Space Science Data Archive: Case Study

William Underwood

Information and Telecommunications Laboratory
Georgia Tech Research Institute
Atlanta, Georgia, USA
William.underwood@gtri.gatech.edu

Abstract

The Chain of Preservation Model developed by the InterPARES Project is a model of the activities and information needed for records creation and maintenance, records appraisal, and archival preservation. However, it needs to be demonstrated that this general preservation model applies to specific bodies of records that archival institutions must preserve. A walkthrough of an activity model is one way of reviewing a model in order to validate or falsify it. This paper discusses a case study of a space science data archive that was designed to collect information that was needed for a walkthrough of the model.

1. Background

The Chain of Preservation model is a functional model of the activities and information needed for records creation and maintenance, records appraisal and archival preservation. It provides a general preservation framework than can be used by archival institutions to develop their own preservation strategies depending on their institutional requirements and the specific bodies of records they must preserve. Figure 1 shows the four high-level activities of the model.

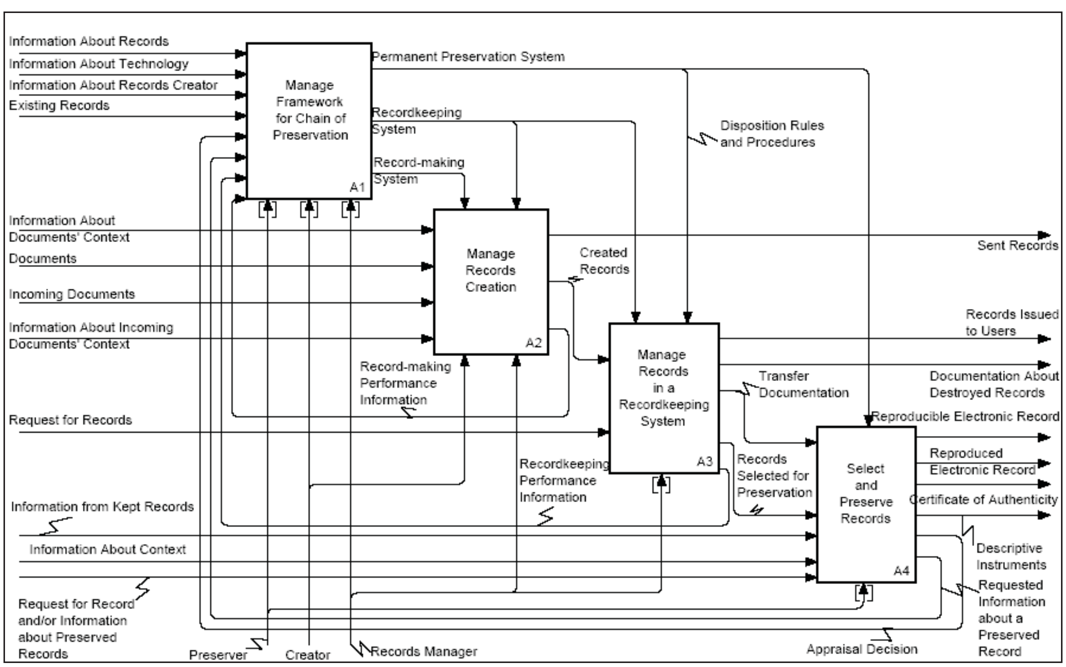

Figure 1. Chain of Preservation

In this notation, the labelled boxes represent activities (functions), the labelled arrows entering the boxes from the left and the arrows exiting the boxes on the right represent inputs and outputs, respectively. The arrows entering the boxes from the top represent controls on the activity, while the labelled arrows entering the boxes from below represent mechanisms for accomplishing an activity.

Activity A1, *Manage the Framework for the Chain of Preservation*, is decomposed into the activities of planning and developing a Record Making System, a Recordkeeping System, and a Permanent Preservation System.

Activity A2, *Manage Records Creation*, is decomposed into the activities shown in Figure 2.

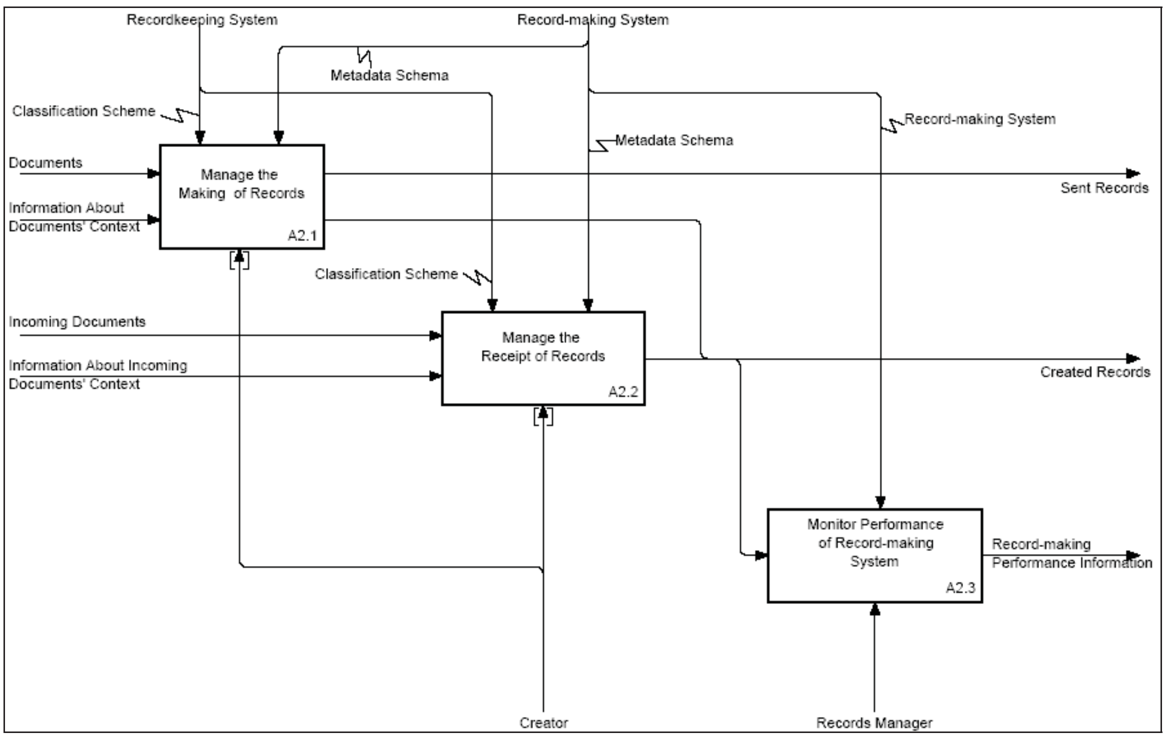

Figure 2. Manage Records Creation.

In Figure2, Activity A2.1, *Manage the Making of Records*, uses information about the context of the record to create values for metadata attributes of the records.

In Figure 1, Activity A3, *Manage Records in a Recordkeeping System*, is decomposed into the activities shown in Figure 3. The activity *Maintain Records in a Recordkeeping System* is decomposed into *Managing Information about the Records*, *Attaching Integrity Metadata*, *Managing Storage*, and *Updating the Records*.

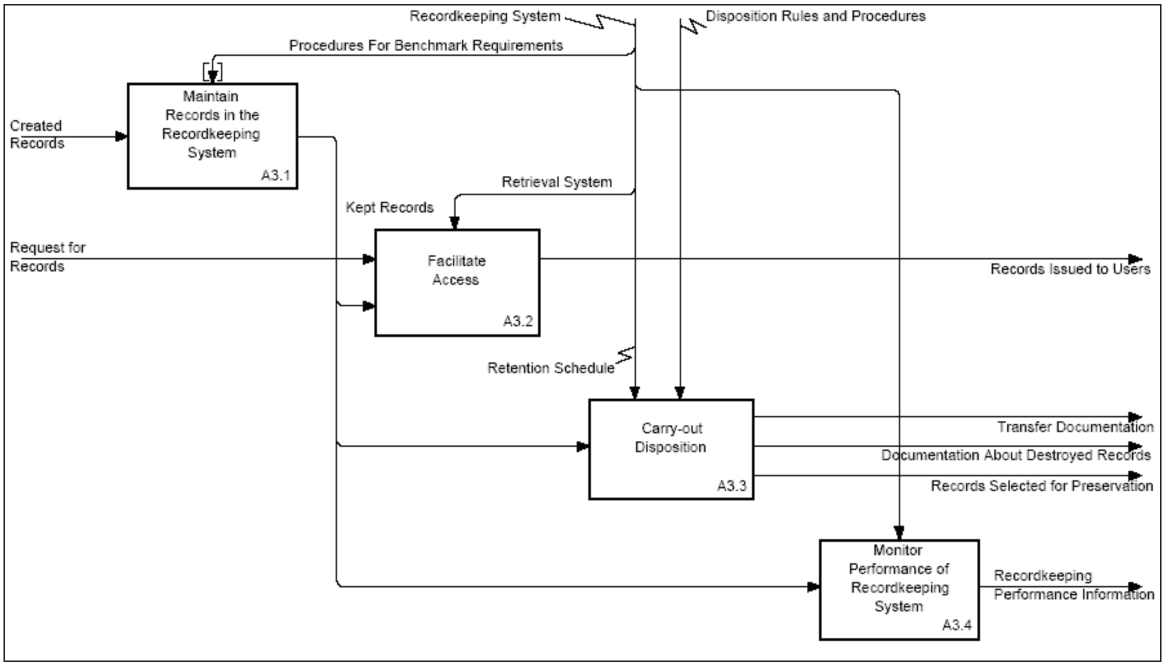

Figure 3. Manage Records in a Recordkeeping System.

Activity A4, *Select and Preserve Records*, is decomposed into the activities shown in Figure 4.

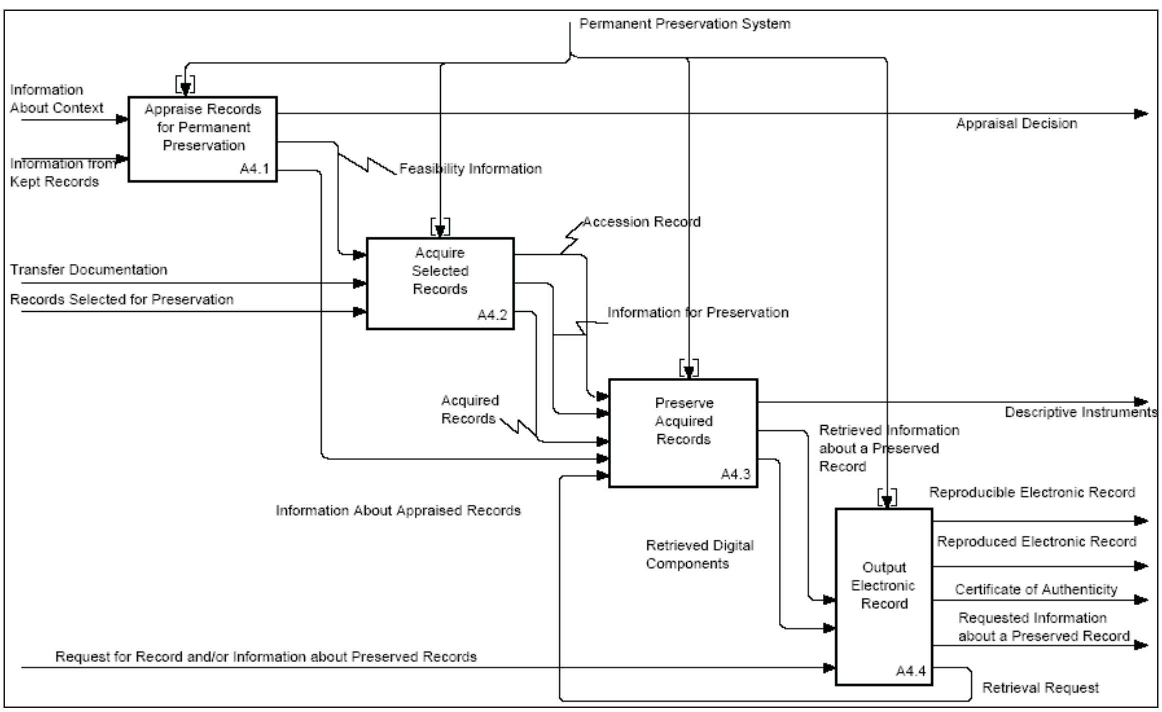

Figure 4. Select and Preserve Records.

Activity A4.1, *Appraise Records for Permanent Preservation*, is decomposed into the activities *Compile information about Records*, *Assess Value of Records*, *Determine Feasibility of Preservation*, *Make Appraisal Decision*, and *Monitor Appraised Records*.

Activity A4.2, *Acquire Records*, is decomposed into the activities *Register Transfer*, *Verify Authorization for Transfer*, *Verify Transfer*, *Confirm Feasibility of Preservation*, and *Accession Record Copies*.

Activity A4.3, *Maintain Electronic Records* is decomposed into (1) the activity *Manage Information about Records*, which can be solved through the use of a database management system that supports storage, update and retrieval of information about accessioned electronic records; (2) the activity *Manage Storage of Digital Components*, which can be solved with an archival storage system that supports storage and retrieval of the digital components of accessioned electronic records; and (3) the activity *Update Digital Components*, which has as its goal that records be reproducible from their digital components. However, the obsolescence of the file formats of the digital components due to new computer hardware, system software or application software places the records at risk of not being reproducible.

Activity A4.4, *Output Electronic records*, was decomposed into the subproblems, *Manage the Request for Information*, *Review Retrieved Components and Information*, *Reconstitute the Record*, *Present Record*, and *Package Output*.

2. The Planetary Data System

The Space Science disciplines are Astrophysics, Space Physics and Planetary Physics. The Astrophysicists primarily use the Flexible Image Transport System (FITS) format for representing astronomical data. The Space Physicists primarily use the Common Data Format (CDF) for representing solar wind and charged particle data. The Planetary Physicists primarily use the Planetary Data System (PDS) for representing data about the solar system. PDS uses a self-describing data format known as the Object Description Language (ODL). The National Space Science Data Center (NSSDC) is the Office of Space Science permanent archive for Space Science data sets.[1]

A case study of a space science data archive, the Planetary Data System (PDS), was designed to collect the information that was needed for a walkthrough of the model. The PDS archives and distributes scientific data from NASA planetary missions, astronomical observations, and laboratory measurements. The PDS is sponsored by NASA's Office of Space Science. Its purpose is to ensure the long-term usability of NASA data and to stimulate advanced research. The PDS is an active archive, not a permanent archive.

One of the issues to be addressed is whether the PDS is performing the functions of record making and a recordkeeping system or an archives for long-term preservation. The

1 D. Sawyer. NSSDC Role and OAIS Implementation, June 2003.
http://ssdoo.gsfc.nasa.gov/nost/isoas/presentations/oais_nssdc_implementation_200306.ppt

PDS is referred to as an active archives. Given the emphasis on associating metadata with data products and creating data sets, and the emphasis on maintaining the data sets and providing them to users, the information collected in the case study has been organized around record making and recordkeeping functions.

2.1 *Manage Records Creation*

The *PDS Data Preparation Workbook* (DPW) serves as a guide for the organization and preparation of data sets intended for submission to the Planetary Data System (PDS).[2] For active projects, archive planning consists of identifying the data to be archived, developing a detailed archiving schedule, and defining an end-to-end data flow through the ground system. A Project Data Management Plan (PDMP) is required by NASA for all new projects. This plan provides a general description of the project data processing, cataloging, and communication plan.[3] The Archive Policy and Data Transfer Plan (APDTP) provides a detailed description of the production and delivery plans for archive products for a project. A Data Product Software Interface Specification (SIS) is a document that describes the format and size of the individual data products.

All data incorporated into the PDS archives must undergo a peer review. The purpose of the review is to determine that:

- The data is accurate, complete and reliable.
- The data are suitable for archiving.
- The PDS standards have been followed.[4]

Figure 5 shows at a high-level the Planetary Science Data Model. The primary classes in the data model are Mission, Spacecraft, Instrument and Target.

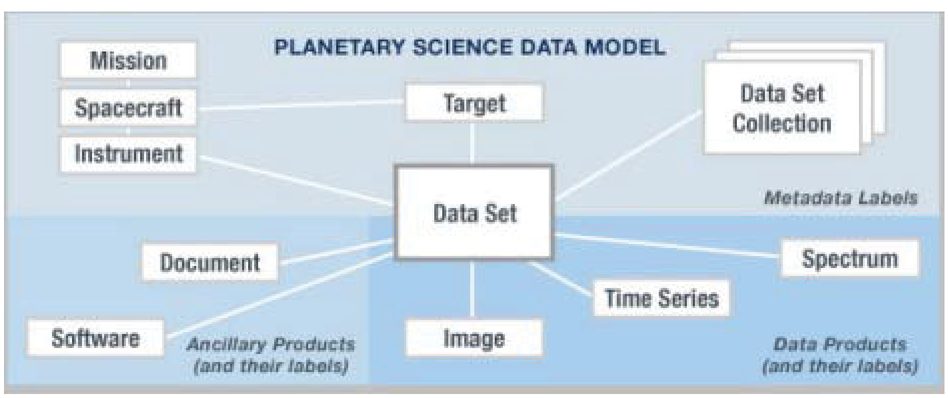

Figure 5. Planetary Science Data Model.

2 Planetary Data System Data Preparation Workbook, Version 3.1, JPL D-7669, Part 1, Feb. 1995, Jet Propulsion Laboratory, Pasadena.
3 PDS Guidelines for Project Data Management Plans, JPL Document D-5111.
4 Planetary Data System Standards Reference, Version 3.6, JPL D-7669, Part 2, Aug. 2003, Jet Propulsion Laboratory, Pasadena.

A data set is an accumulation of data products, supplemental data, software, and documentation, that will completely document and support the use of those data products. For example, the images of Jupiter taken by both Voyager spacecraft comprise a single data set.

The object (or data) types in a data product include those shown in Fig. 5 (image, time series, spectrum) as well as array, cubes, spreadsheets, histograms, tables, text and others.

The Object Description Language (ODL) is used to create labels (data descriptions) for data files and other objects. Figure 6 shows a PDS Label for a data product.

```
/* File Format and Length */
       RECORD_TYPE           = FIXED_LENGTH
       RECORD_BYTES          = 800
       FILE_RECORDS          = 860
/* Pointer to First Record of Major Objects in File */
       ^IMAGE                = 40
       ^IMAGE_HISTOGRAM      = 840
       ^ANCILLARY_TABLE      = 842
/* Image Description */
       SPACECRAFT_NAME       = VOYAGER_2
       TARGET_NAME           = IO
       IMAGE_ID              = "0514J2-00"
       IMAGE_TIME            = 1979-07-08T05:19:11Z
       INSTRUMENT_NAME       = NARROW_ANGLE_CAMERA
       EXPOSURE_DURATION     = 1.9200 <SECONDS>
       NOTE                  = "Routine multispectral longitude
                                coverage, 1 of 7 frames"
/* Description of the Objects Contained in the File */
       OBJECT                = IMAGE
        LINES                = 800
        LINE_SAMPLES         = 800
        SAMPLE_TYPE          = UNSIGNED_INTEGER
        SAMPLE_BITS          = 8
       END_OBJECT            = IMAGE

       OBJECT                = IMAGE_HISTOGRAM
        ITEMS                = 25
        ITEM_TYPE            = INTEGER
        ITEM_BITS            = 32
       END_OBJECT            = IMAGE_HISTOGRAM

       OBJECT                = ANCILLARY_TABLE
        ^STRUCTURE           = "TABLE.FMT"
        END_OBJECT           = ANCILLARY_TABLE
       END
```

Figure 6. PDS Label.

2.2 *Maintain Records in a Recordkeeping System*

Figure 7 shows the major subsystems of the Planetary Data System.

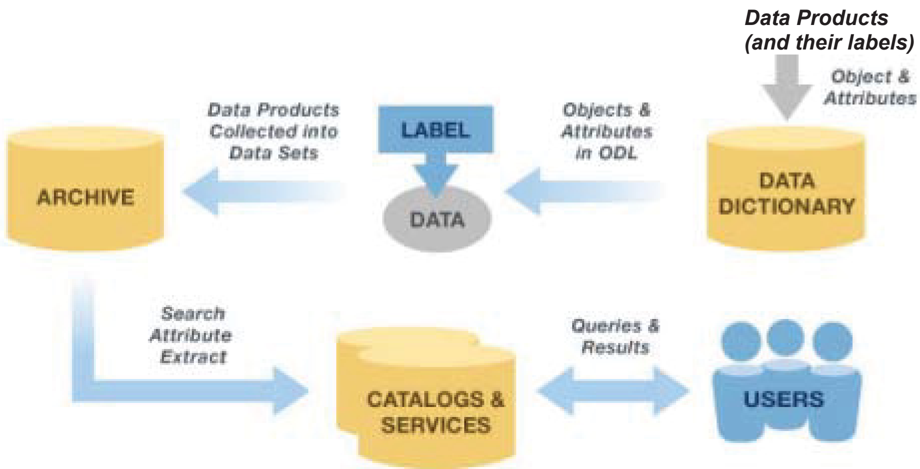

Figure 7. Planetary Data System.

The data product labels and ancillary product labels are included on data volumes and also entered in the PDS high-level catalog. Standard terminology is maintained in the Planetary Science Data Dictionary (PSDD), which is jointly maintained by the PDS and the multi-mission ground data system. The metadata values for new data products are carefully compared with the PSDD and existing values used wherever possible.

The main validation tool of the PDS is the Volume Verifier. This program is run by the Central Node data engineers on each product delivered from a project or a data restoration (in InterPARES terms, a reconstitution). It validates the format and content of all product labels, and validates data files using checksums.

The PDS Archive currently contains about 1000 datasets and about 15 terabytes of data. There are as few as 2-3 products per data set and as many as1000 products per data set. A product is from a few kilobytes to gigabytes in size.

The Archive shown in Fig. 7 performs the function of *Manage Storage* in the InterPARES Model. The Catalog in Fig. 7 performs the function of *Manage Information about Records* in the InterPARES model.

The PDS has been operational since 1990 and it has not been necessary to update (convert or migrate) any of the data products to current or standard data formats. This is because the PDS preservation strategy is what has come to be called *persistent object preservation* (POP). POP is a technique to ensure electronic records remain accessible by making them self-describing in a way that is independent of specific hardware and software.[5]

5 R. Pearce-Moses. A Glossary of Archival and Records Terminology. http://www.archivists.org/glossary/index.asp; A. Rajasekar, R. Marciano and R. Moore. Collection-based Persistent Archives. San Diego Supercomputer Center. www.sdsc.edu/NARA/

2.3 ***Facilitate Access***

NASAView is a PDS archive product display program that runs on multiple platforms in a GUI environment. This application was built using Label Library Light (L3)[6], Object Access Library (OAL)[7] and the XVT Development Solution for C package. Label Library Light parses PDS ODL labels and creates an in-memory representation of the label information. The object access library uses the parse tree and accesses the actual PDS object. The XVT Development solution supplies the cross platform GUI and an object-oriented environment.

OAL allows the inclusion of standard routines for processing specialized representations. For example, the decompression routines for JPEG, GIF, and other standard compression schemes have been included. PDS labels have also been created to describe non-PDS data formats such as FITS and VICR labeled images. Using these "detached" PDS labels to describe non-PDS formatted data, standard OAL function calls can be used to access the data.

3. Walkthrough of the Model using Case Study Data

The InterPARES Chain of Preservation model is a generic model of the process of creating, maintaining, selecting and preserving authentic electronic records. If the model included specific archival decisions, the generality of the model would be compromised. On the other hand, it is intended that it provide a framework for making and carrying out archival decisions. How can archivists know that it is an effective framework for guiding management decisions and implementing preservation strategies?

Walkthroughs using case data are an effective way to test whether a model, design, program code, or user interface achieve what is intended and to improve the quality of the product.[8] A walkthrough is a peer group review of any information system product. A walkthrough of an activity model, such as the chain of preservation model, is concerned with the functionality of the system. Walkthroughs can also be used to determine whether an activity model or design meets functional requirements. To demonstrate that the Chain of Preservation model applies to specific cases of electronic records and to refine and validate the model, a series of walkthroughs is being conducted.

The walkthrough team consists of a *presenter*, who “puts on the table” the model being reviewed; *reviewers*, who have a good understanding of the model, ask questions of the case study expert to identify data corresponding to inputs and outputs of the activities, and raise issues and suggested solutions to problems; a *case study expert*, who answers

---

6 S. Hughes, D. Bernath, S. Monk. Planetary Data System Label Library Light (L3) - Version 1.1, User's Guide, February 3, 1998.
7 R. Davis and S. Monk. Planetary Data System Object Access Library User's Guide, OAL Version 1.2, Laboratory for Atmospheric and Space Physics, University of Colorado, December 1997
8 E. Yourdon. *Structured Walkthroughs*, 4th Ed., Englewood Cliffs, NJ: Yourdon Press, 1989. E. Freedman and G. Weinberg. *Handbook of Walkthroughs, Inspections and Technical Reviews*, 3rd Ed., New York: Dorsett Home Publishing, 1990.

questions posed by the reviewer about data from the case study; and a *secretary*, who records the discussed facts and issues and distributes the minutes.

The method used in the walkthrough is to iteratively step through each of the lowest-level activities in the model:

(1) Reviewing the activity definition and the input, output and control definitions.
(2) Identifying data elements of labels on input and output arrows.
(3) Defining the transformation of inputs to outputs.
(4) Determining values of the data elements that are related to the specific body of records in the case study.
(5) Recording the results and any problems or issues that arise and suggesting possible solutions.

4. Summary

The case study data collected on the Planetary Data System will be used in a walkthrough of the InterPARES Chain of Preservation model to support refinement and validation of the model. Because the walkthroughs provide concrete examples of the application of the models, the walkthrough will also provide archival institutions with knowledge of how they might apply them in their own institutions.

The PDS activities of data preparation and management of data sets in the PDS Archive appears to be more similar to the activities *Managing Records Creation* and *Manage Records in a Recordkeeping System* than to the activity *Select and Preserve Records*. The management at the NSSDC of scientific data sets from different space science disciplines appears to be more similar to the activity *Select and Preserve Records*. However, the Planetary Data System was used as an example of the Open Archival Information System (OAIS) Reference Model.[9] The OAIS may be generic enough to be both a functional model for active archives (recordkeeping systems) and permanent preservation systems.

9 CCSDS 650.0-B-1. Reference Model for an Open Archival Information System (OAIS). Blue Book. Issue 1, January 2002. (IS) 14721:2002) http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf