FORUM

Lost Science: Protecting Data Through Improved Archiving

PAGES 323–324

Mike Carr wrote recently of the wonderful planetary geology accomplished by NASA over the past 50 years of space exploration, with images and descriptions from the early Venus, Mars, and lunar missions as well as more recent missions, which evoked good memories (M. H. Carr, Geologic exploration of the planets: The first 50 years, *Eos, 94*(3), 29–30, doi:10.1029/2013EO030001, 2013). Unfortunately, memories are not science data, and those 50 years have not been as kind to the scientific information collected by these pioneering missions.

In 1970, the ultraviolet spectrometer data collected around Mars by the Mariner 6 and 7 spacecraft were delivered to the National Space Science Data Center (NSSDC) on a Control Data 6400 7-track tape, and the punch card copy of that data was put into storage. In 2006, a discussion of results from the Japanese Mars probe instruments prompted a search for the Mariner 6 and 7 spectra. When I received a copy of the file of the binary data that NSSDC had converted in 1970 from our 7-track tape to an IBM-compatible format, the file conversion appeared to have lost some of the floating-point data. Still needing the original data, I went to reclaim the Mariner punch cards from the storage facility where they were being held. That facility, however, was being closed, and the contents were being sent to the trash. By a quirk of timing, I was just able to reclaim the Mariner 6 and 7 cards, along with data from Mariner 9 (1971 to 1972), Pioneer Venus (1978 to 1992), and Voyager (1977 to 1989). Thus began work, with a NASA Data Restoration grant, to recover, reanalyze, and rearchive data from these missions. During this research, a number of difficulties were encountered that illustrate important data stewardship concerns. These challenges apply not only to NASA's decades-long archiving commitment but also to data creation and curation efforts across many scientific disciplines.

NSSDC was the repository for the early planetary mission data, and in the 1980s, the Planetary Data System (PDS) was established with discipline-specific nodes and stricter archive requirements. It is recognized that the early archiving activities of NSSDC may have resulted in the loss of some data, but some of the responsibility clearly rests with the mission science teams and the NASA funding priorities. Whether due to lack of funding or archiving expertise, usable science data and metadata from the early days of planetary exploration are now missing from the archives.

A Mission Document Archive

The details of spacecraft hardware and software reside in mission documents that are essential not only to implementing the recovery of archived data but to understanding how it can be used. These mission documents contain spacecraft clock definitions, data handling and timing specifics, and configuration details for the spacecraft and science instruments, plus engineering and computer subsystems specifics. For example, it was not possible to correlate the ultraviolet spectrometer data with two other Mariner 6 and 7 data sets because each of those records used different time tags. Science teams routinely discard these mission documents as members retire or move and institutions clean house. The NASA Solicitation and Proposal Integrated Review and Evaluation System (NSPIRES) list could be used to solicit submissions of old mission documents. If scanned, this proposed Mission Document Archive would be a useful addition to NASA's History Office site (http://history.nasa.gov/spdocs.html).

Migration Issues

Hardware and software migration issues are the bane of the archivist and are well recognized. As hardware grows old and systems change, the migration of non-ASCII science data—things that are not just letters, numbers, and some special characters—is of considerable concern. Computer system changes affect archived software operability. Proprietary software and upgrades may not be backward compatible or even available in the future. Short of storing all science data in ASCII, ASCII samples of a limited number of representative data records from archived products should be included for use in later migration verification.

Science Observation Designs

Science observation designs, whether planetary or otherwise, describe the science to be achieved by detailing parameters such as instrument configuration, timing and sampling, target object, and pointing strategy. These documents often contain additional graphic products that visualize the observation and provide essential information for interpreting the science data. Unfortunately, missions have traditionally coded their design products in proprietary or unique software. These files, even if saved, cannot be read once the project ends. It would be easy to archive scanned documents, with available metadata, as PDF-formatted products.

The Need for a Trained Archivist on the Team

The exhilaration of "first data," the verification of instrument operation, and public relations efforts in the early stages of a mission are frequently poorly recorded, usually because time and energy demands on the science and engineering teams result in few document trails. Spacecraft glitches, computer memory corruption caused by random high-energy particles, fault protection, and contingency planning can all contribute to an undocumented trail of relevant events. Furthermore, instrument microprocessors have significantly improved flexibility and science return, but software changes can affect science and engineering formats, timing, data stream contents, ground handling, and archive product formats. These changes are rarely documented in detail or described in archived metadata. If some member of the team were also trained in the archive procedures and formats that are used where data are to be deposited, this person would be able to recognize, document, and save important information.

How Well Does the Repository Work?

Real-time archiving of downlinked raw data is now required by all NASA projects. High-level calibrated data sets, known as level 1A, may not be available until there is a sizable set from which to analyze results, detect inconsistencies, and remove noise or errors and recalibrate. Although most level 1A data get archived, some products may come later, sometimes after the mission has ended. PDS catalog files describe the entire mission, so these catalog products may be incomplete until the derived results are published and archived.

The problem comes from not having enough funding for postmission archiving. This is a significant cause of the lack of delivery to the archive of data and catalog products. NASA's PDS4 catalog version—the most recent update of the Planetary Data System—should improve ingestion response and metadata handling, but the science teams still must provide the data. NASA's postmission Data Analysis Program for analysis of science data after a mission's funding has expired could provide a program for accumulating and archiving all the remaining products from a mission, project-level documents included.

Metadata Need Attention

We have all become accustomed to searching the Internet when looking for information. The PDS node data are not

© 2013. American Geophysical Union. All Rights Reserved.

available as a web style searchable archive. Although there has been some discussion of making PDS data web searchable, there would be problems even if it were: much of the metadata, both catalog files and internal label descriptions used to organize and describe the products in the archive, do not use a consistent terminology throughout that is well defined and accepted by the search community. Science teams, assisted by a trained archivist, would be able to refine the text as well as incorporate additional metadata that is not currently archived. Librarians routinely catalog materials based on a strict set of classification nomenclature. Should we tag our products similarly?

Finally, because publication-quality data analysis is frequently done late in the funded mission period and metadata preparation can be time-consuming, the PDS node to receive these derived data frequently has newer missions whose data have higher priority to be ingested. This newly submitted, best-quality data from retired missions then falls to the end of the PDS ingestion line and can take significant time to become available to the science community. Some method is needed to provide quicker access to the data submitted after a mission is formally completed. The postmission archiving grant described above might provide this needed priority.

Historic Items

Aside from data and documents, what else is important to save? The loss of the early lunar landing movies is well known. Have you seen a photograph of Robert Goddard with his rockets or Bill Braddock with his digitized geologic maps? How about Torrence Johnson presenting an image, generated from Pioneer Venus Ultraviolet Spectrometer data, of Halley's comet to President Reagan? It's important to ask: What historic things have been produced during any of these missions that should be saved? Scientists' and engineers' notebooks, correspondence and documents, photographs, video presentations and press release articles, key equipment, and logo paraphernalia? All of these should arouse interest. The value that historians of science and technology may recognize in these objects may not be obvious to a science team. The local institution normally archives the scientists' papers, but is it fully capable of cataloging and retaining these products? How can we preserve this history without levying additional work on a science team? What part could the NASA History Office play?

Although these examples come from NASA missions, they illustrate data stewardship concerns resulting from the growth of big data as well as the presence of born-digital data. Many U.S. and international scientific and public agencies are engaged in codifying issues of data curation, especially the knowledge and skill requirements needed in the kind of multidisciplinary data stewardship that allows users to find things inside, and outside, their own discipline. For archiving to be successful, there must be interplay between data creators, data users, and archivists. But, in the end, it is imperative that each data creator understand that how something is saved is just as important as what is being saved.

—Karen E. Simmons, Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder; Email: Karen.Simmons@lasp.colorados.edu

© 2013. American Geophysical Union. All Rights Reserved.