Briefing Paper: the OAIS Reference Model

Alex Ball
UKOLN, University of Bath

February 2006

1 Introduction

The purpose of this document is to provide an introduction to the Open Archival Information System (OAIS) Reference Model [CCSDS 2002], and to explore some issues with regard to its implementation. This is offered as a possible context for discussing repositories of Engineering information, and to this end some links between the OAIS Reference Model and the Engineering domain have been identified.

Sections 1 to 6 introduce the OAIS Reference Model itself, in particular the Information Model and Functional Model, and present its discussion on the topics of preservation activities and compliance. The remaining sections deal with issues relating to implementing an OAIS. Section 7 looks at a certification project aiming to produce the criteria for a reliable or trustworthy OAIS implementation. Section 8 looks at various content packaging techniques that can be used to create OAIS information packages. Section 9 looks at two metadata schemata that can be used to provide some of the metadata required under the OAIS Information Model. Section 10 briefly describes some existing OAIS-based implementations. Finally, section 11 looks at areas within the Engineering domain that have synergies with OAIS.

2 Background

The OAIS Reference Model was developed by the Consultative Committee for Space Data Systems (CCSDS) as a first step towards generating formal standards for the long-term archiving of Space Science data. It is a *conceptual framework* rather than a fully-fledged, prescriptive standard, and is intended to identify the necessary features of an archival information system rather than recommend any particular implementation.

The phrase 'open archival information system' is perhaps a misleading title, as 'open' refers to the fact that the model was developed in open fora, so as to include insights from a wide range of communities, rather than describing any particular feature of the system itself. An archival information system is 'an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community,' the latter being 'an identified group of potential Consumers who should be able to understand a particular set of information.' [§1.7.2] The OAIS model is set in the context of producers (who generate the information to be archived), consumers (who retrieve the information) and management (the wider organisation hosting the OAIS).

In the model, a person with an appropriate *knowledge base* (e.g. the ability to read English) extracts and understands the *information* carried by *data*; hence information is seen as knowledge

in an exchange format, manifested physically by data such as a bitstream or a string of printed letters. It is recognised that the knowledge base required to decode the data and understand the information may not be widely available or persistent among consumers, hence the requirement for additional *representation information* to bridge the gap. The model thus introduces the terminology of a *data object* (e.g. a bitstream) which, when interpreted using representation information (e.g. the ASCII standard), becomes an *information object* (e.g. a text file).

The obvious problem arising from this is the recursive nature of representation information, which is most likely also carried by its own bitstream. To prevent an infinite regress, some minimum knowledge base among the consumers must be assumed. The idea of a designated community comes in as a way to ensure that this minimum knowledge base is in fact maintained.

3 Information Model

Information objects move through an OAIS in an encapsulated form known as an *information package*.[1] The information package includes both the data object and the representation information, together referred to as the content information; in addition, it contains preservation description information (PDI), which comes in four flavours: provenance (a detailed history of the content information), context (rationale, relationships to other information), reference (identifiers such as ISBNs) and fixity (checksums, etc. to monitor degradation or alteration). The whole package should be wrapped in packaging information (e.g. manifest and package identifier), and the OAIS should hold descriptive information about the package to facilitate search and retrieval. For a graphical representation of this, see fig. 1.

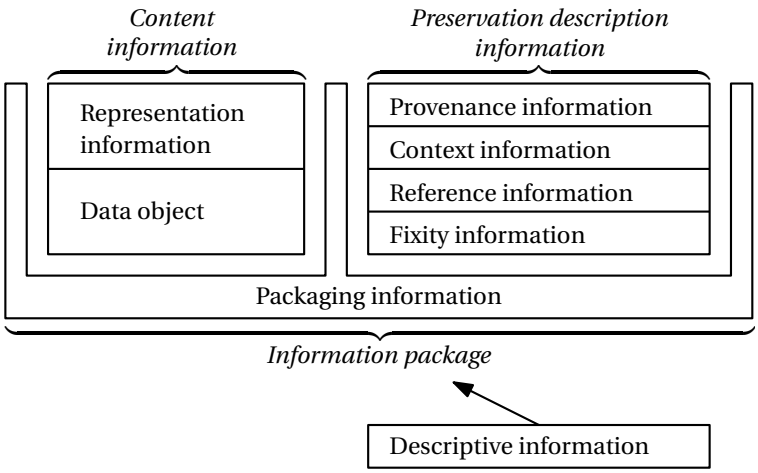

Figure 1: Information Package Model

The OAIS Information Model recognises three variants of information packages. The *archival information package* (AIP) is the variant that the OAIS actually preserves, and of the three variants it is the one likely to have the most detailed PDI. Information is submitted to the OAIS in *submission information packages* (SIPs), which may be structured differently from the AIPs and contain insufficient PDI; however, some minimum standard may be imposed on SIPs. Lastly there are *dissemination information packages* (DIPs), which are versions of the AIPs tailored to consumer requirements.

[1] The model is careful not to require literal encapsulation, such as in a ZIP or TAR file.

All variants of information packages may contain other information packages; SIPs may include content intended to be split into several AIPs, and various AIPs may be bundled into a single DIP for dissemination to a consumer. It may also be convenient to store several AIPs within a larger AIP.

4 Functional Model

The main functions of an OAIS are modelled as seven functional entities:

- *Ingest.* This entity represents the incorporation of submitted information into the archive. The functions of the entity are: to receive SIPs from the producer and subject them to quality assurance; to generate appropriate AIPs and descriptive information; and to co-ordinate the requisite updates to the Archival Storage and Data Management entities.

- *Archival Storage.* This entity covers the storage of the AIPs. The functions of the entity are: to receive AIPs from the Ingest entity; to manage the storage hierarchy (i.e. put the AIP on the appropriate storage medium); to replace media as necessary (which may involve repackaging the content data and PDI); error checking; disaster recovery; and providing copies of AIPs to the Access entity on request.

- *Data Management.* The functions of this entity are: the maintenance of the database of descriptive information and system information; answering queries passed by the Access entity; generating reports as requested by the Ingest, Access or Administration entities; and updating the database with descriptive information from Ingest, and system and review updates from Administration.

- *Administration.* This entity covers the activities needed to run the OAIS smoothly. The functions of this entity are: negotiating a submission policy with producers; managing the system configuration; performing archival information updates (by retrieving DIPs from Access, modifying them, and submitting them back to Ingest as SIPs); physical access control; establishing archive system policies and standards; auditing submissions to ensure at least minimum standards are maintained; activating requests (automatically generating dissemination requests from saved searches); and customer service.

- *Preservation Planning.* This entity ensures that the policies and procedures in place at the OAIS adequately protect it from issues arising from technological changes. The functions of this entity are: monitoring the Designated Community for changes in requirements; monitoring technology, standards and platforms, to track the emergence of new ones and the decline of older ones; developing preservation strategies and standards; and developing packaging designs and migration plans.

- *Access.* This entity covers the search and retrieval of archived information. The functions of this entity are: co-ordination of access activities into a single user interface, including methods for search queries, report requests and orders for DIPs; generation of DIPs from AIPs; and delivery of result sets, reports, DIPs and assistance to consumers.

- *Common Services.* This entity underlies all the others, and includes operating system services, network services and security services.

The relationships between the functional entities can be seen in diagrammatic form in fig. 2. The arrows represent flow of information; grey arrows have been used for clarity rather than

for any particular significance. Common Services are not shown on the diagram, but should be thought of as connecting with the other six entities by means of two-way information flows.

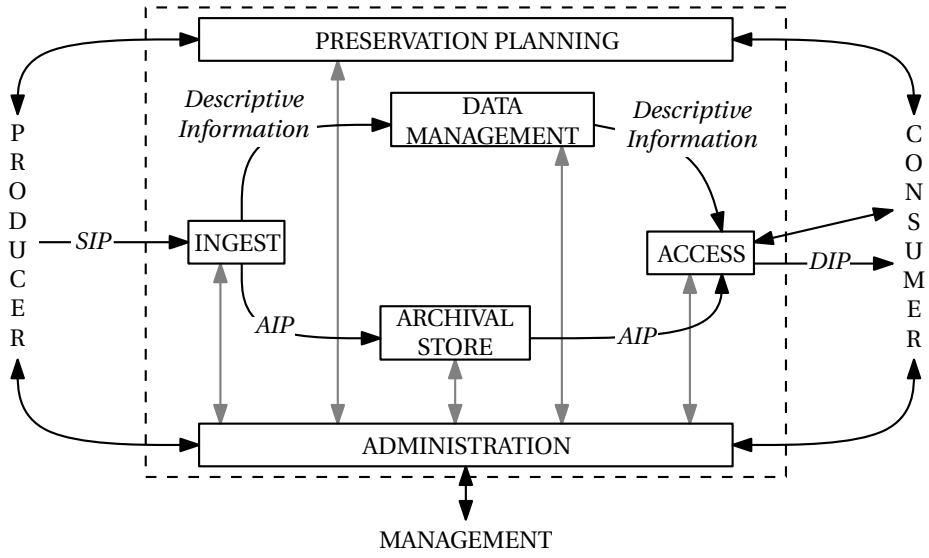

Figure 2: OAIS Functional Model

While much of the functional model could have been produced by abstracting the processes and workflows of traditional libraries and archives, the distinctive flavour of the model comes from its heavy emphasis on preservation, in particular of digital data.

5 Preservation perspectives

The OAIS Reference Model notes that there is a cost-benefit analysis to be made concerning the migration of digital information. On the one hand, it is costly both in terms of time and money, and there is a significant danger of losing data during the migration process. On the other hand, migration may help to improve cost effectiveness, to respond to evolving consumer requirements, and to offset the effects of media decay.

Four types of migration are outlined, in increasing order of risk to data:

- *Refreshment.* This is where a new copy of the AIP is made using the same type of medium as before. It is a special case of replication.
- *Replication.* This is where a new copy of the AIP is made, leaving all the information unchanged, but not necessarily on the same type of medium as before.
- *Repackaging.* This is where a new copy of the AIP is made, but changing some of the packaging information. The content information and PDI are left unchanged.
- *Transformation.* This is where a new AIP is created with different content information or PDI, but which attempts to preserve the full information content of the original AIP,

Of these four types, only transformation is acknowledged to create a new *version* of the AIP, and the PDI of the new version should be updated to reflect this (by referring to the original AIP and detailing what sort of transformation was used). This is in contrast to a new *edition* of an AIP, which is usually intended to improve on, rather than merely preserve, the information in the original AIP.

6 Points of conformance

Although the OAIS Reference Model does not prescribe any particular implementation, it does list some responsibilities that all OAISes should discharge. Each OAIS should:

- negotiate for and accept information;
- obtain sufficient control of the information to ensure long-term preservation;
- determine which groups of consumers should be considered the Designated Community;
- ensure that the information provided by the OAIS is understandable by the Designated Community without recourse back to the producer of the information;
- follow documented policies and procedures;
- make the information available to the Designated Community.

The other point of conformance that is specified in the model is 'support' of the information model, although no indication is given of what form such support should take, beyond the avoidance of conflicting terminology.

7 RLG/NARA Digital Repository Certification Project

The points of conformance listed within the OAIS Reference Model standard are rather brief, and cannot be used to judge the trustworthiness of any particular implementation to preserve information [RLG & OCLC 2002]. In order to fulfil this need, the Digital Repository Certification Project[2] has been set up jointly by RLG, a US-based organisation made up of research libraries, museums and archives, and the [American] National Archives and Records Administration (NARA). The purpose of the project is to formulate a way of identifying which digital repositories can be trusted to store, migrate, and provide access to digital collections. This formula will be codified as a set of certification requirements. The project also wishes to outline a certification process, and to identify agencies to carry it out, on which points it is collaborating with the Center for Research Libraries' Audit and Certification of Digital Archives Project.[3]

The project has recently delivered a draft certification checklist [RLG & NARA 2005], based around the areas of governance and organisation, repository functionality, relationships with the designated community and technological infrastructure.

On organisation, the checklist requires that a repository should have a written commitment to long term operation and have contingency measures in case of closure. Staff should be well trained and should participate in continuing professional development. The repository should have in place a comprehensive set of policies and procedures that ensure that it is kept up to date, accountable, well-documented and responsive to provider and consumer needs. It should have a sustainable business model, and actively address issues of intellectual property rights and copyright.

The checklist also provides a list of requirements for five of the OAIS functional entities. Concerning *Ingest*, the repository should have a policies for the SIPs it is willing to preserve, and a robust system for tracking the progress of the transformation of SIPs to AIPs. It should also have formal agreements with producers concerning rights and responsibilities with respect to the information handled. Concerning *Archival Storage*, the repository should document each type of AIP held, the procedure used to create AIPs from SIPs, the naming convention for ensuring unique package identifiers, and a technique for mapping any unique SIP identifiers to the assigned unique AIP identifiers. It should also be able to verify the fixity of AIP, and provide an

[2] URL: ⟨http://www.rlg.org/en/page.php?Page_ID=580⟩.

[3] URL: ⟨http://www.crl.edu/content.asp?l1=13&l2=58&l3=142⟩.

independent technique for auditing the entire stock of AIPs. Concerning *Preservation Planning*, the repository should have documented preservation strategies, implement them, have a procedure for changing them if necessary, and have evidence that they are successful. It should make use of international representation information repositories, contributing back to them wherever appropriate, and should track whether representation information is approaching obsolescence. It should preserve the content information of AIPs, acquire sufficient preservation description information, and monitor AIP integrity; it should also make and retain contemporaneous records of all activities performed. Concerning *Data Management*, the repository should create descriptive information, ensuring that referential integrity between it and the corresponding AIPs is created and maintained at all times. Concerning *Access Management*, the repository should have documented access policies and implement them. It should be able to demonstrate that all requests are dealt with (whether accepted or rejected) and that all denials of access are logged, with suspicious denials flagged up as potential security problems. All DIPs produced by the repository should be checked for completeness and correctness, and should be authentic copies of the corresponding AIPs.

With regard to the designated community, the checklist requires that a repository should document its designated community (including its knowledge base and service level expectations), make this documentation available, and commit to an operational definition of understanding (which informs the extent of the representation information held). The repository should provide enough descriptive information to fulfil the designated community's search and retrieval needs. It should document and implement access policies consistent with deposit agreements, respect all access agreements, publicise access and delivery options available to the designated community, and record any queries, requests and orders that the producers need to know about. It should also implement a procedure to test that the DIPs can be understood by the designated community (as defined above) and to take remedial action should any DIPs fail the test.

Concerning technological infrastructure, the checklist requires that a repository should use well-supported systems, make regular backups of data, document all identical copies held and keep them synchronised, detect and report any data corruption/loss and attempt to repair it, define a media migration process, define a change management process which tests the effects of all critical changes, and keep all systems up to date with security fixes. The repository should use hardware and software appropriate to the services provided, and monitor any changes in the needs of the designated community. It should systematically analyse its environmental conditions, address all security needs, delineate roles, responsibilities and authorisations for all staff, maintain and test written disaster recovery and service continuity plans, and maintain at least one off-site store of backed up data.

As an appendix, the draft checklist contains some discussion of the OAIS requirement that information stored by a repository should be independently understandable by the designated community, and provides examples of how compliance with the requirement might be achieved in practice.

In parallel with the work done by the RLG/NARA Digital Repository Certification Task Force, a related project is being undertaken by the [German] Network of Expertise in long-term STOrage of digital Resources (nestor) Working Group on Trusted Repositories Certification.[4] This project aims to produce a criteria catalogue which can be used to generate context-specific standards against which to judge repositories; thus different (but equivalent) standards would be generated for, say, archives in the UK and data centres in Germany [Dobratz & Schoger 2005]. Given its emphasis on specific criteria, the nestor criteria catalogue is not expected to make extensive use

[4] URL: ⟨http://nestor.cms.hu-berlin.de/tiki/tiki-index.php?page=Working+Group+on+Trusted+Repositories+Certification+(nestor)⟩.

of the generalised terminology of the OAIS Reference Model.

8 Content packaging techniques

While the OAIS Reference Model does not specify any particular method for packaging information, various implementations have been proposed to fill this role. Examples of such implementations include METS, XFDU, MPEG-21 DIDL and IMS Content Packaging. Each of them is based on the idea of a central XML manifest file that either references or contains the data files that make up the package.

8.1 METS

The Metadata Encoding and Transmission Standard (METS)[5] is developed by the Digital Library Foundation[6] and maintained by the Library of Congress. It is an XML format that may be used as a manifest document referring to linked files, or as a literal container document with files embedded as XML or in Base64 encoding.

While METS is presented as a possible format for OAIS information packages, the structure of METS documents does not mimic the structure of the OAIS Information Model. METS documents have seven major sections:

1. The *METS Header* contains the dates when the document was created and last modified, and details of the agents that have been involved in its history (e.g. creator, editor, disseminator). This partially corresponds with OAIS provenance information.

2. The *Descriptive Metadata* section essentially provides a catalogue record for the document. This corresponds with OAIS descriptive information (although in an OAIS this is stored separately from the information package). It may also contain OAIS reference information.

3. The *Administrative Metadata* section provides four types of metadata: technical, intellectual property rights, source and digital provenance. Technical metadata cover how the document is formatted, and how it should be used, thus corresponding with OAIS representation information. Source metadata cover the analogue source of a digital document, if applicable, while digital provenance metadata cover the relationships between files within and external to the document, specifically where one file is a derivative, manipulation or transformation of another. Along with intellectual property rights metadata, these metadata correspond with OAIS provenance information and possibly context information.

4. The *File* section typically contains links to the file(s) making up the content of the document, although it may include the actual file(s) themselves. This corresponds with the OAIS data object and partially corresponds with packaging information.

5. The *Structural Map* section provides the logical structure of the document, and allows for the referencing of individual portions of the file(s) and different versions of the same logical section (e.g. audio file and transcript). This partially corresponds with OAIS context information and possibly representation information.

6. The *Structural Links* section details any hyperlinks that may exist between items in the structural map. This allows the hypertext structure of websites to be recorded, for example.

[5] URL: ⟨http://www.loc.gov/standards/mets/mets-home.html⟩.
[6] URL: ⟨http://www.diglib.org/⟩.

7. The *Behaviour* section associates the file(s) with the executable code needed to read or manipulate them. This corresponds with OAIS representation information.

8.2 XFDU

The XML Formatted Data Unit (XFDU) standard[7] [CCSDS 2004] is being developed by CCSDS as a new information packaging standard to replace the Standard Formatted Data Unit (SFDU) [CCSDS 1992]. It is similar to METS, but is designed to reflect more closely the OAIS Information Model.

An XFDU is a logical unit made up of a primary Package Interchange File (PIF), also known as the XFDU package, plus any other files, PIFs or repositories referenced by the primary PIF. The PIF is a physical container file such as a ZIP or TAR file, containing an XML manifest document and none, some or all of the files referenced by the manifest (although it is intended that the majority of the files should be included). The manifest document itself has five sections:

1. The *Package Header* contains metadata concerning the XFDU as a whole. Specifically it includes *environment information*, such as the hardware and software platform on which the XFDU package was created, *behaviour information*, explaining mechanisms needed to understand the package, and *transformation information*, specifically how to undo the transformations used in the package. This corresponds with OAIS representation information and partially with OAIS packaging information.

2. The *Metadata* section records metadata for all items in the XFDU. While any suitable metadata model can be used, the XML schema for this section natively provides categories of descriptive, representation and preservation description information, the latter being subdivided into classes of provenance, context, reference and fixity information.

3. The *Information Package Map* organises all the content units (files) in the XFDU into a logical hierarchy, with itself as the highest level content unit. Multiple, alternative maps may be provided if desired. This section partially corresponds with OAIS context information.

4. The *Data Object* section contains details of the files that make up the package (MIME type, checksum, etc.), and for each file, a link to the file (whether within the package or external to it) and/or the file itself encoded as a Base64 bitstream. Instructions on how to decode any encoded or encrypted bitstreams should also be included. This corresponds with the OAIS data object along with some packaging, representation and fixity information.

5. The *Behaviour* section associates the files with the executable code needed to read or manipulate them. This corresponds with OAIS representation information.

8.3 MPEG-21 DIDL

MPEG-21 is a standard, open multimedia framework, the purpose of which is to 'enable transparent and augmented use of multimedia resources across a wide range of networks and devices used by different communities.' Part 2 lays out a mechanism for declaring the structure and make-up of digital items, namely the Digital Item Declaration (DID), by defining the Digital Item Declaration Language (DIDL) XML schema [ISO/IEC 21000-2:2005].

[7] URL: ⟨`http://sindbad.gsfc.nasa.gov/xfdu/`⟩.

Unlike METS and XFDU, a DIDL document is a single hierarchical XML tree of digital resources, with metadata nested within the objects to which they refer. Digital resources are specified by a <resource> element; normally this is done with a link, but the resource may be included directly as XML or a Base64 encoded bitstream. The element also specifies any encoding performed on the resource (whether included or referenced) and the order in which encodings were performed (corresponding with the OAIS data object and some representation information). A <component> element contains a single resource (although multiple <resource> elements are allowed so that multiple copies can be referenced). An <item> element contains one or more components, and if these components are different versions of the same content, further elements are provided for choosing between them. Items may be grouped inside a <container> element. The top level element is <DIDL> which must contain exactly one container or item.

Metadata concerning the DIDL document as a whole are provided by the <DIDLInfo> element (corresponding with parts of the OIAS packaging and fixity information). Containers, items, components, etc. can be described by a <descriptor> element nested immediately within them. A descriptor can either contain a component (for non-textual descriptions or summaries) or a <statement> element containing textual metadata (which could be more XML). Descriptors can also be made to refer to fragments of a resource by being wrapped in an <anchor> element nested within a component. If inserting a <descriptor> element into a certain (sub-item) element is not possible or desired, it is possible to link a descriptor to that element by means of an <annotation> element nested immediately within the parent item.

Much of the work in getting MPEG-21 DIDL to reflect the OAIS information model must be done by the <descriptor> elements. Statements containing XML from elsewhere in the MPEG-21 standard can be used to create reference, representation and provenance information [Bekaert et al. 2003], although there is no restriction in the DIDL standard on the metadata schemata that can be used.

8.4 IMS Content Packaging

Developed by the IMS Global Learning Consortium,[8] the IMS Content Packaging Specification[9] was initially designed for use with physical file packages (typically in ZIP format), but from version 1.2 will be suitable for logical file packages as well.

The IMS Manifest document is made up of a <manifest> element with the following child elements:

1. The <metadata> element is optional, and contains metadata pertaining to the entire package.

2. The <organizations> element is mandatory, and contains a structural map, or 'organization', of the resources listed in the following section. The structure is represented by a tree of <item> elements. Multiple, alternative organizations are allowed. This section partially corresponds with OAIS context information.

3. The <resources> element is mandatory, and contains links to the files that make up the resources. It is possible to include one resource within another by means of a dependency declaration; this makes it possible to reuse the same file links across several resources. This section corresponds with the OAIS data object and partially corresponds with OAIS packaging information.

[8] URL: ⟨http://www.imsglobal.org/⟩.

[9] URL: ⟨http://www.imsglobal.org/content/packaging/index.html⟩.

4. Optionally, a manifest may contain one or more subsidiary manifests, allowing the contents of the package to be divided into more manageable sections. A subsidiary manifest may be associated with an item in an organization in place of a resource.

The work needed to make an IMS Manifest document reflect the OAIS Information Model must be done by `<metadata>` elements, which may be located within `<organization>`, `<item>`, `<resource>` and `<file>` elements as well as `<manifest>` elements.

9 Preservation metadata schemata

There are several published metadata schemata which have been designed with the preservation of digital documents in mind, and these have varying degrees of correspondence with the OAIS Information Model. For example, PREMIS and the National Library of New Zealand both use their own data model and metadata structure, but both map fairly well on the OAIS Information Model, with PREMIS being a little more comprehensive.

9.1 PREMIS

An OCLC/RLG working group entitled Preservation Metadata Implementation Strategies (PREMIS) has produced a preservation metadata schema, supported by a data dictionary, intended for use in a variety of digital preservation situations [OCLC & RLG 2005].

The work of PREMIS builds on a report published by the OCLC/RLG Preservation Metadata Framework Working Group entitled *A Metadata Framework to Support the Preservation of Digital Objects* [OCLC & RLG 2002]. This earlier document forged together several existing preservation metadata schemata into a single, widely applicable schema based squarely on the OAIS Reference Model. This schema was only expressed in terms of a list of human-readable metadata elements, however, and needed to be formalised to allow it to be put into a machine-readable format such as an XML.

The work of PREMIS has been to come up with a more formalised schema suitable for encoding in XML, for example. While covering the same basic ground as the earlier framework — representation and preservation description information — it has a different approach and uses a completely new structure.

The core of the PREMIS schema is a data model and a data dictionary. The data model has five entities:

- *Intellectual Entities* are coherent sets of content. The name emphasises the content rather than the format; thus a TIFF scan and an HTML transcription of a book are both representations of the same Intellectual Entity. No particular granularity is implied, so both an album of photographs and a single photograph may be considered intellectual entities.
- *Objects* are chunks of digital content, specifically bitstreams (in PREMIS, this means the ones and zeroes representing some *content*), files (bitstreams along with the headers that describe the format, access permissions, date last modified, etc.) and representations (all the files and additional metadata needed to reconstruct an intellectual entity). Files embedded in larger files are known as filestreams.
- *Events* are actions or processes performed on objects: integrity checks, deletions, ingestions, transformations, etc. Transformation Events are considered to create new objects rather than modify existing ones.

- *Agents* are the people, organisations and software that are involved in the preservation process. They may have a role in events or hold rights.
- *Rights* within PREMIS are statements of permission, granted by an agent, allowing a repository to perform specified preservation activities within a given time period.

The data dictionary contains a metadata tree for each of these entities. The tree is made up of *semantic units;* these semantic units can either contain further semantic units or take a value (but not both). Fig. 3 shows the metadata tree for the Rights entity by way of example.

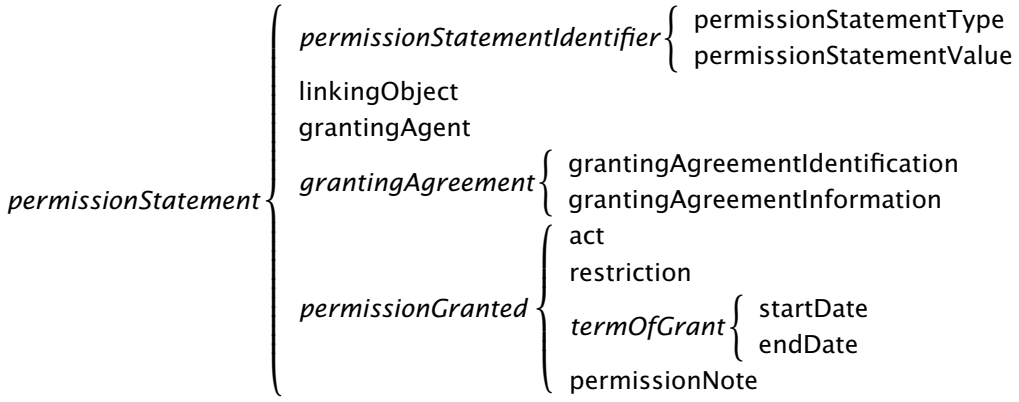

Figure 3: Semantic units for the Rights entity. Valueless container units are shown in italics.

The semantic units related to the Rights and Events entities collectively correspond with OAIS provenance information, while the semantic units of the Object entity cover the rest of the OAIS representation and preservation description information. Reference information is covered by the objectIdentifier semantic unit; context information is covered by the relationship, linkingEventIdentifier, linkingIntellectualEntityIdentifier and linkingPermissionStatementIdentifier semantic units; and fixity information is covered by the signatureInformation and fixity semantic units.

Significant additions to PREMIS from the previous framework include explicit provisions for referencing file format registries and for digital signatures, and the semantic units associated with the Agent entity. On the other hand, PREMIS is more streamlined with regards to recording software and hardware requirements and object characteristics.

9.2 National Library of New Zealand Preservation Metadata

The National Library of New Zealand is developing a metadata schema for use in its own digital library collection. The first phase of work centred on resource discovery metadata, and resulted in a framework made up of existing metadata standards [National Library of New Zealand 2000]. This work was followed by a custom schema for preservation metadata [2002]; the current version of this schema is the 2003 revision.

The preservation metadata schema has its own data model, based around four entities: Object, File, Metadata Modification and Process. The latter three may only be associated with one Object, but an Object may have more than one of each of the other entities. A digital object may be described as *simple* (containing exactly one file), *complex* (containing multiple files in a particular structure, making up a single logical object) or as an *object group* (containing several independent files); 'empty' objects are not considered.

The schema itself is a flat hierarchy, with the possible exceptions of the File entity, which has specific blocks of metadata for different types of file (images, audio, video, text). In terms of the OAIS Information Model, the Process entity metadata covers provenance information, while the reference information is covered by the Object entity. Representation information is split between the Object entity (hardware/software environment, access issues, known problems) and the File entity (format, size). Context information is provided by Object elements such as Logical composition, Is part of group and Structural composition, and also by the File element Target indicator. No provision is made for fixity information, however.

One significant feature of this schema is the Metadata Modification entity, the elements of which record any changes made to the object's metadata. The use of specific blocks of metadata for specific types of files is also a notable feature of the schema; while this makes it more useful for digital libraries, it could potentially limit its usefulness in other areas unless appropriately extended.

10 Examples of systems using the OAIS Reference Model

10.1 Digital repository systems

aDORe

The aDORe digital repository system was developed by the Research Library of the Los Alamos National Laboratory (LANL)[10] to store local, digital copies of scholarly publications, etc. It uses the MPEG-21 DIDL method of content packaging. LANL's own implementation of aDORe held about eighty million packages in February 2005 [van de Sompel et al. 2005].

DAITSS

The DAITSS digital repository system[11] was developed by the Florida Center for Library Automation as a generic back-end for digital libraries or institutional repositories. It uses the METS method of content packaging, and uses the PREMIS metadata schema for internal preservation metadata. Its architecture is based on the OAIS Reference Model, although the Access functionality is minimal since it is intended to be a 'dark archive' [Caplan 2004].

DSpace

The DSpace digital repository system[12] was developed by MIT Libraries and Hewlett-Packard Laboratories, initially to serve as an institutional repository for research output. It uses the METS method of content packaging. DSpace's use of the OAIS Reference Model is patchy but increasing [Bass et al. 2002; Celeste & Branschofsky 2002; DSpace 2005].

Fedora

The Fedora digital repository system[13] was developed by Cornell University and the University of Virginia, initially to store multimedia digital library collections. It uses the custom Fedora Object XML (FOXML) method for content packaging, although it can ingest and disseminate packages in

[10] URL: ⟨http://lanl.gov/⟩.

[11] URL: ⟨http://www.fcla.edu/digitalArchive/⟩.

[12] URL: ⟨http://dspace.org/⟩. A list of institutions using the software can be accessed at ⟨http://wiki.dspace.org/DspaceInstances⟩.

[13] URL: ⟨http://fedora.info/⟩.

METS and MPEG-21 DIDL formats. Fedora's claim to OAIS compliance mainly centres around its ingest and dissemination functions [Fedora Development Team 2005].

10.2 Custom repositories

CDPP

The *Centre de Données de la Physique des Plasmas* (CDPP)[14] is a data centre that primarily archives information and data sets from French missions relating to space plasma physics. It has established strict standards for the metadata that producers supply in their SIPs, uses the EAST data description language [CCSDS 2000], and has a functional structure that maps easily onto the OAIS Functional Model [ERPANET; Sawyer et al. 2002].

MathArc

MathArc[15] is a collaborative project between Cornell University Library and Göttingen State and University Library to set up a long-term repository of electronic journal articles in the field of mathematics. The repository will be based on the OAIS Reference Model, and is investigating the use of the METS method of content packaging coupled with the PREMIS metadata schema.

ESA MMFI

The European Space Agency (ESA) Multi-Mission Facility Infrastructure (MMFI) is a system designed to handle the data from all ESA missions and thereby achieve an economy of scale. It has been designed to map transparently on to the OAIS Reference Model, and uses a specialised form of XFDU for content packaging. Its distinctive feature is that it is a distributed system, with only data management and consumer access being centralised [Pinna et al. 2005*a*;*b*].

NOAA CLASS

The [American] National Oceanic and Atmospheric Administration (NOAA) Comprehensive Large Array-data Stewardship System (CLASS)[16] is a digital repository of NOAA and US Department of Defense satellite data, and will eventually hold all of NOAA's datasets. The system itself was designed using the OAIS Reference Model, and the Model also underlies the data submission guidelines and agreements associated with CLASS [Rank & McDonald 2005].

NSSDC DIOnAS

The [American] National Space Science Data Center (NSSDC) used the OAIS Reference Model to structure its migration of legacy data sets from old magnetic tape onto digital linear tape (DLT). The legacy tapes (Archival Storage) were read and analysed (Access) to form DIPs; the metadata gleaned from analysis were supplemented by metadata from the NSSDC Information Management System, NIMS (Data Management). These metadata were passed to an Offline Transition to Online (OTTO) database, which informed the conversion of the data into canonical format (Administration), and the packaging of the data and metadata into SIPs. The SIPs were ingested by the Data Ingest and Online Access System (DIOnAS) and converted to AIPs (Archival Storage).

[14] URL: ⟨http://cdpp.cesr.fr/english/⟩.
[15] URL: ⟨http://www.library.cornell.edu/dlit/MathArc/⟩.
[16] URL: ⟨http://www.class.noaa.gov/⟩.

The benefit of using the OAIS Reference Model was the focus it gave to preservation planning, especially the requirements for adequate metadata, enabling future migrations to be accomplished with greater ease [Sawyer et al. 2005].

11 Application to Engineering

There is only one notable Engineering research project using the OAIS Reference Model at the moment. This project, LOTAR, nevertheless reveals a potential synergy between this model and the STEP standard.

11.1 LOTAR

The Long Term Archiving and Retrieval in the Aerospace Industry (LOTAR) project,[17] undertaken by the ProSTEP iViP Association and AECMA-STAN, is an attempt to specify standards for the long term archiving of (3D) CAD models and PDM documents for Aerospace projects. In particular, the project hopes to specify standards for archiving, methods, scenarios, detailed process descriptions and process modules, suitable data schemata, a system architecture framework, and recommended practices.

The LOTAR White Paper [2002] makes extensive use of the OAIS Reference Model. It uses the Functional Model as the foundation for a set of scenarios summarising the requirements for an industrial repository [§7], the OAIS high-level view of external interactions for data exchange guidelines [§8], and the OAIS perspective on digital migration motivators as the basis for a discussion of system architecture [§8.2.2]. When proposing a technical solution [§9], the White Paper uses the Environment Model for recommended roles, the Information Model for the LOTAR data concept, and the Functional Model for recommended processes; a combination of all three is used for a recommended system architecture.

LOTAR will use STEP [ISO 10303] as its recommended platform-neutral archival format, in particular Application Protocol 214 [ISO 10303-214:2003].

11.2 STEP

The Standard for the Exchange of Product model data (STEP) is a large standard made up of multiple parts (currently about 350) in ten categories. There are a number of parts in the standard that are of interest with respect to content packaging.

Part 21, 'Clear text encoding of the exchange structure' [ISO 10303-21:2002], defines a method of recording data in an ASCII text STEP File using the EXPRESS data modelling language [ISO 10303-11:2004]. A STEP File consists of a header section and one or more data sections. The header consists of a file description, a file name (consisting of a name, date/time stamp, author, organisation, the preprocessor which wrote the file, the system that created the data in the file, and person authorising the file), file schema used, file population (for covering data from various data sections by a single schema), section language (for specifying the language of free-text fields) and section context. The data section lists entities and values according to the specified schema(ta).

Application Protocol 203, 'Configuration controlled 3D design of mechanical parts and assemblies (modular version)' [ISO/TS 10303-203:2005], and Application Protocol 214, 'Core data for automotive mechanical design processes' [ISO 10303-214:2003], provide schemata for detailing

[17] URL: ⟨http://www.prostep.org/en/projektgruppen/lotar/⟩.

the design phase of a product, specifically cars in the case of AP214. AP203 specifically deals with three-dimensional models of mechanical parts and assemblies, while AP214 is broader, covering simulation data, identification of standard and bespoke parts/tools and their properties, design change documentation, suppliers and contracts, and release and approval documentation.

Application Protocol 232, 'Technical data packaging core information and exchange' [ISO 10303-232:2002] provides a schema for technical data packages, which typically include drawings and associated lists, and are used for various purposes, from submitting product definition concepts for evaluation to full product disclosure. In addition to specifying the documentation that should be included and its format, AP232 also requires certain pieces of context information (relating the product and processes to existing specifications and standards, identifying the relationships between the various elements of the package) and representation information (file formats, etc. for the various elements, storage media).

12 Conclusion

The OAIS Reference Model is a useful vocabulary for discussing the preservation of digital objects in a repository context. Within KIM, the issues discussed in this paper may be helpful when considering: how to package together the component parts of the extended product model (T1.1); how to structure a set of documents without altering them (T1.2); and how to add other forms of metadata to a set of documents without altering them (T1.2, link to WP2).

13 Acknowledgement

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) and the Economic and Social Research Council (ESRC) under Grant Numbers EP/C534220/1 and RES-331-27-0006.

References

Bass, Michael J.; David Stuve; Robert Tansle; Margret Branschofsky; Peter Breton; Peter Carmichael; Bill Cattey; Dan Chudnov; & Joyce Ng. 2002. 'DSpace internal reference specification: Technology and architecture.' URL: ⟨`http://www.dspace.org/technology/architecture.pdf`⟩.

Bekaert, Jeroen; Patrick Hochstenbach; & Herbert van de Sompel. 2003. 'Using MPEG-21 DIDL to represent complex digital objects in the Los Alamos National Laboratory Digital Library.' *D-Lib Magazine* 9(11). ISSN 1082-9873. doi:10.1045/november2003-bekaert. URL: ⟨`http://www.dlib.org/dlib/november03/bekaert/11bekaert.html`⟩.

Caplan, Priscilla. 2004. 'DAITSS overview.' URL: ⟨`http://www.fcla.edu/digitalArchive/pdfs/DAITSS.pdf`⟩.

CCSDS. 1992. 'Standard Formatted Data Units structure and construction rules.' Blue Book CCSDS 620.0-B-2, Consultative Committee for Space Data Systems. Also published as ISO 12175:1994. URL: ⟨`http://www.ccsds.org/documents/620x0b2.pdf`⟩.

———. 2000. 'The data description language EAST specification (CCSD0010).' Blue Book CCSDS 644.0-B-2, Consultative Committee for Space Data Systems. Also published as ISO 15889:2003. URL: ⟨`http://www.ccsds.org/documents/644x0b2.pdf`⟩.

———. 2002. 'Reference model for an Open Archival Information System (OAIS).' Blue Book CCSDS 650.0-B-1, Consultative Committee for Space Data Systems. Also published as ISO 14721:2003. URL: ⟨http://www.ccsds.org/documents/650x0b1.pdf⟩.

———. 2004. 'XML Formatted Data Unit (XFDU) structure and construction rules.' White book, Consultative Committee for Space Data Systems. URL: ⟨http://sindbad.gsfc.nasa.gov/xfdu/pdfdocs/iprwbv2a.pdf⟩.

Celeste, Eric & Margret Branschofsky. 2002. 'Building DSpace to enhance scholarly communication.' In: *E-Serials: Publishers, Libraries, Users and Standards*, ed. Wayne Jones, 2nd ed., pp. 239–247. New York: Haworth Press. ISBN 0-7890-1229-4. URL: ⟨http://www.dspace.org/news/articles/celeste.pdf⟩.

Dobratz, Susanne & Astrid Schoger. 2005. 'Digital repository certification: A report from Germany.' *RLG DigiNews* 9(5). ISSN 1093-5371. URL: ⟨http://www.rlg.org/en/page.php?Page_ID=20793#article3⟩.

DSpace. 2005. 'Asset store.' Wiki page. URL: ⟨http://wiki.dspace.org/AssetStore?action=recall&date=1123280360⟩.

ERPANET. 2003. *ERPANET OAIS Training Seminar Report*. København, Denmark. 28–29 Nov. 2002.

Fedora Development Team. 2005. 'Fedora open source repository software.' White Paper. URL: ⟨http://www.fedora.info/documents/WhitePaper/FedoraWhitePaper.pdf⟩.

ISO 10303. 'Industrial automation systems and integration – Product data representation and exchange.' Multipart standard.

ISO 10303-11:2004. 'Industrial automation systems and integration – Product data representation and exchange – Part 11: Description methods: The EXPRESS language reference manual.'

ISO 10303-21:2002. 'Industrial automation systems and integration – Product data representation and exchange – Part 21: Implementation methods: Clear text encoding of the exchange structure.'

ISO 10303-214:2003. 'Industrial automation systems and integration – Product data representation and exchange – Part 214: Application protocol: Core data for automotive mechanical design processes.'

ISO 10303-232:2002. 'Industrial automation systems and integration – Product data representation and exchange – Part 232: Application protocol: Technical data packaging core information and exchange.'

ISO/IEC 21000-2:2005. 'Information technology — Multimedia framework (MPEG-21) — Part 2: Digital Item Declaration.' URL: ⟨http://standards.iso.org/ittf/PubliclyAvailableStandards/c041112_ISO_IEC_21000-2_2005(E).zip⟩.

ISO/TS 10303-203:2005. 'Industrial automation systems and integration – Product data representation and exchange – Part 203: Application protocol: Configuration controlled 3D design of mechanical parts and assemblies (modular version).'

Lavoie, Brian F. 2004. 'The Open Archival Information System reference model: Introductory guide.' DPC Technology Watch Series Report 04-01, Digital Preservation Coalition. URL: ⟨http://www.dpconline.org/docs/lavoie_OAIS.pdf⟩.

LOTAR. 2002. 'Long term archiving and retrieval of product data within the aerospace industry (LOTAR): Technical aspects of an approach for application.' White paper, ProSTEP iViP Association. URL: ⟨http://www.prostep.org/file/14037.wp_v10⟩.

National Library of New Zealand. 2000. 'Metadata standards framework for National Library of New Zealand.' URL: ⟨http://www.natlib.govt.nz/files/4initiatives_metafw.pdf⟩.

———. 2002. 'Metadata standards framework: Preservation metadata.' URL: ⟨http://www.natlib.govt.nz/files/4initiatives_metaschema.pdf⟩.

———. 2003. 'Metadata standards framework: Preservation metadata (revised).' URL: ⟨http://www.natlib.govt.nz/files/4initiatives_metaschema_revised.pdf⟩.

OCLC & RLG. 2002. 'Preservation metadata and the OAIS Information Model: A metadata framework to support the preservation of digital objects.' A Report by the OCLC/RLG Working Group on Preservation Metadata. URL: ⟨http://www.oclc.org/research/projects/pmwg/pm_framework.pdf⟩.

———. 2005. 'Data dictionary for preservation metadata.' Final report of the PREMIS Working Group. URL: ⟨http://www.oclc.org/research/projects/pmwg/premis-final.pdf⟩.

Pinna, Gian Maria; Vincenzo Beruti; Stephane Mbaye; Mathias Moucha; Valter Spaventa; & Davide Castellazzi. 2005*a*. 'From HARM to SAFE: the ESA's proposal for a standard archive format for Europe.' In: *PV 2005: Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data*. Edinburgh. 21–23 Nov. URL: ⟨http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/018.pdf⟩.

Pinna, Gian Maria; Eberhard Mikusch; Manfred Bollner; & Bernard Pruin. 2005*b*. 'Earth observation payload data long term archiving: the ESA's Multi-Mission Facility Infrastructure.' In: *PV 2005: Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data*. Edinburgh. 21–23 Nov. URL: ⟨http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/005.pdf⟩.

Rank, Robert & Kenneth R. McDonald. 2005. 'A NOAA/NASA pilot project for the preservation of MODIS data from the Earth Observing System (EOS).' In: *PV 2005: Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data*. Edinburgh. 21–23 Nov. URL: ⟨http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/037.pdf⟩.

RLG & NARA. 2005. *An Audit Checklist for the Certification of Trusted Digital Repositories: Draft for Public Comment*. Mountain View, CA: RLG. URL: ⟨http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf⟩.

RLG & OCLC. 2002. *Trusted Digital Repositories: Attitudes and Responsibilities*. Mountain View, CA: RLG. URL: ⟨http://www.rlg.org/en/pdfs/repositories.pdf⟩.

Sawyer, Donald; H. Kent Hills; Pat McCaslin; & John Garrett. 2005. 'Performing a migration in the framework of the OAIS Reference Model: NSSDC case study.' In: *PV 2005: Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data*. Edinburgh. 21–23 Nov. URL: ⟨http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/042.pdf⟩.

Sawyer, Donald; Lou Reich; David Giaretta; Patrick Mazal; Claude Huc; Michel Nonon-Latapie; & Nestor Peccia. 2002. 'The Open Archival Information System (OAIS) Reference Model and its usage.' In: *SpaceOps2002*. Houston, TX. 9–12 Oct. URL: ⟨http://www.ccsds.org/documents/so2002/spaceops02_p_t5_39.pdf⟩.

van de Sompel, Herbert; Jeroen Bekaert; Xiaoming Liu; Luda Balakireva; & Thorsten Schwander. 2005. 'aDORe: A modular, standards-based digital object repository.' Preprint. URL: ⟨http://arxiv.org/abs/cs/0502028⟩.

All links were correct on 31 January 2006.

This work is licensed under the Creative Commons Attribution-ShareAlike 2.0 England & Wales Licence. To view a copy of this licence, visit ⟨http://creativecommons.org/licenses/by-sa/2.0/uk/⟩ or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.