

**Framework for Facilitating Communication at
“Science Archives in the 21st Century”
23 February 2007**

Whenever a diverse group gets together to do real work, experience shows there is always a significant period needed to bridge gaps in understanding as a result of unstated context and the use of terms in different ways. Many organizations, addressing issues surrounding archives, have reported notable success at quickly bridging these gaps through use of the terms and concepts contained in the ISO standard reference model for an archival system (<http://public.ccsds.org/publications/archive/650x0b1.pdf>).

The purpose of this small document is to summarize a few key ISO concepts associated with information modeling and with archival functions so they can provide context for the material submitted as abstracts, presentations, and posters. Likely acceptance of an abstract will significantly increase by relating your material to these terms and concepts. In the following text, the term ‘archive’ is taken to refer to all types of systems and associated personnel who manage a set of scientific data and associated metadata for periods of several years or more. In other contexts they may be referred to as data centers or repositories.

Archive Functional Model

Archival Functional Entities

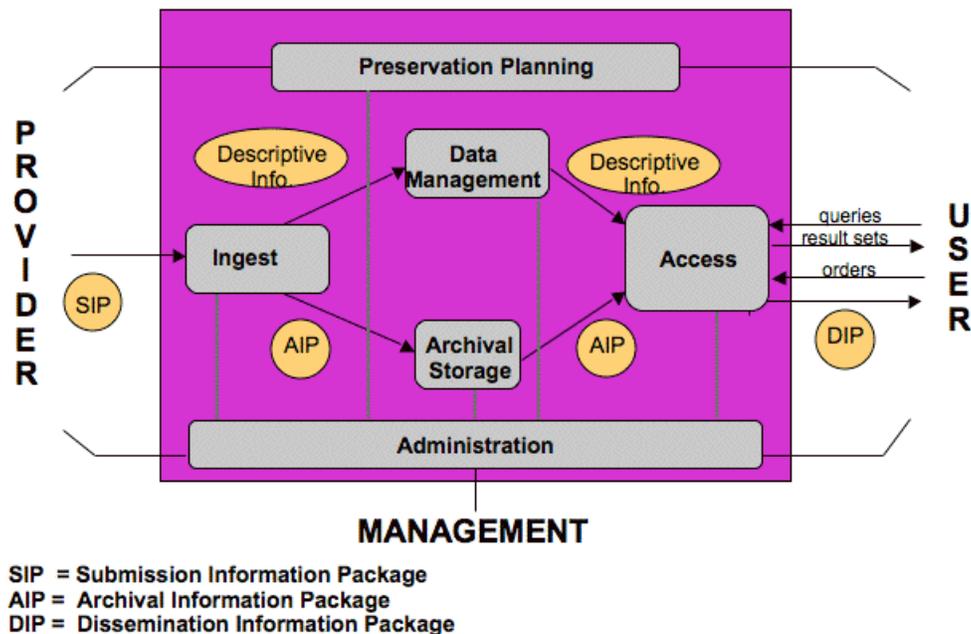


Figure 1: An archive’s environment with data flows

An abstract view of an archive is given in figure 1. It shows a breakout of archival functions along with interfaces to three external roles played by other persons or systems. These roles are the **Provider**, the **User**, and **Management**. The role of **Provider** is to submit science data and documentation (i.e., science information) to the archive. The role of the **User** is to search for and utilize the science information of interest. The role of **Management** is to oversee the operation of the archive on a part time basis. This is distinguished from daily oversight which is accomplished by the **Administration** functional entity ‘inside’ the archive.

The conceptual archive is composed of six functional entities that accomplish all the activities needed to acquire the information, manage it for the long-term, and make it available in forms understandable and useable to its user community. One way to begin to understand the six archival functions is to examine what happens in a typical data flow scenario.

The data **Provider** sends information to the archive’s **Ingest** function in the form of **Submission Information Packages (SIPs)**. These are recognizable packages as agreed between the data **Provider** and the archive. The **Ingest** function processes these packages as needed to meet its internal storage forms, called **Archival Information Packages (AIPs)**. There may be any number of **SIPs** needed to form a complete **AIP**. The **AIPs** are given unique identifiers and are sent to the **Archival Storage** function for long-term preservation. At the same time, **Descriptive Information** is sent to the **Data Management** function. This information is used to aid in searching and finding the **AIPs**.

The **Archival Storage** function is responsible for ensuring the long-term preservation of the **AIPs**, and it may perform media refreshes to aid this process. It responds to requests for **AIPs** based on their unique identifiers, and it provides reports on its holdings.

The **Data Management** function stores the inventory and search information associated with **AIPs**, and it also stores various administrative information such as user IDs, request processing history, etc. It supplies the search and inventory information to the **Access** function, and it provides reports on its holdings.

The **Access** function interfaces with the **User** and provides access to **AIP** inventory and search information. **User** searches will result in the identification of one or more **AIPs** and these will be requested from **Archival Storage**. **Access** may then do processing on the **AIPs** to meet the requirements of the **User**’s request. The results are provided to the **User** as **Dissemination Information Packages**. These packages may have any form the archive has agreed to support.

The **Preservation Planning** function is not directly in the normal data flow of the archive. Its role is to track technology and standards evolution, track the needs of the data **Providers** and **Users**, and to make recommendations to the **Administration**

function on internal policy and implementations. The perspective is to ensure good preservation practices while also meeting data **Provider** and **User** needs.

Finally, the **Administration** function is responsible for day-to-day coordination of the other functions and has interfaces to both data **Providers** and **Users**. It also takes on the responsibility for coordinating data migrations that are more involved than the basic media refresh. Conceptually, such migrations involve the extraction of data using the **Access** function and re-submission of data to the **Ingest** function. Processing during the migration may be located in **Access**, **Administration**, and/or **Ingest** as seems most appropriate for the archive.

Information Modeling

While there are many information modeling concepts and diagrams in the ISO standard, two are of particular use in this context. These are ‘**Information**’ and the ‘**Archival Information Package**’.

The archive is seen to ingest, preserve, and serve information. **Information** is understood as ‘data together with its **Representation Information**’. Here ‘data’ starts out as a sequence of bits, and then additional **Information (Representation Information)** is needed to understand those bits. Thus **Representation Information** includes format information, but may also include as much additional information about the meaning of fields and how the field values were obtained as is deemed appropriate. Since **Representation Information** is also **Information**, it may have its own data bits that need additional **Representation Information**. This recursion is typically terminated by the existence of hardcopy documentation or software what displays **Information** in a human readable manner. Figure 2 shows a simple view of **Information** without the recursion.

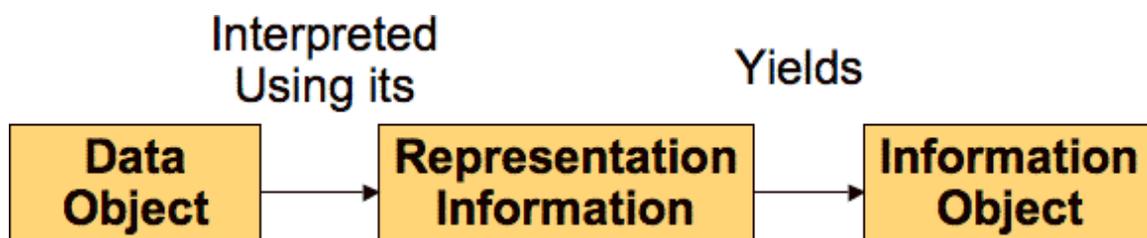


Figure 2: Relationship of a Data Object and its Representation Information to obtain an Information Object

The **Archival Information Package (AIP)** is a concept that addresses the need to not only associate **Representation Information** with the data bits, but it also addresses the need for additional information useful in maintaining the preservation of that Information. In particular, these are referred to as **Preservation Description Information (PDI)** and **Packaging Information**.

Their relationships are shown in Figure 3 using a simplified UML diagram. This says that an **AIP** is composed of two **Information** objects referred to as **Content Information**

and **PDI**. The **Content Information** is that information that the archive is tasked to preserve and examples are given in the figure.

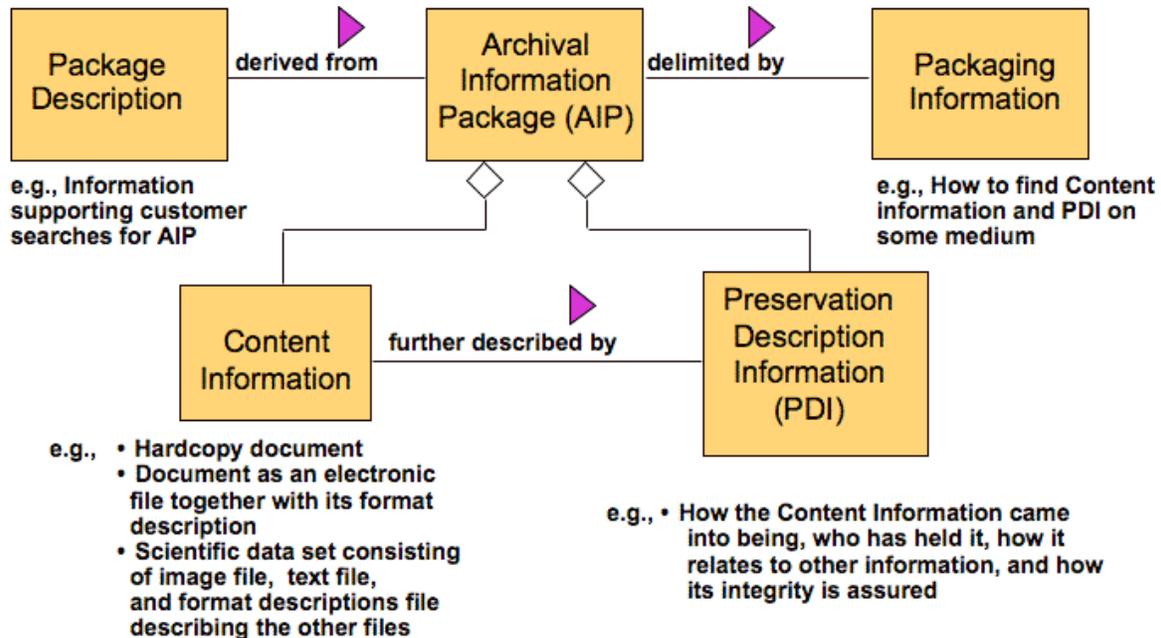


Figure 3: Archival Information Package is composed of a Content Information object and a PDI Information object, and it has two related Information objects.

Preservation Description Information (PDI) is information that is needed to help use and maintain the **Content Information**. Once the **Content Information** has been clearly defined, then the **PDI** can be assessed. The **PDI** may be broken into four categories referred to as **Context**, **Provenance**, **Fixity**, and **Reference**.

Context: relates the **Content Information** to other information outside the **AIP**. This provides **Users** with an understanding of how the information being preserved relates to a wider environment.

Provenance: describes the history of the **Content Information**, including the chain of custody, so that **Users** can better judge how much to trust the information.

Fixity: helps ensure that the **Content Information** is not altered in an undocumented manner. For example, this might include checksums and digital signatures.

Reference: provides one or more systems of identifiers by which to identify the **Content Information**. For example, this might include bibliographic attributes and/or a digital object identifier.

Figure 3 also shows that an **AIP** has two associated information objects called '**Package Description**' and '**Packaging Information**'. The **Package Description** is information

supporting user searching, and it is typically found in **Data Management**. The **Packaging Information** is that information used to bind the **Content Information** and the **PDI** into a recognizable and retrievable entity. It may make use of directory structures, file structures, and containers of various types such as a tar file, along with infrastructure implementing pointers.

The scope of information addressed by the **Content Information** within an **AIP** may be very broad. To provide a finer granularity, the **AIP** has two recognized forms referred to as an **Archival Information Unit (AIU)** and an **Archival Information Collection (AIC)**. The **AIU** is an **AIP** whose scope is such that it is the smallest granule of information that the archive holds with full **PDI** and can be readily disseminated, while the **AIC** is an **AIP** composed of one or more **AIUs** and other **AICs**. This allows for describing collections of collections.

Designated Community and Federation of Archives

The concept of **Designated Community** has been found to be useful in a number of contexts. For a given archival collection, consisting of one or more **AIPs**, it is expected that a **Designated Community** will be defined. This **Designated Community** is composed of all those **Users** who are expected to be able to use and understand the **AIPs**. A given archive may have many associated **Designated Communities**, or it may have only one. For example, the Planetary Data System performs extensive reviews of its **Collections** to ensure adequate understanding. They have some **Designated Community** in mind for this review.

The ISO standard also addresses the concept of the **Federation** of archives. Largely independent archives may share some common infrastructure, or some common standards for exchange, and this can be represented using the above functional and information model concepts by showing multiple archives and the ways in which they interact.