Senior Review 2009 of the Heliophysics Data Archives

September 14, 2009

Submitted to:

Richard R. Fisher, Director
Heliophysics Division
Science Mission Directorate

Submitted by:

Bruce Berriman, Reiner Friedel, Frank Hill, Susan Niebur, Alexei Pevtsov, Mark Sykes, and Cecil Tranquille.

Introduction:

As a matter of policy, NASA's Science Mission Directorate (SMD) periodically conducts comparative reviews of Mission Operations and Data Analysis (MO&DA) programs to maximize the scientific return from these programs within finite resources. The acronym "MO&DA" encompasses operating missions, data analysis from current and past missions, and supporting science and data archive centers and services.

SMD is undertaking a survey of its holdings in the Heliophysics data environment in the context of the Senior Review process. To this end, the Heliophysics Data Environment Senior Review peer process was held July 21-23, 2009, in Washington DC. The review was to take into consideration the fact that the scientific community is moving to a more integrated approach to the research and analysis of scientific questions; the use of diverse datasets in multiple regimes to perform said analyses; the growing accessibility of NASA data assets on the Web; and the community's and the general public's expectations to find these data on the Web. NASA's Heliophysics operating missions are now entering a new era of Legacy datasets that are proving of inestimable value, as well as new and important missions that will contribute significantly to NASA's data holdings in Heliophysics.

The Heliophysics Division recognizes that, in order to make the Heliophysics data environment a lasting and useful resource, a change in paradigm in the funding of these assets must be undertaken. Over the last few years, the Heliophysics Division has undertaken an extensive series of discussions with the community, and on the basis of these discussions, has issued a community-endorsed Science Data Management Policy. This new policy is to be taken as the guide for data related issues. SMD understands that archiving, while essential, is not the final be-all and end-all: to make all of these data products a living asset, one must curate and develop the concept of archival data centers (in the broadest concept), as opposed to the current model of the more simple data archives ("spinning bits").

Purpose of the Senior Review:

The purpose of this Senior Review is to assist NASA in maximizing the scientific productivity of the Heliophysics data environment program. NASA will use recommendations from the Senior Review to

- Define an implementation strategy;
- Give programmatic direction to the projects concerned for 2010 and 2011; and
- Issue preliminary guidelines for 2012 through 2014 (to be reviewed again in 2013).

The Data Environment Senior Review:

To maximize the scientific return from its programs and projects, NASA routinely seeks input from the scientific community. Working groups and user groups deal with NASA's Space Science program by focusing on discipline- or theme-wide, sub-discipline, or project-specific issues. The Heliophysics Data Environment Senior Review under this call for proposals, is held every four years, and complements the standing working groups and other peer reviews, which conduct independent, comparative evaluations of the various projects. The Data Environment Senior Review evaluated proposals for continued and augmented funding for a number of significant and important projects.

Previous Senior Reviews, as well as standing advisory groups, have recommended that performance factors for this review should include scientific productivity (which would include but is not limited to: the numbers of refereed papers produced, completeness of data holdings, the conformity of the data and metadata held to accepted standards, and the ease of access and use of data), technical status, future plans and expectations, and budget reasonableness.

Charter for this Senior Review:

In the following descriptions, "project" will denote an archive project with a specific Work Breakdown Structure (WBS) element in the Agency's budget, either as a directed project or as one selected through various elements of the ROSES NRA.

NASA's charter to the Data Environment Senior Review panel was to:

(1) In the context of the science goals, objectives, and research focus areas described in the Science Mission Directorate's Science and Strategic Plans, assess the scientific merit on a "science per dollar" basis – based upon the expected returns from the projects reviewed during 2010 through 2014, and how the various project have, or are implementing, the Heliophysics Science Data Management Policy.

(2) Assess the cost efficiency, as a secondary evaluation criterion, after science merit/usefulness.

(3) Based on (1) and (2), provide findings to assist with an implementation strategy for Heliophysics Division data curation, dissemination and archiving, which will include an appropriate mix of

- Continuation of projects as currently baselined;
- Continuation of projects with either enhancements or reductions to the current baseline; and
- Consolidation of projects and activities to enhance efficient management of limited budgetary resources.

Panel Review Process:

The Senior Review panel met for three days:

Day 1:
Morning: Charter; discussion of conflicts of interest and procedures to minimize their impacts; logistics (writing assignments, etc.), background, comparisons, metrics and criteria.
Rest of day: Program presentation, plus questions and answers.

Day 2:
Morning: Continue program presentation.
Rest of day: Senior Review Panel begins Charter Tasks (1) through (3).

Day 3:
Senior Review Panel completes Charter Tasks (1) through (3).

Presentations to the Review Panel:

Each proposing project was allotted 60 minutes for an oral presentation to the senior review panel. To minimize the burden on projects, no more than a total of four persons may represent any one of the projects. During each project presentation, the project representatives were to have planned on using no more than 30 minutes for their prepared presentation, and reserving the remaining 30 minutes for questions and answers (as needed). The primary purpose of the oral presentations was to provide a forum for questions from panelists and answers from the programs. Secondarily, this was an opportunity for programs to provide any significant updates, *e.g*. data holdings or science results obtained since proposal submission. Lastly, and with lowest priority, this was an opportunity to repeat highlights of the proposals, which had been read by all the panelists, and provide live demonstrations where appropriate.

Individual archive comments and findings:

SDAC evaluation.

The SDAC serves data and provides analysis tools for the solar physics research community. Since most of the SDAC data are essentially images, the basic tools and processes are mostly generic and fairly complete. The SDAC combines several roles from the Heliophysics Science Data Management Plan (HSDMP), and in its current form serves the solar community very well. The SDAC provides data access and mission critical services for SOHO, STEREO, Yokhoh, and Hinode, and leads the way in the development of virtual observatories in NASA's Heliophysics Directorate with the VSO. It supports the development of the SolarSoft package for solar analysis, and makes high-quality solar data easily available to all segments of society. The SDAC is also developing a system for distributed serving of the forthcoming SDO data.

However, the panel felt that there are a number of areas where the SDAC should redirect their efforts. Specifically, the development of the distributed SDO service may not be the most cost effective solution to that problem, and SDAC resources may be better spent on further integration into an overall heliophysics data system, as well as improvements to the user interface. In addition, if SDAC is to be the final archive for solar physics data, then such planning needs to begin now.

Science strengths and weaknesses:

The SDAC and the associated VSO, are useful tools for solar physicists to locate a variety of data for their research. Since almost all observational solar physics studies use multiple data sets, the SDAC greatly enhances the productivity of solar physicists and leverages the scientific return of NASA's solar missions.

The data sets are of crucial importance to the heliophysics community and this is reflected in the citation rate in the literature. The SDAC effectively reaches a large number of researchers, as evidenced by the 282 publications that cite data obtained in FY07 and FY08.

The strategy and implementation of using the VSO not only as a "front end" for distribution of all SDAC online datasets beginning with *SOHO,* but also for the data sets served by other VSO data providers, is a sound one, and one that could feed forward into any future heliophysics data archiving systems without much trouble.

SDO services:

- The construction of an SDO "satellite server" system will partially overcome the bandwidth limitations of the active archive at Stanford. This will increase the scientific productivity of SDO. However, the panel felt that the need for such satellite servers is not convincingly justified.
- The data volume is sufficiently high that this approach may not really solve the bandwidth problem. Download speeds are mainly limited by the end user's access to bandwidth at their home institutions, and there is nothing SDAC can do about this. The Stanford bandwidth bottleneck is likely to be a temporary situation as the Internet

infrastructure continues to improve. Thus, a more cost-effective near-term solution may be to simply ship loaded hard drives to users as needed (as described by the SDAC in their presentation for one case).
- The maintenance demands on the satellite server system were never adequately described.
- One major concern of the review panel is the SDAC role in SDO data dissemination, which is strictly a SDO project role and appears to be an unfunded role for SDAC. This appears to be a way to spread the SDO cost, which should be strongly resisted, so that SDAC resources can be used for more of its prime mission.

The SDAC scientific strength would be increased by a tighter integration with other VxOs. The SDAC appears to be reluctant to work on the SPASE standard. The panel feels that this should be one of the primary roles of the archive. The removal of meta-VxO development from the baseline budget (Section 7.1) is felt to be a loss.

Providing greater leadership for the Solar VO, rather than providing access to SDO data, would allow SDAC to provide best value to the community. The current VSO services are rudimentary, and further development is needed – *e.g.* rather than direct users to sites that serve data, it should aim for "one-stop" shopping for access to data, and develop science-driven services identified by the community. Some additional desirable features would include:

- a graphical user interface,
- integrated analysis services,
- data quality flags.

The SDAC serves as an active archive for several NASA solar data sets (SSM, OSO-7, SOHO, STEREO, Yohkoh) and provides links to archives outside of SDAC (TRACE, Hinode). However, the Yohkoh data set in SDAC is not the most current one; there is no clear explanation of what version of the data is provided and where the most up-to-date data can be accessed (i.e. the Yohkoh Legacy Resident Archive at http://solar.physics.montana.edu/ylegacy/).

The panel is concerned that the level of effort to maintain existing SDAC resources may suffer as the project takes on more responsibilities. For example, the SDAC web page would benefit from a better layout. The front page contains many links secondary in importance to the purpose of SDAC. The "Solar monitor" is an excellent tool, but it does not always work as intended (*e.g.* it may show images as not available, even if there is data in the SOHO/EIT data base). The concern is that some of these issues are neglected as the team is committed to new projects.

The SolarSoft package is powerful and of great value to the Heliophysics community, but is the SDAC the best place for data analysis tool development? Should the SDAC be the repository of a community wide effort to develop the best tools? It is currently the only data center strongly involved in analysis – something done by the VxOs elsewhere. The current analysis role is historical, but is this the most efficient way of doing it?

Relevancy strengths and weaknesses:

The SDAC is highly relevant to NASA's heliophysics data system. It has pioneered the development of virtual observatories in heliophysics.

The SDAC risks becoming irrelevant to NASA's heliophysics data system unless it increases its efforts to become integrated into the system, and improves its access and analysis tools.

Some functions are clearly duplicated by the SDAC and the VSO. For example, the provision of links to data sets outside SDAC (e.g., Hinode and TRACE) duplicates the search capabilities of VSO.

The panel was concerned with repetition of the “Slashdot situation.” With no clearly defined “for scientists” label on the web site, the public could easily erroneously conclude that this is a good place to play with data and/or make movies to show their friends. This is clearly a problem in terms of bandwidth, but could become one of interpretation as well. A straightforward solution to this is to redesign the front page, adding a prominent link to one or more E/PO sites, and possibly include a low resolution JPG movie of an interesting data cycle. Moving the referenced data offsite is a solution for a one-off situation, but it is hoped that sun data will continue to be interesting to the public – and so this may happen again. The suggested strategy could alleviate future problems.

Applying the principles of Search Engine Optimization (SEO) might help: it is an excellent point that SDAC comes up first or second on a “solar images” Google search, but it does not immediately come up with other variations. For example, a Google search on “sun pictures” yields SDAC is #7, and “pictures of the sun,” #9. If site popularity or relevance is a metric, the results can be easily improved.

Organizational strengths and weaknesses:

The SDAC is well organized and is able to store and serve the portions of solar data it holds. SDAC has done an outstanding job of monitoring advances in hardware and selecting the most cost-effective options when making capital purchases.

The web page was noted to be unclear *vis-à-vis*:

- No statement of what SDAC is or does,
- No organization of services under menus or tabs; the page appears as if new info was simply added to the page with little thought to organization,
- No reference to SolarSoft,
- No reference to a number of the SDAC’s major projects.

The proposal was unclear as to whether the VSO work described was separable from SDAC final active archive work. The panel did not see a commitment from SDAC for providing long-term data availability for the Resident Archives. The current relationship between the Resident Archives (those not run by SDAC) and the SDAC as a Final Active Archive was unclear. If SDAC is to be the final archive for solar physics, development and planning needs to start now.

Overall assessment and findings:

Overall, the SDAC is doing an excellent job on a very small budget. Since the budget is restricted, the panel felt that SDAC should not use its limited resources to solve the SDO/Stanford bottleneck problem, which can be addressed by the SDO program shipping media to users until the bandwidth is increased, and which costs the SDAC 1 FTE in the near term. Instead, the SDAC should use the resources to accelerate the integration of solar data into the heliophysics data system and to develop new tools, such as a graphical user interface.

As SPASE is adopted for all heliophysics data as the metadata standard, the SDAC should take the leading role in developing the descriptive fields appropriate for solar data for SPASE. To become a true archive for solar data, SDAC needs to develop a plan for transitioning Resident Archives into the SDAC final archive.

HDMC evaluation.

The HDMC acts as the “glue” in binding together the separate efforts needed and described in the NASA Heliophysics Science Data Management Policy (HSDMP). Efforts include:

SPASE data model definition;
Funding and management of VxOs (except VSO);
Funding and management of Resident Archives;
Management tasks to oversee / implement HSDMP across SDAC, SPDF, the NSSDC, and the VxOs.

The HDMC will provide the central structure uniting the various VxOs and archives of the heliophysics data systems. The establishment of this system will revolutionize the field of heliospheric science by revealing previously unknown correlations between physical processes. However, this promise of a “new science” is not well supported in the proposal; at this time there appears to be no strong push from the research community for a unified heliophysics archive and or a single interface. If the promise of new discoveries is fulfilled, then the science per dollar will be high.

Science strengths and weaknesses:

The HDMC could provide the framework for a unified heliophysics data system that could provide new scientific insights.

There is no guarantee that new scientific discoveries could actually result, as there is no community outcry for the system

There are no provisions for data quality variables in SPASE Data model.

Inventory development and SPASE keywords are a vital part of increasing the VxO’s search capabilities. The panel was concerned to see that there was little enthusiasm for the SPASE data model from some archives. This needs to be addressed quickly, perhaps by more scientific involvement in the development. To be useful, SPASE development should involve researchers from all sub-fields of heliophysics.

The release of the SPASE-QL would also motivate usage by others. This development should be accelerated.

The development of a FITS2SPASE tool would also help uptake by SDAC specifically, since most of the SDAC data is in FITS format.

Relevancy strengths and weaknesses:

The HDMC has facilitated a good deal of convergence of technologies and tools across the VxOs.

The HDMC is highly relevant to NASA’s heliophysics data system. It is the framework for the system. However, the HDMC risks becoming irrelevant to NASA’s heliophysics data system, unless the SPASE data model is adopted by all data providers and archives.

The HDMC provides support to aid setting-up and maintaining existing and emerging VxOs, and indeed is key to providing the key role in the definition and development of SPASE and SPASE-QL in consultation with SPDF and SDAC.

The HDMC provides tools to facilitate the generation of SPASE description XML files from other formats (*i.e*. from CDF with CDF2SPASE). However, there is a serious gap, in that there is no

FITS2SPASE tool. The development of tools to read and manipulate format independent data will greatly facilitate science analysis from a much wider community.

Organizational strengths and weaknesses:

It is not clear from the proposal what actual authority HDMC has. In particular, Resident Archives as an activity appear to be voluntary. If a project does not propose for an RA at end of mission, there is no guarantee that any of the data will ever make it to a Final Archive or the NSSDC. Does the HDMC have oversight on RA products and ways of doing business? Quality control? Standards for files, software used?

There seems no clear strategy on how to enforce or ensure the adoption of SPASE by SDAC, SPDF, NSSDC: – after all, SPASE is *central* to all proposed additional Search and List management functions.

The panel feels there is too much duplication of activities with other data centers that should be rationalized and redefined if necessary. For example, ‘Browsing and visualization’ (Section 4 under Proposed Work) might be best left to SPDF and SDAC Many of the proposed efforts by the HDMC are also done or covered by SDAC, SPDF, and the NSSDC.

The web page is rudimentary but what is there is useful, well written, and nicely optimized for search engines. The HDMC interaction with the VxOs to date has been valuable. HDMC will need to revise their work plan after new VxOs are selected to adjust to management changes and any evolution of the VxO structure.

Overall assessment and findings:

Overall, the HDMC is doing an excellent and essential job. However, there must be uptake by SDAC, SPDF and all of the VxOs for the effort to pay off. There needs to be clarification of the overall structure of the HDS to avoid VxOs, SDAC and SPDF playing similar and overlapping roles. The HDMC could be made “responsible” for the overall structure given the role of the “glue that holds together all other pieces.”

Much of the budget is pass-through; approximately half goes to the VxOs. This direct involvement with the VxOs increases cohesiveness and effectiveness in cross-VxO activities such as SPASE.

The tasks planned for the additional $500K optimal budget are worthy and will add value to the HDMC work. The urgency, however, was not well supported in the proposal, particularly with respect to the increased user buy-in that would result.

SPDF evaluation.

The SPDF is doing an excellent job of serving in-situ heliophysics data from a wide variety of missions. The SPDF, through CDAweb, also provides some very useful analysis tools for plotting and concatenating data. The SPDF has already set up a VSPO and provide a large amount of SPASE descriptors for their own (and linked) data holdings. The current VSPO does a good job in pointing the user either to the source of the data (link) or provide a “get data” button for those data held within the SPDF.

The SPDF has shown some success in implementing its new role, that of Final Archive for non-solar image heliospheric data. It is proactive with RAs; offers mirror services to current and recent missions, and in the absence of clear guidelines, encourages missions to “be as good as the last known good solution“.

The SPDF is a strong player within the HSDMP. It should pursue more vigorously integration with the heliophysics data system via championing of the SPASE data model, and concentrate on a more limited role within the HSDMP that allows it to optimally serve the science community, and to optimally utilize its resources. It offers science-enabling services for NASA’s heliophysics community, with an emphasis on cross-discipline services. These services include: the Heliophysics Resource Gateway; CDAWeb and OMNIweb data information services; acting as the final archive for non-solar heliophysics data; and providing leadership for the development and maintenance of the CDF and format library.

Science strengths and weaknesses:

The SPDF is an extremely useful tool for heliosphere researchers to locate and plot a variety of data. The SPDF greatly enhances the productivity of scientists and leverages the scientific return of NASA’s heliospheric missions. SPDF has the buy-in of the user community, as shown by its heavy citation rate (20% of JGR SP citations), and heavy use of its services by a worldwide user community.

The SPDF is well advanced in the adoption of the SPASE metadata model and has provided access to most of its data through the SPASE descriptors on the VSPO gateway.

There is an impressive collection of web-based services for accessing and visualization data sets, although some do seem dated. The 4D orbit tool is very useful and is a good first step towards a GUI for the virtual observatories. SPDF is making a genuine effort to modernize its services and take advantage of the modern platforms (such as JAVA).

It was noted that the SPDF provide useful and historic web resources: CDAWeb, SSCWeb, and OMNIweb.

The SPDF has implemented a data quality flag in its CDF model, and has developed a plan for transitioning data to NSSDC for deep archiving.

However, the panel thought that providing a seamless metadata to SPASE translation should be the highest priority for the inventory effort. Such a tool already exists for the CDF conversion (CDF2SPASE), but conversions from other data formats (*i.e*. FITS, CEF) would simplify and reduce effort.

There is no defined timeline for transition of data, metadata and products from/to RA’s and a deep archive.

Relevancy strengths and weaknesses:

The SPDF is highly relevant to NASA’s heliophysics data system, and the data inventory it has developed will be a valuable service that can be utilized by the relevant VxOs.

CDF as become the de-facto standard for *in-situ* data thanks to SPDF’s efforts. Indeed this is the first “standard” in the field. The panel feels that the time is ripe for CDF to become a NASA mandated standard (as FITS is for astrophysical images and data).

The SPDF is well organized and is able to store and serve the data it holds.

It is not clear what is available or done with non-solar image data; – only IMAGE data can be found via VSPO. What of other such data; for example, aurorae? There seem to be no tools that allow images to be put in context with other time-series data.

The 4D data tools are good for ‘show and tell’, but to be useful for actual analysis, these tools should be connected to actual data.

The current in-house API developments seem to be “closed shop” developments not making use of open software development tools such as wikis, Sourceforge, etc., that would allow a clear path for user input and feature requests/bug reports (*e.g.* see autoplot.org environment)

Organizational strengths and weaknesses:

The SPDF has multiple roles that are not clearly defined or mandated. Should a “Final Archive” also be the prime mission archive for some missions?

The current “quality review” is limited to making sure data can be read. There is no process in place to review whether data actually makes sense beyond an informal response to issues brought up by end users.

The topic to “ensure data preservation” raised a number of concerns in the panel. Shouldn’t the job of data preservation lay with the NSSDC? It was stated that processes are underway to do this through the NSSDC (which is good), but this should be the *only* data preservation effort to avoid duplication of effort.

The SPDF has multiple roles that are not clearly defined or mandated, and overlap exists with other data centers in many activities that should be clarified: SPASE work, RAs, etc.

It is not clear to the panel that the SPDF the best place for API development. If not, should this be solicited work via proposals, with the SPDF mainly responsible for implementing and maintaining the best APIs coming out of such solicitations? There are a very large amount of parallel efforts undertaken at basically all of the major institutions hosting NASA missions.

Overall assessment and findings:

The SPDF is doing an excellent job, and has gone a long way in adopting the SPASE model and is substantially fulfilling its role within the HSDMP. SPDF should be encouraged to provide additional translation services of legacy metadata formats to SPASE.

The SPDF responds to community needs outside its core responsibility (e.g. active mission archive, data preservation efforts). A re-worked HSDMP structure should allow SPDF to concentrate on its core mission.

User interface work and API development needs to be improved and transitioned to an open-source, large community effort. Tighter integration of existing tools (*i.e.* the orbit viewer) to data is thought to be desirable.

SPDF needs to play a larger role in ensuring data quality and help to work towards SPASE standards that capture and enforce data quality standards.

Submissions for final data sets to the NSSDC from either a mission or a Resident Archive could be facilitated as a central “service” by the SPDF in its role as a final archive. The panel feels that this would be an ideal and very valuable role for the SPDF.

NSSDC evaluation.

The NSSDC serves a wide range of archival functions and customers. As a primary archive for some missions, NSSDC is the principal interface for the science community interested in accessing that data. Such heliophysics missions include, but are not limited to, AIM, ISIS-Aloutte-2, and TWINS. In this capacity, NSSDC must provide some level of knowledgeable support to its users and an interface to which VxOs link (and provide metadata for the VxOs to identify desired data within these holdings). There are some plans for NSSDC to assume the primary archive responsibility from SDAC for RHESSI at some point in the future. There are no criteria articulated in the proposal under which the NSSDC assumes primary archive responsibility for a given mission's data.

Archive backup services are also provided for some Astrophysics missions or Active Archives. In this case, data are provided on media and stored as received.

A major role of the NSSDC is to be a deep archive for other Active Archives (such as the NASA Planetary Data System, some astrophysics archives and some heliophysics missions). In this case, NSSDC accepts the data in electronic format with an 'electronic wrapper' (an AIP) that is either generated by the provider or by NSSDC, depending upon prior agreement. NSSDC provides legacy tape migration services. It manages and enhances the storage infrastructure for all of its media and analog holdings. It monitors and evaluates changing technology to ensure long-term presentation of digital data. NSSDC is *not* the deep archive for all space science data.

NSSDC is also responsible for significant analog holdings, which it curates. It expends some level of effort to migrate high-priority analog data to digital format. This prioritization is accomplished for different data at different times by individuals within NSSDC, in response to requests, and evaluated by groups organized for that purpose by NSSDC. It also "transforms" or "modernizes" selected data in older electronic formats to formats used today.

NSSDC populates and maintains a catalog of its holdings. It supports Web services for general users to identify data and to download or submit requests to transfer the data by other means.

NSSDC provides guidance to some flight projects, primarily in heliophysics, in planning their data products and formats, creating Project Data Management Plans, and assisting in the creation of some heliophysics Resident Archives. Its support for heliophysics Resident Archives includes "preservation guides" and workshops to establish a "Trusted Repositories Audit and Certification" for providing long term access to managed resources now and into the future. It is not clear that this guidance is provided to all flight projects.

NSSDC supports the development of SPASE, providing its core budget, hosting the SPASE website, supporting data model development and populating SPASE fields for appropriate NSSDC holdings. NSSDC is also involved with the development of standards regarding the implementation and management of archives as a participant in CCSDS. It also maintains web pages that locate data, fact sheets and white papers on the archive process and data widely used formats.

NSSDC assigns and disseminates unique international satellite designators for all spacecraft that attain orbit. It also generates and maintains brief descriptions and other limited information about these satellites.

Science strengths and weaknesses:

NSSDC plays a critical role in the continued enabling of NASA mission data to generate science in all areas of the NASA space science enterprise, primarily in its function as a deep archive. While some data, such as laboratory calibration spectra of well-characterized samples, are reproducible, observations of the universe and solar system are intrinsically time-variable and unique and do not lose their value. Therefore, capability for long-term maintenance of data holdings is essential, and a deep archive against loss of the active archives interfacing with the science community is required to allow for such long-term maintenance. Also, this is a well-defined task that is not suited for distributed implementation. There is expertise in areas such as storage media and migration, as well as the cumulative knowledge of data formats that may come in and out of fashion in the user and producer world. Plus, this is a common functionality in its execution across the accumulated data sets of all NASA space science disciplines.

Maintaining a deep archive is not a static process. The ingestion process, particularly when bringing in large legacy holdings, can be significant. Media is not permanent and so holdings must be regularly refreshed. As holdings continue to expand, this process increases. NSSDC does this on an 8-year cycle for magnetic media.

NSSDC also performs a critical function in curating legacy data from older missions whose data may be in analog form and may be unique. It provides ad hoc support to the user community to restore and digitize non-digital data on request as resources allow. In this curatorial function for older data sets, NSSDC maintains knowledge of no longer standard data formats necessary to convert them to modern standards.

SPASE is viewed as a positive activity, which allows for visibility of the NSSDC holdings in the context of the VSPO inventory. NSSDC plans to add SPASE metadata for all of its holdings.

The NSSDC maintains a duplicate copy of its archival holdings 15 miles away in a location called “Iron Mountain” near Laurel, Maryland. This is not considered a distributed location.

NSSDC performs functions that are duplicated by other archives (e.g., SDAC), such as being a primary archive for missions and providing user services.

The NSSDC’s mission is to undertake the long-term archiving and preservation of all space science data. Yet, it has neither a plan nor the capacity to accommodate future large data sets on petabyte scales, such as from SDO.

Overall, the NSSDC is doing an essential job. However, under the present schema, access to its holdings is difficult, often requiring manual staging or shipping.

Relevance strengths and weaknesses:

The NSSDC plays a role in data archiving that is necessary for SMD to achieve its strategic goals. As the deep archive, it holds the basic historical record of the science, which is necessary for its advancement.

The NSSDC has a long track record of executing its baseline activities of deep archiving and other support and is able to accurately scope the level of effort required to undertake its necessary tasks.

In its requested/enhanced budget, NSSDC proposes to convert high-value microfiche and microfilm copies of Solar/Space Physics archived data to electronic form. It is not clear who prioritized these data. Was it a Data Review Group, or a couple of individuals? If it was a couple of people, the prioritization is not necessarily adequately informed. The Data Review Group was exclusively GFSC, and the determination of what conversions should be priority would benefit from involving a broader segment of the Heliophysics community outside Goddard.

NSSDC's backup function to several mission archives does give value, however it begs the question of why this is not consistent across missions and what other missions or active archives are doing for backup. It is not explained why this is optimally executed as an NSSDC task or why it is cost-effective over other alternatives.

NSSDC does not plan for the possibility of being the deep archive for planned heliophysics missions that could expand its holdings by more than an order of magnitude in storage capacity.

Overall assessment and findings:

The panel finds that the NSSDC is doing excellent work in a number of very diverse and varied science fields, as well as preserving and keeping current with the evolving science of data management, curation and archiving. Having said this, the panel feels that the NSSDC could profitably limit its activities to become more effective in its primary role as the NASA space science deep archive.

All of the NSSDC's holdings of analog data have value and should be digitized, and a plan should be developed to achieve this.

NSSDC plays a vital role in the heliophysics community as a deep archive from which data can be retrieved after a disaster or the loss of access to an active archive. Transition to current media and some file format updates are incredibly important tasks that will ensure that the data is readable and usable by future generations not currently involved with the missions in their active phase. These are vital core competencies of the NSSDC that should be maintained and supported.

NSSDC's main role should be to provide a deep archive for all NASA space science missions.

The diversity of tasks that NSSDC has undertaken, including being a primary archive for some missions, arises from lack of systematic coverage of data archiving and distribution responsibilities within NASA Heliophysics for its missions. This diversity distracts from NSSDC's core mission role of deep archive.

The substantial interaction of NSSDC and the Resident Archives overlaps that of the Final Active Archives and the Resident Archives, though not consistently. Functional redundancies across Heliophysics data archive and distribution process need to be identified and removed, perhaps through a reorganization of tasks. In such a reorganization, NSSDC as the deep archive should only interact with the entity having the active archive role (perhaps SDAC or SPDF) to whom researchers go to identify and collect mission data (and NSSDC should not be the active archive).

In the context of a restructured heliophysics data archive system, NSSDC user community interface and E/PO activities should be transferred to the entity(s) that is (are) the standing active archive(s).

A more distant location should be found than Laurel, Maryland, for the second instance of NSSDC holdings against a disaster scenario, preferably in a well-separated part of the country. The option of the San Diego Supercomputer as the backup location should be further investigated.

NSSDC is the logical center for accumulating knowledge about old data formats no longer in use (*e.g.*, EBCDIC, VAX binary) and it should continue to provide the tools and applications for data conversion (*e.g.*, to ASCII, IEEE 754]).

NSSDC should be the default deep archive for all missions across all SMD space science divisions. There are some missions, such as SDO, that currently have no deep archive plans, which is a major concern. This would be a major unplanned expansion of NSSDC holdings. NSSDC needs to create a plan and budget for the option of being the deep archive for all space science mission data.

General Comments to NASA on the Heliophysics Archives.

The Program Executive for Heliophysics MO & DA solicited the Senior Review panel for general findings of the overall scope and organization of the Heliophysics data environment. These findings a given below.

The panel finds that the organization of the heliophysics data system is needlessly complex and confusing (see Figure 1). The roles and responsibilities of the mission archives, resident archives, final archives, and the deep archive are inadequately defined and not clearly separated. This can mean that data are not guaranteed to arrive at the final and/or deep archive in usable form and has led archives to duplicate effort and to fail to work together. The net effect is to reduce the science value of the archives to NASA (see Figure 1 below).

The panel's remarks are not intended as criticisms of the current archives. We recognize that there are historical reasons for the current organization, that the archives we have reviewed perform well and that they are widely used and cited by their user communities.

The panel's professional opinion is that there is an urgent need for a detailed follow-on Heliophysics Science Data Management Policy, having the following components:

- Plans for data management and archiving need to be proposed at the AO phase of the mission. New missions should be required to develop and maintain a data management and archive plan for implementation during the active mission phase (prime and extended). The appropriate final active archive should be involved in the development of these plans, which should be documented in a Project Data Management Plan (PDMP). Compliance with the plan should be determined by a review panel.
- Resident archives should remain "live" for an appropriate period after the end of the mission, say 3 years. NASA should consider requiring all resident archives to deposit usable annual data in the final discipline archive, to be assessed by external peer review, before annual renewal.
- Before the resident archive closes, the data should transfer to a discipline archive. The resident and final archives should be jointly responsible for the transition of the data. The transition should be as transparent as possible to end-users, who should expect no interruption of service.
- Upon receipt of the dataset a the discipline archive, the discipline archive is responsible for delivering it to the deep archive at NSSDC in usable and appropriate archival form, as determined by consultation between NSSDC and the appropriate final active archive, with only the necessary consultation by the mission PIs. This delivery should be performed by the final archive.

There are some outstanding issues that must be addressed in the areas of data standards and accessibility:

- The data quality must be described. The data quality measures will be dependent on the mission of course, but generally these will include statistical measures of uncertainty in data ("error bars"), and data fidelity measures such as quality flags. All these measures must be fully documented. The mission should be responsible for determining quality measures and should be subject to external peer review.

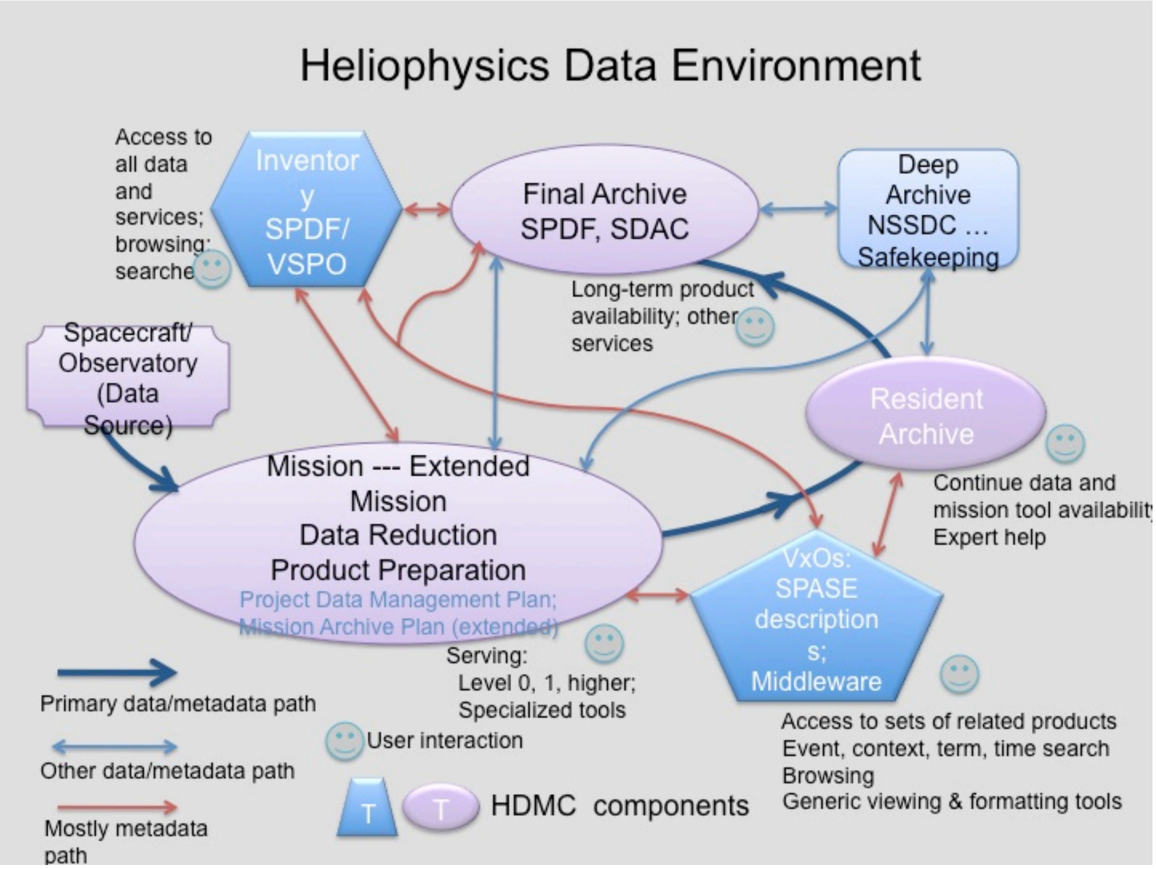
- There needs to be a policy of what data should eventually be in the final archive. Data volumes may preclude long-term curation of all levels of the data, from raw to most highly processed, and decisions need to be based on maximizing the long term science value of the data. Where all levels of data cannot be archived, the mission should be responsible for recording the provenance of the products to enable re-derivation of data products.
- Data products must include not just the science data sets, but product descriptions and metadata, technical descriptions of processing pipelines, and as needed, the pipeline code itself.
- The archives must enable interoperability between data sets, and this will require the development and buy-in of a data model. SPASE represents the first step in this direction, but must have widespread scientific input to ensure adoption. The release of SPASE-QL is a second essential item for buy-in. th panel believe that both these items should be a priority.
- The implementation of the VxO initiatives as short term "proof-of-concepts" is a good one. Selected VxOs should be expected to interoperate with other data sets via standard data model and contribute to the further development of the data model. Future options remain wide open: VxO's could be recompeted after 3 years, HDMC could assume responsibility for continuation of their services, by mutual agreement, or the final data archives could assume the VxO interface responsibilities. However, there needs to be a well defined path forward.

The panel is concerned that the two-party heliophysics discipline model; SPDF and SDAC, is not an optimal solution. The panel urges consideration of a single final archive or a set of discipline archives that would together form a final archive.

Further, HDMC appears to not have buy-in of the SDAC. Both SDAC and SPDF are developing inventories with only SPDF having gone forward in adopting SPASE, and we did not see a commitment to cooperate from SDAC. The panel suggests the following:

1. Expand SDAC to be become a proper final archive.
2. Consolidate SPASE work at HDMC.
3. NSSDC should act as a Deep Archive only, with discipline archives (SPDF/SDAC at this time) as their prime customers.
4. Thoroughly review other roles and responsibilities of the current SDAC, SPDF, HDMC, and NSSDC groups at GSFC. These groups appear to be overcommitted and possibly overwhelmed by the unconstrained expectations placed on them. If extraneous responsibilities can be removed – across the board, we do not intend to single any group(s) out – then each group can focus on the role in which they excel, and duplication of effort is minimized.
5. Streamline, communication between the current SDAC, HDMC, SPDF and NSSDC groups at GSFC. For instance, missions should communicate regularly with the final archives, but NSSDC requests should largely be dealt with by the final archives, not the mission archives. VxOs should interface with HDMC and the other archives in support of the SPASE definitions and inventory. Communication and reporting lines must be more clearly drawn (see Figure 2 as a possible example).
6. Develop more powerful and unified user interfaces. Current interfaces seem dated and limited. They do not include interactivity with the data, such as including images, time series graphs etc. The SPDF 4D orbit tool, for example, would be made much more powerful if it was integrated with and provided access to data sets. A uniform "look and feel" interface would be highly desirable.
7. To aid (6), move development of services and interface to an open collaborative model that encourages user input and allows open-source orientated development. This would include making source JAVA code available via Sourceforge (or similar systems that

allow multiple user code tracking, feature requests, bug reports etc) and maintaining public discussion wikis on these services



Heliophysics Data Environment
Access to all data and services; browsing; searche
Inventor y SPDF/ VSPO
Final Archive SPDF, SDAC
Deep Archive NSSDC … Safekeeping
Long-term product availability; other services
Spacecraft/ Observatory (Data Source)
Resident Archive
Continue data and mission tool availabilit
Expert help
Mission --- Extended Mission
Data Reduction
Product Preparation
Project Data Management Plan; Mission Archive Plan (extended)
VxOs: SPASE description s; Middleware
Serving:
Level 0, 1, higher;
Specialized tools
Primary data/metadata path
Other data/metadata path
Mostly metadata path
User interaction
T
T
HDMC components
Access to sets of related products
Event, context, term, time search
Browsing
Generic viewing & formatting tools

Figure 1: The current “simplified” model of the Heliophysics data environment. Note the overlapping lines of communication, and the lack of lines of responsibility/authority.

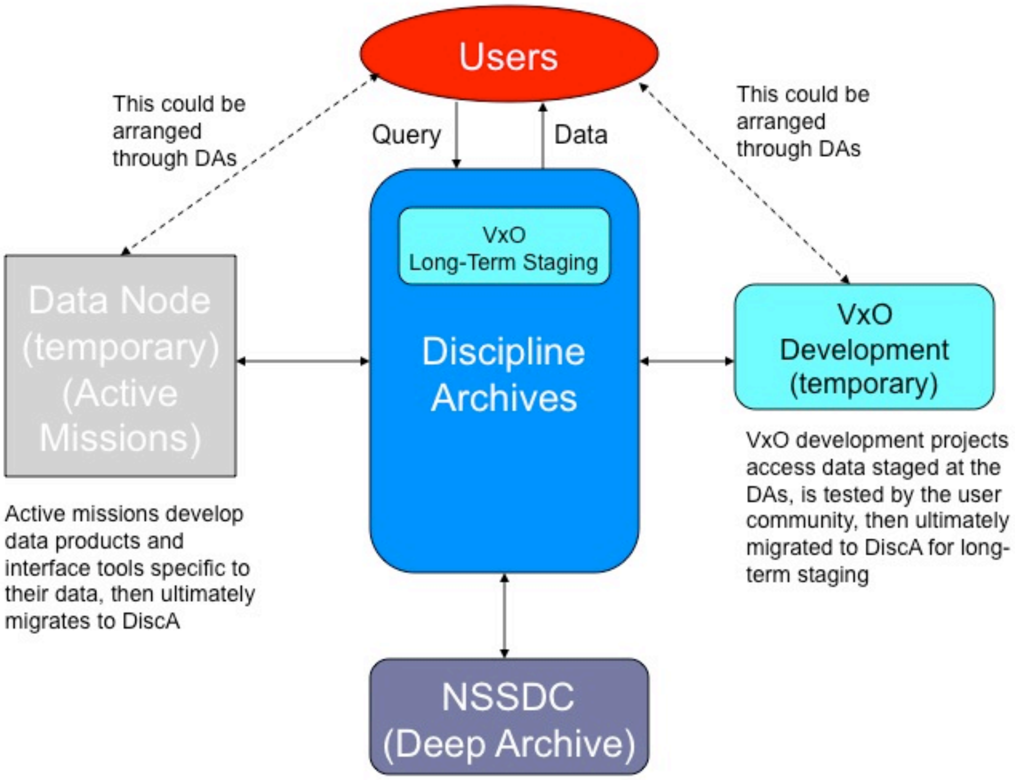

Users
This could be arranged through DAs
Query
Data
This could be arranged through DAs
VxO Long-Term Staging
Discipline Archives
Data Node (temporary) (Active Missions)
VxO Development (temporary)
VxO development projects access data staged at the DAs, is tested by the user community, then ultimately migrated to DiscA for long-term staging
Active missions develop data products and interface tools specific to their data, then ultimately migrates to DiscA
NSSDC (Deep Archive)

Figure 2: A possible model of the Heliophysics data environment, which shows discipline archives, the relationship to the mission archives, Resident Archives and the VxO. Note that in the model the NSSDC acts only as a deep archive and not an active one.