MAKING NASA'S ARCHIVED SCIENCE DATA MORE USABLE

A BDTF white paper with an accompanying recommendation

Ad Hoc Task Force on Big Data

November 3, 2017

All studies, findings and recommendations in these deliverables have been submitted to the Science Committee and to officials at NASA HQ. The opinions expressed in these materials do not reflect NASA's concurrence, approval, or indicate steps to implementation.

A BDTF White Paper

Making NASA Science Data More Usable

Data Discovery

Since the first satellites had orbited, almost fifty years earlier, trillions and quadrillions of pulses of information had been pouring down from space, to be stored against the day when they might contribute to the advance of knowledge. Only a minute fraction of all this raw material would ever be processed: but there was no way of telling what observation some scientist might wish to consult, ten or fifty, or a hundred years from now. So everything had to be kept on file, stacked in endless air-conditioned galleries, triplicated at three centers against the possibility of accidental loss. It was part of the real treasure of mankind, more valuable than all the gold locked away in bank vaults.

Arthur C. Clarke, 2001

I. Executive summary

The national policy to make data "open" to the public has placed significant demands on the NASA science data archives. They are mandated to acquire the data and make sure it is of the highest quality possible while dealing with very large and increasing data volumes containing ever more complex data. They are challenged additionally to meet ever-increasing demands from the science data user community. The four divisions of NASA Mission Science Directorate (SMD) each have specific challenges that result from the types of science they support, the types and complexity of the data, the difficulty of acquiring observations and the amount of data. In this report, we review the current state of the data activities in each of the disciplines and the challenges each faces. For example, in Planetary Science and Astrophysics the data are sparse and mission/instrument specific. Earth Science and Heliophysics have very large (multiple petabyte) data sets that are too large for users to efficiently download. The archives are challenged to provide a way for users to efficiently find and access just the needed data. In Heliophysics and Planetary data complexity is an important issue.

Experienced scientists usually can locate and access the data available through the archive systems. In general the archives have been proactive in working to include much of the NASA data and in some cases data from non-NASA sources necessary for the interpretation of the NASA data. However, it is still not possible for even the most experienced users to determine if the data they want doesn't exist or simply is not in the archive. It frequently is much harder for non-experts such as students or those from other disciplines to locate and access the data. Non-expert users frequently want to search on physical parameters or a given topic rather than mission and instrument. Support for broader access to the data based on queries that do not require domain expertise should receive high priority. Not all data are calibrated into physical units and applying calibration to raw data is frequently left to the user. This can be a major task that is beyond the scope of typical research grants. Therefore the requirement for access should

extend to calibrated data either by requiring it to be made available by the data providers or providing processing on demand at the server. Even when calibrated data are available, the complex data can be difficult to use even by experts.

A major step toward alleviating these problems can be accomplished by assuring that optimal metadata is provided and that user's guides, which clearly define the process by which the data were acquired and how they were reduced and that specifies detailed information of data coverage and quality, are generated. If a standardized approach that requires assessment of the quality of archived data near the end of primary missions were implemented and a specific budget item to produce user's guides was made available at that time, this would make the data far more available and would assure the impact of the mission would extend well beyond its lifetime.

II. Introduction

Currently, the demand that all federally funded data be "open" to the general public is placing demands on NASA science archives. Among the basic challenges for maintaining NASA's archives are: 1) data acquisition and quality assurance, 2) limited budgets, 3) dealing with increasing data volume and complexity, 4) utilizing developing technology and 5) meeting user's expectations. Nevertheless, the NASA Science Mission Directorate (SMD) is making an ongoing effort to improve user's access to archived data. A major problem that SMD faces when dealing with preservation of scientific data is the diversity of the data. This stems from the fact that NASA's involvement in the sciences spans a broad range of disciplines across the Science Mission Directorate: Astrophysics, Earth Sciences, Heliophysics and Planetary Science. As the ability of some missions to produce large data volumes has accelerated, the range of problems associated with providing adequate access to the data has demanded diverse approaches for data access. Although mission types, complexity and duration vary across the disciplines, the data can be characterized by four characteristics: velocity, veracity, volume, and variety. The rate of arrival of the data (velocity) must be addressed at the individual mission level, validation and documentation of the data (veracity), data volume and the wide variety of data products present huge challenges as the science disciplines strive to provide transparent access to their available data.

The quality and quantity of archived SMD data has been determined by 3 factors: available power and mass that can be allocated to a specific instrument, the maximum rates of downlink to retrieve data and the difficulty and time required to reach the desired target. These constraints will continue to influence the nature and collection rate of the data in the four science areas and to generate unique problems for each area. In the case of Astrophysics and Planetary Sciences problems stem from the fact that the data are scattered and in many cases sparse and the collection is mission/instrument centered while user's searches tend to be topical or physical parameter oriented. Where temporal and spatial coverage is more available in Earth Science and to a limited extent in Heliophysics, problems are associated with efficiently accessing appropriate amounts of data to address widely varying problems.

a. Astrophysics

Astrophysics (https://science.nasa.gov/astrophysics/astrophysics-data-centers/) supports an integrated system of data archives containing more than 525 terabytes of data, based in part on frequencies covered (ie. UV, visible, IR, etc.) or subject areas (extrasolar planets, extra galactic, etc.). The astrophysics archives are composed of a set of coordinated archives. A tool that provides a powerful preview of the status of a specific field is supplied by the Astrophysical Data System (ADS) (http://adswww.harvard.edu), an online bibliographic database that allows novice users to access current research in their area of interest and identify specific data sets associated with publications. This is supplemented by the resource: Set of Identifications, Measurements, and Bibliography for Astronomical Data (SIMBAD). The individual centers (https://science.nasa.gov/astrophysics/astrophysics-data-centers/) are High Energy Astrophysics Science Archive Research Center (HEASARC), Mukulski Archive for Space Telescopes (MAST), NASA Exoplanet Science Institute (NExSci), NASA/IPAC Extragalactic Database (NED) and NASA/IPAC Science Archive (IRSA). Each center provides its introduction to access of the data, requiring the user to develop an understanding of the individual centers and the characteristics and documentation associated with individual data sets. A number of individuals (8 FTE equivalent) threaded within these groups are funded by a separate intra-NASA entity called the NASA Astronomical Virtual Observatory (NAVO), the U.S. funded arm of the International Virtual Observatory Alliance that seeks to 'federate' dispersed and heterogeneous astronomical databases from thousands of instruments, space-based and ground-based. The IVOA/NAVO software protocols are mostly used internally by the world's astronomical archive centers, but the software is also available to individual researchers. Among its current capabilities is the ability to search globally to find 'if the data needed for a given study' exists. More sophisticated capabilities (e.g. Common Archive Observation Model) are coming on board, emphasizing the synergistic use of multi-wavelength databases and surveys from different missions and observatories.

Although, a preliminary survey of users indicated that they tend to access specific areas of the system and are familiar and satisfied with the archival support they receive, astrophysics is faced with a considerable challenge in the near future. This involves dealing with the output of the Large Synoptic Survey Telescope (LSST) that will produce 15 terabytes per night. The project proposes a system that would automatically process the data and issue alerts to worldwide participating observatories for follow-up observations. Although LSST is funded by the National Science Foundation (NSF), NASA will be faced with combining the results from the LSST with NASA's Wide-Field Infrared Survey Telescope (WFIRST) and ESA's Euclid infrared and visual mission. This effort has the potential of providing multi-wavelength high-resolution images of galaxies and broadband data covering much of the stellar energy spectrum and will involve dealing with a database several times larger than the current astrophysics holdings.

b. Planetary Science

The PDS (https://pds.nasa.gov) is the main archive for NASA mission data and supporting ground-based data. It consists of science discipline nodes (Atmospheres, Geosciences, Cartography and Imaging Sciences, Planetary Plasma Interactions, Ring-Moon Systems, and Small Bodies) and two supporting nodes (Engineering and the Navigation and Ancillary Information Facility (NAIF)). The science discipline nodes are charged with curating the data. This includes acquiring, documenting, validating, distributing and preserving the data.

Engineering provides system-wide engineering support, controls standards, develops system- (IPDA)
(https://planetarydata.org).
(https://naif.jpl.nasa.gov/naif/spiceconcept.html)

In addition, the IAU Minor Planet Center (http://www.minorplanetcenter.net/iau/mpc.html) is linked to the PDS Small Bodies Node and the Planetary Cartography Program (https://astrogeology.usgs.gov/groups/nasa-planetary-cartography-planning) is associated with the Node. Physical sample returns are independently curated by the Astromaterials Curation Facility (https://curator.jsc.nasa.gov/curation.cfm); however, a project is underway to re-engineer sample catalog(s), increase online accessibility, and link to the PDS.

Historically, collection of planetary data has been highly impacted by limitations of available power, downlink rates and cruise times to reach the wide range of available targets: inner terrestrial planets, outer planets with their numerous satellites and the myriad of asteroidal and cometary bodies. In addition, extensive use of PI led missions designed to meet specialized goals, the broad range of disciplines and environments encountered in planetary exploration, management of missions by different agencies and formulation of data structures and pipelines independent of the PDS has led to a large variety of data types and metadata of varying quality. Although the current archive contains more than 1.3 petabytes of data, these constraints have led to decadal gaps in the data for individual targets and the need to access data from widely different missions. These constraining factors will continue to limit the nature and collection rate of the data.

The need to compare reasonable samples of small solar system bodies frequently demands extraction of data from ground-based facilities in combination with detailed mission data. In the inner solar system, the quest for life and planning human exploration has strongly influenced the program, resulting in a stress on Mars and the Moon, leaving Venus to the European and Japanese space agencies.

These circumstances have influenced the manner in which planetary data are archived. In 1982 the National Academy of Science carried out a study that resulted in recommendations for preserving scientific data. Included in the recommendations were that there should be scientific oversight and involvement in archiving the data and that standards of usable formats, documentation and ancillary data be established. This led to the development of the Planetary Data System (PDS) (a discipline based system), the current update of the PDS to an XML-based data model structure (PDS4) and the growth of the International Planetary Data Alliance (IPDA) (using the PDS4 standards and developing international access, which is highly desirable considering the international achievements in characterizing Venus, Moon and Mars).

PDS has been proactive in assuring access to data from NASA missions and through IPDA for missions from other countries. As noted above, those data have been archived to PDS metadata

standards (now PDS4). However, the data products can be very complex and difficult to use by novice users and even by experienced users when they need data not in their precise area of expertise. Following the example of the Cassini mission, the planetary missions have begun to write detailed user's guides for the data. These guides provide detailed instructions for users on how to use the data and point out limitations. They can be accessed through the online PDS system. They have been so successful for Cassini data that hopefully they will be part of the documentation for present and future missions in addition to the PDS4 XML metadata.

c. Earth Sciences

Earth Sciences provides a single portal, called EarthData (https://earthdata.nasa.gov/), to access NASA's Earth science data holdings and software tools supported by the Earth Observing System Data and Information System (EOSDIS; https://earthdata.nasa.gov/about). The EOSDIS system manages all the Earth science satellite data and provides including scheduling, data capture and Level 0 processing. It supports 12 Distributed Active Archive Centers (DAACs). Currently it contains about 17.5 petabytes of data, growing at a rate of 12.1 terabytes per day. Unlike planetary exploration, which is motivated to explore numerous bodies, Earth science deals with varying disciplines applied to one body and has concentrated on acquiring longer and more continuous time-lines of well-specified data, requiring a limited set of data types and development of tools to enable access to the data.

The investigations that these data support and are primarily focused on improving understanding of how and why the global Earth system is changing, including changes in atmospheric composition, the Earth's radiation balance, air quality, ozone layer, ecosystems, biogeochemical cycles and water cycle, and the dynamic surface and interior of the Earth. NASA Earth science also seeks to improve the capability to predict weather, extreme weather events, and climate changes by improving understanding of the roles of and interactions among the ocean, atmosphere, land surface and cryosphere. An important function of Earth system science conducted and supported by NASA is to inform decisions and provide benefits to society (https://science.nasa.gov/earth-science/focus-areas). Specific applications include the development of sophisticated data assimilation methods and first-guess models by the Goddard Modeling and Assimilation Office (GMAO; https://gmao.gsfc.nasa.gov/) and global change simulations and projections produced by the Goddard Institute of Space Studies (https://www.giss.nasa.gov/projects/gcm/). The GMAO produces the Modern-Era Retrospective Analysis for Research and Applications, now in version 2 (MERRA2; https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/), which is a reanalysis of the observations of Earth's atmosphere taken during the modern satellite era that generates a regular four-dimensional state estimate of the global atmosphere. The Goddard provides several tools for manipulating its data holdings, including a subsetter for the MERRA2 data set and a web-based interface, called Giovanni (https://giovanni.gsfc.nasa.gov/giovanni/) that provides users with capabilities for deriving other quantities from the Earth science data archives and visualizing the results, without transporting the data.

Currently EOSDIS houses data from 148 instruments; however, there are only 9 instrument types (https://earthdata.nasa.gov/user-resources/remote-sensors). The reanalysis products (MERRA, MERRA2) represent another data type that assimilates data from all the sensors. Specialized

tools include the Global Imagery Browse Services (GIBS; https://earthdata.nasa.gov/about/science-system-description/eosdis-components/global-imagery-browse-services-gibs), the Land, Atmosphere Near-real-time Capability for Earth Observing System (LANCE; https://earthdata.nasa.gov/earth-observation-data/near-real-time), and a Global Change Master Directory (GCMD; https://gcmd.nasa.gov/). The GCMD provides for inter-DAAC searches. Significant infrastructure supports and enables data ingest via the Science Data Processing Segment (SDPS), configuration control, and metrics tracking via the Configuration Management EOSDIS Tool (COMET), and dedicated metrics tracking via the ESDIS Metrics System (EMS). These systems are managed top-down by the Earth Science Data and Information System (ESDIS) Project at the Flight Projects Directorate of GSFC.

d. Heliophysics

The Space Physics Data Facility (SPDF) (https://spdf.gsfc.nasa.gov/) supports the heliophysics community. Like the planetary sciences, complexity of the data is a major issue. The data include in situ observations of the local charged particles and fields plus remote sensing observations such as images of the auroral ionosphere and neutral atom images (ENA). Observations from more than one instrument are most frequently used and observations from multiple instruments are combined to make higher-level products. The data span interplanetary to interstellar space. Specifically, the SPDF is the final archive for solar wind, magnetospheric and ionospheric data. Heliophysics uses the SPASE (Space Physics Access Search and Extract) metadata standard (http://spase-group.org). Its primary function is to help science data users find and access the data they need for a given study. Up until now the SPASE metadata have been used to locate and access data which frequently are from widely distributed data sources (one recent study found over 100 sources of heliophysics data) but not extract the data. Over the past few years the heliophysics community has developed the Heliophysics Application Programmer's Interface (HAPI) which is designed to aid in extracting the data. HAPI is a data access specification and streaming format specification for time series data (https://github.com/hapi-server/data-specification). Many of the heliophysics repositories are actively creating HAPI interfaces. All heliophysics missions except solar missions use the SPASE standard. SPASE metadata also have been written for observations made on the ground (e.g. from magnetic observatories, radar data, auroral imagers) that are needed to interpret the spacecraft data. Much of the data in the SPDF is available in the Common Data Format (CDF). Models and simulations are an important part of heliophysics research. The SPASE metadata standard is currently being enlarged to include results from models and simulations.

Most solar data are available through the Virtual Solar Observatory (VSO) (http://virtualsolar.org) which is coordinated by the Solar Data Analysis Center (SDAC) (https://umbra.nascom.nasa.gov/index.html). Almost all solar data are stored in files using the Flexible Image Transport System (FITS) as images or spectra and, at least since the SOHO mission, have reasonably consistent metadata that make cross-instrument analysis straight forward. The VSO uses its own metadata system to manage distributed datasets at the file level that are primarily hosted by NASA mission PI teams. The SDAC itself hosts about ~120 terabytes of data, up from ~5 terabytes a decade ago. The largest solar data set resides at the Joint Science Operations Center (JSOC) for the Solar Dynamics Observatory (SDO). The JSOC hosts several petabytes of data at Stanford University. SDO and related missions also supported efforts to provide higher-level descriptions of the data resulting in the Heliophysics Events

Knowledgebase (http://www.lmsal.com/hek). The HEK search tool (http://www.lmsal.com/heksearch) provides a coordinated search interface for finding multi-mission datasets that capture specific phenomena (flares, coronal holes, etc.). In addition to observations from spacecraft solar science requires observations from ground telescopes. Studies of the solar wind, magnetosphere and ionosphere interaction frequently use solar observations.

The actual consumption of solar data in research falls into two categories. First, the majority of solar data analysis and much of science data processing are conducted using SolarSoft/IDL on local workstations and mission servers. The SolarSoft system supports both functions as part of a set of distributed software distribution trees and with a wide range of web services. On the other hand, helioseismology studies typically require datasets that are more easily analyzed by running the analysis code on compute clusters tightly connected to the large data stores. The JSOC, as hosts to the largest helioseismology datasets, hosts both types of interactions.

The divisions of the mission directory are striving to respond to the needs of their user communities. However, the demands for open data, especially derived data to fold into sophisticated modeling efforts, are increasing and will continue to challenge SMD. Data discovery will become an increasing concern as NASA funded research produces interdisciplinary derived data that cuts across disciplines, missions and instruments.

III. Statement of the Problem

Users want data searches to be intuitive. Novices want to search for data based on a theme, target or discipline. More informed users expect to locate data based on instrument type and frequency spanned, possibly mission or instrument name or by a citation reference. Neither wishes to encounter difficulty in reaching a site that provides the desired data. Expectations based on technology development are placing increasing demands on accessibility and data discovery while budgetary restrictions, lack of control of data generation and limitations of supporting documentation restrict development. These constraints present a supreme challenge for optimal development of the archives.

The typical user is often not blessed with high BAWD rates that would enable data transfer. Once users access the desired datasets, they want services that allow them to determine the quality of the data, to select specific characteristics that will limit data transfer or to determine that there is access to local customizable processing. The challenge is striking a balance between allowing easy access to the data and assuring the user is aware of the existence of and need to utilize available metadata. An ongoing effort should be made to expand search capabilities while assuring correct use of the data.

Presentations in a recent symposium concerning open science by the National Academy of Science stressing computers and information technology addressed data access. Among work cited was a study of skill gap analysis by the Belmont Forum, an international partnership dealing with environmental change that strives to remove critical barriers to sustainable e-infrastructures for global change research. The opening question in a usage survey asked, “How would you describe the largest challenge you encounter in your data use?” 73% of the respondents cited one of 4 issues: 1) Data complexity, 2) Lack of data standards and exchange

standards, 3) Finding relevant existing data – knowing what's out there and 4) Data management and storage. The remainder of significant topics is shown in Figure 1.

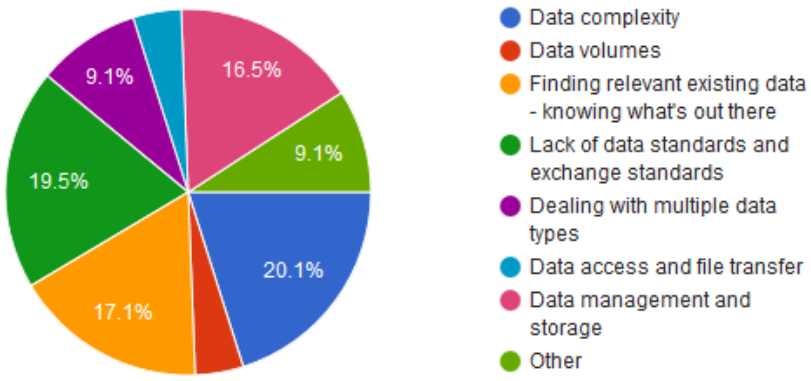

Figure 1: The largest data use challenge encountered by respondents – Vicky Lucas, Belmont Forum, Skill Sap Analysis, e-infrastructures and data management in Global Change Research.

These studies and others are framing the needs of current users, many who expect to utilize limited google-like searches, which require commitment of considerable expertise and upkeep to provide.

IV. Ongoing Development

The 4 divisions of SMD provide individual master websites, Astrophysics https://science.nasa.gov/astrophysics/astrophysics-data-centers/, Earth Science https://earthdata.nasa.gov/) Heliophysics with the Space Physics Data Facility (https://spdf.gsfc.nasa.gov/) and the Solar Data Analysis Center (https://umbra.nascom.nasa.gov/index.htm) and Planetary science (https://pds.nasa.gov), that identify and link to the available components of their archive systems, leaving the development and maintenance of user assistance to the individual components. For many discipline- oriented users this approach is adequate; however, the extent to which the archives are mission/instrument oriented versus identified physical parameters that are represented in the data varies considerably. Presentations and demonstrations by members of SMD and the data centers demonstrated that there is awareness that the current archival structures will need considerable modifications and, within the current budgetary limitations, efforts are being made to improve specific areas. Use of the cloud technology and concerns about access limited due to costing are being studied.

V. Recommended Approach

The following recommendations are geared toward users who are not directly involved in a mission.

Frequently asked questions by the user are: Does the data exist, Did I get it all, How do I transport it or can I access it on site, Are there tools to assist in accessing this data, Where can I go for help in understanding the data. Efforts should be made to assure that the individual components of 4 SMD archives regularly assess their systems based on user needs.

Even when the data are available and can be accessed they can be very complex and therefore difficult to use even for expert users. Detailed user's guides have proven to be a good way to help the address this problem. Efforts should be made to include detailed usage guides with the archived data.

Current concepts of data mining tend to assume that the data is homogenous. Unfortunately, this is not the case. However, in many cases appropriate indexing of the data based on observational parameters should be implemented to allow for efficient retrieval of the data at the file level.

VI. Conclusion

On February 22, 2013 in response to increasing congressional pressure to make public funded research more open to the public, John Holdren, Director of the Office of Science and Technology Policy issued a memo (See the Appendix) stating that federal agencies investing in research and development must have clear and coordinated policies for increasing such access. "Open" is a tall order, requiring making data findable, accessible and usable.

- In general NASA data systems allow experienced science users to find and access the data available in their systems.
- It can be difficult for those who are not domain experts to find and access data. Improved search and access methods with emphasis on the needs of the non-domain expert are needed.
- It can be difficult for even experienced users to determine whether the data they desire for a study exist.
- Access to calibrated data in physical units is not universal within NASA science domains. Frequently the data require extensive processing to be scientifically useful. NASA data systems need to work toward providing the data in forms that do not require a great deal of processing by the user.

Data volumes, the variety of data products and the number of data providers are all increasing. These data come from instruments of increased complexity. As a result it is more important than ever to pay close attention to providing effective access to the data. The data management structure supporting access to these data needs to be extremely robust. Today NASA science data systems allow a user to access data in the system especially if they are familiar with the research discipline and NASA missions. It is much harder for those who are unfamiliar with the science domain or the mission. However even the experienced science user finds problems. Even with detailed documentation and calibration data, the time and effort required to obtain the data products, understand how to use them and make them directly useful make it difficult to use the data. The trade-offs between processing the data on request and requiring data providers to include data processed into products that are immediately useful need to be addressed.

In many cases, it is very difficult to know if the data needed for a given study exist. When a researcher's queries to one of the data systems receives a null result does it mean that the data do not exist or that they simply are not in the archive? It is important that the data systems work to capture information about all of the relevant data.

Once the data are found it is not always possible to understand their status, structure and scope. A major step toward reducing this problem can be accomplished by assuring that optimal metadata are provided and that users guides, which clearly define the process by which the data were acquired and how they were reduced and that specifies detailed information of data coverage and quality, are generated**. If a standardized approach that requires assessment of the quality of archived data near the end of primary missions were implemented and a specific budget item to produce user's guides was made available at that time, this would make the data far more available and would assure the impact of the mission would extend well beyond its lifetime.**

VII. Acknowledgements

The BDTF appreciates the many opportunities to interview scientists and managers at NASA HQ and at the archive centers. These exchanges were open, succinct, and extremely valuable in preparation of this study**.**

VIII. List of Acronyms

ADS	Astrophysical Data System
CDF	Common Data Format
COMET	Configuration Management EOSDIS Tool
DAAC	Distributed Active Archive Centers
DISC	Data and Information Services Center
EMS	ESDIS Metrics System
ENA	Energetic neutral atom data
EOSDIS	Earth Observing System
ESA	European Space Agency
ESDIS	Earth Science Data and Information System
FITS	Flexible Image Transport System
GCMD	Global Change Master Directory
GES	Goddard Earth Sciences
GIBS	Global Imagery Browse Services
GAMO	Goddard Modeling and Assimilation Office
GSFC	Goddard Space Flight Center
HAPI	Heliophysics Application Programmer's Interface
HEASARC	High Energy Astrophysics Science Archive Research Center
HEK	Heliophysics Events Knowledgebase
IDL	Interactive Data Language

IPAC	Infrared Processing & Analysis Center
IPDA	International Planetary Data Alliance
IRSA	NASA IPAC Infrared Science Archive
IVOA	International Astronomical Virtual Observatory
JSOC	Joint Science Operations Center
LANCE	Land, Atmosphere Near-real-time Capability for Earth Observing System
LSST	Large Synoptic Survey Telescope
MAST	Mukulski Archive for Space Telescopes
MERRA	Modern-Era Retrospective Analysis for Research and Applications, version 1
MERRA2	Modern-Era Retrospective Analysis for Research and Applications, version 2
NAIF	Navigation and Ancillary Information Facility
NASA	National Aeronautics and Space Administration
NAVO	ASA Astronomical Virtual Observatory
NED	NASA/IPAC Extragalactic Database
NExSci	NASA Exoplanet Science Institute
NSF	National Science Foundation
PDS	Planetary Data System
PDS4	XML-based data model structure
SDAC	Solar Data Analysis Center
SDO	Solar Dynamics Observatory
SDPS	Science Data Processing Segment
SIMBAD	Set of Identifications, Measurements, and Bibliography of Astronomical Data
SMD	NASA Science Mission Directorate
SOHO	Solar and Heliospheric Observatory
SPASE	Space Physics Access Search and Extract
SPICE	Spacecraft, Planet, Instrument, Orientation, Events information
SPDF	Space Physics Data Facility
VSO	Virtual Solar Observatory
WFIRST	Wide-Field Infrared Survey Telescope

Appendix

Excerpts from a Directive from the Office of Science and Technology Concerning Increasing Access to Results of Federally Funded Scientific Research

The directive is accessible via:
https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research

In February 2013 Director John Holdren issued a directive to the heads of executive departments and agencies on the subject of Increasing access to the results of federally funded scientific research stating that, "The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data…….".

Regarding digital archives, the directive required that each agency develop a plan to support increased public access to the results of research funded by the Federal Government that contained the following elements:

- a strategy for leveraging existing archives, where appropriate, and fostering public- private partnerships with scientific journals relevant to the agency's research;
- a strategy for improving the public's ability to locate and access digital data resulting from federally funded scientific research;
- an approach for optimizing search, archival, and dissemination features that encourages innovation in accessibility and interoperability, while ensuring long-term stewardship of the results of federally funded research;
- identification of resources within the existing agency budget to implement the plan;
- a timeline for implementation.

The objectives affecting Digital archives in this directive included, "To the extent feasible and consistent with applicable law and policy[1]; agency mission; resource constraints; U.S. national, homeland, and economic security; and the objectives listed below, digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding should be stored and publicly accessible to search, retrieve, and analyze. For purposes of this memorandum, data is defined, consistent with OMB circular A-110, as the digital recorded

factual material commonly accepted in the scientific community as necessary to validate research findings including data sets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer review reports, communications with colleagues, or physical objects, such as laboratory specimens." Each agency's public access plan shall:

a) Maximize access, by the general public and without charge, to digitally formatted scientific data created with Federal funds;

b) Ensure that all extramural researchers receiving Federal grants and contracts for scientific research and intramural researchers develop data management plans, as appropriate, describing how they will provide for long-term preservation of, and access to, scientific data in digital formats resulting from federally funded research, or explaining why long- term preservation and access cannot be justified;

c) Promote the deposit of data in publicly accessible databases, where appropriate and available;

d) Develop approaches for identifying and providing appropriate attribution to scientific data sets that are made available under the plan.

---

[1] These policies include, but are not limited to OMB Circular A-130, Management of Federal Information Resources, available at: http://www.whitehouse.gov/omb/circulars_a130_a130trans4

BDTF Recommendation for Making NASA's Archived Science Data More Usable

Background: The Big Data Task Force review of the NASA science archives found that while the archives were in general proactive in encouraging missions to submit data, the quality of the metadata describing the data and the calibration in some cases were inadequate for successful analysis of the data. Science data volumes from ever more sophisticated instruments are growing. While users can find the data products using the online systems at the archives many are very difficult to use. Even domain experts often find the data difficult to use. A major part of the problem with using the data results from the complexity of the instruments which leads to very complex data products. The key to having archival data products that the science community can readily use is well calibrated data and metadata that clearly describe the data products in the archive. Recently the Cassini mission to Saturn has augmented the metadata by including detailed user's guides. These text documents have been very successful in aiding users as they work with the very complex data from Saturn.

Recommendation:
The BDTF recommends that near the end of the prime mission NASA conduct a review of data entering the archives including the quality of the calibration and the metadata describing the mission. Spacecraft and instrument status may have changed during the mission and should be updated. In addition, we recommend that at this time the missions prepare or update user's guides for the data products from each instrument detailing their use. These are important steps in making the data truly useable and essentially extending the effective life of the mission.

Rationale for the recommendation:
As missions age instrument states change and poorly calibrated data can get into the archives. Including calibration reviews in a major review such as that at the end of the prime mission will improve this by catching errors and updating documentation and calibration tables. Space instruments have become so complex that using the data can be very challenging especially for those with limited resources from small grants. User's guides have proven to be a straightforward and effective way to make the data more useful and essentially extend the missions beyond their active lifetimes.

Consequences of no action:
Without the calibration review poorly calibrated data will continue to be mixed into the archives. Without the user's guides more scientist time and effort is needed to learn the complex instruments and use the data. Some studies for which a given data product is appropriate will not be feasible given limited resources.