Heliophysics Decadal Survey 2022 White Paper

Data Needs to be...

A. Halford[1], E. J. Rigler[2], L. Rastaetter[1], T. Y. Chen[3]

[1]NASA/GSFC, [2]USGS Geomagnetism Program, [3]Columbia University

Synopsis

Findable, Accessible, Interoperable, and Reusable (FAIR) data are essential to heliophysics, indeed all scientific research. We make recommendations intended to prioritize resources needed to satisfy FAIR data principles, treating them as a fundamental research infrastructure, rather than a simple research product.

1. **Data collection, preparation, archiving, and accessibility needs to be a priority**
2. **Data collection, preparation, archiving, and accessibility needs dedicated, sustained funding support.**
3. **Data needs to be accessible through investment in infrastructure, tools to access and read the data and personnel to maintain this data and IT infrastructure.**
4. **Data needs to be collected in sustained ways to enable further science, and specifically, model validation efforts**

More details, and specific suggested actions, are provided in the main body of this position paper.

1 Introduction

Data are the foundation of good science. This is true whether data are observations of the physical world, or output from computer simulations. Data must be collected carefully, prepared consistently, interpreted in appropriate context, archived and made easy to access for reproducibility and subsequent research, and, more-so in recent years, made interoperable with other data sets as the most cutting-edge science is interdisciplinary and spans multiple physical domains.

The growing awareness of Findability, Accessibility, Interoperability, and Reuse (FAIR) data principles (https://www.go-fair.org/fair-principles/) is an acknowledgment of these truths within the broader scientific community, but applying these principles consistently is a scientific challenge of its own, and different disciplines have made more and less progress toward these important goals.

The following is a list of recommendations we believe will, if followed, bring heliophysics up-to-par with other fields who have been working longer to implement FAIR data principles (we should not assume they've solved all challenges [see Kelbert, 2014 https://doi.org/10.1002/2014EO490006]). Each recommendation includes several suggested actions intended to help realize the stated goals.

2 Recommendation 1: Data needs to be a priority

Historically, the expense and time that goes into the production of data has led to a relatively low priority being placed on what are now known as FAIR data principles. It is so often more cost-effective to maintain datasets than to launch new satellite missions, to develop new supercomputer models, or re-run existing models. Additionally, there are long-period oscillations in solar activity (11-year sunspot cycle), and even slow secular variations in Earth's magnetic main field where we have yet to obtain or prioritize longitudinal sets of these and similar data. This lack of data creates roadblocks for researchers interested in space climatology. Finally, while sensors and models will usually improve with better engineering and more sophisticated algorithms, good data stewardship is essential to track the evolution of data and model quality over time, and thus assess the quality of the science.

> Heliophysics-related data stewardship must be made a priority in its own right, on par with, but independent from, data collection and interpretation.

There are many scientists and programs dedicated to making scientific data FAIR. However, these efforts have been hampered by outdated academic expectations and practices such as "publish or perish", not to mention insufficient and unreliable funding. Heliophysics-related data stewardship must be made a priority in its own right, on par with, but independent from, data collection and interpretation. Below we offer possible actions which will mitigate some of the current challenges:

Action 1 towards FAIR data being a priority:: We also acknowledge that this landscape changes quickly and rapidly over the course of a decade. Thus we encourage that *this topic be re-visited on a more regular basis than once every 10 years.*

Action 2 towards FAIR data being a priority:: *Make Space Physics Data Facility (SPDF) a Goddard Heliophysics Tier 1 capability* to match the CCMC priority and thus easier to ensure funding and connection between model and observational data validation. The SPDF and similar archives/portals housed at NASA Goddard have become large nodes where the community can find much space physics data - especially satellite, rocket, and balloon based data. SPDF currently has personnel to help ensure all data follow a basic set of standards, e.g. file type and metadata. It has been so successful that other groups also have data access through them including NOAA, and ground based indices. We suggest the further funding of SPDF to expand to provide portal access to other data centers and archives such as SuperMag and Madrigal.

Action 3 towards FAIR data being a priority:: *Similar funding/resources for other agencies with a portability/portal to connect all agency and industry/academic data.* While SPDF covers much of NASA heliophysics data, we see a symmetric need at other agencies. Other centers/facilities need sufficient and reliable funding to maintain their own data archives and accessible portals to access the data. Multiple repositories build in redundancy in case of loss.

3 Recommendation 2: Data needs dedicated funding

Dedicated funding is necessary to ensure the lasting infrastructure necessary for data archival and access. Data archival can be achieved through multiple platforms, and, ideally, will use multiple platforms to help ensure redundancy. All options take expertise, equipment, and upkeep.

Dedicated funding is necessary to ensure the lasting infrastructure necessary for data archival and access. Data archival can be achieved through multiple platforms, and, ideally, will use multiple platforms to help ensure redundancy. All options take expertise, equipment, and upkeep.

Action 1 towards FAIR data being funded:: *Similar funding/resources for agencies with a portability/portal to connect all agency and industry/academic data.* As we raise data stewardship as a priority across all agencies, this effort needs to be supported through funding the infrastructure and maintenance, as well as community access capabilities, e.g. development of APIs or standard code libraries for data access, load, and piloting routines.

Action 2 towards FAIR data being funded:: Ideally the government will fund *multiple data repositories for redundancy.* We recommend that at least one repository should be government based as the government, more than academia or industry, has an obligation to ensure data archive purposes as well as data accessibility to the public ensuring open science is achieved. The multi platform approach would make use of both government owned and commercial platforms including cloud technologies. While we expect much innovation from commercial entities, we recommend

that the government also consider providing a services simiar to basic cloud based platforms and other collaborative tools which would ensure further accessibility to data and computing to the public (e.g. git.gov), specifically for those who can not afford the costs of commercial tools, cloud computing, and storage.

Action 3 towards data being funded:: We see a growing need for *funding for small projects to port data to the archives.* We acknowledge that this in part already exists. However, it is still not well-known throughout the community. One such expansion of the current HDEE calls could be a dedicated call for successfully flown LCAS missions. Once the mission flies, the team could then propose to make the data more easily accessible and available to the public. While many LCAS missions work to make their data public, they are run on shoestring budgets. A GI type program where data stewards, data specialists or data historians apply and work with the instrument team to further process the data, include proper metadata, and basic read, load, and plot routines would greatly enable accessibility and further science returns from these bargain basement missions. Larger missions may also benefit from similar expansions of their calls to include specific proposals for data specialists to work with mission teams to help ensure the accessibility and usability of the data, ensuring adherence to best practices and current standards.

Action 4 towards data being funded:: While the above has focused on observational data, *the same recommendations hold true for model data.* Running a large simulation is costly financially, computationally, and has an impact on our carbon footprint. Retaining model runs and improving the accessibility of the data will enable better, and faster model data comparisons, validation activities, as well as including more model data and global views for data driven studies.

4 Recommendation 3: Data needs to be accessible

Software libraries, interactive applications, and web-based application program interfaces (web-APIs) are different tools necessary for accessing data. As popular tools change and update, expertise and resources are needed to ensure continued access to the data. Many groups have identified an additional need for repositories or portals which may not house data, but point users, or grab data for users through their interface. We can look to other fields for successful examples, such as IRIS in seismology (https://www.iris.edu/hq/about_iris).

As popular tools change and update, expertise and resources are needed to ensure continued access to the data. We can look to other fields for successful examples, such as IRIS in seismology.

Action 1 towards data being accessible: : There is a strong need to *better connect the different data portals/observatories/viewers/archives* (e.g. SPDF, Helioviewer, madrigal, Supermag, TREX,

AuroraX). Many of the data archives and portals currently in existence were built around one sub field of heliophysics, e.g. solar, ionosphere, or magnetosphere, and sometimes even smaller sub fields, e.g. radiation at aviation altitudes. As we look towards a future with more transdisciplinary science, we see a growing need to have these data portals more interconnected. However, this recommendation can only come once the data archives and repositories and their respective are funded and sufficiently operational.

Action 2 towards data being accessible: : *Standardized data formats* enable easier access and an increased use of data. ASCII, Comma Separated Variables, are human readable data types which will always be useful but often take up a lot of space. Compressed file types or binaries are vital for our ever growing data. We suggest moving towards Hdf5 or NetCDFs as they have a larger user database and thus a wider developer base for read routines. This wider user base also increases the sustain support and updates into the future.

Action 3 towards data being accessible: : With historical data sets and the likelihood of new data formats in the future, there is a clear need for *continued development and maintenance of converters between data types.* Maintaining converters between data types ensures the continued use and science gains from the data into the future. We also encourage *data archives to update data into new currently standard data types.*

Action 4 towards data being accessible:: *Maintain data readers for multiple code languages* is necessary to ensure the continued use within the research community. The research community uses multiple different types of coding languages, and ensuring easier access to data will enable further science returns. Providing readers also reduces potential errors in correctly reading in data and thus correctly processing and interpreting it.

Metadata is necessary to ensure the findability of data. However, producing and developing metadata is a very time consuming task, requiring skills which not all researchers or data providers continually cultivate. As our field grows and expands, it is necessary to ensure metadata standards are maintained and improved so that our data can be easily accessed and used. This may also necessitate the encouragement of new professional titles and the cultivation of new skill sets in our field including data curators and data specialists.

Action 5 towards data being accessible: : *Simplify metadata with clear ways to crowd-source and improve it with continued use.* Simplifying the initial basic metadata needed reduces the barrier of entry for making data public. Providing a way for others to then contribute to the further updates of metadata will allow the metadata to be updated, as well as provide insights into what science phenomena the data may be used to study which may not have been considered by the data providers.

Action 6 towards data being accessible:: *Develop wrappers or code packages to help produce basic metadata and identify what metadata/common format/etc is missing*. This would help data providers create and know what information is needed, ease the transition of the data from the data provider to data archives/repositories such as SPDF. For example, SPASE exists as the protocol for metadata, but it is not easy to use. There is a large learning curve for new people to make their datasets SPASE compliant. It would be fantastic to have a tool to help someone conform to the SPASE protocol.

Action 7 towards data being accessible: : The work of maintaining, archiving, and ensuring accessibility to data takes time and effort which must be recognized. We encourage the *development of new metrics or use of current metrics to give credit to these activities.* As hiring, promotion, and awards often rely in part on quick metrics such as the H-index, there is a growing need to find ways to give credit to this unseen and often unacknowledged labor and skills. We encourage institutions to help promote and aid in the following:

- Further the culture to accurately cite data (e.g. SPDF providing DOIs on the screen where data is chosen to be plotted/downloaded.
- Encourage more papers on data descriptions. We suggest a journal or journal article type like those in JOSS (https://joss.theoj.org/), approximately 1 pagers describing the data and metadata to advertise the new data availability. This type of peer review journal article for the data can check for such items as the following; does the data exist, does it have appropriate metadata, and is it accessible.

5 Recommendation 4: Data needs to be collected with global model validation and systems science as priorities

Although we have been visiting space for over 70 years, and Earth at ground level and the lower atmosphere for a few hundred, we have not adequately sampled all regions of space which our models and theory cover. To advance our understanding of the dynamics and coupling across the entire heliosphere, we need more data (and supporting missions) that prioritize systems science and model validation. As access to space is cost prohibitive, we should ensure that we make the most of historical datasets.

We have not adequately sampled all regions of space which our models and theory cover. To advance our understanding of the dynamics and coupling across the entire heliosphere, we need more data (and supporting missions) that prioritize systems science and model validation.

Action 1 Towards more data: : *Bench marking activities need to become more formalized and standard within the research community.* Time intervals for different phenomena intervals, such as quiet, solar storm, geomagnetic storm, ionospheric impacts, atmospheric impacts, need to be identified and agreed upon by the community to provide a standard set of data for model validation activities.

Action 2 Towards more data :: *Expand or create new mission lines to focus on questions around increasing data for model validation.* Current mission lines prioritize fundamental science questions. While needed validation for models can be shoehorned into this framework, we believe better validation efforts will be achieved if this is prioritized. We envision this as a new funding line with a budget of at least that of explorers and potentially much larger as multiple well instrumented platforms may be necessary to gather sufficient data for the validation efforts.

Action 3 Towards more data:: *Increase incentives to have model outputs in similar physical units to observational data.*

- Parameters such as Phase Space Density, L*, and others are intrinsic to models but not to our observational capabilities. Thus for better calculation of errors within our measurements, observations, and analysis, we encourage transforming the model outputs into observational parameters instead of converting observational data into model outputs.
- Likewise, we need to treat model data similarly to observational data by working towards collecting and analyzing the model data in a similar fashion to the observed data. This may include:
  - Integration time to collect particle observations, field-aligned current (FAC) maps, etc.
  - Integration of climatology vs single model run instance. (chorus wave occurrence, particle loss)

Action 4 Towards more data: : *New research line with similar funding levels of the Guest Investigator Program with a focus to use historical data to address ongoing science questions.* Historical data is incredibly valuable to both address the primary science questions of the original mission as well as new science questions long after the mission itself has ended. A new funding line focused on using historical data will further extend the scientific gains from these past missions and extend the usefulness of the data archives and portals.

6 Conclusion

As we move into the coming decade, data needs to be a priority, with dedicated funding and accessible. It needs to be collected with validation and system science as a priority. These four

recommendations and the suggested actions will help to ensure data is available to answer the science questions, space weather tools, and validation needs for the coming decade.

Recommendations

- Data needs to be a priority
- Data needs dedicated funding
- Data needs to be accessible
- Data needs to be collected with global model validation and systems science as priorities

7 Acknowledgments

This work has been completed as part of the GEM MMV resource group and the HISFM SPI group. The authors would like to thank overleaf which aided in our ability to collaborate. This LaTeXposition paper template was originally created by Alexa Halford and generalized for the Heliophysics Decadal Survey by Ryan McGranaghan.