InterPARES 2 Project
International Research on Permanent Authentic Records in Electronic Systems

Domain 3 Research Questions

Case Study 26:
Microvariability & Oscillations of Stars (MOST) Satellite Mission – Preservation of Space Telescope Data

Sherry Xie, UBC

August 2006

1. What types of entities does the diplomatic analysis identify in this case study? (i.e., records, publications, data, etc.)

The diplomatic analysis (DA) of this case study identifies the following scientific datasets as records generated in the process of conducting scientific activities; that is, to monitor variations of the brightness of stars for the purpose of studying the structure and evolution of stars:

- the sds raw data (directly captured by the CCD in the space telescope on the satellite)
- the sds 1 and sds 2 files (processed by the software on the satellite and prepared to be transmitted back to ground stations)
- the FITS file (created based on sds 2 files), and
- other data products resulting from data-reduction and analysis techniques (e.g., light curves).

1a. If there are no records, should there be records? If not, why not?

Not applicable.

1b. If there should be records, what kinds of records should be created to satisfy the creator's needs (as defined by an archivist)?

Not applicable.

1c. What characteristics of records (as defined by an archivist) are missing yet necessary to preserve these entities?

The identified records satisfy all requirements of being considered as records. In other words, they possess all record components, namely,

- a fixed documentary form,
- a stable content,
- an archival bond with other records generated in the same activity, and
- identifiable contexts.[1]

2. Are the entities reliable? If not, why not?

Yes. MOST is Canada‘s first and (as of mid-2005) only space telescope with the most advanced technologies available.

It was developed as a joint effort of the Canadian Space Agency, Dynacon Enterprises Limited, the University of Toronto Institute for Aerospace Studies and the University of British Columbia.

The processes of capturing raw data, creating sds1 and sds2 file, and transmitting both to ground stations are controlled by computer programs without any human interventions.

The data reliability is therefore assumed based on the scientific authority of the MOST project and the process of capturing datasets.

3. Are the entities accurate? If not, why not?

Yes. The raw data and sds files are accurate to the degree that the satellite’s Science CCD (Charge Coupled Device) electronic detector can capture. According to the project overview in the case study proposal, it is designed with “unprecedented precision.”

The FITS files are accurate to the degree that the software used to create them can express.

One of the MOST team members, an astronomer, is assigned with responsibilities to control the integrity of the data transmitted from the satellite. Mostly due to technical problems, the transmitted data are possibly not integers. This integrity check is based on a technical analysis (done by software) and on an intellectual analysis (done by the instrument scientist).

“Throughout the process of transmission of data, some check sums are in place to ensure that the data have not been changed. This is a technical issue, and has to be done for all raw data files to ensure that they are reliable. In addition, the instrument scientist controls the data in an intellectual way, by looking to individual files or sequences of files. From the moment these checks have been done, the MOST researchers consider the data as good and reliable.”

“In the process of creating these FITS files, there are various checks to assure that the information that is put in, is good. If errors occur, the researchers typically will examine the problem and recreate the FITS files.”

[1] Please note that, in the diplomatic analysis of this case study, five elements (fixed content and form, embedded action, archival bond, persons and contexts) are identified.

4. To what degree can the entities be presumed to be authentic, and why?

Benchmark Requirements Supporting the Production of Authentic Copies of Electronic Records (these apply to the creator**)**:

1. **Expression of Record Attributes and Linkage to Record**

1.a Identity of the record

1.a.i Names of the persons concurring in the formation of the record:

- **name of author:** The MOST Project
- **name of writer:** The MOST Project
- **name of originator:** Each ground station participating in the Project has its own Web site for public communication purpose, which include some of the records generated from the Project. The complete sets of data, after the one year proprietary period, are hosted by the Canadian Space Agency's Canadian Astronomy Data Center (CADC)
- **name of addressee:** Within the first year after data creation: The MOST Project. After the first year: All members of the astronomical community and the general public.

1.a.ii Name of action or matter:

- The study of stars.

1.a.iii Date(s) of creation and transmission:

- **chronological date:** Captured by software automatically.
- **received date:** Captured by software automatically.
- **archival date:** Not applicable.
- **transmission date:** Captured by software automatically.

1.a.iv Expression of archival bond:

- The archival bond among the sds files are determined by the software that transformed the sds raw data to sds files on the satellite. They are organized according to the name of the targeted star and the capturing dates.[2]

1.a.v Indication of attachments:

- Not applicable.

1.b Integrity of the record

- **name of handling office:** The MOST Team.
- **name of office of primary responsibility**: The MOST Team.
- **indications of types of annotations added to the record:** Not applicable.
- **indication of technical modifications:** Not applicable.

2. **Access Privileges**:

It is not clear that whether the MOST researchers (each of whom has distinctive job tasks) have their own passwords to log into their computers. Nor is it clear whether researchers have access to other researchers' records generated from their

[2] The time period for capturing data can be as long as 60 days.

job tasks. They all have the password to the internal Web site that hosts sds data and FITS files. In the sense of external security system, the compartment that hosts the data server in the MOST project office can be locked. There is, however, no information in the report on how the keys are assigned.

3. **Protection Procedures: Loss and corruption of records:**
Data are backed up on DVDs on a regular basis. Two sets of back-ups are made for protection purposes; one is kept onsite and one is kept elsewhere.

Both physical and technological measures are in place after data transmission and creation. They are physically kept in the secured Astronomy Building on the UBC campus, and the compartment in the office that hosts the main computer and ground station devices can be locked. Data can only be accessed through a password-protected internal Web site, and the password is only distributed to the MOST team members.

4. **Protective Procedures: Media and Technology**:
No information was found about procedures established against media deterioration. Since the MOST researchers mostly work with in-house-made software, the software is preserved for future use. They have also done migration from the previous back-up medium, CD, to the currently used medium, DVD, but, again, no written procedures are in place regarding such migration.

5. **Establishment of Documentary Forms**:
The file format of the FITS file determines the documentary form of FITS files. FITS (flexible image transport system) is a standard format for astronomical data. Each FITS file typically represents one image (or spectrum) containing header information (date, exposure time, telescope, etc) and the image itself. These are created from the sds files, with additional input from the streamID files, timing ticks, TLEs, etc.

6. **Authentication of Records:**
There is no individual in the Project having the power or responsibility to authenticate the records in the sense of archival science.

7. **Identification of Authoritative Record:**
The MOST researchers consider the original sds data and the FITS files as the most important entities. The sds data files are never deleted. The FITS files are considered as important because they are the first instantiations in which all relevant data are captured. But FITS files can be re-created at any given time. The official datasets, in the view of the MOST researchers, are datasets stored in the designated data server (including data backed up on DVDs). One MOST research is assigned with responsibility to back up data on DVD.

8. **Removal and Transfer of Relevant Documentation**:
Not applicable.

Based on the above examination against the benchmark requirements, the authenticity of records can be assumed to some extent.

Baseline Requirements Supporting the Production of Authentic Copies of Electronic Records (these apply to the preserver):

Not applicable.

5. For what purpose(s) are the entities to be preserved?

The sds files are preserved for scientific research activities, and the FITS files are available for public consultation through the Canadian Space Agency's Web site when the one-year data proprietary period has passed.

6. Has the feasibility of preservation been explored?

For sds files, a complete set of observations of one target star as well as the metadata (capturing time, information about the devices used, etc.) associated with the dataset need to be preserved.

For FITS files, the transformed data and information from other sources (e.g., data about the orbit) that are necessary for creating them need to be preserved. Such information and other metadata are captured in the fields of the FITS file, hereby the FITS format is important for future use of the FITS files.

For data reductions generated from FITS files, further examination about the data products is need for identifying the elements and components to be preserved. The current strategy used by the creator is to preserve the software used for the reductions.

6a. If yes, what elements and components need to be preserved?

Not applicable.

7. Which preservation strategies might most usefully be applied, and what are their strengths and weaknesses, including costs and degree of technical difficulty?

The creator now maintains the original technology (software) used for transformation and reduction. They also dealt with problems caused by the size of the datasets (approximately 250,000 files per target). The huge data size makes it impossible to store data on the computer hard drive and DVDs. Their solutions were to reformat the Windows OS file management from FAT (File Allocation Table) 32 to NTFS (New Technology File System) and, for burning DVDs, zipping data before the back-up process.

They have also done the migration of data from the previous back-up medium, CD, to DVD.

For long-term preservation of sds files and FITS files, encapsulation (i.e., Preservation Strategy B1.2) might be suitable since the datasets only require specific software to be accessible. If the data and software can be stored on the same media (e.g., DVD for now), future access and use could be achieved.

The advantage of this recommendation is that the MOST team has its own computer programmer taking care of software development. And once the data and software are on the same media, data can be accessed regardless of time and location. But the use of commercial software (as stated above, the use of zip software to reduce data size) causes problems that need to be addressed.

Another consideration worth mentioning here is the hardware needed to play the preservation media. This means that migration of the medium in accordance with the development of computer hardware will be needed.

Have no idea about the cost.

7a. Which alternative preservation strategies might be applied? What are their strengths and weaknesses, including costs and degree of technical difficulty?

B1.4. Conversion
For FITS files, file format conversion/migration will be needed if the preservation decision is to migrate file format for future access.

This recommendation is based on the fact that FITS file format is a widely used file format in the astronomy community, meaning its maintenance, upgrade, and advance might be easier for the Project to handle.

Have no idea about the cost.

8. What additional information does the preserver need to know to facilitate appraisal and preservation?

There is no need for appraisal for the sds files and FITS files because of the nature of scientific data. It's meaningful and desirable to preserve all of them permanently because the observations about natural phenomena are one-time events. They could be preserved for historical values, and also for possible new contributions to our understanding about the natural world when more advanced technologies have been developed and used for their interpretation.

More detailed information about the files and the nature of the research activities could be useful for the preserver. In any events, long-term preservation requires close collaboration between the creator and the preserver.

8a. If required information is missing, where should it come from and how should it be made manifest?

More communications between the creator and the preserver could be helpful.

9. Are there any policies in place that affect preservation?

[Not answered.]

9a. Are there any policies in place that present obstacles to preservation?

Not found.

9b. Are there any policies that would need to be put in place to facilitate appraisal and preservation?

The MOST Project has assigned preservation responsibility (to back up data to DVDs) to one researcher (who is actually the software developer). And one internal document titled "Archival Manual" outlines preservation procedures, which, however, are not always followed in practice.