



Digital**Preservation**Coalition



Technology Watch Report

The Open Archival Information System Reference Model: Introductory Guide

Brian F. Lavoie

**Office of Research
OCLC Online Computer Library Center, Inc.**

6565 Frantz Road, Dublin OH USA 43017-3395

lavoie@oclc.org

DPC Technology Watch Series Report 04-01
January 2004

© OCLC Online Computer Library Center, Inc. and Digital Preservation Coalition 2004

I. Background

The impact of digital information environments has been remarkably universal, extending to industry, government, and the academy; to businesspersons, scientists, engineers, and scholars of the humanities; to the individual in the workplace and the individual in the home. Vast quantities of information in digital form – text, images, audio, video, Web pages, computer programs, databases – are produced, exchanged, and used in a variety of settings, for myriad purposes. These diverse applications of digital technology rest on a common foundation of shared benefits, including powerful search and retrieval capabilities, network delivery, perfect duplication, and interoperability.

Just as the benefits of digital information environments transcend people, systems, and domains, so do the challenges which accompany them. Nowhere is this more evident than in regard to *digital preservation* – securing the long-term persistence of information in digital form. The capacity both to create and consume digital information has advanced steadily; unfortunately, the capacity to manage the long-term stewardship of this information has been comparatively slow to develop. The problem is exacerbated by the relatively brief time horizon beyond which preservation of digital materials becomes an imperative, a consequence of the fragility of digital storage media, as well as rapid obsolescence of storage and rendering environments.

The immediacy of the preservation requirements associated with digital materials has confronted organizations of all descriptions – cultural heritage institutions, businesses, government agencies, etc. – with the need to take steps to secure the long-term viability of the digital materials in their custody. Many of these entities do not perceive an archival function within the scope of their organizational mission. Yet the ubiquity of the digital preservation issue establishes common ground for cross-domain dialog and cooperation in addressing the challenges of digital preservation. Moreover, efforts to create digital preservation solutions in one community often produce a ripple effect impacting a host of apparently unrelated communities. So it was when the space data community began to think about its own digital preservation problem.¹

II. Genesis of the OAIS Reference Model

The Consultative Committee for Space Data Systems (CCSDS), established in 1982, is a forum for national space agencies interested in the cooperative development of data handling standards in support of space research.² In 1990, the CCSDS launched a cooperative arrangement with the International Organization for Standardization (ISO)³, whereby CCSDS Recommendations – i.e., recommended solutions to the data handling problems shared by its membership – would undergo normal ISO review and voting procedures in the course of becoming formal ISO standards.

¹ The following discussion pertains to the *Reference Model for an Open Archival Information System* (Blue Book version), January 2002. See References for access information. The discussion is based on a talk given by the author at the Museum Computer Network Annual Conference in September 2002.

² For more information on the CCSDS, visit their Web site at <http://www.ccsds.org/>.

³ Specifically, this arrangement was made with the ISO's Subcommittee 13 (space data and information transfer systems) under Technical Committee 20 (aircraft and space vehicles).

At the request of the ISO, the CCSDS initiated work aimed at developing formal standards for the long-term storage of digital data generated from space missions. In preparing for this effort, the CCSDS found no widely-accepted framework that could serve as a foundation for standards-building activities: nothing, for example, that established shared concepts and terminology associated with digital preservation; characterized the basic functions constituting a digital archiving system; or defined the important attributes of the digital information objects towards which preservation efforts would be directed. In short, there was no perceived consensus on the needs and requirements for maintaining digital information over the long-term. A unifying framework that could fill this gap would be invaluable in terms of encouraging dialog and collaboration among participants in standards-building activities, as well as identifying areas most likely to benefit from standards development.

In the absence of such a framework, the CCSDS determined that its first step should be to create one. An international workshop convened by the CCSDS in 1995⁴ validated this strategy, and a proposal was advanced to develop a *reference model for an open archival information system*. The reference model would define the basic functional components of a system dedicated to the long-term preservation of digital information⁵, detail the key internal and external system interfaces, and characterize the information objects managed by the system. These descriptions would be expressed in terms of a well-defined set of concepts and terminology transcending, yet mappable to, domain-specific vocabularies. The reference model would also enumerate a set of minimum requirements an archival system is expected to meet. When complete, the reference model would represent a comprehensive and consistent framework for describing and analyzing digital preservation issues, provide a sound footing for future standards-building activity, and serve as a point of reference for vendors interested in building digital preservation products and services.

From the earliest stages of the reference model's development, the CCSDS recognized that its relevance extended well beyond the space data community. The reference model would address fundamental questions regarding the long-term preservation of digital materials – questions which cut across domain-specific implementations. Consequently, the decision was made to make the process of crafting the model open to any interested individual or organization. In adopting this ecumenical approach, the CCSDS reached beyond the relatively narrow purposes of the space data community to engage a diverse collection of organizations in government, private industry, and academia. The reference model represented common ground upon which to consolidate understanding of the needs and requirements of digital preservation: an opportunity to gather the strands of isolated digital preservation activities, merging them into a shared (albeit highly conceptual) characterization of the problem's boundaries.

⁴ Proceedings from this workshop are available at <http://ssdoo.gsfc.nasa.gov/nost/isoas/int01/ws.html>

⁵ Strictly speaking, the reference model could be applied to the long-term preservation of items in any form, including physical artifacts. The authors of the reference model take pains to remind readers that there is no explicit assumption that the items which are the focus of preservation are in fact digital. However, it is in the digital realm that the OAIS reference model has gained its widest visibility and acceptance, and it is within this context that it is discussed in this report.

The reference model was developed through an open, iterative process of drafting, review, and revision; community feedback was provided through face-to-face workshop discussions, and as written responses to formal requests for comment. Draft versions of the reference model were released for review in May 1997 and May 1999; it was approved and published as a draft ISO standard in June 2000. After a final period of review and revision, the reference model was approved in January 2002 as international ISO standard 14721.⁶

III. Open Archival Information System

The central concept in the reference model is that of an *open archival information system* (OAIS)⁷. The term *open* refers to the fact that the reference model was developed and released in an open public forum, in which any interested party was encouraged to participate. An *archival information system* is “an organization of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community.”⁸ This definition emphasizes two primary functions for an archival repository: first, to *preserve* information – i.e., to secure its long-term persistence – and second, to *provide access* to the archived information, in a manner consistent with the needs of the OAIS’s primary users, or Designated Community. The concept of a Designated Community will be discussed in the next section.

On the surface, this definition does little to distinguish the reference model’s use of the term “archive” from its usage in other contexts. However, the definition is supplemented with a list of *mandatory responsibilities* that an OAIS-type archive is expected to meet. In particular, an OAIS must:⁹

- Negotiate for and accept appropriate information from information producers
- Obtain sufficient control of the information in order to meet long-term preservation objectives
- Determine the scope of the archive’s user community
- Ensure that the preserved information is independently understandable to the user community, in the sense that the information can be understood by users without the assistance of the information producer
- Follow documented policies and procedures to ensure the information is preserved against all reasonable contingencies, and to enable dissemination of authenticated copies of the preserved information in its original form, or in a form traceable to the original
- Make the preserved information available to the user community

⁶ All versions of the reference model are available at http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

⁷ The acronym “OAIS” is sometimes confused with “OAI” (Open Archives Initiative), a similarly named but entirely different activity. OAI promotes interoperability through the development of standards and protocols for the dissemination of content. Hirtle (2001) has written a short piece discussing the differences between OAIS and OAI.

⁸ OAIS reference model, p. 1.1

⁹ The responsibilities listed below are paraphrased and edited to avoid the use of terminology that has not yet been introduced, and for brevity. The responsibilities in their original form are listed on page 3-1 in the reference model documentation.

The first responsibility of an OAIS-type archive is to establish criteria for determining which materials are appropriate for inclusion in the archival store. These criteria might be based on such factors as subject, origin, or format. Once the scope of the archival collection is defined, appropriate steps must be taken to motivate the producers/owners of the targeted items to transfer them into the custody of the OAIS for preservation. But it is not enough simply to acquire custody of the items. The second responsibility emphasizes the need for the OAIS to obtain sufficient intellectual property rights, along with custody of the items, to authorize the procedures necessary to meet preservation objectives. For example, if the OAIS must create a new version of the archived item so that it can be rendered by current technologies, it must have the explicit right to do so.

Another responsibility of an OAIS-type archive is to determine the scope of its primary user community. This will be discussed in greater detail in the next section, but the key point is that an accurate characterization of the primary users of the archived information is a pre-condition for meeting another of the OAIS's responsibilities: ensuring that the information is preserved in a form that is independently understandable to these users. The production of information always occurs in some context, and it is often the case that understanding this context is necessary to fully understand the information itself. Given this, the OAIS must not only preserve information, but also a sufficient portion of its associated context to ensure that the information is *understandable*, and ultimately, *useable* by future generations. "Contextual information" that might be preserved includes, but is not limited to, a description of the structure or format in which the information is stored, explanations of how and why the information was created, and even its appropriate interpretation. Delineating the scope of the primary user community is essential for determining how much of this context should be preserved along with the information itself. This in turn has important implications for the metadata requirements needed to support the archived information.

The final two OAIS responsibilities concern the preservation process, and the mechanisms for making the archived information available to the user community. In regard to the former, an OAIS should establish and document clear policies and procedures for carrying out the preservation of the information in its custody; these should be accessible to and understandable by stakeholders in the OAIS, as well as in conformance with a set of clearly defined preservation objectives. Finally, an OAIS should be committed to making the contents of its archival store available to its intended user community, through the implementation of access mechanisms and services which support, to the extent possible, users' needs and requirements.

In summary, use of the term *OAIS*, or equivalently, the term *archive* in the OAIS context, implies an archival system dedicated to *preserving* digital information and *making it available* over the long-term, as well as meeting, in some form, the six mandatory responsibilities listed above.

The OAIS reference model consists of three separate but related parts, each centered around the concept of an OAIS-type archive. The first part describes the external

environment within which the OAIS operates; the second part describes the functional components, or internal mechanisms, which collectively fulfill the OAIS's preservation responsibilities. The third part describes the information objects which are ingested, managed, and disseminated by the OAIS. The next three sections discuss each of these parts in turn.

IV. OAIS Environment

An OAIS-type archive does not operate in a vacuum; rather, it carries out its preservation and access responsibilities in an environment populated by several key external stakeholders. An OAIS must act in cooperation with these external stakeholders in the course of fulfilling its mission. The reference model identifies and describes the external entities constituting an OAIS's environment, and characterizes the interfaces between these entities and the OAIS.

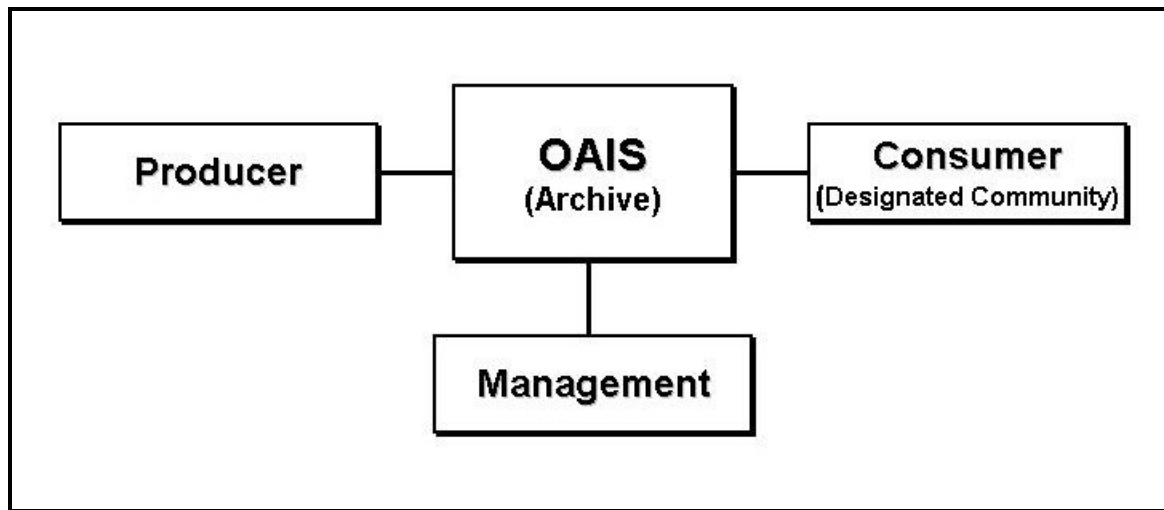


Figure 1: OAIS Environment

Figure 1 illustrates the OAIS environment. The environment comprises three distinct entities external to, and interacting with, the OAIS: Management, Producer, and Consumer. Management's responsibilities include formulating, revising, and in some circumstances, enforcing, the high-level policy framework governing the OAIS's activities. Examples of functions carried out by Management include strategic planning, defining the scope of the OAIS's archived collection, and articulating the preservation "guarantee" associated with items entrusted to the archive. Management may also represent the funding source for the OAIS, and often serves in an oversight capacity, periodically reviewing the OAIS's policies and performance.

It should be noted that Management is *not* responsible for managing the day-to-day operations of the OAIS. This responsibility is handled by a functional component within the archive itself (see section V).

The second external entity interacting with an OAIIS is the Producer (or Producers): the individuals, organizations, or systems that transfer information to the OAIIS for long-term preservation. Producers submit the information to be preserved, along with associated metadata, to the OAIIS via an *ingest process*, which accepts the submitted data and prepares it for inclusion in the archival store. Interaction between the OAIIS and Producers is often formalized and guided by a *Submission Agreement*, which establishes specific details of the interaction such as the type of information submitted, the metadata the Producer is expected to provide, and the logistics of the actual transfer of custody from the Producer to the archive.

In addition to Management and Producers, an OAIIS also interacts with Consumers. As the name suggests, Consumers are the individuals, organizations, or systems expected to use the information preserved by the OAIIS. The reference model goes on to define a special class of Consumers known as the *Designated Community*: the subset of Consumers expected to independently understand the archived information in the form in which it is preserved and made available by the OAIIS. This point was touched on briefly in the previous section: recall that one of the mandatory responsibilities of an OAIIS is to preserve information in such a way that it is independently understandable to its primary users. These primary users are the OAIIS's Designated Community.

If the OAIIS contains scholarly papers and data sets specific to a particular discipline, then the Designated Community might consist of all individuals possessing a certain level of expertise in that discipline, who would use the archived information to inform and motivate basic or applied research. Similarly, if the OAIIS's archived content consists of balance sheets, tax returns, and other financial records pertaining to commercial enterprises, the Designated Community might be government regulatory bodies, accountants, and other financial professionals skilled at synthesizing and interpreting this information. In both of these examples, the contents of the OAIIS may be freely available for use by anyone; in this case, the OAIIS's Consumers would be the general public. But it is only those individuals possessing sufficient specialized knowledge to use the archived information without expert assistance who comprise the OAIIS's Designated Community.

It should not be inferred that the scope of the Designated Community is determined *ex post* by the nature of the archive's contents; rather, it is the scope of the Designated Community that determines both the contents of the OAIIS *and* the forms in which the contents are preserved, such that they remain available to, and independently understandable by, the Designated Community.¹⁰

Determining the scope of the Designated Community is a critical aspect of the preservation process for an OAIIS-type archive. As the discussion in section VI will make clear, the broader the scope of the Designated Community, the greater the metadata requirements necessary to maintain digital materials over the long-term. The Designated Community could extend as far as the public at large, which is tantamount to assuming no particular expertise or specialized knowledge on the part of users of the archived information. But in this case, the task of preserving the information in an "independently

¹⁰ The author thanks an anonymous reviewer for this point.

understandable” form becomes commensurately more difficult. One additional point to note is that the scope of the Designated Community is not necessarily static: there is nothing to preclude the Designated Community from changing over time. Dynamic features of the Designated Community include its extent, as well as the expectations of its members in regard to access and use of the OAIS’s contents.

The concepts of Management, Producers, Consumers, and Designated Community, as well as that of an OAIS, represent *functional* rather than organizational roles. Consequently, all of these roles can be subsumed within a single organizational structure, or distributed across multiple organizations. The key point is not the *physical* separation of one role from another, but rather, the *logical* separation of the decision-making roles and stakeholder interests attached to most digital preservation activities.

An example is useful for understanding the application of the OAIS environment in practice. The National Digital Archive of Datasets (NDAD)¹¹ is a UK-based initiative aimed at preserving computer datasets produced by UK central government departments and agencies. In this scenario, the OAIS is the National Data Repository (NDR) service, a digital preservation and access system operated by the University of London Computer Centre (ULCC). The Management role, however, resides with the UK National Archives, which retains legal custody of the archived datasets and performs a number of high-level functions associated with the NDAD initiative, including the provision of funds and selection of datasets for long-term preservation. The Producers are, of course, the various UK government departments and agencies which, as part of their organizational mission, produce computer datasets. The archived datasets are freely available for use by anyone with Web access, so NDAD’s Consumers appear to be defined in the broadest terms: the general public. A visit to the NDAD Web site suggests that the scope of the Designated Community extends to the general public as well: the Web site notes that apart from Web access, little else is required to use the NDAD database. Moreover, the archived datasets are accompanied by fairly detailed descriptive information, including finding aids that explain “why, how, and when the datasets were created”. In short, no scientific expertise or domain-specific knowledge appears to be required to use the datasets in the NDAD collection; put another way, the datasets are, by and large, “independently understandable” by the general public¹².

V. OAIS Functional Model

The reference model identifies and describes the core set of mechanisms with which an OAIS-type archive meets its primary mission of preserving information over the long-term and making it available to the Designated Community. These mechanisms are summarized by the *OAIS functional model*: a collection of six high-level services, or functional components, that, taken together, fulfill the OAIS’s dual role of preserving and providing access to the information in its custody.

¹¹ <http://ndad.ulcc.ac.uk/>

¹² Since the NDAD Web site is in English, it is more accurate to say that it is independently understandable by the English-speaking public.

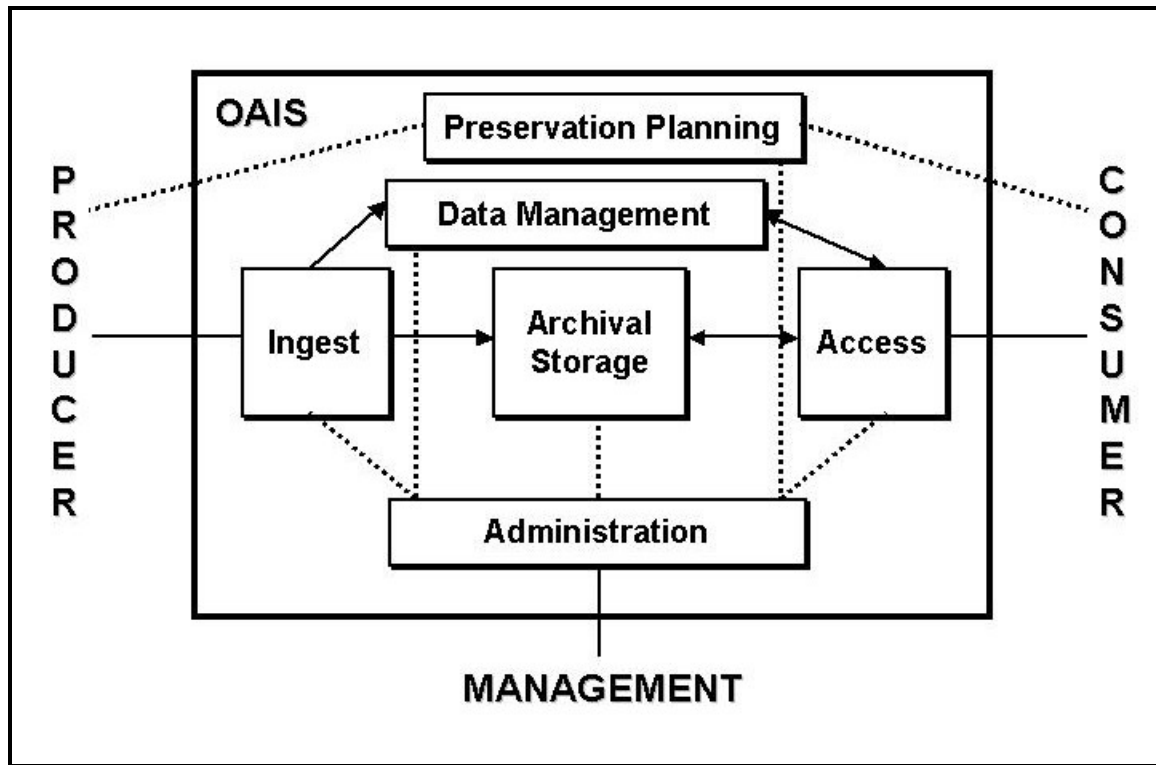


Figure 2: OAIS Functional Model

Figure 2 illustrates the OAIS functional model. The first functional component is *Ingest*, the set of processes responsible for accepting information submitted by Producers and preparing it for inclusion in the archival store. Specific functions performed by Ingest includes receipt of information transferred to the OAIS by a Producer; validation that the information received is uncorrupted and complete; transformation of the submitted information into a form suitable for storage and management within the archival system; extraction and/or creation of descriptive metadata to support the OAIS's search and retrieval tools and finding aids; and transfer of the submitted information and its associated metadata to the archival store. In short, the Ingest function serves as the OAIS's external interface with Producers, managing the entire process of accepting custody of submitted information and preparing it for archival retention.

The second functional component of an OAIS-type archive is *Archival Storage*. This is the portion of the archival system that manages the long-term storage and maintenance of digital materials entrusted to the OAIS. More specifically, the Archival Storage function is responsible for ensuring that archived content resides in appropriate forms of storage – e.g., online, near-line, off-line – and that the bit streams comprising the preserved information remain complete and renderable over the long-term. To meet this responsibility, Archival Storage periodically undertakes procedures such as media refreshment or format migration. The Archival Storage function also implements various safeguard mechanisms, such as error-checking procedures, to evaluate the outcome of preservation processes, as well as disaster recovery policies to mitigate the effects of

catastrophic events. Finally, Archival Storage retrieves items from the OAIS's storage systems in support of access requests by Consumers. Note that the Archival Storage function has no direct external interface; interaction with Archival Storage is confined to the OAIS's internal high-level services.

Data Management is the third functional component of an OAIS. The Data Management function maintains databases of descriptive metadata identifying and describing the archived information in support of the OAIS's finding aids; it also manages the administrative data supporting the OAIS's internal system operations, such as system performance data or access statistics. The primary functions of Data Management include maintaining the databases for which it is responsible; performing queries on these databases and generating reports in response to requests from other functional components within the OAIS; and conducting updates to the databases as new information arrives, or existing information is modified or deleted. In managing these databases, the Data Management function supports search and retrieval of the OAIS's archived content, and administration of the OAIS's internal operations.

The fourth functional component of an OAIS is *Preservation Planning*. This service is responsible for mapping out the OAIS's preservation strategy, as well as recommending appropriate revisions to this strategy in response to evolving conditions in the OAIS environment. The Preservation Planning service monitors the external environment for changes that could impact the OAIS's ability to preserve and maintain access to the information in its custody, such as innovations in storage and access technologies, or shifts in the scope or expectations of the Designated Community. Preservation Planning then develops recommendations for updating the OAIS's policies and procedures to accommodate these changes. The Preservation Planning function represents the OAIS's safeguard against a constantly evolving user and technology environment. It detects changes impacting the OAIS's ability to meet its responsibilities, designs strategies for addressing these changes, and assists in the implementation of these strategies within the archival system.

Access is the fifth functional component of an OAIS-type archive. As its name suggests, the Access function manages the processes and services by which Consumers – and especially the Designated Community – locate, request, and receive delivery of items residing in the OAIS's archival store. Typical services provided by Access in support of the Consumer include processing queries of the OAIS's holdings – specifically, forwarding the request to Data Management and presenting the response (e.g., a result set) to the Consumer; and coordinating the retrieval and delivery of requested content – by forwarding the request to Archival Storage, receiving the requested items, and performing any necessary transformations¹³ that must occur prior to delivery to the Consumer. Access is also responsible for implementing any security or access control mechanisms associated with the archived content. The Access function represents the OAIS's interface with its Consumers (and Designated Community): as such, it is the

¹³ Possible transformations include altering the archived item's format to one more suitable for dissemination, or stripping away unneeded metadata. See the discussion of the Dissemination Information Package in section VI.

primary mechanism by which the OAIS meets its responsibility to make its archived information available to the user community.

The sixth and final functional component of an OAIS is *Administration*. The Administration function is responsible for managing the day-to-day operations of the OAIS, as well as coordinating the activities of the other five high-level OAIS services. Other responsibilities include interacting with Producers (e.g., negotiating Submission Agreements), Consumers (e.g., providing customer service support), and Management (e.g., implementing and maintaining archive policies and standards). The Administration function is also responsible for overseeing the operation of the archiving and access systems, monitoring system performance, and coordinating updates to the system as appropriate. Administration serves as the central hub for the OAIS's internal and external interactions: it communicates directly with the five other OAIS high-level services – Ingest, Archival Storage, Data Management, and Access, as well as the OAIS's external stakeholders – Producers, Consumers and Management.

In summary, the OAIS encompasses six high-level functional components which, taken together, constitute the mechanisms by which the OAIS preserves information over the long-term and makes it available to the Designated Community. An OAIS-type archive will implement each of these services, in one form or another, in the course of building a complete archival system.

The OCLC Digital Archive¹⁴ service provides an illustration of how the six OAIS high-level services might be implemented in practice. The architecture of this service is based on the OAIS reference model: therefore, each component of the OAIS functional model is recoverable from the wide array of OCLC organizational units supporting the Digital Archive. Specifically, the OCLC Digital Archive draws on Batch Services and the Connexion Service (Ingest); Preservation Services, Operations, and Database Support (Archival Storage); Cataloging and Metadata (Data Management); Office of Research and Systems Planning (Preservation Planning); Cooperative Discovery, Regional Network Services, and Support Services (Access); and Corporate Security, Legal, Network Support, and Systems Support (Administration). The OCLC Digital Archive exemplifies the diverse collection of processes, services, and expertise that must be integrated to produce the six OAIS high-level functional components. However, the OCLC approach is but one of many possible strategies for implementing the OAIS functional model.

VI. OAIS Information Model

In addition to describing the functional components of an OAIS-type archive, the reference model also provides a high-level description of the information objects managed by the archive. The *OAIS information model* is built around the concept of an *information package*: a conceptualization of the structure of information as it moves into, through, and out of the archival system. An information package consists of the digital object that is the focus of preservation, along with metadata necessary to support its long-term preservation and access, bound into a single logical package. There are three

¹⁴ <http://www.oclc.org/digitalarchive/>

important variants of the information package concept: the Submission Information Package, the Archival Information Package, and the Dissemination Information Package.

The *Submission Information Package*, or SIP, is the version of the information package that is transferred from the Producer to the OAIS when information is ingested into the archive. The exact form of the SIP may be the result of a negotiated agreement between the Producer and the OAIS, or it may be constructed on an *ad hoc* basis: e.g., the digital object and as much metadata as the Producer is willing or able to supply. The concept of the SIP emphasizes the fact that information may not be preserved in the exact form in which it is submitted by the Producer. For example, the preserved information may be the aggregation of content provided in multiple SIPs; or, the Producer may provide the information in a format not supported by the OAIS, necessitating migration to another format prior to inclusion in the archival store. It may also be the case that the metadata supplied by the Producer is incomplete or inadequate, and must be augmented during the ingest process.

The *Archival Information Package*, or AIP, is the version of the information package that is stored and preserved by the OAIS. The AIP consists of the information that is the focus of preservation, accompanied by a complete set of metadata sufficient to support the OAIS's preservation and access services. The archived information and its associated metadata represent a single *logical* package within the archival system: there is, however, no requirement that any form of physical association be maintained, such as embedding the metadata in the information object itself and storing the combined object as a single bit stream. Arrangements for storing archived information and its metadata are left to the OAIS's implementers; possible solutions might range from complete physical integration, to storage in separate yet logically related databases.

Finally, the *Dissemination Information Package*, or DIP, is the version of the information package delivered to the Consumer in response to an access request. The DIP concept emphasizes the fact that the information package disseminated by the OAIS to the Consumer may differ in form or content to that which resides in the archival store. Points of differentiation between the DIP and AIP may include, but are not limited to, the format of the content (e.g., an image file might be converted from TIFF to JPEG prior to dissemination); the amount of content (a DIP may correspond to one AIP, multiple AIPs, or even part of an AIP); and the amount of metadata supplied alongside the content (it is likely that the DIP will not contain the complete set of metadata associated with an archived digital object, since much of it is of little interest to the Consumer).

SIPs, AIPs, and DIPs represent the information objects deposited into, managed by, and disseminated from an OAIS-type archive. But it is the AIP – the Archival Information Package – that is the focus of preservation: it is the information package variant which the OAIS is committed to perpetuate over the long-term. Given the importance of the AIP in regard to the OAIS's preservation and access responsibilities, it is useful to take a closer look at this information package and examine its key components.

Recall that an information package contains the content to be preserved, along with its associated metadata. The AIP embodies a stricter interpretation of this concept, in that it must include the *complete set of metadata* necessary to support the content's long-term preservation and availability to the Designated Community. The reference model characterizes the types of metadata that should be included with the archived information.

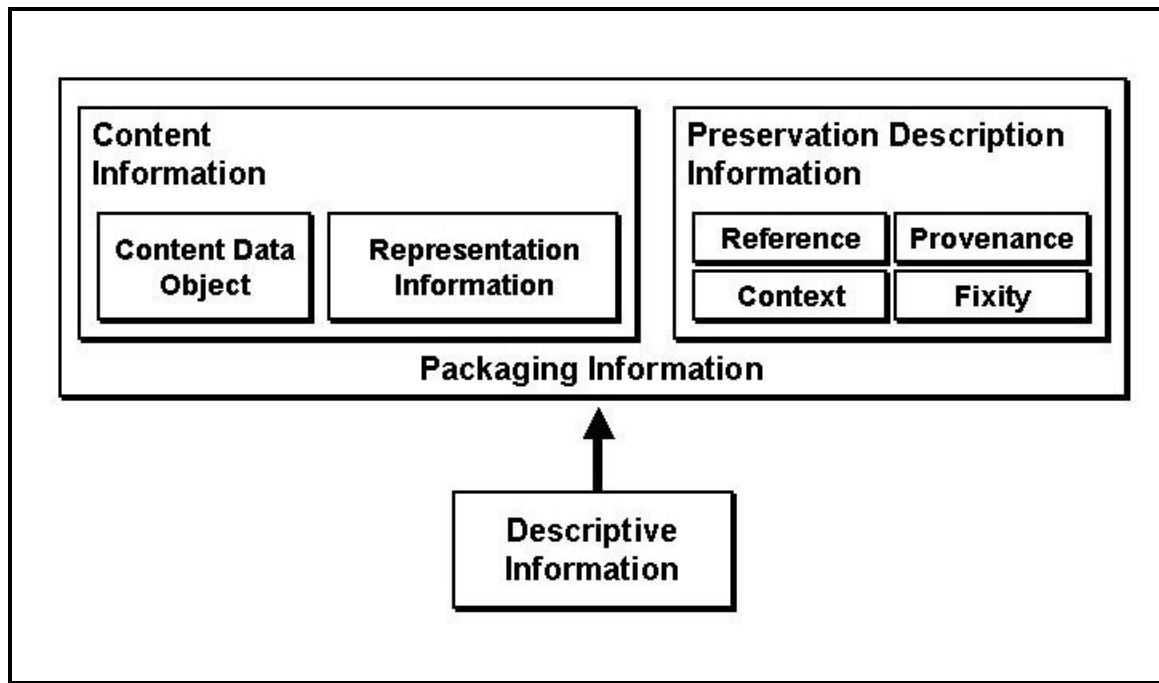


Figure 3: Archival Information Package

Figure 3 illustrates the information components of an AIP. Construction of an AIP begins with the *Content Data Object* – the information that is the focus of preservation. The Content Data Object can take the form of any class of digital material: text, images, video, databases, computer programs, etc. The Content Data Object may be comprised of a single, self-contained digital file – for example, a document in PDF format; it may also encompass multiple files, such as a Web site consisting of text (HTML files) and static images (GIF or JPEG files). The key point is that the OAIS is responsible for preserving the Content Data Object over the long-term, as well as for making it available in a form that is independently understandable by the Designated Community.

In order to meet the second responsibility – to make the Content Data Object available in a form that is independently understandable by the Designated Community – the Content Data Object must be accompanied by an appropriate quantity of *Representation Information*: information necessary to render and understand the bit sequences constituting the Content Data Object. Representation Information might include a description of the hardware and software environment needed to display the Content Data Object and/or access its contents; it might also summarize the appropriate interpretation

of the Content Data Object. For example, if the Content Data Object is an ASCII file of numbers, Representation Information might indicate that the numbers correspond to average daily air temperature readings for Manhattan, measured in degrees Celsius, for the period 1972 – 2000.

It was mentioned earlier that the scope of the Designated Community impacts the amount of metadata required to support the preservation process. It is in regard to Representation Information that this is so. In general, the broader the scope of the Designated Community, the less specialized the *knowledge base* associated with that community – that is, the less information relevant to interpreting and understanding the archived information the OAIS can assume its Designated Community possesses. The less specialized the knowledge base, the more Representation Information is needed to ensure that the preserved information remains renderable and understandable to the Designated Community over the long-term.¹⁵

The Content Data Object and its associated Representation Information are collectively known as *Content Information*. It is the Content Information – the information that is the focus of preservation, along with sufficient metadata to ensure it remains renderable and understandable to the Designated Community – that the OAIS must perpetuate over time.

Long-term retention of the Content Information requires additional metadata to support and document the OAIS's preservation processes. This metadata is called *Preservation Description Information*, or PDI. PDI consists of four components. *Reference Information* uniquely identifies the Content Information within the OAIS's internal systems, as well as to entities and systems external to the OAIS. Examples include a system-generated internal identifier, or an ISBN. *Context Information* describes the Content Information's relationships to other Content Information objects: for example, those that are related to it thematically (e.g., as part of a subject-based collection), or those that represent versions of the same content in alternative formats. *Provenance Information* documents the history of the Content Information, including its creation, any alterations to its content or format over time, its chain of custody, any actions (such as media refreshment or migration) taken to preserve the Content Information, and the outcome of these actions. *Fixity Information* validates the authenticity or integrity of the Content Information: for example, a check sum, a digital signature, or a digital watermark.

¹⁵ In practice, the structure of Representation Information can be extremely complex. A particular set of Representation Information may require additional Representation Information in order to be rendered, interpreted, and/or understood. The second set of Representation Information may itself depend on yet another set of Representation information. This regressive process can continue for an arbitrary number of steps, resulting in a *Representation network*. The reference model (p. 4-21) provides the following example of a Representation network: data on a CD-ROM is stored in an ISO 9660-compliant file system. ISO 9660 serves as Representation Information for the stored data: it must be understood in order to retrieve the data from the CD-ROM. ISO 9660 requires that file names and extensions take the form of ASCII characters; therefore, in order to understand the ISO 9660 file system, one must also understand ASCII text. In this sense, the ASCII standard is Representation Information for ISO 9660, and indirectly, then, for the data residing on the CD-ROM. The Representation network for this data would therefore include (among other things) the ISO 9660 standard and the ASCII standard.

Taken together, the Content Information and Preservation Description Information represent the archived digital content, the metadata necessary to render and understand it, and the metadata necessary to support its preservation. *Packaging Information* is used to bind all of these information components into a single logical package: in other words, Packaging Information serves to associate all of the various components of an AIP, permitting them to be identified and located as a single logical unit within the archival system.

Finally, *Descriptive Information* is information that supports the discovery and retrieval of Content Information by the Designated Community, via the OAIS's finding aids. For example, Descriptive Information might take the form of a Dublin Core metadata record maintained by the OAIS to facilitate resource discovery on the part of the archive's users.

The information components described above – Content Information (the Content Data Object and Representation Information), Preservation Description Information (Reference, Context, Provenance, and Fixity Information), Packaging Information, and Descriptive Information collectively form the Archival Information Package, which in turn represents the combination of the preserved digital information and a complete set of associated metadata.

VII. Beyond the OAIS Reference Model

The OAIS reference model is a conceptualization of the environment, functional components, and information objects associated with a system designed to effect the long-term preservation of digital materials. But the reference model is *not* an implementation: it says nothing about system architectures, storage or processing technologies, database design, computing platforms, or any of the myriad technical details involved in setting up a functioning archival system. However, the reference model provides a starting point for implementation, in the sense that it characterizes the high-level responsibilities, services, and informational requirements that the implemented system must, in one form or another, incorporate.

A number of initiatives have used the reference model as a conceptual foundation for more focused work in digital preservation. Major areas of study include, but are not limited to, developing "OAIS-compliant" repository architectures; establishing OAIS-related standards; adapting the general OAIS model to domain-specific implementations; fleshing out the metadata requirements of the OAIS information model; and developing methods and protocols for encoding and exchanging archived information. The remainder of this section briefly describes several OAIS-related activities that are moving the reference model concepts closer to implementation.

OCLC and RLG have jointly sponsored consensus-building activities in two areas related to the OAIS reference model.¹⁶ The first area addresses the *metadata requirements associated with the long-term preservation of digital materials*.¹⁷ An international

¹⁶ See <http://www.rlg.org/pr/pr2000-oclc.html> for a news release describing this collaboration.

¹⁷ <http://www.oclc.org/research/projects/pmwg/wg1.htm>

working group of experts was convened to consider this issue from the perspective of the cultural heritage community. The group defined the concept of preservation metadata, and described its significance in regard to the long-term persistence of digital materials. The group also reviewed and synthesized a number of existing preservation metadata schema, with the purpose of identifying points of convergence and divergence. One point of convergence was the use of OAIS information model concepts as a starting point for the schema. The working group published its findings in a white paper in January 2001.¹⁸

The white paper provided context for the working group's next task: developing a comprehensive preservation metadata framework that would identify and describe the types of information that could be used to support the preservation of digital materials. The framework took the form of an expanded conceptual structure for the OAIS information model, along with a set of "prototype" metadata elements mapped to the conceptual structure and reflecting the information concepts and requirements set forth in the OAIS reference model. The framework refined and extended the information components constituting an Archival Information Package, and clarified how preservation metadata supports the preservation process. The working group published the framework in June 2002.¹⁹

As a follow-on to this work, OCLC and RLG established a second working group in June 2003. *Preservation Metadata: Implementation Strategies*, or PREMIS²⁰, is a working group comprised of representatives from libraries, museums, government agencies, and the private sector. The organizations represented on the group share the experience of having built, or being in the process of building, a repository for the long-term preservation of digital materials. Using the preservation metadata framework developed by the first working group as a starting point, PREMIS will develop recommendations for an implementable set of "core" preservation metadata elements, supported by a data dictionary, and broadly applicable within the digital preservation community. In concert with this effort, PREMIS also will examine alternate strategies for encoding, storing, and managing metadata within an OAIS-type digital archiving system.

The second OCLC/RLG-sponsored initiative addressed the *attributes of a trusted digital repository*.²¹ A working group comprised of international experts was established to translate the OAIS concepts and models into a consensus statement on the responsibilities and characteristics of a digital repository housing a large-scale, heterogeneous collection of culturally significant materials. A key objective of this effort was to enumerate attributes of a digital repository that, taken together, serve to inspire trust within the Designated Community that the repository is indeed capable of preserving and making available the portion of the scholarly and cultural record in its custody. The working group published its report in May 2002.²²

¹⁸ http://www.oclc.org/research/projects/pmwg/presmeta_wp.pdf

¹⁹ http://www.oclc.org/research/projects/pmwg/pm_framework.pdf

²⁰ <http://www.oclc.org/research/projects/pmwg/>

²¹ <http://www.rlg.org/longterm/attribswg.html>

²² <http://www.rlg.org/longterm/repositories.pdf>

A new activity has been established under the sponsorship of the US National Archives and Records Administration (NARA) and RLG. The Task Force on Digital Repository Certification²³ will identify “certifiable elements” of a digital archiving system, and develop a plan for establishing appropriate bodies, policies, and procedures for certifying digital repositories. This effort builds on the concepts of the OAIS reference model, and in particular, the definition of a trusted digital repository described in the May 2002 RLG/OCLC report. The task force, comprised of individuals from a variety of institutional and geographical backgrounds, intends to submit its work to the standardization process administered by the International Organization for Standardization (ISO).

Finally, an effort is underway to create a standardized description of the *Producer-Archive interface*: i.e., the interactions that take place between Producers and an OAIS-type archive. The standard segments the process of transferring information from the Producer to the OAIS into a set of distinct phases, and provides a detailed description of the anticipated outcome of each phase, as well as the set of actions which must take place to bring about this outcome. This framework serves as a basis for identifying areas within the Producer-Archive interface that would benefit from more focused standards, recommendations, and best practices, and also provides a foundation for the development of automated processes and software tools to support the information transfer process. Finally, the standard offers a more detailed exposition of the responsibilities and functions of the OAIS Ingest and Administration high-level services than what is provided in the reference model.

The Producer-Archive interface standard is an important contribution toward shaping the transfer of information from the Producer to the OAIS into a consistent, well-understood process. It is especially useful in terms of cultivating a mutual understanding between Producers and archives in regard to their respective responsibilities and expectations as participants in the ingest process. The Producer-Archive interface is currently in the form of a draft standard²⁴, and has been made available to interested parties in the digital preservation community for review and comment.

VIII. Conformance

The reference model states that an OAIS-compliant digital archive supports the OAIS information model (described in section VI above). It also is committed to meet the mandatory responsibilities enumerated by the reference model (see section III above). Finally, the reference model notes that standards and other documentation that purport to conform to the OAIS reference model must incorporate relevant OAIS terminology and concepts, applied according to the interpretation and context defined in the reference model.

In surveying current activities directed toward the long-term preservation of digital materials, it is not uncommon to encounter the term “OAIS-compliant” used in reference

²³ <http://www.rlg.org/longterm/certification.html>

²⁴ <http://ssdoo.gsfc.nasa.gov/nost/isoas/CCSDS-651.0-R-1-draft2.pdf>

to a digital archiving system. For example, the Stanford Digital Repository²⁵, the Digital Information Archiving System (DIAS) built by IBM on behalf of the National Library of the Netherlands²⁶, and the OCLC Digital Archive service are all positioned as conforming to the OAIS reference model. The architects of the Metadata Encoding and Transmission Standard (METS)²⁷, an XML document format supporting the management and exchange of digital objects, point out its potential application as an implementation of the Archival Information Package concept.

What validation does OAIS compliance provide to stakeholders in the long-term preservation of digital materials? Because the reference model is a conceptual framework rather than a concrete implementation, the meaning of “OAIS-compliant” is necessarily vague. Conformance to the reference model can imply an explicit application of OAIS concepts, terminology, and the functional and information models in the course of developing a digital repository’s system architecture and data model; but it can also mean that the OAIS concepts and models are “recoverable” from the implementation – in other words, it is possible to map, at least from a high-level perspective, the various components in the archival system to the corresponding features of the reference model. Further ambiguity is introduced when institutions and organizations claim OAIS compliance without defining or clarifying what this means in regard to their particular implementation.

This assessment of the current state of OAIS compliance may be disappointing to those who anticipated a much more precise meaning: e.g., a rigorous application of a well-defined suite of standards, protocols, and best practices. Yet the importance of OAIS-compliant archives should not be discounted, for two reasons. A shared view of the core functional and information requirements of digital archiving systems is essential for creating long-term preservation solutions that are well-understood and accepted by a potentially extended stakeholder community. Moreover, this shared view facilitates the development of an interoperable network of digital archives distributed across multiple institutions. The concepts, terminology, and detailed models of the OAIS reference model represent a common point of reference around which consensus and interoperability can gel.

Second, a more rigorous interpretation of OAIS compliance will likely emerge as a consequence of ongoing efforts to extend and build on the conceptual framework established by the reference model. One of the original motivations for producing the reference model was to put forward a widely-applicable framework that would serve as the starting point for more focused standards- or consensus-building activities. A number of these activities are currently underway, touching on nearly every aspect of the original OAIS reference model; as they are completed (and new activities spring up in their wake), it is likely that OAIS-related standards and best practices will gain a foothold in the digital preservation community, which, in turn, will impart a more concrete interpretation to OAIS compliance. Moreover, digital archives currently incorporating

²⁵ <http://library.stanford.edu/depts/pres/mediapres/digital.html>

²⁶ <http://www-5.ibm.com/nl/dias/resource/overview.pdf>

²⁷ <http://www.loc.gov/standards/mets/>

nothing more than OAIS terminology and concepts will find that even this loose conformance to the reference model facilitates the process of adopting new OAIS-related standards and best practices as they become available.

The significance attached to OAIS compliance, both now and in the future, ultimately rests on whether it produces a tangible impact on stakeholder confidence in a digital archive's ability to meet its preservation objectives. Libraries, museums, and other collecting institutions, for example, are faced with the prospect of entrusting irreplaceable portions of the scholarly and cultural record to digital archiving systems whose capacity to provide effective long-term stewardship is as yet unproven. Does OAIS compliance make this decision easier? At this stage in the reference model's development, a tentative "yes" seems appropriate. The OAIS reference model has been quite successful in consolidating understanding of the fundamental requirements for securing the long-term persistence of digital materials. A shared perception of these requirements is a necessary condition for building well-understood, sustainable, and ultimately, *trusted* digital archiving systems.

Glossary

AIP	Archival Information Package
ASCII	American Standard Code for Information Interchange
CCSDS	Consultative Committee for Space Data Systems
DIAS	Digital Information Archiving System
DIP	Dissemination Information Package
GIF	Graphic Interchange Format
HTML	Hypertext Markup Language
ISO	International Organization for Standardization
JPEG	Joint Photographic Experts Group
METS	Metadata Encoding and Transmission Standard
NARA	US National Archives and Records Administration
NDAD	UK National Digital Archive of Datasets
NDR	National Data Repository
OAIS	Open Archival Information System
OCLC	Online Computer Library Center, Inc.
PDF	Portable Document Format
PDI	Preservation Description Information
PREMIS	Preservation Metadata: Information Strategies
RLG	Research Libraries Group
SIP	Submission Information Package
TIFF	Tagged Image File Format
ULCC	University of London Computer Centre
XML	Extensible Markup Language

References

Consultative Committee for Space Data Systems (2002) "Reference Model for an Open Archival Information System (OAIS)" CCSDS 650.0-R-1 – Blue Book.

Available at: <http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf>

Hirtle, P. (2001) "OAI and OAIS: What's in a Name?" *D-Lib Magazine* 7,4.

Available at: <http://www.dlib.org/dlib/april01/04editorial.html>

Sawyer, D., and Reich, L. (2002) "ISO 'Reference Model for an Open Archival Information System (OAIS): Tutorial Presentation'" PowerPoint presentation

Available at: http://ssdoo.gsfc.nasa.gov/nost/isoas/presentations/oais_tutorial_200210.ppt