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

Overview

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

Peter Gagné, Université Laval

May 2006

The Creator Context / Activity

Creator: MOST (Microvariability and Oscillations of STars) research group.

Creator type: Scientific focus / Mixed sphere (university research groups & laboratory). The creator can be considered a mixed creating body because of its particular context, which is that of a non-profit research group.

Juridical context: MOST is a non-profit research group established in 2003. Intellectual property rights are not an issue for the MOST team once data are made available, although they request that acknowledgement be given to them if data are used for outside research and scholarship.

Activity: A satellite was launched in 2003 to monitor and record variations in the brightness of stars in order to study their structure, age and evolution, to understand the effects of magnetic fields on our Sun and other stars and to determine the nature of planets around other stars. The satellite transmits data to three ground stations and will continue operating indefinitely, in theory.

Specific activities include:

- Data input
- Transformation of data into new file types
- Reduction (analysis) of data
- Publication of data

Nature of the partnership

The project is managed as a partnership between three university research units, a large corporation, and a research centre supported in part by a public agency. It is led by the Physics and Astronomy laboratory at the University of British Columbia (Vancouver, Canada), where one of the three ground stations for the reception of satellite data is located. The two other ground stations are located at the Space Flight Lab of UTIAS (University of Toronto Institute for

Aerospace Studies) and at the University of Vienna (Austria), although these sites are not case study subjects and were not interviewed. Dynacon Enterprises was under contract with MOST to build and maintain the satellite and is legally responsible for all technical matters.

MOST is funded by the Canadian Space Agency and must, therefore, abide by guidelines and standards applicable to the CSA and Canadian scientific missions. The contract between MOST and CSA is always limited to one year, because the funding agency is not willing to commit for longer periods. Every year, MOST and the CSA make a new scientific support agreement. Although not articulated in the contract between the CSA and MOST, the research team has an agreement that MOST observations are subject to a one-year proprietary period, after which all scientific data are to be made available to the astronomy community and the general public. The MOST project also has an unwritten agreement with the Canadian Astronomy Data Centre to submit data files that have passed the one-year proprietary period. The CSA is willing to continue funding the project on an annual contractual basis as long as the satellite/cameras function and researchers make use of the collected data.

The University of Vienna has a 'special' institutional status as a partner. It became a partner after the project was started by building a third ground station, which resulted in the opportunity to transmit more data from the satellite. As a non-Canadian partner, it does not receive CSA funding, although it is considered as a full partner in the scientific sense by the two Canadian research units.

Bureaucratic/Organizational Structure

From an organizational perspective, the MOST project can be divided in two separate parts: the technical aspect and the scientific aspect. The technical aspect is the responsibility of Dynacon Inc., which signed a contract with CSA to build and maintain the satellite, telescope and camera. All technical matters and issues related to the satellite are thus the unique responsibility of Dynacon Inc. The scientific aspect includes the reception, treatment and analysis of the data at the three ground stations (UBC, UTIAS and University of Vienna). Dynacon does not receive the scientific data.

Although there is no official written organizational structure and no administrative responsibilities have been explicitly assigned or articulated in writing, specific tasks are assigned to one or more team members, presumably by inference or oral communication. All researchers have a specific field of expertise, such as the instrument scientist and the software developer. The principal scientist is accountable to the funding agency.

Digital Entities Studied

There are two types or levels of digital entities:

- SDS (science data stream) files. These raw data files are transmitted from the satellite to the ground stations. SDS files have several "levels": SDS raw, SDS1, and SDS2 files. SDS2 files contain all critical and some supplementary data, while SDS1 files are compacted version of SDS2 files.
- FITS (Flexible Image Transfer System) files. The SDS files (above) are converted into FITS files by C++ and are the basis for all scientific analysis and data. The FITS files are considered the basis for further work, because they contain the data and the descriptive data (metadata) related it.

In addition, reductions (analyses) and interpretations of the reductions are presented and published in textual documents with graphs.

“The scientific data consist of series of nearly uninterrupted measurements of star fields lasting up to two months, sampled at rates of 1 - 8 times per minute, which can result in up to about 500,000 individual files for a single target.” (FR 1)

Documentary Practices Observed

This project has brought to light some problems associated with **nascent business practices** and serves as a point of reference for such problems in the astronomy field. “Because the MOST research is pioneering in the sense that it accumulates a specific type of data over time, new astronomical projects have contacted the MOST researchers with questions related to handling data.” (FR 3)

No formal **records management** program or written policy has been established to date. This is said to be due to time constraints and focusing priorities elsewhere. “There are hardly any procedures beyond the MOST Archiving Manual due to the organizational culture, the size of the research team and resources available.” (FR 13) It should also be noted that “the researchers use the word ‘record’ in another sense than archivists.” (FR 12)

Records Creation and Maintenance

As part of the university community, the MOST researchers are operating in an academic environment. The MOST research team reports to the Natural Sciences and Engineering Research Council of Canada (NSERC) and CSA. There are few formal **rules, norms or standards** that apply to the research.” (FR 3) However, “as part of the astronomical research community, the MOST researchers use some technical guidelines and norms in regard to the data. For instance the choice of the Flexible Image Transport System (FITS) file format for storing the astronomical data is based on best practice in the field.” (FR 3)

Digital entities are **uniquely identified** by file names managed first by primary target (star) and secondly by date. The files are organized according to this naming convention.

Various **metadata** are included with the digital entities studied. These include contextual information about the satellite camera, information about how the image is cut into relevant pieces (captured in a RASTA file), timing information for each exposure and orbit information downloaded from the Norad Web site. “The metadata or descriptive fields that are attached to the FITS files were partly imposed by the file format, and partly chosen by the MOST researchers.” (FR 3) “In general, no metadata standards are used, the MOST researchers created their own scheme of important descriptive fields…[which] are based on experience and best practice in the astronomical research community, and on the foreseeable use of the records. There is an internal MOST document that describes the descriptive fields of the FITS files.” (FR 13)

With regards to **changing** or modifying records in the system, “the measurement values in the raw data should never be updated, although the format and supporting information could be augmented.” (FR 1) “In general, digital entities are not changed: sds files are never changed, or deleted…FITS files might be recreated if they contain errors…There is no **version** control on recreated FITS files. By recreating FITS files, the old ones are supersceded." (FR 17) Although "there is no audit trail" (FR 12) in the system, the creator claims that "it is always possible to recreate the superseded FITS files using with the earlier, preserved versions of the application." (FR 11)

Recordkeeping and Preservation

A basic **recordkeeping program** using Microsoft Windows tools is currently used. Existing record maintenance activities include **backup** practices similar to those seen in the artistic focus: One person is responsible for the preservation of official data sets, including a routine backup onto two DVDs (originally CDs), which are then held in different locations. A second "unofficial" set of the data is stored on one of the researchers' computer. The files that were previously saved on CDs have yet to be migrated to DVD.

Some internal **guidelines** and written documents concerning the preservation or "archiving" of files have been created, such as naming conventions, metadata included in the FITS files and an "archiving manual." However, these guidelines and manuals are the result of experience and perceived best practice, rather than archival science. The MOST Archiving Manual indicates what should be done on a daily and weekly basis for successfully ""processing, archiving and backing up [of the] MOST data."" (FR 4) However, "in practice, the duties as described in the 'archiving manual' are not always executed as stated in the manual." (FR 5)

Files with scientific data (datasets generated by the satellite camera) are always preserved, even if they are corrupt or false, as corruptions can later be filtered out without having to delete the corrupt file. Other entities are removed from the system when they are no longer up-to-date.

Only the FITS and SDS files are routinely **captured** (and backed up). From the moment other entities are not up-to-date (because for instance there is a better reduction) they can be removed. Other than the Microsoft Windows tools, there is no formal capture system in place.

The use of the FITS file format for data **storage** is part of the methodologies related to academic research best practice of the astronomical research community. It is recognized as "the most widespread standard in the global astronomical community." (FR 1) The repository for the scientific data is currently UBC, where the raw data from the satellite are transformed into FITS (Flexible Image Transport System) format. The research team at UBC has one large office with several computers that store data. One computer is specifically assigned to serve as the main storage space of the data. The office is located in the UBC Astronomy Building and is secured. The main computer and ground station devices are kept in a locked compartment.

Originally, the hard drive was in FAT 32 (File Allocation Table) format. However, this drive type can only handle approximately 64,000 files in all the subfolders. Because the MOST research generates approximately 250,000 files per target, it was impossible to manage, use and preserve the data. The hard drive was reformatted into NTFS (New Technology File System) and the restriction of 64,000 individual files was thus removed.

For the DVDs, a similar problem occurred: a DVD with approximately 250,000 files is problematic, either during the burning process (errors occur) or during retrieval (files cannot be opened). The preservation strategy has been changed and the files are zipped in a single file and then burned.

With regards to the issue of **interoperability** and technological **obsolescence**, backups are made of the custom-made software programs that are used in the project. Old versions of the programs are preserved whenever anything in the software is altered or updated, so that researchers are able to recreate results previously produced. Also, as mentioned above, the FITS format is the most widespread standard in the global astronomical community.

As previously stated, there is a one-year proprietary period for the data generated by the project. During this time, **access** to the raw and reduced data is restricted to the team members via an internal password-protected Web site. After the one-year proprietary period, the data are placed in a public archive accessible to all members of the astronomical community and the general public.

Accuracy, Authenticity and Reliability

Accuracy

"The accuracy of some of the information (e.g., exposure times and duration, detector temperature, etc.) is much more critical to extract scientific information from these data than in many other astronomical catalogues." (FR 1)

"Accuracy and reliability are important issues for the FITS files.... In the process of creating FITS files, there are various checks to assure that the information that is input is good. If errors occur, the researchers typically will examine the problem and recreate the FITS files." (FR 10-11) However, "a researcher noted that values may need to be added or corrected in the FITS file. In this sense, the researchers do not appear to use the words accurate and reliable in an absolute sense.

Authenticity

Authenticity does not seem to be a pressing problem for this project. Files with scientific data are always preserved, even if they are corrupt or false, as corruptions can later be filtered out without having to delete the corrupt file. In addition (or as a result), "There is no reason for the creator to assume that the digital entities are not authentic." (FR 11)

Reliability

In this case study, the notion of reliability seems to be closely related to the integrity of the data used. One astronomer is responsible for monitoring the integrity of data sent by satellite through daily technical analyses, because technical difficulties can sometimes disrupt the integrity of transmitted data. "This is a technical issue and has to be done for all raw data files to ensure that they are reliable. Besides checksums, the instrument scientist looks at individual files or sequences of files. From the moment these checks are completed, the MOST researchers consider the data to be reliable." (FR 10)