POWRR Institute, Day One

Technology Instruction Module – OAIS Speakers Notes

# Slide 1 - Title Slide

* This section will provide an overview of the Open Archival Information System Reference Model.
* It will introduce key terms and concepts as well as giving guidance on how to practically implement aspects of the model.

# Slide 2 – Why Do We Need Models?

* Before getting into the detail of OAIS it is important to understand why having models for activities like digital preservation is important.

Models:

* Provide a high level conceptual map of the generic activities that need to be undertaken, providing a view of the bigger picture.
* Allow us to start understanding our needs and set requirements for the policies, documentation, tools and systems we will look to develop.
* Provide a consensus between different organizations and practitioners allowing them to identify and develop standards that will benefit their efforts.
* Can be a basis for comparing and assessing the approaches taken by different organisations. In digital preservation, this has seen the development of audit methodologies.

# Slide 3 – What is OAIS?

* The acronym OAIS can be used to refer to the Reference Model for an Open Archival Information System standard itself, or a repository that adheres to the standard (an OAIS).
* The standard originally was developed out of the space data community through the Consultative Committee for Space Data Systems.
* It has since been established as an international standard, the 1st edition published in 2006, and a 2nd edition in 2012. A 3rd edition is currently in development.
* The full text is still available for free via the CCSDS website in the form of the ‘Magenta Book’
* OAIS is particularly important as it provides a basic common framework for communicating about digital preservation activities and has provided much of the vocabulary used by practitioners.
* An important caveat: while understanding OAIS will facilitate your work in digital preservation it is important to remember that you may not want to use all of the standard, instead using what is useful/relevant to your organization.

# Slide 4 – Basic Definition of an OAIS

* The basic definition of an OAIS is included here
* It is important to note that is mentions both systems and organizational activities
  + Not just technology
* It also establishes that the OAIS assumes a responsibility for the information within it, to be preserved for a particularly
* Designated community is a key term from OAIS which is important to understand

# Slide 5 – Designated Communities

* The designated community is an identified group of primary users for the OAIS’ content.
* The size of this group may depend on both the content/collection and the organization.
* It can be anything from one person to the whole world
* It is important to define your designated community as their needs should shape decisions about how the content of the OAIS will be managed and preserved.
* The designated community must be monitored on an ongoing basis to capture changing needs.

# Slide 6 – What an OAIS Encompasses

* An OAIS covers all activities between the creation of the data to its use by the designated community.
* This encompasses activities such as:
  + Acceptance into the repository
  + Cataloguing
  + Storage
  + Preservation action
  + Access provision
* Important to note: it is intended that the model defined by the standard is meant to be relevant for both physical and digital collections.

# Slide 7 – OAIS Mandatory Responsibilities

* The OAIS model includes a list of 6 mandatory responsibilities that a repository must fulfil to be an OAIS.
* This slide includes an edited list of these responsibilities.
* They emphasise:
  1. Only accepting appropriate content: which can relate to subject matter but also taking data you have confidence you have the capability to preserve
  2. The importance of appropriate deposit agreements and licences
  3. That the repository must monitor its community
  4. That documentation must be collected that allows data reuse
  5. That policies and procedures are created and followed
  6. That data is made available to the identified designated community

# Slide 8 – Functional Model

* The OAIS standard includes within it a number of models describing aspects of an OAIS.
* The main two are the functional model and the information model.
* Pictured here is the high level functional model which shows:
  + The OAIS **functional entities**
  + The **information objects** that they interact with
  + The **actors** who interact with the OAIS
* The following slides will examine these in more detail

# Slide 9 – Actors

The standard defines 3 actors who interact with the OAIS:

* Producer – The individuals, organizations, or systems that transfer information to the OAIS for long-term preservation. Interactions include negotiation, completing agreements and transfer of data.
* Management – The individuals or organizations responsible for formulating, revising and enforcing the high-level policy framework governing an OAIS. They may also provide the funding required to run the OAIS.
* Consumer – The individuals, organizations, or systems that use the information preserved in the OAIS. Interactions may include searching catalogs and making requests for access. The designated community is a special subset of potential consumers.

# Slide 10 – Objects

* The objects shown in the functional model here is the information to be preserved in various states as it is processed in the OAIS.
* These objects are referred to as Information Packages and there are three types:
  + Submission Information Packages
  + Archive Information Packages
  + Dissemination Information Packages
* Information Packages contain the original information object as well as any necessary documentation and metadata.
* We will examine the structure of information packages further when we look at the Information Model.

# Slide 11 –Functional Entities

The functional entities defined by the OAIS standard are the core set of mechanisms that allow it to deliver it primary mission of preserving information. They are:

* 🖰 **Ingest:** this includes the set of processes responsible for accepting information submitted by producers and preparing it for archival storage. This may include: checking data is uncorrupted and complete, virus checking, and creation of finding aids/catalog information.
* 🖰 **Archival Storage:** performs the storage function of the OAIS, including ensuring information is stored in the appropriate location, error-checking and refreshing storage as required.
* 🖰 **Data Management:** this function maintains the descriptive and administrative data required for the management of the OAIS.
* 🖰 **Administration:** undertakes the day-to-day management of the OAIS and coordinates the activities of the other functional entities. This includes varied tasks such as negotiating with producers, managing policies, and customer service.
* 🖰 **Preservation Planning:** maps the preservation strategy for the OAIS, including monitoring changes in technology and the designated community, as well as developing preservation plans.
* 🖰 **Access:** manages the processes and services which provide the consumers with access to the information held by the OAIS.

# Slide 12 - Full Functional Model

* OAIS also includes a more detailed breakdown of the individual functions within each entity
* This model better shows the interactions between the entities and the processes required to operate an OAIS
* Will not look at this in detail but it can be useful when designing processes and procedures

# Slide 13 - More on Information Packages

* Introduced earlier the main information objects with the OAIS model – Information Packages
* They are the originally digital objects plus relevant documentation and metadata and come in three forms:
  + Submission Information Package
  + Archive Information Package
  + Dissemination Information Package
* Remaining slides in this section will examine the structure of IPs in more detail and how you can begin designing/constructing them

# Slide 14 – The Basic Model

* On the screen now if the basic model for an information object within an OAIS, which is a sub-section of the full information model.
* The model shows:
  1. 🖰 The actual **data object** which contains the information to be preserved. This can be a **physical object** or a **digital object** which can be composed of one or more bitstreams or files.
  2. 🖰 The **data object** is accompanied by **representation information**, which is any documentation or metadata required to provide access to the data object.
  3. 🖰 The **representation information** is used to interpret the **data object** (be it physical or digital) to provide access to an **information object** which is understandable by a user (in OAIS the Consumer).

# Slide 15 – Rendering the Data Object

* In OAIS: The **Data Object** is interpreted by the **Representation Information** to yield a useable **Information Object**
* A comparable example in the physical world: A film reel is interpreted by a project and yields on screen a movie.
* The **Representation Information** is the mechanism for providing access to the preserved information.

# Slide 16 – Representation Information

* Two types of representation information:
  + Structure information – about the technical requirements of rendering the data object, includes information such as file format and software package.
  + Semantic information – how to interpret and understand the data object, can include user documentation on a software package or a data dictionary identifying columns in a database.
* Representation information can be simple or very complex, and may include things like full software packages. Some representation information may require its own representation to interpret it.
* How much representation information is saved is determined by the needs of the designated community. Less may be needed for a specialist user group and more for the general public.
* Needs tend to become more complex over time as users become less familiar with the data objects and environments they are created in.

# Slide 17 – Full Information Model

* The full information model shows what is included within a complete **Information Package**.
* This includes:
  + 🖰 The **Data Object** and **Representation Information** as shown in the basic model
  + 🖰 As well as **Preservation Description Information**, which is other documentation and metadata required to understand data object and manage it over time
  + 🖰 Along with the **Information Package** there is also a **Package Description** (descriptive information about the package, e.g. a catalog entry) and **Package Information** (how/where the information contained with the information package is stored)
* This all seems very complicated can the amount of information captured can vary widely depending on the needs of the designated community and the capacity of the preserving organisation.

# Slide 18 – Preservation Description Information

* **Preservation Description Information** in the additional metadata that is required to support and document the OAIS’s preservation processes.
* This includes:
  + Recording preservation activities that have been undertaken
  + Retaining metadata to demonstrate the continued authenticity of the data objects
  + Providing metadata to facilitate the dissemination of the data objects
* The OAIS models describes this as metadata that “is specifically focused on describing the past and present states” of the data objects.
* There are five component of **Preservation Description Information**:
  + **Reference Information** – a unique identifier for the data object within the OAIS and, potentially, out with it. Examples may include an ISBN or a persistent identifier like at DOI.
  + **Context Information** – describes the data objects relationship to others within the OAIS. For example, if it belongs to a particular collection, or to other versions of the same document.
  + **Provenance Information** – captures information about the history of the data object, which can include who created it, chain of custody and any actions to preserve it (e.g. migration).
  + **Fixity Information –** documents that no changes have been made to the data object through mechanisms such as checksums, digital signatures or watermarks.
  + **Access Rights Information –** captures any restrictions relating to the data objects in relation to preservation and access. This may include information on licences, lists of those with permission to access, and information on agreed preservation options from the depositor agreement.

# Slide 19 – PREMIS

* PREMIS is a metadata standard for preservation metadata and has become the de facto for digital preservation.
* It covers many of the elements required within the OAIS standard’s **Representation Information** and **Preservation Description Information** but is specifically focused on the needs of those managing the repository and the repository systems themselves.
* It does not claim to be comprehensive and use of additional standards or custom metadata elements may be required.
* A key example is that the PREMIS standard does not include elements for descriptive metadata. Users are instead advised to use existing standards such as MARC, ISAD(G) or Dublin Core.
* The PREMIS model gathers metadata into five entities. These are:
  + **Object** – information on the individual data objects, e.g. size or file format
  + **Intellectual Entities** – Metadata about the relationships between objects, e.g. if a number of objects combine to make a book
  + **Rights –** information on digital rights and permissions, e.g. who can access the object or license details
  + **Events** – records of events that involve the objects, e.g. when it was ingested or details of a preservation action
  + **Agents** – Details about people, organizations or software who relate to rights or events relevant to the object, e.g. the copyright holder or the member of staff who carried out the preservation action.
* It is important to always consider what is necessary and what is feasible when deciding what metadata to capture. It is advisable to carefully consider how much of PREMIS you might use as it is a large and complex standard, only adopting the elements that are necessary.

# Slide 20 – Information Package Structures

* How information packages are implemented will differ from organization to organization.
* Decisions on how to structure the packages will depend on a number of factors. Need to consider:
  + What are the needs of your designated community? This will influence choices in relation to what metadata is kept and where as well as driving all decisions about Dissemination Information Packages.
  + Do you have any existing systems that you will need the information packages to be compatible with? Will these determine the formats and/or technology you will use?
  + What resources do you have available? Consider skills as well as costs.
  + Have you already made decisions for preservation? This may determine the amount of metadata needed and what form of the data object to store in the Archival Information Packages.
* The options for formats are virtually endless and can be tailored to specific organizations. Common choices include:
  + Storing the data objects in specially ordered folder structures.
  + Capturing metadata in a database with links to the location of the data objects
  + ‘Wrapping’ the data objects and their metadata using an XML file that includes file locations and some or all of the metadata.
* There are a number of tools available for managing the creation of information packages. Most commercial repository systems include this functionality but there are also free or opensource tools available, the most popular being BagIt.

# Slide 21 – METS

* METS – Metadata Encoding and Transmission Standard is an XML-based standard for the encoding of the metadata necessary for the management of digital data object.
* It also helps ensure information packages are transferrable between different repository systems.
* It is one of the most commonly used structures for creating information packages and is compatible with the PREMIS standard.
* A METS document contains seven major sections:
  + **METS Header** – information about the METS document itself e.g. creator
  + **Descriptive Metadata** – may include a pointer to descriptive metadata held elsewhere (e.g. in a database) or contain it within the METS document. Multiple descriptions can be included.
  + **Administrative Metadata** – information required to manage the data object, e.g. provenance information.
  + **File Section** – a list of all of the individual files that make up the data object, if there are several it is possible to group these if needed.
  + **Structural Map** – The heart of the METS document, stores the structure of the data object, including both how files relate to each other and how metadata relates to the parts of the object.
  + **Structural Links** – allows the storage of links between different parts of the object as defined in the Structural Map. Particularly useful for websites as hyperlinks between pages can be stored.
  + **Behavior** – captures information on behaviors of the data object, including mechanisms for their execution.
* Many organizations use both METS and PREMIS metadata standards as between them they cover most of the metadata and packaging requirements of OAIS.

# Slide 22 – E-ARK Project

* As mention before, there are many different approaches to the creation of information packages, varying from organization to organization.
* The key aim of the E-ARK project was to develop a standardized structure for information packages to allow them to be easily transferrable between different repository systems.
* The project developed a Common Specification for Information Packages based on the METS standard.
* The specification has additional elements relating to each type of information package described in OAIS.
* The project developed a number of open source tools for the creation and management of information packages, details of which are available on the project website.
* All of the specifications and tools produced by the project were tested in real-world pilots carried out by a number of organizations around Europe using their own data.

# Slide 23 – OAIS in the Wild

* Final few slides in this section will provide a short case study of how OAIS was used by an archive to carry out a gap analysis of their digital preservation capabilities and needs.

# Slide 24 – An Introduction to RCAHMS

*Acronym commonly pronounced ‘R-CAMS’*

* The Royal Commission on the Ancient and Historical Monuments of Scotland (now part of Historic Environment Scotland) was a medium-sized organization based in Edinburgh, Scotland.
* Its mission was to maintain a record of Scotland’s built heritage through architectural and archaeological surveys and the collecting and management of its archive collections.
* RCAHMS collections were built from the outputs of the organization’s survey work and from material collected from external depositors (primarily architectural firms, archaeological companies and individuals with an interest in built heritage).

# Slide 25 – First Digital Archive

* RCAHMS received its first deposit containing digital data in 1992, from an archaeological dig. At this time (and for the next decade) the digital data was stored on the original media with physical archives.
* In 2002, it was decided that steps needed to be taken to manage this information and, in partnership with the UK’s Archaeology Data Service, a report was produced detailing RCAHMS’ digital preservation needs.
* A 6-week contract was awarded in to 2003 to an external contractor to develop solutions for digital archiving. At that time, limited tools and standards were available.
* The solutions that were developed were:
  + An area in the RCAHMS catalogue database to record limited technical metadata
  + A dedicated ‘digital archive’ storage area was created on the organization’s network.
  + A tool was created for batch processing digital images, including automated generation of metadata.
* These solutions only address a limited number of issues relating to ingest and archival storage. No consideration was given to solutions for preservation and dissemination.

# Slide 26 – Motivations for Redevelopment

* In 2006, RCAHMS hired a new Digital Archivist and this was the initial catalyst to a review of the organization’s digital preservation activities.
* Over the previous 5 years RCAHMS had also seen an exponential growth in the number of deposits they received containing digital material, with:
  + Over 500,000 digital objects now in the archive
  + New and more complex data types being received, particularly things like CAD, GIS data and laser scanning data
  + The current systems were not coping with the strain of the increased input
* During this time, a number of new standards and tools had been developed meaning:
  + There were opportunities for new and enhanced systems and procedures
  + The current systems were far from adhering to ‘best practice’ and did not meet minimum standards for preservation.
* It was agreed that a more strategic approach to the management and development of the digital archive was required to make better use of the limited resources that were available.

# Slide 27 – A Plan of Action

* Realised it was important to approach the redevelopment carefully, thinking about the process and ultimate aims. Questions considered included:
  + What might success look like? And how could it be measured?
  + How could buy-in from stakeholders both inside and outside the organization be achieved?
  + How could the best and most relevant standards and tools be identified?
* Quickly established that OAIS would be core to the redevelopment process. The standard helped:
  + As a benchmark for setting aims and objectives for what was to be achieved.
  + Provide a framework for planning what standards and tools were needed and how they should fit together.
  + As the basis for a gap analysis comparing the existing systems. Current workflows were mapped against proposed/ideal workflows and gaps identified. This was a particularly successful process.

# Slide 28 – Gap Analysis: Ingest

* Using the full functional model of OAIS as the basis, ideal workflows were created for Ingest, Archival Storage, Preservation, and Dissemination. The slide shows the Ingest workflow.
* Each element of the workflow was labelled with the number of the corresponding function from the full functional model.
* A comparison was then made between the existing workflows and the ideal workflows. Each element was color-coded depending on if:
  + The element existed and was fit for purpose – the black circles
  + The element existed but was not fit for purpose and required upgrading – beige circles
  + The element did not exist – white circles
* It can be easily seen that a very large gap exists, covering almost the entire workflow.
* These graphics proved very useful for planning developments but also for advocacy purposes. They were instrumental in gaining senior manager support and having a digital repository programme established as a key aim of the organization.
* A fuller account of this work can be found in an article in the Journal of the Society of Archivists, listed in the course resources.

# Slide 29 – Questions?

Potential discussion points:

* Are there aspects of the OAIS standard you think would help with your digital preservation work?
* Have you considered using any other standards?
* Do you capture any metadata? What parts of the information model does it cover?