Use of Archival Information Packages at the National Space Science Data Center (NSSDC)

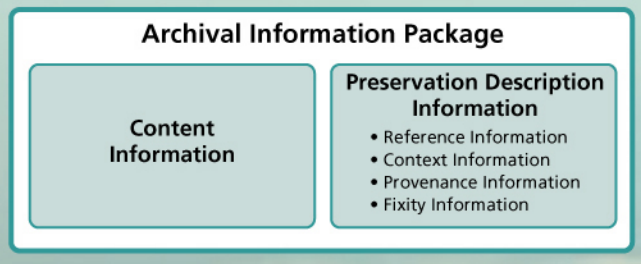
Patrick McCaslin, Perot Systems Government Services (PSGS)

The Archival Information Package

The National Space Science Data Center (NSSDC) has adopted concepts from the "Reference Model for an Open Archival Information System (OAIS)" as a framework for the evolution of its systems and processes. A key element of the OAIS model is the Archival Information Package (AIP). The AIP is a concept for identifying a collection of information for long-term preservation.

The AIP contains two information objects:

- Content Information – the data to be preserved.
- Preservation Description Information - the information necessary to adequately preserve the Content Information including:
  - Reference Information: Provides assigned identifiers for the Content Information.
  - Context Information: Documents the relationships of the Content Information to other information outside the AIP.
  - Provenance Information: Documents the history of the Content Information including origin, changes since it originated, and who has had custody of it.
  - Fixity Information: Provides the data integrity checks keys used to ensure that the Content Information not been altered.



NSSDC's Experience

2000 First generation of AIPs at NSSDC. Data residing on a VMS-hosted near-line archive packaged into AIPs. Record-formatted data were captured in a canonical form in the AIPs eliminating the need for a proprietary system for long-term preservation.

2001 First AIPs generated by an external Data Provider. IMAGE project uses NSSDC-provided packaging software to package daily UDF files into AIPs which are delivered to NSSDC for permanent archival.

2002 All data delivered electronically to NSSDC are packaged into AIPs for preservation.

2004 First NSSDC AIPs containing multiple data files. Bundling multiple files into a single AIP allows NSSDC to migrate data from legacy media while keeping original file groupings intact.

2006 NSSDC develops AIP form that can package large directory structures (e.g. CD/DVD volumes)

2007 NSSDC and Planetary Data System (PDS) prepare for delivery of PDS volumes in AIPs generated at the PDS nodes

AIPs Generated by Data Providers

NSSDC encourages Data Providers to use NSSDC-provided AIP packaging softwareto generate AIPs for delivery to NSSDC. Advantages to Providers are:

Integrity - Fixity information is captured at the provider's site and incorporated into the AIP affording the maximum possible assurance that the data is preserved at the NSSDC exactly as the provider intended.

Long-term usability - Critical low-level attributes, required for long-term preservation and use of the data, are extracted automatically from the data at the provider's site and incorporated into the AIP.

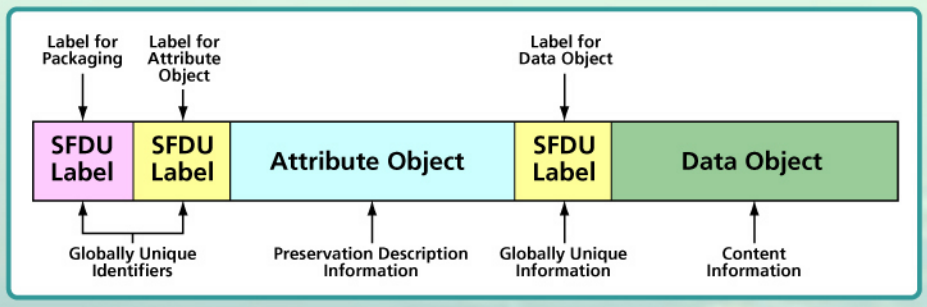
Automation - AIP delivery and ingest is automated, eliminating human errors and delays in the process at both the Provider's site and the NSSDC.

Multiple packaging options - Contents of AIPs may be selected in several ways (individual file, multiple files, directory structures) allowing the provider to bundle related data in the preservation package.

Large capacity - AIPs can accommodate large data volumes (currently tested to 150 GB), again facilitating the bundling of related data in the preservation package.

NSSDC's AIP Implementation

- A single file in a media independent canonical form to facilitate long-term preservation across multiple media types.
- Holds one or many science data files or entire directory structures.
- Contains a number of attributes of the data files that point to supporting documentation.
- No reformatting of the science data unless record boundaries need to be retained and are not already in the byte stream.
- Files that are transformed can be returned to original state.
- Contains a data object and an attribute object packaged within a Standard Formatted Data Unit (SFDU).



Example AIP Attribute Object Generated by IMAGE Project

```
BEGIN_OBJECT = NSSDC_ATTRIBUTE_OBJECT;
   OBJECT_TYPE_VERSION = "1";
   BEGIN_OBJECT = PACKAGE_IDENTIFICATION;
      ARCHIVAL_STORAGE_ID = "imag0000024944";
      PRIMARY_COLLECTION_ID = "00-017A-04B";
      CRC_TYPE = "NSSDC_A:V0";
      ATTRIBUTE_OBJECT_CRC = "87bce815"; /* hex format, covers all following
      complete PVL objects */
   END_OBJECT = PACKAGE_IDENTIFICATION;

   BEGIN_OBJECT = STREAM_STRUCTURE;
      BEGIN_OBJECT = ORIGINAL_STREAM_STRUCTURE;
         STREAM_TYPE = "BINARY";
         ORIGINATING_SYSTEM = "UNIX: Dec Alpha";
         DATE_TIME_CREATED = "2003-03-28T00:51:17";
         DATE_TIME_LAST_MODIFIED = "2003-03-28T00:51:17";
         FILE_ORGANIZATION = "sequential";
         RECORD_FORMAT = "undefined";
         RECORD_CONTROL = "none";
         STREAM_SIZE_BYTES = "8388153";
         MAXIMUM_RECORD_LENGTH_BYTES = "0";
         FILE_NAME = "IMAGE.IMAGE1.UDFH.200308300.tgz";
         CRC_TYPE  = "NSSDC_A:V0";
         CRC = "45a04cff"; /* hex format */
         DATE_TIME_OF_ATTRIBUTE_VALUE_CAPTURE = "2003-03-28T06:20:49";
      END_OBJECT = ORIGINAL_STREAM_STRUCTURE;
      BEGIN_OBJECT = CANONICAL_STREAM_STRUCTURE;
         STREAM_TYPE = "BINARY";
         STREAM_RECORD_DELIMITER = "NONE";
         STREAM_SIZE_BYTES = "8388153";
         MAXIMUM_RECORD_LENGTH_BYTES = "0";
         CRC_TYPE = "NSSDC_A:V0";
         CRC = "45a04cff"; /* hex format */
         RECOMMENDED_FILE_NAME = "IMAGE.IMAGE1.UDFH.200308300.tgz";
         PROCESSING_REPORT = "FGET_FN-P_UNB PASS: found BINARY with no carriage
         control undefined records";
         DATE_TIME_OF_GENERATION = "2003-03-28T06:20:49";
         FORMAT_IDENTIFIER = "NSSD0375";
      END_OBJECT = CANONICAL_STREAM_STRUCTURE;
   END_OBJECT = STREAM_STRUCTURE;

   BEGIN_OBJECT = SUPPORTING_ATTRIBUTES;
      ORDERED_APPLIED_ENCODINGS = "tar,gzip"; /* example:  (base64, gzip) */
      NDADS_VMS_PROJECT_ID = "none"; /* default is 'NONE' */
      NDADS_VMS_DATATYPE = "none";
      NDADS_VMS_ENTRY_ID = "none";
      NDADS_VMS_SUPER_ENTRY_ID = "none";
      DATA_BEGIN_DATE_TIME = "2003-03-24T00:00:00";
      DATA_END_DATE_TIME = "none";
   END_OBJECT = SUPPORTING_ATTRIBUTES;
END_OBJECT = NSSDC_ATTRIBUTE_OBJECT;
```