Restructuring Science Data Archive Systems to Be More Usable

Thomas C. Stein [(1)]

[(1)] ***Washington University***
Earth and Planetary Remote Sensing Laboratory
Campus Box 1169, 1 Brookings Drive, St. Louis, MO 63130 USA
EMail: stein@wunder.wustl.edu

ABSTRACT

NASA's planetary science data are archived for long term usability. Archives are created under two limiting factors: the technology of day, and the expectations and abilities of the end users. Over time, advances are made in both of these areas, necessitation restructuring of the archive system by reformatting the archives, providing value added services, or both. The implications of such changes must be considered carefully.

Keywords: data usability, software tools, user expectations

INTRODUCTION

Planetary science data archives are created with the primary goal of preserving data for future study. In addition to data, varying amounts of supporting documentation are included in the archives to ensure long term usability.

In the 1980s, these archives were collections of files following a standard format. Late in that decade and into the 1990s, simple search forms were developed to allow finer grain selection of data. Up to the mid 1990s, data search engines were at the forefront of Internet-based services technology and at or ahead of the standard user's computer ability. Interface improvements stalled at this time, though, because archives were viewed as passive, static data holdings.

Since that time, new techniques and technologies in other areas of daily computing [1] have had two distinct impacts: users have become more sophisticated in their ability and more demanding of the interface. The distance between the data archives and the science user has continually shortened. Even though usability has improved due to additional supporting data and documentation within the archives, the critical link between the data and the user has become stagnant.

To correct for this disconnect, data archivists must think beyond their traditional boundaries to understand the needs of their customers and to deliver useful and usable products. In short, archivists must think of themselves as marketers and sellers of a service-oriented business product. This is a challenging task, because of the increased scope of such an undertaking. Ideally, the archivist will provide a means for their user community to interact through the use of forums or social media in a meaningful and non-trivial manner. Consideration must be given to the way users currently interact with computer systems, such as expected behaviors of web site elements. Despite the resources needed to create such a system, it is necessary that data archivists take this step in order to avoid becoming inconsequential to their community.

EARLY PLANETARY SCIENCE ARCHIVES

Changes in information technology over the past 25 years have changed the way scientific planetary data are archived as well as users' expectations of the archives. The technical expertise required to access and use scientific data has decreased while the pervasion of technology into everyday components of our lives has increased. In addition, the gap between the technicological understanding by professional scientist, the science hobbyist, and the general public has decisively narrowed.

Because of these changes, the nature of the science data archive has changed. Early generation archives contained peer reviewed data and the documentation required to understand the data. For example, the process for finding Viking Lander data of interest was limited to scanning through paper catalogs and then locating the correct binder or bin containing hardcopy prints.[1]

A game of catch up

NASA's planetary science data are archived within its Planetary Data System (PDS), a federation of discipline and support nodes that provides expertise to guide and assist missions, programs, and individuals to organize and document digital data that can be used to support NASA's goals in planetary and Solar System exploration (Figure 1). The PDS was formed in response to the 1986 findings of the Committee on Data Management and Computation's (CODMAC's) Space Science Board. These finding concluded that NASA should "establish a standing data advisory group composed of experienced space scientists…as well as experts in the relevant technologies" in order to provide advice on "matters of data policy, together with computation and data management practices." [2]

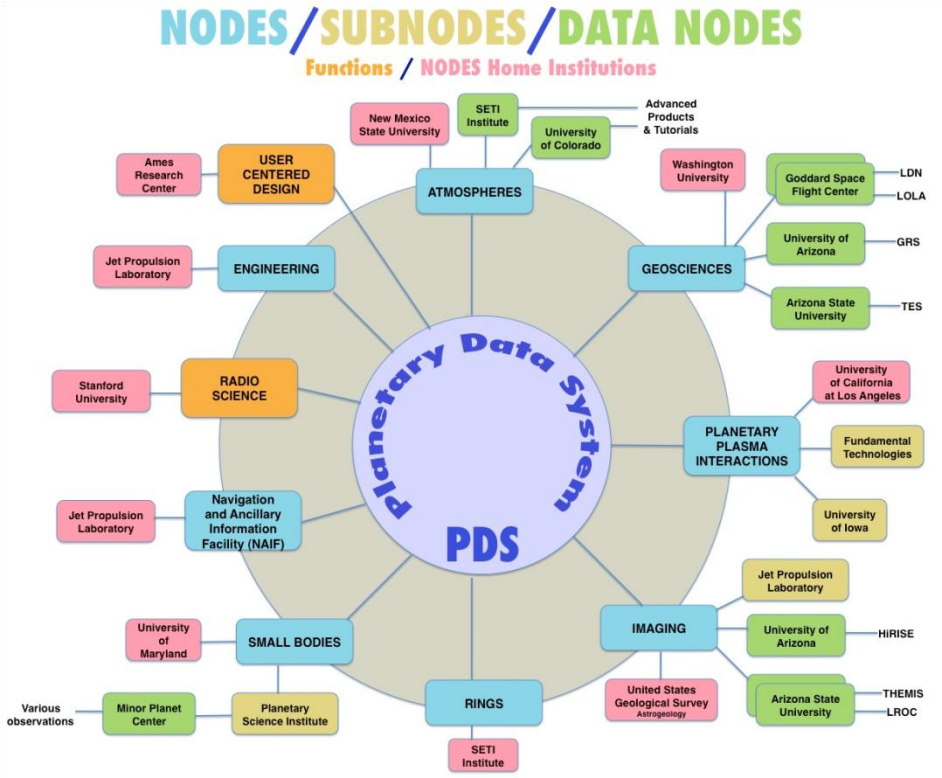

Figure 1. The distributed nature of NASA's Planetary Data System (PDS)

For NASA missions prior to these findings, most gave little thought to the importance of data accessibility after the mission. Many scientists expected a continual succession of missions that would return data sets markedly improved so as to make current data irrelevant. In addition, data archiving was regarded as a future task, and few resources, if any, were dedicated for doing so. Long term data management was not at the forefront of anyone's agenda.

[1] It is noteworthy that for Viking Lander, the printed catalogs contain copious amounts of documentation, such as skyline drawings showing the orientation of the spacecraft and location of acquired image data on each sol [3]. The digital images, originally distributed as photographic prints, were considered to be the primary data. As a result, when the paper catalogs were converted to an online data base, the value added lander diagrams were not part of the archive.

The CODMAC recommendations that led to creation of the PDS were evident in NASA's Magellan mission to Venus. This was the first planetary mission to have a data management and archiving plan prior to launch. Nevertheless, funding for producing the data archive was low, as was the desire for members of the mission archiving team in carrying out the task. Numerous errors in the data were discovered by peer reviewers, both errors in the data and in the delivery mechanism[2]. The mission archiving team worked frantically at the end of the mission to complete their task of delivering data to PDS, and the quality showed. To complicate matters, Magellan was, in one archivist's description, the first mission to produce more data than the science team could look through. Some errors in data products persisted into the archives that likely would have been caught during better scrutiny.

The changing nature of science archives

At the beginning, the primary focus of the PDS archive process was on long term preservation. This included sufficient documentation to understand and work with the data, but less emphasis on adding value to the data. Some effort was made to increase the usability of the archives. For example, in 1990 Magellan MIDR images were archived as an 8 by 7 grid of 1 MB tiles (56 files) for use with computer graphic displays.

Research scientists who made use of the archived data were well versed in the instrument characteristics, often preferring to start with raw data to ensure proper calibration. At this time, however, catalogs of archived data were largely accessible only at archive centers. Thus users routinely requested assistance in locating data of interest that were subsequently copied to tape and delivered via mail.

With the advent of the world wide web in the early 1990s, data accessibility changed. "High speed" internet connections were limited to places of business, but that allowed users to start downloading some data on demand. This resulted in searchable catalogs of data, including browse versions, to facilitate data location.

In addition, computing power (and accessibility to it by the user) were increasing sufficiently that users requested local copies of the data archive. In the case of Mars Viking Orbiter, this meant a 42-voume CD-ROM set of data. No longer did users have to make individual product requests and wait for delivery of tapes.[3]

The changing nature of science archives is part of a self-propagating system (Figure 2). The process begins when value is added to the archive, such as a tool supporting search or visualization. One example of this is the PDS Analyst's Notebook (Figure 3) that adds value to data archives from landed missions to the Moon and Mars. Rather than a directory structure of the archive files, the Notebook presents an interface providing search, browse, transformation, and download functions, among others.

[2] More than 10% of the 21,000 9-track tapes delivered for archiving were unreadable.

[3] The use of CD-ROM made sense at this point: very few users had access to 27 GB of hard drive space for storing the data. Subsequent follow-up, however, showed that many of the CD-ROMs in users' possession remained in shrink wrap years after delivery: scientists were interested in specific (but different) regions of the planet.

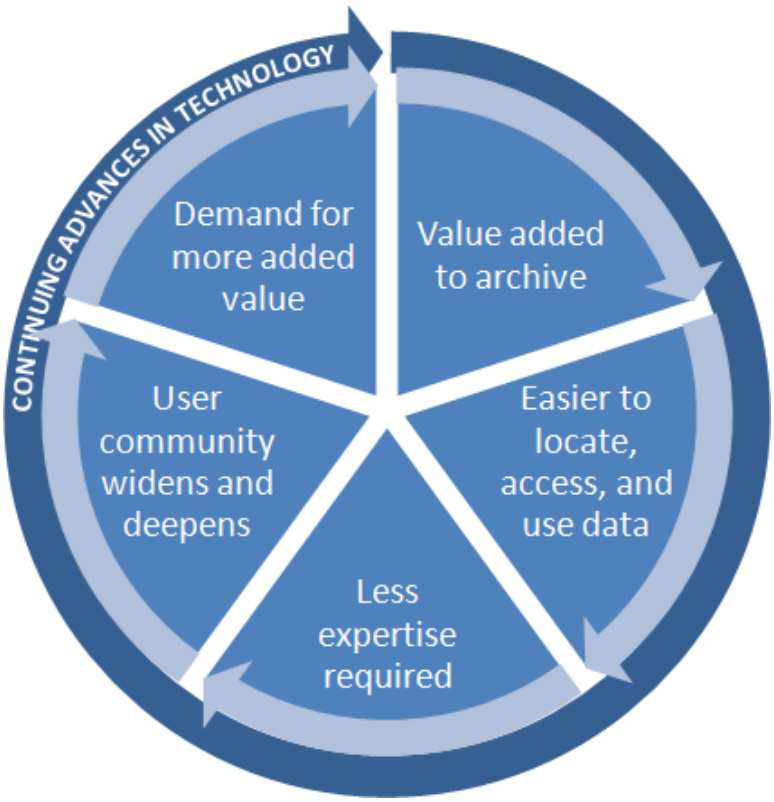

Figure 2. Self-propagating nature of adding value to science data archives.

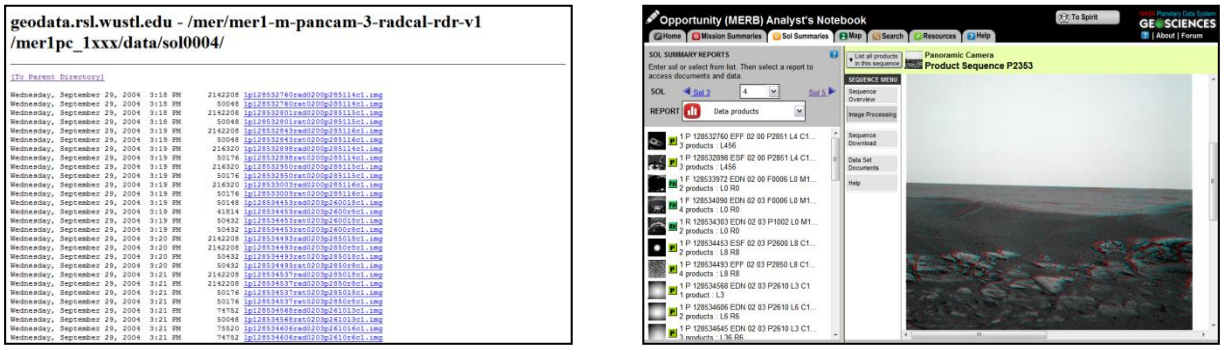

Figure 3. Two methods for accessing MER Opportunity rover sol (Martian day) 4 Panoramic Camera image data from PDS: a directory listing of files (left), and from within the Analyst's Notebook (right). In the Notebook view, browse versions of data from all instruments for the selected sol are listed. In this example, an anaglyph stereo image—not an archive product—has been generated on the fly at the user's request as is available for download.

The value added by the Notebook makes the science data easier to locate (discovery and search), access (preview and download), and use (custom transformation), which in turn reduces the specialized knowledge required on the part of the end user. The result is two types of gains in the user community. First, the community increases in size because less specialized knowledge is required to work with the science archives. Second, the community increases in productivity because less time is spent getting the data. Naturally, the demand for more added value follows, and the cycle begins again.

A companion catalyst in the process is the continuing advancement in information technology. Increased processing power, data storage capabilities, and network transfer rates positively affect all science archive stakeholders. Just as important, science users' experiences in other areas of computing—e.g., social networking, online transactions, etc.—color their expectations of both the science archives and value added tools.

An observable result of the value added tools is the existence of two distinct user populations of planetary science. The first is the research scientist who has or seeks an intimate knowledge of the science data content and formats, and one who expects to invest resources to achieve his or her scientific goals. The second comprises amateur scientists and the interested public who expect quick results with minimal work. The most complaints about the science data and the archives originate from this latter group.

PROBLEMS WITH TOOLS THAT ADD VALUE

There is an inherent challenge with software tools created to add value to scientific data. Notably, software is expensive to create and maintain, and is not generally transportable. As such, scientific analysis software has never been a deliverable item, although various algorithms and software tools have been made available in association with science data archives

Consider the Mosaic Viewer interface to Mars Exploration Rover (MER) image mosaics. Roughly 10,000 image mosaics, ranging from 25 MB to greater than 1 GB per file, have been created by the MER science team. Locating, accessing, and viewing mosaics of interest in a timely fashion is impractical without a dedicated software tool.

The Geosciences Node created Mosaic Viewer (http://an.rsl.wustl.edu/mv) allows users to begin with a base map of each rover's traverse on Mars (Figure 4). As users zoom into the map, higher resolution map tiles in the area of interest are read from a data base, streamed in real time to the client, and seamlessly displayed. Pop up windows display available mosaics at a given location on the user's request. In turn, the user can select an individual mosaic for further inspection.

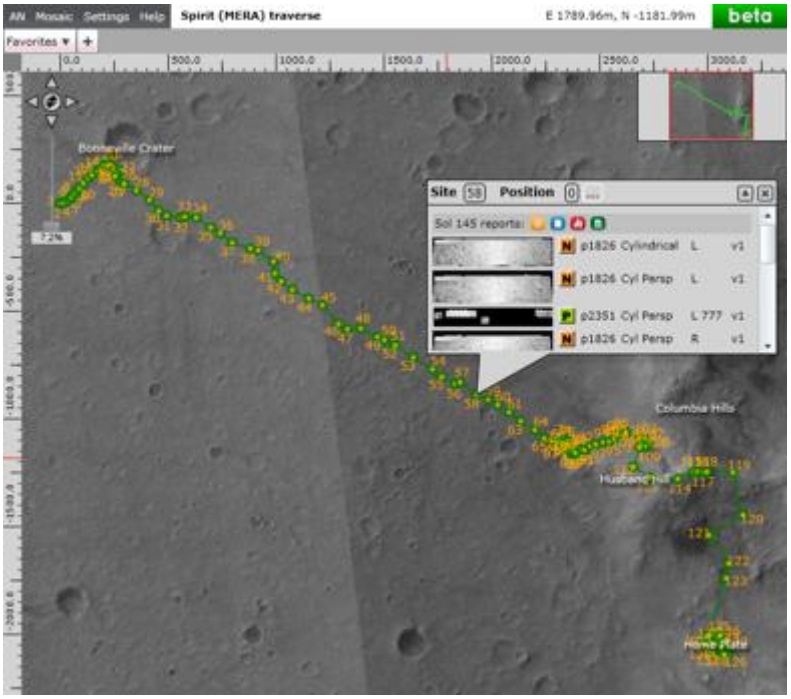

Figure 4. Mosaic Viewer tool showing traverse map for Spirit rover.

Using the Mosaic Viewer, users can zoom and pan around a mosaic without first downloading the data. Each mosaic is made up of two or more source frames—individual images taken by one of the rover cameras stitched together (Figure 5). The software tool provides a listing of the source frames that shows a thumbnail image and the archive product ID of each source frame along with a link to further details available in the Analyst's Notebook. Users also can display "footprints" of the source frames on the mosaic. These footprints show the location of individual frames within the mosaic. Finally, users can download mosaic and source frame data and documentation from a simple order form (Figure 6).

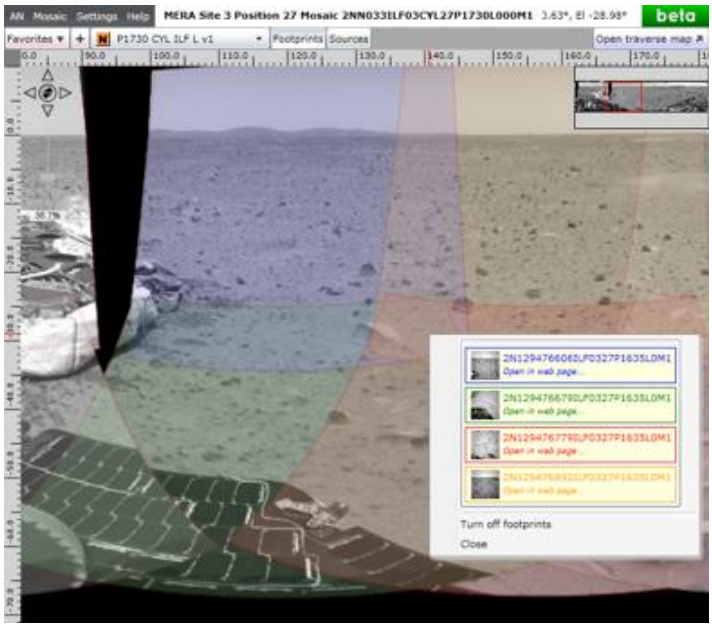

Figure 5. Mosaic Viewer tool showing source frames used to create one of the mission mosaic images.

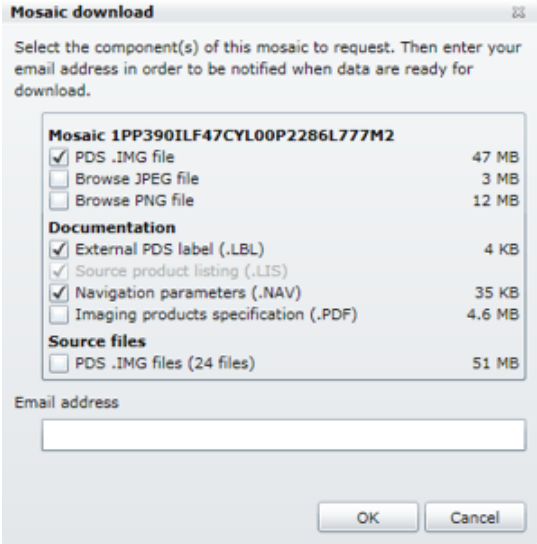

Figure 6. Order form within Mosaic Viewer to provide users with download access.

The Mosaic Viewer is not an archive product per se, but is an extremely powerful tool that certainly adds value to the MER archives. On the up side, the tool runs within a web browser on almost all Windows and Mac/OS platforms. The load time is minimal, response time is quick, and the interface uses common metaphors that are highly intuitive for most users. On the down side, the tool is based on the Microsoft Silverlight browser plugin and its longevity is unknown.

The Mosaic Viewer, is part of the Analyst's Notebook. However, the two applications have different implementation strategies. The Analyst's Notebook is written to maximize the user experience with minimal use of browser plug-ins and operating system dependent code. As a result, the Analyst's Notebook has been publicly available since 2004 without any major changes to the code. The Mosaic Viewer is highly dependent on Microsoft Silverlight and its embedded multiscale image support required to provided to view such large images. Although the development time for the Mosaic Viewer was relatively short, its usable lifetime likely will be shorter than that of the Notebook.

At the close of the MER mission, the Geosciences Node is working to maintain the functionality of many specialized science team tools to work with the data. Decisions are pending with regard to the balance among end user requirements and program functionality and longevity.

What happens to non-archive data that add value to the archive?

A difficulty arising from software tools that add value to archives is the long term fate of non-archived metadata used by the tools. For example, the MER Analyst's Notebook contains a mission summary table—a single line description of the rover activity for each sol (Martian day) of the mission (e.g., "Targeted remote sensing"). This table is searchable within the Notebook, and contains link into data for each sol, yet the table is not an archive product.

There are additional examples within the Analyst's Notebook:

- Daily mission manager and documentarian reports that describe science activities.
- The integrated science plans give detailed sequences of science activities for each sol.
- The start UTC time and L_s for each sol is displayed on sol overview reports.
- Video of LCROSS lunar impact showing image data and target locator (Figure 7).

All of these add value to the archived data, but none are, themselves, archived, nor is there a plan to do so at this time. Beyond these examples is the knowledge inherently present with investigators that collect and produce science data archives. Future growth in archive systems will have to address these cases to preserve the value added information in a meaningful and useful way.

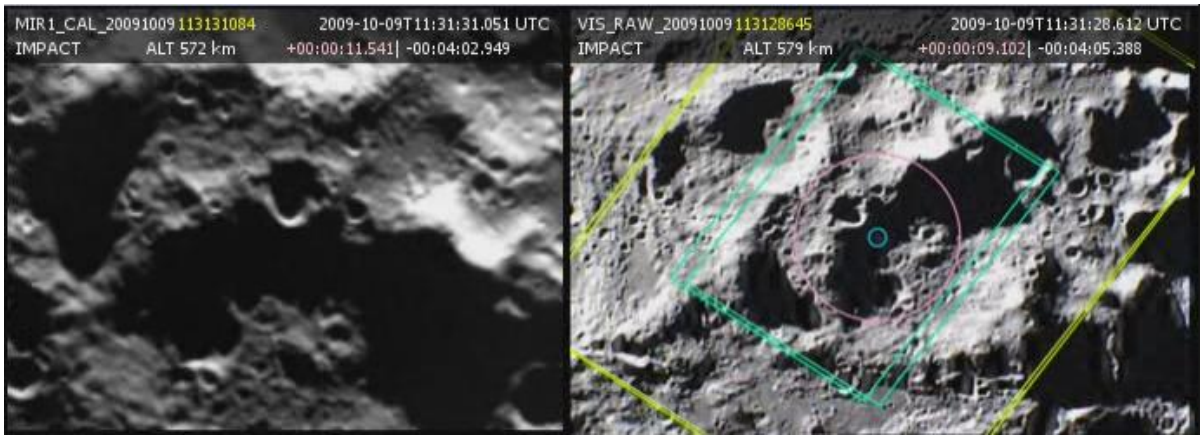

Figure 7. A single image from the online video of the LCROSS mission lunar impact. The movie was created as a value added product to provide location context for LCROSS data. The various color outlines in the right hand image show field of view and targeting information for different instruments.

CONCLUSION

Adding value to science data archives is necessary to facilitate user access. Some enhancements, such as browse images and download carts, are standard enough to have become hygiene factors. The benefits of such added value has two challenges that must be considered and made clear by all parties: the added value will be considered as part of the archive by the user, and the added value may not be usable for the entire life of the archives. In addition, changes in computing capabilities as well as user expectations may require changes to the archives, either to the format of archived data, or via value added tools.

REFERENCES

[1] - Preserving Scientific Data on Our Physical Universe: a New Strategy for Archiving the Nation's Scientific Information Resources. Washington D.C.: National Academy Press (1995)

[2] - Committee on Data Management and Computation Space Science Board: Issues and recommendations Associated with Distributed Computation and Data Management Systems for the Space Sciences. Washington DC : National Academy Press (1986)

[3] - R.B. Tucker: Viking lander imaging investigation: Picture catalog of primary mission Experiment Data Record. NASA Reference Publication 1007 (1978)