind preservers of additional functionality that can be developed in the future.

Current Applications of the OAIS model – Scientific Data

A conforming OAIS archive implementation is described in the reference model document as supporting the model as outlined and fulfilling the list of responsibilities. A roadmap also identifies areas suitable for OAIS-related standards development. However, the model does not define or require any particular method of implementation and does not specify implementation mechanisms.

The OAIS responsibilities include the negotiation and acceptance of Information Packages from information producers, sufficient control over these to ensure long-term preservation, determination of which designated communities need to be able to understand the preserved information, ensuring the information to be preserved is independently understandable to the designated communities, following documented policies and procedures to guard against all reasonable contingencies, and making the preserved information available to the designated communities in forms understandable to them.

The NSSDC (National Space Science Data Centre) project DIONAS (Data Ingest and Online Access Sub-System) uses OAIS concepts and attempts to bring existing data into an OAIS system. It produces AIPs based on an ISO standard format data unit. The authority and data identifier (ADID) is an international object-type-unique identifier that can point to where the metadata is.

NSSDC also provided software to the IMAGE (Imager for Magnetopause-to-Aurora Global Exploration)¹¹ project to convert their 'Level-0.5' data products into AIPs. Consumers will be provided with UDF (Universal Data Format) files as well as software to read them. (This may change in the future, however, as UDF files fall out of use.)

Another example provided was the CDPP (Centre de Donnees de la Physique des Plasmas), which was formed in 1998 and is the general archive for all French Space Agency missions. This archive contains information from about 10 missions, 30 experiments and about 100 data sets, and is predominantly measurement data. The CDPP forces producers to provide the necessary metadata, which can be problematic for older projects and related

¹⁰ <http://dublincore.org/>

¹¹ <http://image.gsfc.nasa.gov/>

data. The designated community has been identified as having a scientific background and knowledge of plasma physics, and data is made available to them over the web. The main problems faced by the CDPP in implementing OAIS include the diversity of the experiments and data produced, the late intervention of Archive in the data production process, the absence of standard archiving structure format in plasma physics, and no imposed schedule.

The final example provided was that of LOTAR (Long Term Archiving and Retrieval and Product Data within the Aerospace Industry)¹². This project is concerned with the storage of a very large number of three-dimensional CAD (Computer Aided Design) models which are in a range of formats and software dependant. In their current state they are not safe for legal and product liability. An archive must ensure that a part produced by the aerospace industry can be verified as conforming to its documentation, data security and protection of data privacy over the life cycle of the archives, and the possibility of auditing the process of archiving and retrieval. In addition, they need to ensure that the information is understandable without the assistance of the information producers. They want to be compliant with the OAIS reference model.

OAIS at Edinburgh University Library

Edinburgh University Library (EUL) is at present in the process of implementing OAIS procedures and is still working on its preservation metadata schema. It was a system that they were keen to proceed with despite its very limited application and implementation within the UK academic library community.

EUL provides administrative services for the university, and must preserve official materials produced. The research departments of the university also create important digital materials, but because of the size of the institution, the library has little control over creation and management of these resources. A survey is currently underway of existing digital resources that may have long-term research value. The Library has no formal remit to archive materials as the result of agreements with producers.

Edinburgh University Library began with a pilot project of a small test-bed archive of the University of Edinburgh calendar, a volume that the university is legally bound to publish outlining courses, terms, and tutors. The plan is to make this document fully electronic with the electronic version the *de facto* one. The decision was made to use the OAIS reference model as their guide for digital preservation practices. The CEDARS and NEDLIB project resources proved very useful, providing models and guidance stressing that the library should build on what it already has in place.

EUL analysed their processes in line with the OAIS model and broke them down into pre-ingest, ingest, archival storage, data management, access, and administration. At the pre-ingest stage, data are received by the library as a submission information package which contains a range of types of information, including some bibliographic metadata, file type information, and other technical details, retention and access details. Most information needs to be sought from the source, however, and as the Calendar is created in a range of different departments, obtaining this information can prove complicated.

At the ingest stage the team allocates a unique identifier, as suggested by the CEDARS project, containing the name of the institution, the name of the archive, unit, and file type. A checksum is allocated (MD5). A byte stream is created using tar, and the AIP is created on ingest. The AIP contains the content information, digital object, and representation information as an XML file, the PDI as an XML file (a copy of which is sent to the data management), the specification of the packaging tar is contained as a text file, and the unique identification is also kept as a text file.

The archival storage is performed by both a service provider and in-house. The library uses OCLC/DOMS for day-to-day usage on server storage, and in-house on optical media with contact only via FTP (file transfer protocol). However, with the system at the university there can be no deletion or alteration of files, and therefore, no updates if this becomes necessary. This optical read only storage is unsuitable for data management, but a copy of the PDI is kept separate to allow for this. Updates will cause a problem, but a new AIP will need to be entered to resolve this for each updated copy, with a reference to the old copy.

Data management proves the most time consuming element of OAIS implementation. Eventually the Systems team in the library will be able to deal with this, and it will be mapped with the user dissemination programme in their digital object management system. Metadata is kept in a database in XML files. There the XML file will be stored, indexed, and linked to the library's OPAC, so that it is fully searchable by the users. The user is given access to the DIP that contains the digital object, software to automatically render the file, and some fields from the PDI such as rights information and access details.

Metadata creation is both manual and harvested. The manual solution is not ideal as it is a very time consuming process. The Library examined the CEDARS, National Library of New Zealand and OCLC/RLG Working Group metadata schemas. Files were created in XML using the CEDARS DTD (Document Type Description). There were a large number of fields that could have been included, but it was key to consider what was important and feasible for the library. The pilot project consisted of relatively simple file formats and so the preservation

¹² <http://www.prostep.org/en/arbeitsgruppen/arbeitsstruktur/lotar/>

metadata would have to become even more detailed and challenging for more complex files such as multimedia.

The OCLC/RLG explanation of Representation Information in which the findings of CEDARS, NEDLIB, and the NLA were synthesised proved to be a very useful tool, as there is little else in the way of published information on this topic. The representation information is divided into two components – structural and semantic, as outlined by the OAIS reference model. The structural information provides a technical description of the organisation of the digital objects, and the semantic information provides information about the making the information renderable to the user in terms of intellectual content. Semantic information was hard to come by essentially because the files to be archived were relatively straightforward.

Fields do overlap in the PDI and CI and they are not mutually exclusive, although it should be remembered that the PDI does strictly manage the preservation process, whereas the RI translates the intellectual value of the object. The New Zealand metadata set has metadata modification fields that could prove to be very useful, and the other detailed fields for datasets, video, audio should also be taken on board.

EUL is now looking forward to implementing their new Digital Object Management System (DOMS). In a joint partnership with the National Library of Scotland, they are currently in the process of putting it out to tender. Overall, the requirements for the system indicate that they are looking for a robust and comprehensive system that looks after the institution's digital assets. These include licensed resources, bibliographic databases, third-party electronic journals, internally developed teaching resources, departmental websites, images and multimedia objects. It is expected that there will be over 20,000 digital images created in the next two years by the university.

Development of a system similar to the D-Space initiative, where departments can deposit their own research data and corporate material, is the ideal for the Library. The DOMS could manage the metadata with pointers to the digital objects stored in local repositories. There would be federated searching of these distributed archives, as well as the ability to search the catalogue, commercial resources, and even some Internet search engines. This system would also need to be able to communicate with the many other systems in place to provide a seamless interface to all the different types of digital information available.

Another concern of the EUL is the emerging importance of METS (Metadata Encoding and Transmission Standard)¹³ as a standard that is inextricably linked to OAIS. EUL would like to use METS to implement an OAIS compliant archive. The use of METS would

enable the inclusion of other metadata standards as well as implement an easier cross-repository search throughout the university.

The future use of OAIS will need to include a lot more guidance if it is to be implemented at a very local level. There needs to be many more pilot projects and increased amounts of literature written about the practical implementation of this model and the problems encountered. Librarians and archivists need to work together with information technology professionals to aid implementation. All agreed that implementation examples were necessary to be able to develop the model and apply it to the range of different workflows found in archives, libraries, and with scientific data centres. However, it was pointed out that there would never be two identical OAIS archives, as each is developed for individual needs.

Again, some participants voiced concern about the necessary resources to implement the OAIS model, pointing out that each of the presentations were made by institutions which had sizeable resources to undertake such a commitment of time and money.

OAIS, Boundaries, and Issues

The OAIS Reference Model provides different communities with a common reference model for preservation, and deals with the preservation of digital objects in a general sense, maintaining a predominantly technical view of them. The OAIS is 'an archive consisting of an *organisation* of people and systems, that has accepted the responsibility to preserve information and make it available for a 'designated community'', and here the concept of the designated community is a very important one. However, Hans Hofman points out that the model does not have a dedicated preservation function, but instead locates this somewhere between preservation planning and administration.

An important question to ask when faced with the OAIS reference model is, what is the digital or information object, and what is it exactly that we are trying to preserve? Terms like digital object, information object, or information resource are often used, but what do they mean? In digital form there is no physical entity anymore. He stressed the need to stop looking at information in such a physical way.

Discussions about the central concept of the information object questioned whether the data file and the representation information can be preserved to provide access to the intellectual content, or whether the intellectual entity should be preserved which has to be recreated every time when accessed. In addition, to what extent do we know that what comes

¹³ <http://www.loc.gov/standards/mets/>

out of an archive is the same as what went in? This recreation of the intellectual entity is a problem and requires a process to move a concept like a record through time and recreate that in the future. This is a problem that is not addressed in the OAIS model. Hofman emphasised that we should not lose sight of the intellectual object we are trying to preserve.

Questions of authenticity are important when discussing digital preservation. Information needs to be trustworthy if we want to use it. We need to be able to establish where it came from and assess that it is what it says it is. One criterion for authenticity is the establishment of the identity of the object, including information about origin or provenance. This relates to preserving the integrity of the data, the integrity of the representation (the intellectual entity), and determining who or what determines authenticity. There is no original state in the digital world, and a record, for example, has evidential characteristics, which the preservation of a byte stream alone would not be enough to recreate. Is authenticity determined by the creator of the digital object, the user, or both? We need to understand what the intention of the creator was, what is essential in the message the creator wishes to convey. The OAIS model looks backwards, but it is essential to look from the other perspective, that of the creator.

From the interpretation of the model through to implementation, every designated community has to go through the identification of concepts such as authenticity, accessibility, identification of requirements, identification and analysis of the nature of the objects, and the functionality needed in certain contexts. Any preservation strategy has to encompass the creator, preserver, the user, and the digital object.

Participants were then presented with the preservation function model as developed by the InterPARES¹⁴ project, which serves as an example of functionality for preservation systems for the archival community. The InterPARES model is an application of the OAIS model, but goes further to take into account the different aspects of intellectual and physical components, from the perspective of the preserver. Hofman described the InterPARES preservation function model as an archivist's interpretation of the OAIS model.

Discussion also focussed on the subject of metadata, with several comments about the amount that certain institutions were using. Is it necessary to create and keep such a range of metadata? References were made to these questions of authenticity, stressing the need for the creator to provide as much of the necessary metadata as possible. In addition, there needs to be as little manual metadata creation as possible, and to make it increasingly automated. A common sense approach was needed for metadata, with suggestions of examining how much metadata can be inherited down. If metadata is set at a higher level it could be applied to subsets of data as well. Having a flexible metadata definition was

considered useful, and allowing it to be dynamically extensible over time. The bigger the metadata set used, the more difficult the data would be to obtain and the process would be more expensive. When defining metadata fields as mandatory, it was suggested that we question whether we are prepared to reject a document if the field is not provided in the submission.

Participants also questioned the nature of OAIS compliance, debating whether compliance is actually possible or plausible. Would compliance simply entail compliance to an interpretation of the model, given that the model itself is only a guidance document? Participants felt that the only plausible method would be to judge compliance on a range of levels. As there are different levels in which the model can be used, there should be different levels of compliance. Parallels with the ongoing discussion of certification of digital repositories were mentioned, and there were calls for some future document that would clarify some of these issues, so that levels of certification and compliance would be the goal. In practical terms this would result in a different level of compliance being awarded for the amount of metadata brought in at the ingest stage, for example. Until there is more implementation and cooperation, it was felt that the idea of compliance is not a very useful one.

It was also agreed that certification could be difficult without a clear understanding of what OAIS compliance would mean (i.e. whether there would be levels of compliance). The example of the ISO 9000 certification process was discussed, but concerns were raised because this is a self-certification process, and some felt that there is a danger in allowing the self-certification of digital repositories. Certification bodies could also prove problematic if there was more than one in existence. There needs to be meaning in certification, and issues of trust should be clarified. Certification can have both a legal purpose as well as an economic one, where users might be encouraged because it conforms to a certain level of standard. Products developed based on OAIS need to be of a certain standard and of a certain compliance level. Certification, after some debate was seen as not easily obtainable, and the focus of activity should be on developing tools to facilitate risk analysis. As a reference model, OAIS is set at too high a level to allow for certification.

¹⁴ <http://www.interpares.org/>

Conclusions

Key to the success of interpreting and implementing the OAIS Reference Model are more working examples and implementations that need to be made available. There remains much that an organisation will need to establish, especially the boundaries of the system, as well as many arbitrary judgements to which the reference model can provide no guidance. Despite many reservations and remaining questions that were highlighted at the seminar, participants felt that there were a number of key benefits to considering the OAIS reference model as a guide to developing an information management and preservation system:

- OAIS is generic enough to handle any kind of information that needs to be preserved;
- OAIS is a natural system for archivists;
- OAIS encourages the right planning for an archival system;
- OAIS concepts can be retrofitted to existing systems;
- implementers become part of a community where similar issues and concerns can be discussed, and there is a common language in which to communicate needs;
- it has a widespread interest and a widespread dissemination which means that as products start to emerge they are more likely to fit with what is already in place;
- there is credibility in trying to adhere to standards, which gives confidence to producers and consumers; and
- there is little in the way of alternatives.

These benefits, highlighted by the seminar participants, and the existence of a standard can help build a business case for even smaller institutions to use to sell the system to their organisations. There was no major opposition to the model *per se* at the seminar, but the main problem lies in implementing a digital archives – its hard to sell the need of a true electronic archive to an organisation that doesn't think this is a priority.

Appendix One: Speakers at the OAIS Training Seminar

Robin L. Dale has been a Program Officer for Member Initiatives with RLG for the past 6 years. In that position, she leads one of RLG's key initiatives, the Long-term Retention of Digital Research Materials. Her current work focuses on trusted digital repositories, preservation metadata, and digital repository certification. Prior to joining RLG, Robin was Head of the Preservation Reformatting Department at Columbia University and worked in the Preservation Replacement Department at the University of California, Berkeley. She is the Associate Editor of RLG DigiNews, is active in digital preservation standards-building activities, and is an adjunct faculty member teaching preservation at a local graduate school.

Dr David Giaretta has worked for many years in the field of Space Data archives, in particular those involving data from Astronomical satellites. He is chairman of CCSDS Panel 2, the standards group under which the OAIS Reference Model was produced, and he played an active role in its development. He is currently involved in the development of several standards which follow on from the Reference Model.

Hans Hofman is co-director of the European project ERPANET (www.erpanet.org) and senior advisor for the government program 'Digital Longevity' concerning information management at the National Archives. On the international scene he is co-investigator and representative of the National Archives of the Netherlands in the Inter Pares 2 research project (2002-2007) and he represents the Netherlands in the ISO TC46/SC11 on records management (e.g. ISO RM standard 15489). He liaises with the Dutch Archives School in teaching the management of electronic records.

David Holdsworth is a Consultant in Information Systems Services working for Leeds University's Information Systems Services. He has worked in IT since the mid 60s and seen digital storage technology evolve over many generations. He is the architect of the demonstrator system produced by the CEDARS project, and has also built long-term storage facilities used at Leeds University and elsewhere.

Julien Masanès has Master degrees in Philosophy and Cognitive Science. Since autumn 1999 he has worked at the Bibliothèque nationale de France on Web archiving. He is also Project leader of the initiative launched at the end of 2000 to prepare the extension of legal deposit to the Internet.

Najla Semple is a qualified librarian with a background in Arabic studies, and has been involved in digital preservation since she came to Edinburgh University Library in 2001 after working at an Oxford University library. In addition to implementing digital archiving procedures for Edinburgh University, she has been involved in a number of digital preservation training events both internally and externally, as well as initiating a project to archive Arabic websites.

Derek Sergeant is a formally trained computer scientist. In 1998 he joined the University of Leeds Computing Service to modify the local Data Storage Archive (LEEDS). From this he moved on to work for the Cedars project, to investigate practical solutions to the digital preservation problem. Currently Derek works as a senior researcher on the CAMiLEON project, and is implementing an emulator for the BBC Domesday system.

Deborah Woodyard is the Digital Preservation Coordinator at The British Library in London. She began working on the preservation of digital materials in 1996 at the National Library of Australia and joined the British Library in January 2001. Deborah conducts internal and external digital preservation work such as participating in the core development team for long term digital storage.

Appendix Two: Participants at the OAIS Training Seminar

Name	Position	Institution	Country
Aas, Kuldar	Digital Archiving Specialist	Estonian Historical Archives	Estonia
Aka, Kyrian Ezebunwa	Remote Sensing Information Analyst	Shell	Nigeria
Andersen, Bjarne	IT Developer	Statsbiblioteket	Denmark
Andersen, Jakob	Deputy Director	Danmarks Pædagogiske Bibliotek	Denmark
Andersson, Jörgen	University Archivist	Lunds Universitet	Sweden
Andersson, Ulf		AstraZeneca R&D	Sweden
Aschenbrenner, Andreas	Content Editor	ERPANET	The Netherlands
Bausenhart, Ursula	Archivist	Staatsarchiv Basel-Stadt	Switzerland
Bogaarts, Jacques	Senior ICT Advisor	National Archives	Netherlands
Brown, Jane	Senior Inspecting Officer	National Archives of Scotland	United Kingdom
Dale, Robin		Research Libraries Group	USA
Dam, Claus	Curator	Danish Cultural Heritage Agency	Denmark
Enseñat, Luis	Archivist	Archivo General de la Administración	Spain
Fogelvik, Stefan		City Archives of Stockholm	Sweden
Frid, Anders		AstraZeneca R&D	Sweden
Fritzon, Johan	Archivist	AstraZeneca R&D	Sweden
Giaretta, David		CCLRC Rutherford Appleton Laboratory	United Kingdom
Greenan, Monica	Content Editor	ERPANET	United Kingdom
Gross, Jennifer	Library Assistant	University of Heidelberg	Germany
Hallgrimsson, Thorsteinn	Deputy National Librarian	The National and University Library of Iceland	Iceland
Hansen, Jytte	Web Consultant, Librarian	Danish Bibliographic Centre	Denmark
Hanssen, Erling Midtgaard	Head of National Newspaper Collection	State and University Library Aarhus	Denmark
Henriksen, Birgit	Head of Digitisation	Royal Library	Denmark
Heuscher, Stephan	Datenarchitekt	Schweizerisches	Switzerland

		Bundesarchiv	
Hibberd, Lee	Digitisation Officer	National Library of Scotland	United Kingdom
Hoel, Ivar	Chief Librarian	Royal School of Librarianship Library	Denmark
Hofman, Hans		ERPANET, Nationaal Archief	The Netherlands
Høgås, Hilde	Adviser, IT Department	National Library of Norway	Norway
Holdsworth, David		CEDARS	United Kingdom
James, Hamish	Collections Manager	Arts and Humanities Data Service (Kings College London)	United Kingdom
Johansson, Lars-Åke		AstraZeneca R&D	Sweden
Juliussen, Jon		National Tax Board	Sweden
Kaiser, Max		Austrian National Library	Austria
Kann, Bettina	System Librarian	Austrian National Library	Austria
Kansy, Lambert	Archivist	Staatsarchiv Basel-Stadt	Switzerland
Kejser, Ulla	Head of Preservation	Det Kongelige Bibliotek	Denmark
Lecher, Hanno	Librarian	University of Heidelberg	Germany
Ling, Per-Åke		AstraZeneca R&D	Sweden
Locher, Hansueli	Integration Manager	Swiss National Library	Switzerland
Madsen, Kirsten	Archivist	Danish State Archives	Denmark
Masanès, Julien		Bibliothèque nationale de France	France
Mikkelsen, Hans Kristian	Research Librarian	Royal Library	Denmark
Müller, Eva	Head of Electronic Publishing Centre	Uppsala University Library	Sweden
Müller, Paul	IT Officer	Staatsarchiv Basel-Stadt	Switzerland
Nielsen, Lise Qwist	Archivist	Danish National Archives	Denmark
Nondal, Lars	Project Leader, Digital Library	Copenhagen Business School Library	Denmark
Öhrström, Bo	Deputy Director	Danish National Library Authority	Denmark
Pétillat, Christine	Director	National Archives of France	France
Poulsen, Jens Christian	Research Librarian	Royal Library	Denmark
Ross, Seamus	Director, ERPANET and HATII	University of Glasgow	United Kingdom

Schomacker, Tommy		Dansk BiblioteksCenter	Denmark
Semple, Najla		Edinburgh University Library	United Kingdom
Sergeant, Derek		CAMiLEON	United Kingdom
Slabbertje, Martin	Project Manager	University Utrecht Library	Netherlands
Sørensen, Arne	Head of IT	Statsbiblioteket	Denmark
Spiby, Phil		AstraZeneca R&D	Sweden
Teil, Jean-Pierre	Constance Program Manager	National Archives of France	France
Thorborg, Susanne	Database Consultant	Danish Bibliographic Centre	Denmark
Westcott, Keith	Curatorial Officer	Archaeology Data Service	United Kingdom
Woodyard, Deborah		British Library	United Kingdom
Zabel, Walter	Head of IT Department	Austrian National Library	Austria