Report on a Workshop for Heliophysics Infrastructure

Virtual NASA Workshop
May 17-19th, 2021

Draft: Sep 6, 2022

Contact: brian.a.thomas@nasa.gov

Executive Summary

We report observations and findings from a three-day virtual workshop held May 17-19, 2021 which examined the current heliophysics research infrastructure to determine which elements were most utilized, what gaps exist in these elements between current utility and desired capability and, from a user standpoint, what a future state for the infrastructure might look like. Approximately 40 subject matter experts (SMEs) with backgrounds in heliophysics research, computer science and research infrastructure were gathered to consider this topic.

The workshop identified existing key research capabilities which included capabilities to browse, search and deliver scientific data products. Importantly, not all needed capabilities were resident at NASA, and at least some in the community use capabilities provided by other institutions including commercial entities like Twitter and Google.

Gaps and concerns which were identified indicated that the research process still requires significant work and can be improved. In some cases helpful capabilities were not known of but in other cases improvements to existing capabilities are needed. Interestingly gaps indicate new needed capabilities in areas previously unaddressed by the infrastructure which include processing / compute for big data, addressing gaps in forming research teams and otherwise facilitating collaboration, and enabling Open Science.

Imagined new capabilities included shared software environment for doing research, a shared analytics platform in the cloud which has access to PB of science data, novel means of searching such as by phenomena or researcher, ability to browse collections of interlinked research artifacts (data, code, models, publications and so forth) and tools to help users publish these research artifacts in a manner which is consistent with Open Science needs for reproducibility and transparency.

1. Introduction: Rapidly Advancing Infrastructure Needs and an Opportunity for Effective Change

Heliophysics, the science of the physical connections between the Sun and the Solar System, has been changing rapidly over the past few decades. There are significantly more data and need to understand the interrelationships between data than previously appreciated. These drivers flow from several things.

First, and simply, data volume and type has grown dramatically. Recent and past missions (such as SDO, MMS, THEMIS, ACE, SOHO, Wind, Polar) have so far contributed more than 20 Pb of data over the past few decades. These data comprise an admixture of data products -- there are both data products provided by the missions as well as the higher level data products produced by NASA-funded researchers. As such both the data volume and complexity of archived heliophysics data have grown.

Mission and mission-derived data are not the only data which have evolved. Other related capabilities and data products are now widely utilized in heliophysics research. Models, which provide understanding of the underlying physics across the entire analysis domain, are critical for heliophysics research and have long been important. What has changed is that over the past decade NASA has led the charge to make both the models and model outputs more available and immediately usable by the community (CCMC, ref).

Change is not isolated to just data products. There has also been a conceptual shift in the field to recognize that heliophysics is a ‘system science’ (McGranaghan, R., Borovsky, J. E., and Denton, M., 2018) -- the best way to understand the complex physics at work researchers must consider the system components, from the solar interior, through to the solar surface, and onwards to the corona, the solar wind, interplanetary conditions, planetary magnetic environments and ultimately the upper atmosphere of the planets. Complexity refers to the fact that these components do not combine in a linear manner and that the system is more than the sum of its parts. Systems research requires a greater variety of measurements to advance understanding of heliophysics phenomena. As a result the coordination and integration of data products have become inextricable from the progress of Heliophysics knowledge.

Summing up from the preceding statements, it is clear that there has been significant change in how the community conducts heliophysics research. There are more data, more variety of data products and new, needed capabilities to better integrate and understand the data. Taken together these facts indicate a need to reevaluate our ‘research infrastructure’ – the capabilities which support heliophysics researchers ranging from data archives, scientific source code, models and binaries (such as libraries), computing/processing resources, and the services and/or APIs (Application

Programming Interface) that allow the community to access these resources. While it is useful for NASA to know where we are providing good value to our users, it is critical that we understand the gaps and areas where improvement must be made to better serve the research community in heliophysics.

We have therefore opted to 'crowd source' the problem by engaging a select number of heliophysics researchers and related infrastructure SME. As will be seen in this report, we hope to create a common understanding of the current infrastructure, a common agreement on the important gaps which currently exist and capture suggestions on how the community thinks we may potentially solve these issues to better enable heliophysics research for all.

2. Workshop Organization

The objectives for the workshop were three-fold and we tackled each on a separate day of the workshop. Our goals were the following:

- *Identify key infrastructure capabilities for research scientists in heliophysics. [day 1]*

- *Identify gaps in these key infrastructure capabilities. [day 2]*

- *From a user perspective, capture thoughts on hypothetical future research capabilities. [day 3]*

Because of the COVID19 pandemic the workshop was held as a virtual meeting over three days: May 17-19, 2021. We selected participants who either were experts in heliophysics research, computer science which supports this research or in non-NASA heliophysics infrastructure.

In order to ensure effective participation by all attendees, we kept the workshop participation small (~40 attendees, see Appendix A) and limited the meeting to partial days (~4 hrs per day). Each day comprised 3 segments which included a shared briefing with all participants to frame the goals for that day. This was followed by simultaneous breakout sessions in groups lead by designated chairs. The last daily segment was to reconvene all participants to discuss and digest results from breakout sessions. Our five breakout session chairs were Monica Bobra (Stanford University), Neal Hurlburt (LMSAL/Lockheed Martin Solar and Astrophysics Laboratory), Alexa Halford (Goddard Space Flight Center/NASA), Larisa Goncharenko (MIT/Haystack) and

Will Barnes (NRL/Naval Research Laboratory). Members of the organizing committee also served as moderators to ensure that all voices were heard in breakout sessions.

Each breakout group discussed the daily goal and recorded information from their discussion. The breakout chair was responsible for leading and synthesizing their group discussion and presented their findings at a plenary session which followed the breakout session. In the plenary session, after reporting out their individual group synthesis, the breakout chairs then responded to each other's findings and observations, with particular attention given to overlapping and contrasting this information. All participants were allowed to then raise questions and provide feedback commentary at the end of the plenary session. In all segments of the meeting, we used collaborative shared note taking by using online tools which allowed all participants to simultaneously provide edits and content. Both Google Docs (https://docs.google.com) and Miro (https://miro.com) were utilized for this purpose.

The workshop agenda appears in Appendix B.

3. Workshop Results

As will be seen below, there were a large number of findings and in order to improve their presentation here we have taken the liberty to combine similar findings from the workshop together. Findings are grouped by goal and then similar information grouped together. This information is presented in no particular order of importance.

3.1. Key Research Capabilities

The workshop conducted a census of capabilities needed for heliophysics research. These capabilities included both technical and non-technical items.

3.1.1. Online Discovery Resources

Participants indicated many online resources which were valuable for research. Table 1 summarizes the resources identified in the workshop. While there is little doubt that the resources in table 1 are not exhaustive, they are an interesting sampling of what the participants felt were useful discovery resources. There were many NASA-supported resources identified but this is unsurprising considering this was a NASA workshop drawing upon its user community for participation. What is more interesting is that participants also found other governmental resources valuable as well as commercial assets (e.g. Google) including the use of social media (Twitter) by some. The type of information being discovered included data products and the research (literature).

As a final comment on this section, it was noted that an important aspect of these resources is that they are generally ‘platform independent’.

Discovery Resource Name	Discovery Type	Supported by	Notes
Heliophysics Data Portal	Data Products	NASA	HDP; https://heliophysicsdata.gsfc.nasa.gov/
Virtual Solar Observatory	Data Products	NASA	*VSO; https://virtualsolar.org*
Heliophysics Events Knowledgebase	Data Products	NASA	HEK; https://www.lmsal.com/hek/
CDAWeb	Data Products	NASA	https://cdaweb.gsfc.nasa.gov/WebServices/
SPEDAS	Data Products	NASA	*Space Physics Environment Data Analysis System*; https://link.springer.com/article/10.1007/s11214-018-0576-4
CEDARWeb	Data Products	NSF	*Coupling, Energetic, and Dynamics of Atmospheric Regions*; http://cedarweb.vsp.ucar.edu/wiki/index.php/Main_Page
Planetary Data System	Data Products	NASA	PDS; https://pds.nasa.gov
Astrophysical Data System	Literature	NASA	ADS; https://ui.adsabs.harvard.edu/
Google search	Everything?	Commercial	https://google.com
Google Scholar	Literature	Commercial	https://scholar.google.com
Twitter	Everything?	Commercial	https://twitter.com

Table 1. *Sample heliophysics discovery resources identified.*

3.1.2. Online Data Product Access

Participants indicated many online resources contain vital information for heliophysics research. Again, there is every expectation that if we spent more time considering resources we would turn up a much larger list and we also suffer from bias of participants of the workshop being a largely NASA-based community (for example resources in Asia are absent from the list).

Nevertheless we can draw some interesting conclusions. First there are several different means of accessing these data which include working with a website, scripting access using an API, using a prepared software client to access an API. Also of interest is that NOAA, NRL and ESA supported resources appear on this list but don't appear in Table 1. This is an indicator of the incompleteness of these lists as well as a potential indication of greater needed coordination between data discovery resources and data

access resources. Finally, we note that some of these APIs supply more than data retrieval functionality. There exist some APIs which supply additional on-the-fly data analysis and/or processing (SPDF SSCweb, https://sscweb.gsfc.nasa.gov/WebServices/ for example).

Access Capability Name	Access Type(s)	Supported By	Notes
SDAC	Website, API	NASA	*Solar Data Analysis Center;* https://umbra.nascom.nasa.gov. API access via VSO.
SPDF	Website, API	NASA	*Space Physics Data Facility*; https://spdf.gsfc.nasa.gov
CCMC	Website	NASA	*Community Coordinated Modeling Center;* https://ccmc.gsfc.nasa.gov/
JSOC	Website, API	NASA	*Joint Science Operations Center,* http://jsoc.stanford.edu/
Madrigal	Website	NSF?	*Madrical Database website,* http://millstonehill.haystack.mit.edu/register
AMPERE	Website, API?	NSF	*Active Magnetosphere and Planetary Electrodynamics Response Experiment;* https://ampere.jhuapl.edu/
DMSP	Website, API?	NOAA	*Defense Meteorological Satellite Program,* https://www.ngdc.noaa.gov/stp/satellite/dmsp/
GOES	Website, API	NOAA	*https://www.ncei.noaa.gov/products*
ESAC	Website, API	ESA	*European Space Astronomy Center; https://www.cosmos.esa.int/web/esdc. API access via VSO.*
Code Repositories	API	Commercial, NASA, Others	Git-based repositories either hosted by a servie such as GitHub (https://github.com), or GitLab (https://gitlab.com) as well as older repositories such as SVN, CVS. In some cases repositories may be hosted on a non-commercial platform.
Distributed Client	Software Client	NASA, ESA, NRL	SolarSoft (https://www.lmsal.com/solarsoft); sunpy fido https://sunpy.org; pySat https://github.com/pysat/pysat; VirES, https://earth.esa.int/eogateway/tools/vires-for-swarm
Distributed Client-Server	Software Client and Server	NASA	HAPI https://github.com/hapi-server

Table 2. *Sample data product access resources identified.*

3.1.3. Scientific Software & Code development

Many code libraries were discussed and considered important (ex PyHC-supported projects, *https://heliopython.org*; the CDF library, https://cdf.gsfc.nasa.gov/, and the aforementioned SolarSoft library). Python was the most popular software language mentioned for developing scientific software although IDL is still used by many and there was mention of software to call functions/software written in one language from other languages (pybind11, f2py, etc).

Sites which answer questions such as StackExchange (https://stackexchange.com; an online forum for the worldwide community of developers, including Stack Overflow) or code package tutorials (e.x. https://whpi.hao.ucar.edu/whpi_showandtelllibrary.php) were valuable for supporting scientific research software development. Furthermore, many participants mentioned best practices in software development being useful, in particular the practices of unit testing, capturing versions of 3rd party software used, use of virtual programming environments (conda, https://docs.conda.io/en/latest/; venv, https://docs.python.org/3/library/venv.html), and utilizing containers (Docker, https://www.docker.com/).

3.1.4. Data Standards

Shared data file formats and metadata standards were discussed and considered important for enabling research. Items mentioned included CDF[1], ISTP Metadata Guidelines[2], SPASE[3], DOIs[4] (used for citation) and FITS[5].

3.1.5. Collaboration Capabilities

Collaboration capabilities were discussed in detail. Email remains important and some participants mentioned the use of notebooks (e.g. Jupyter notebooks[6]) for sharing results. Discussion boards and team collaboration/chat software (for example Slack[7]; MS Teams[8], Zoom[9], Discord[10], GatherTown[11], Linkedin[12], GitHub[13] lobbies) are *some* of new capabilities which were mentioned as important for science collaboration. Virtual meetings were also deemed to be important, particularly with regard to allowing greater participation in meetings, seminars (because attendees are not constrained by the need to travel or commit significantly more time beyond the interval in which the meeting may occur).

---

[1] Common Data Format, https://cdf.gsfc.nasa.gov

[2] https://spdf.gsfc.nasa.gov/sp_use_of_cdf.html

[3] Space Physics Archive Search and Extract, https://spase-group.org/

[4] Digital Object Identifiers, https://www.doi.org

[5] Flexible Image Transport System, https://fits.gsfc.nasa.gov/

[6] https://jupyter.org

[7] https://slack.com

[8] https://www.microsoft.com/en-us/microsoft-teams/group-chat-software

[9] https://zoom.us

[10] https://discord.com

[11] https://gather.town

[12] https://linkedin.com

[13] https://github.com

3.1.6. Outward Communication / Engagement

A variety of items were mentioned which were non-technical in nature. These may be summed up as things which involve extended interaction between community members. Activities such as mentoring, professional networking and in-person meetings (not necessarily formal gatherings, could be 'hallway encounters') were deemed to be particularly important. These encounters could be about science research but also important were meetings about infrastructure capabilities (ex. 'ask-me-anything sessions with model/data developers'; quick question-and-answers to learn about the model/data/infrastructure). Podcasts and other online media (ex youtube[14]) which capture meetings or presentations on relevant topics were valuable too.

3.1.7. Publications

Both refereed and non refereed publications were considered important and there were many participants who called out open access journals being valuable. Publications of scientific papers but also information on methodologies, codes, catalogs, etc. contained within publications and supplementary information were considered important.

3.2. Gaps & Concerns

A discussion among participants reached general agreement about the following identified gaps. Some of the original gaps have been edited to group, merge and combine issues in order to improve readability. The order of presentation of these gaps does not indicate priority.

3.2.1. Science Data Products and Data Management

There were several gaps expressed which related to output science data products (images, spectra, catalogs, models, etc). The participants felt that access to current offerings were 'working' but could be improved.

- Better interlinking. Many voiced a desire for more/better interlinking of data products. Interlinking between domains, other types of data products (ex. scientific publications using data products, mission/instrument documentation). This interlinking would have additional value in support of Open Science (*ref*) as it would help make more transparent how the science result was achieved, particularly with regard to the related data products.

There has been a recognition that NASA needs to provide a more integrated approach towards science data delivery. HSO Connect is a NASA program to support integrating science activities for a virtual environment called the "Heliophysics System

[14] https://youtube.com

Observatory." The HSO consists of all operating missions that provide data relevant to heliophysics, plus ground-based observing, modeling, and analysis activities.

The goal of the HSO Connect program is to enhance the scientific return of the HSO by supporting investigations that innovatively connect observations from one or more HSO missions with spacecraft or ground-based observations from other SMD Divisions, and/or other agencies within or outside the U.S.

- High level data products. There was a strong desire to have 'analysis-ready' data products available. These would be well-tested, easily understood for use in science, reliable (won't disappear from the source of the data) and could be easily integrated into the users research (e.g. some pre-processing from raw state had been performed and in a form which the user toolset can readily read it[15]). Some expressed the opinion that NASA should change expectations on what is archived (in many cases lower level data products that require additional processing) to also archive non-trivially derived data products (such as require High End Computing or otherwise hard to replicate pipelines). This is also in alignment with a desire for 'machine learning ready' data products although the distinction between 'analysis read' and 'machine learning ready' is not easily distinguishable.

- Weak data management practices. There were many aspects mentioned which included:
    - Often no requirement for archiving and public distribution of higher level data products.
    - Can lack a plan to address the long term viability and utility of these data sets ("what happens when the PI dies?").
    - Some data are being embargoed even if created with public funds. Data access restrictions exist, sometimes data are 'open' but in practical terms cannot be retrieved because of physical limitations (poor network connectivity for large datasets or simply 'big data' which only a sliver can be retrieved).
    - Metadata are irregularly implemented in datasets and may be missing or done in a manner which is inconsistent with typical use.

[15] There probably are additional requirements on 'easily integrated' and the level of pre-processing desired will differ among users, but it is safe to say that something beyond level 0 (and probably beyond 'level 1') is required for most datasets.

    - Lack of community best practices for metadata, such as a need for versioning of all data products (code/models/data) which would improve knowledge of what is the 'latest' data product are not uniformly followed.
    - No strong requirements on NASA funded research data management plans. There is wide variation and inconsistencies which limit data utility for others.

- Insufficient data standards across missions. There is a gap in continuity for science data products which makes the analysis more difficult when utilizing data from multiple missions / sources. This gap flows from a lack of shared definitions in heliophysics for the data product levels and some metadata.

- Associate data products with observed phenomena. Hard to search for relevant observations by phenomena. Related to this issue is the problem that common definition/semantics of phenomena is lacking. To further illustrate the lack of a definitive community source of semantics for phenomena participants bemoaned that in many cases Google would provide Wikipedia definition over one available from a federal government website.
- Preservation and obsolescence of high value data products and software. Concern was voiced that there is a lack of ability to preserve old datasets (pre-2000s) in hands of PIs who are retiring and we are losing knowledge/access to the code. Maintenance of important software was also a concern. It is not clear when data and software are 'out of date'.

- Support structure for 'data' is missing/not funded on missions. NASA incentives are forcing bad-behavior: for cost-capped missions the data management plan usually gets cut/short-changed.

3.2.2. Social/Cultural Issues

- Lack of attribution for professionals which support science. We regularly fail to give people credit for creating the datasets and code used in research. There are many folks engaged in these activities such as satellite operations, members of the instrument team, data scientists, data processing specialists and modelers. Currency of credit is an issue when it is done (for example, as done in astrophysics in *beginning* instrument data analysis papers, but does not help if the instrument team changes). In either case, we lose valuable talent because they find it difficult to advance in their career without this needed professional credit.

- Mentoring / supporting young researchers - The participants spent some time considering how to better mentor new researchers into the field. Important questions for any new / young researchers which are difficult and may be potentially addressed by our research infrastructure include: “How do I build my network?”, “How do I establish my reputation/research?”, “How do I become part of a team?”, “Where do I find folks to work with in research?”, “How is the best way to do X?”. Solving these interrelated issues will, in many cases, take cultural as well as technical change. Nevertheless, we should ask ourselves which parts of the research infrastructure could be developed to help promote / facilitate change here.

3.2.3. Collaborative Analytics and Research

- Lack of support for ‘modern’ collaboration technologies in collaborative setting. Common capabilities such as notebooks, code repositories, and so forth means we are moving to a different scientific workflow. Modern scientific workflows exist, but are not adopted widely sometimes because of technical barriers, other times because of cultural ones.

- Create an ecosystem of reproducibility and openness. Open science offers a framework for this cultural shift. Activities should target the development of curricula for data science and best software and data practices, and for the living practice and refinement of the curricula through communities of practice.

- Open Science not easily accomplished. Easy reproducibility by others takes considerable time from researchers (preparing notebooks, versioning and cleaning up software, versions of data and correct citations/interlinking to these supporting documents/data products). The research infrastructure is missing critical support in this area.

3.2.4. Research Artifact Access Issues

Research artifacts are any object created or used in the research process and are frequently digital in nature these days. These include digital assets such as mission data at all levels, user generated datasets, software / code, executable analysis artifacts such as binaries, containers and notebooks, drafts and final versions of research documents (howto guides, research papers, scientific articles, etc), email, bulletin board conversations, and so forth.

- Barriers in scientific publishing. Lack of infrastructure to promote ‘open science’ publishing.

    - Better support for the ability to publish in journals and evolving requirements by journals (e.x. be able to provide data, code to reproduce plots in article)
    - Data/model availability & citation (DOIs).
    - Lack of traceability/transparency/reproducibility of research. Lack of clarity of how outputs of research were created/obtained.

- Big Data Science. Science which requires analysis of large data volumes (~>100 TB) are restricted to only a small subset of the community which have access to the original data repositories.

- Barriers for External and International Collaboration / Sharing. Sharing data, code, and knowledge between institutions and international collaborators is regularly met with prohibitive regulatory environments, even when performing NASA business. It is difficult to get foreign nationals onto many NASA (on premise) computing environments, for example.

3.2.5. Resource Discovery and Information Dissemination

- Documentation of Infrastructure Capabilities. During the meeting a variety of gaps were called out for missing capabilities in the infrastructure; more in depth discussion indicated that in many cases the capability existed already but was not (generally) known. This points to a gap in advertising and/or making it easier to discover the existing capabilities by the community.

- Legacy Knowledge Capture. When a researcher retires, much critical information is currently lost.

- Difficult to locate a community of interest. Where are the ('best') chat/discussion boards, code repositories and other community shared knowledge? This is also related to the issue of mentoring for young / new investigators.

- Improve discovery and browsing of data product offerings. Finding proposal data products is not optimal and can be difficult (for non-mission data?). There is a lack of a centralized database of resources/knowledge of "where to go" (what do you do if you don't know what to 'google'). A fully-populated Heliophysics Data Portal (HDP) registry of world-wide SPASE data and model descriptions will greatly help with this.

- Improved documentation. What data and/or problems are there in the community? Who do I talk to about this data? Who has used it? Create a knowledge base of ‘solved problems’.

3.3. Envisioned Future State

On the last day of the workshop we asked participants to consider the gaps and imagine what fixing the gaps might look like to an end user. The following is a merged summary of their contributions.

3.3.1. Shared Analytics Platform / Environment

Heliophysics would benefit from a shared analytics platform which may be accessed via the internet. In the future all heliophysics researchers will have the option to utilize this capability. Critical characteristics of this platform include:

- Easy to utilize, no specialized software or hardware beyond a browser and an internet-enabled computer.

- Capabilities which enable scientific analysis across all areas of Heliophysics; these capabilities should be equal to, or ideally, greater than most investigators have access to today. Increased capabilities include ability to work with very large volumes of scientific data (>100 Tb) and access to High End Computing (GPUs, multiple threaded processes >~ 20, >~ 100 Gb RAM, >~ 20 Tb of storage scratch space).

- Supports remote team collaboration well. Teams may easily share member results, add members with little overhead/paperwork and team members, who may reside anywhere in the world, may have access to the same capabilities regardless of location.

- Support for open science -- the publication of results, data, notebooks, code and other research artifacts is enabled and made easier. Other researchers may obtain and easily replicate results with the artifacts using the platform.

- Not specialized for a particular sub-domain of heliophysics (for example, “ITM” or “solar”), equally useful for science research across all heliophysics sub-domains.

A future state would have heliophysics researchers worldwide having access to these common heliophysics software environments for doing research. Online platforms would host these for use by researchers via a browser or other common internet enabled software. The online platforms would also have access to large volumes of scientific data. Alternatively, the environment could also be utilized locally within a department or on a researcher's machine.

3.3.2. Enhanced Discovery

There was strong agreement that we need to streamline the process of finding information for research. This information is not limited to mission data, but also includes finding other research artifacts up to and including the research capabilities of the infrastructure, other researchers, important problems, digitized conversations and so forth.

An envisioned future state is one in which the researcher has powerful, yet simple to use interface for the discovery of interesting (to them) research artifacts. The exact interface was not spelled out, but it was agreed that it work beyond simple keywords associated with the paper and dataset. The search should turn up ‘unexpected but useful’ materials which may not be explicitly labeled with the keyword or phrase initiating the search (examples include searching by parent class of phenomena, discovering information based on researcher field of study, meetings attended or relationship to other researchers, and/or a suggested subgraph of terms and desired relationships which must be matched) -- in other words the types of searches which are enabled by technologies such as a graph database.

It is also important for the discovery of the information that it be presented in a fashion that makes it understandable for the searching person. The interface needs to do a good job at synthesizing and presenting the results so as to not overwhelm the human with less relevant detail.

A desired future state would have discovery service(s) which overlay many different sources of information in heliophysics (spanning all sub-domains) and provide succinct, but useful results. Each of these discovery services has an API which allows it to be called by other machines and thereby facilitate cross-communication and indexing of the results between them. Results from the discovery in any one of these services may redirect the user to another service for a deeper search of a particular facet.

3.3.3. OpenScience and Interlinked Data Products

The need for interlinked data products appeared in numerous conversations, particularly with reference to the needs of Open Science. The following points summarize the discussion from a researcher perspective:

- Researchers need a means to more readily keep track of useful things developed along the way during research.

- Researchers need better publication support : infrastructure is needed to aid in interlinking of data products created by missions and researchers with other published datasets. We should have citable code, data products, models, papers all interlinked.

- Researchers need better support for creating the documentation to enhance reproducibility of research.

This challenge cannot be met by technical change alone. Changes to policy, culture and funding will also be important. Examples of each of these include:

Policy: we should have a journal requirement for referencing datasets to help drive this change.

Culture: Communities of practice (COPs) are also needed to help with the effort to disseminate best practices borrowed from software development and data science. These COPs should also promote diversity, equity, inclusion, and access (DEIA), that crystallize, interact with, and evolve knowledge.

Funding: Open Science, even with good tooling, will not come ‘for free’. There should be funding to support additional work needed by the publishing research team to document and create requested artifacts.

For a future state then, the workshop expects to see a cultural change which affects the way in which research in heliophysics is done. Open Science requirements for transparency and reproducibility become a natural part of the research ‘workflow’ and are aided and abetted by commensurate changes in policy, funding, technical support and community practices as mentioned above. Researchers will come to view these changes as both natural and needed -- a shared responsibility for the heliophysics community but also a means to make their own research more relevant and cited by others.

3.3.4. Fewer Barriers to Using Data

It takes significant effort for a researcher to ‘come up to speed’ with using an unfamiliar data product. Data products which are ‘well processed’ and ‘ready’ for the science project are needed and it is important that the infrastructure be able to answer many of the questions which currently require asking the mission/instrument pi/other researchers.

In the future a researcher will often take less than a day to prepare most datasets using other data of interest. The meta-data, tooling and documentation to adequately capture the complexities of the instruments, techniques employed in creating it and mission-specific details will exist and this information will adhere to community-driven shared international standards. The metadata will go beyond simply detailing the processing level of the files, it will also capture the relationship to other research artifacts such as the raw dataset, published papers, related catalogs, source code known to work with the data, notebooks, visualizations and whatnot. This ecosystem of artifacts will more readily allow the researcher to gauge how to approach using the dataset in the best manner. Nevertheless, when even this enhanced amount of information is not sufficient, in the future the researcher can easily find curation staff or a community of practice who are familiar with the data product, and its issues, and can help resolve difficulties using the data.

Furthermore, ‘big data’, whether by volume or variety, will be easier to work with as well. Researchers will be able to utilize cloud-based environments with standardized software to process the large volumes of data they are unable to pull back to their local machines and they will be able to easily combine, merge data from multiple instruments because they are provided in an ‘analysis ready’ state whose provenance is easily understandable and more quickly processed to a state they need for their science.

3.3.5. Accelerated Collaboration and Communication in the Research Community.

In the future researchers will increasingly rely on forming collaborations beyond the boundary of their local community and will be enabled to find collaborators wherever the interested party may be. Infrastructure will aid in forming these collaborations and characteristics which would be considered for a future infrastructure include:

- Asynchronous capabilities for communication and collaboration so that a lot of the pressure off of always being present for all events all the time is removed. This is also needed to lower the issues arising from time zone differences between collaborators and will aid in changing the expected time scale for progress on a project from a one hour synchronous meeting to a full day asynchronous session.

- Collaboration communication is thought of in a holistic sense – it involves all aspects of how a research team communicates from chats / messages to (virtual) meetings to the actual writing of papers and sharing of research artifacts for the project.

- High quality Communication. Little off-topic ‘chatter’ to balance against overwhelming fire hose of information and potential chaff. Good discovery and filtering tools will exist to help the researcher pinpoint information and collaborators of interest.

- Cross-domain science. Participation in meetings outside of the sub-field should be enabled and encouraged and platforms for communication will not be tied to specific sub-communities of heliophysics.

- Training for researchers to handle/practice this environment will be readily available and easy to understand.

Examples of collaborative efforts/tools which the workshop wanted to call out which might be future exemplars of this type of approach include:

- Overleaf - is a tool for cross-disciplinary collaborative paper writing.

- FDL is an example of the collaborative approach. Teams participate to formulate questions and others apply to work on the problem.

The future state will include asynchronous/synchronous collaborative platforms such as Slack or MSTeams which allow for researchers to interact with each other on and discuss heliophysics. The platforms will be lightly curated so as to prevent non-science topics or spamming and ensure high quality content. Furthermore innovative funding will be applied to better encourage research collaboration between government, academia and industry around important research ‘questions’ such as the FDL effort. Participants hypothesized the existence of an international platform where folks can propose a topic/problem and associated funding would be available and would broaden and democratize the number of researchers who could participate in heliophysics research.

3.3.6. Big Data Science

Big data science is a problem which involves both access to High End Computing (HEC) and access to at least modestly large (> ~10 TB) datasets. The workshop participants were not in uniform agreement about access to HEC being insufficient, but there was agreement that the community did not have sufficient access to the large datasets already being produced by some missions.

In a future state, internet-available platforms (perhaps cloud-hosted) will host modestly large (or larger) scientific datasets and provide the needed processing power to analyze these data. Details of this platform look much the same as the platform projected for the future in section 3.3.1. but may involve other novel access, in particular the ability to open a terminal / shell on the remote system. In all cases, users will readily be able to obtain access to the platforms, be able to use research funding (from their grant, or institution) to defray the cost of the computing, and work with known software tooling such as PyHC libraries or SolarSoft as well as advanced HEC / HPC tools which are

readily available on the platform (CUDA, MPI, GPU acceleration, Slurm, etc) to allow the user to get the most out of the computing power available.

4. Analysis: common themes and interrelationships

Looking at the suggested future state solutions several themes or interrelationships are apparent.

Collaboration is a theme for both how research should be conducted but also is an underlying theme as to how the infrastructure should be evolved. Shared research environments should enable greater cross-domain science and discovery of colleagues. Shared communities of infrastructure professionals should co-develop standards for data products, shared dictionaries of defining terms, shared analysis environments, platforms and tooling for doing research and a shared culture of best practices for doing heliophysics research. There was an important understanding that in all cases DEIA[16] should be an important requirement for how the solutions are realized.

While it may not be apparent in the proceeding text, Open Science came up as a common theme in the context of many discussions. There was a recognition that Open Science provided, in part, the driver for much of the change as its requirements of transparency and reproducibility are not currently practiced and are difficult to currently implement by researchers. Solutions should ultimately provide a means to lower this burden on researchers -- ideally by reducing both the ‘upfront costs’ of research (discovery, access, understanding of data products, forming collaborations) or the backend work of actual publication of interlinked research artifacts for others to consume (interlinked data products, software / models, etc with the paper itself).

Interestingly participants raised cultural/social issues as being important for infrastructure. They mentioned the need to better draw together research teams, creating a better ability to bridge institutional and scientific (sub-domains such as ITM, solar, etc) to create collaborations. There is also an expressed need by the workshop participants to also better integrate our new members (young researchers) and get them productively researching faster. How the infrastructure tackles these issues is a novel problem and can be addressed in part by making existing information easier to find and understand, but also perhaps to provide new types of information as well (such as an explicit search to discover who is working in a particular field). Creating virtual meeting grounds using high-quality bulletin boards / team chat or other platforms may also help address these challenges.

Looking over the material from the workshop, it is clear that the community will not find a new technical widget that can solve all of their problems. Solutions must come from

[16] Diversity, Equity, Inclusion and Accessibility

not only NASA efforts to implement and lead the community, but also from partnering with outside entities/institutions to make effective change. These solutions will encompass change in many areas including not only technical, but also cultural, procedural, and programmatic change. Policy change at NASA is also needed to both help accelerate and ensure the aforementioned changes occur as well (for example, better handling of open source data and code).

Finally, funding was also often mentioned. Innovative funding structures and means to deliver research funding where it would best be utilized repeatedly were mentioned.

3.3.7. Incentivised Resources

Solutions discussed included:

- *taking advantage of solutions that already exist (those that we as a community have not had the funding/incentive/time to implement;*
- *funding dedicated to making a result reproducible;*
- *guest investigator programs to produce infrastructure for past and future missions; short-term grants for implementing best practices;*
- *grants for working with groups external to NASA (e.g., industry, professional, computer scientists); funded positions to bridge the data and research best practices; expanded programs like NASA's Advanced Systems Information Technology (AIST) for Heliophysics;*

representation/recognition structures to create a cultural shift (e.g., credit for creating datasets of value, community awards for exceptional contributions). A line that stood out was 'doing best practices for free is not helping the community.'

5. Summary of Findings and Observations

We have tried to distill the large amount of contributed material from the workshop – significant gaps and envisioned solutions where found (sections 3.2 and 3.3). In review of the material of this workshop it is clear that NASA infrastructure provides critical capabilities to the heliophysics research community but much remains to be done.

Some research activities are still hard for many in the community as the research process still takes a great deal of work and with the increasing amount of readily available information and types of research artifacts it has become even more

challenging. In cases the desired capabilities already exist, but are not well known by the community – better advertising of NASA capability is required.

New challenges are also being asked of the infrastructure such as enabling Open Science and there are new cases, such as doing research with big data, where interesting research is sometimes not done for lack of resources / capabilities. NASA needs to expand its portfolio of capabilities to address these needs. The participants of the workshop imagined a future where wrangling the data of interest was much faster and easier to work with. Novel means of discovering and browsing data are part of making things easier and there was an expressed desire for a more sophisticated system in which new information, such as phenomena can be used to find data of interest. Also mentioned as being helpful at the workshop were cloud-based platform(s) with associated compute which would provide the needed processing power, large data volumes and tooling to enable new science and collaboration across traditional boundaries. Improved tooling, including a basic research software environment where things 'just work', was also part of the picture of making things easier for researchers. Open science publications are both a source of concern and hope for the community and we need to develop the tooling which better accelerates the adoption of this approach to research. At an infrastructural level, these things imply a need for improved metadata and handling of the data (e.g. interlinking in new ways with new types of information).

From a user standpoint these new capabilities should not target narrow areas of research but should be shared across the NASA and broader communities and should be extensible to allow for creative application in research. The workshop has also made clear that the NASA infrastructure does not stand alone. The community that relies on our capabilities also rely on capabilities resident at other institutions or at commercial entities. We must seek solutions which co-leverage each other's capabilities effectively.

More broadly, the discipline of Heliophysics must respond to trends in science publication and research, including adopting FAIR data principles. The concept of Open Science provides a framework for our community whereby researchers publish code alongside data products with research papers. FAIR data principles [*Wilkinson et al.*, 2016] and Open Science [e.g., *Gentemann et al.*, 2021] enable advanced, collaborative, interdisciplinary science.

References

Gentemann, C. L., Holdgraf, C., Abernathey, R., Crichton, D., Colliander, J., Kearns, E. J., et al., 2021, Science storms the cloud. *AGU Advances*, 2, e2020AV000354. https://doi.org/10.1029/2020AV000354

McGranaghan, R., Borovsky, J. E., and Denton, M., 2018, in https://eos.org/meeting-reports/how-do-we-accomplish-system-science-in-space

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al., 2016, The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

Appendices

A. Workshop Participants

Name	Session	Role
Monica Bobra	A	Session Chair
Michael Kirl	A	Moderator
Asti Bhatt	A	
Chiu Weigand	A	
Hazel Bain	A	
Liam Kilcommons	A	
Patrick Koehn	A	
Ricky Egeland	A	
Alexa Halford	B	Session Chair
Barbara Thompson	B	Moderator
Darren De Zeeuw	B	
Dustin Kempton	B	
Heather Futrell	B	
Jesper Gjerloev	B	
Katya Verner	B	
Nathalia Alzate	B	
Robert Allen	B	
Larisa Goncharenko	C	Session Chair
Shing Fung	C	Moderator
Aaron Ridley	C	
Alisdair Davey	C	
Alvin Robles	C	

Rafal Angryk	C	
Ryan Timmons	C	
Sarah Vines	C	
Will Barnes	D	Session Chair
Maria Kuznetsova	D	Moderator
Ayris Narock	D	
Greg Lucas	D	
Jenny Knuth	D	
Mark Cheung	D	
Susanna Finn	D	
Neal Hurlburt	E	Session Chair
Lan Jian	E	Moderator
Jonathan Niehof	E	
Julie Barnum	E	
Raphael Attie	E	
Rebecca Ringuette	E	
Reinhard Friedel	E	
Sara Jennings	E	

Members of the organizing committee who served as moderators appear in the table above. Other organizers included Robert Candey, Jack Ireland, Ryan McGranaghan, Aaron Roberts and Brian Thomas.

B. Workshop Agenda

- **Day 1:** [4 hrs] Goal/Theme: Examination of Current Infrastructure
    - [30 min] **Daily Briefing**
        - [10 min] Introduction to workshop [Brian]
            - Why? What are goals of workshop

        - 20 questions exercise
    - Structure of workshop
        - Ground rules, expectations, agenda, etc
- [20 min] Setting the Stage - Examples of Infrastructure [Aaron/Lan/Jack/Barbara/Masha]
    - What do we mean by infrastructure?
    - Overview of some NASA research infrastructure
    - Your responses/thoughts

- [~1.5 hrs] ***Breakout Sessions***. Infrastructure Discussion - split into working groups; each has a moderator to lead discussion.
    - [~10 min] Instructions to breakout teams [Brian]
    - [1.5 hrs] Divide breakout groups (see table of assignments) - Try to capture how folks are doing research today, based on the questions presented. What services, tools, etc help them now. Indicate consensus (or not) on thoughts. Prepare talking points.

- [20 min] - ***Coffee Break***

- [1.5 hrs] **Plenary/Moderated discussion**.**(Starting at 1:30 pm EDT)** Start to develop some initial themes/commonalities from these presentations.

    - [30 min] Session leads give presentations on what folks came up with [Session Chairs]
    - [30 min] Panel Discussion. Session chairs reflect on output in a moderated discussion [Brian, Session Chairs]
    - [30 min] General Q&A. Participants reflect on content.

- [30 min., Optional] **Virtual icebreaker social hour** [Host: Barbara]. BYOB(everage). https://tinyurl.com/HelioWonder Password: HelioInfra

- **Day 2:** [4 hrs] Identifying gaps in current infrastructure.

    - [30 min] **Daily Briefing.** [Brian]
        - Summary of 1st day's output, Key infrastructure identified, presentation of some apparent commonalities seen from yesterday's outputs.

- [~1.5 hrs] ***Breakout Sessions***. Infrastructure Discussion - split into working groups; each has a moderator to lead discussion.
  - [~5 min] Instructions to breakout teams [Brian]
  - [~1.5 hrs] Divide breakout groups (see table of assignments) - Try to capture what gaps exist in the key, and other, research infrastructure. Prepare talking points.

- [20 min] - ***Coffee Break***

- [1.5 hrs] **Plenary/Moderated discussion**. Develop some initial themes/commonalities of the gaps in the research infrastructure from these presentations.

  - [30 min] Session leads give presentations on what folks came up with **[Session Chairs]**

  - [30 min] Panel Discussion. Session chairs reflect on output in a moderated discussion **[Brian, Session Chairs]**

  - [30 min] General Q&A. Participants reflect on content.

- **Day 3** [4 hrs]: Synthesis and user-facing solutions. What would solutions to identified gaps 'look like' to the user? Which has the most "Value for customer"?

  - [30 min] **Daily Briefing.**
    - Kick off with presentation of some apparent overlaps, commonalities in gaps.
    - Instructions/examples to users on how to create user-facing solutions.

  - [~1.5 hrs] ***Breakout Sessions***. Infrastructure Discussion - split into working groups; each has a moderator to lead discussion. Discuss and catalog ideas on user-facing infrastructure solutions.

  - [20 min] - ***Coffee Break***

- [1.5 hrs] **Plenary/Moderated discussion**. Develop some initial themes/commonalities of the gaps in the research infrastructure from these presentations.

    - [30 min] Session leads give presentations on what folks came up with **[Session Chairs]**

    - [30 min] Panel Discussion. Session chairs reflect on output in a moderated discussion **[Brian, Session Chairs]**

    - [30 min] General Discussion.
        - Participants reflect on content in Q&A.
        - Poll participants on their priorities, is the ‘consensus’ on best things for infrastructure change? Review poll and get feedback (with comments from participants)