A White Paper on
Essential Best Practices for the Geospace Community
Concerning
Reproducible Research, Open Science, and Digital Scholarship

Asti Bhatt*, SRI International (asti.bhatt@sri.com)
Ryan McGranaghan, NASA Jet Propulsion Lab
Tomoko Matsuo, University of Colorado Boulder
Yolanda Gil, University of Southern California

The American Geophysical Union[1] (AGU), in a recent effort to advance open and fair data standards in earth and space sciences noted that "Open, accessible, and high-quality data and related data products and software are critical to the integrity of published research. They ensure transparency and support reproducibility and are necessary for accelerating the advancement of science." In this white paper, we outline why the geospace community needs a coordinated effort aimed at advancing open and fair data and software standards and enabling the use of modern data science approaches to solve critical challenges in geospace science.

The National Space Weather Strategy and Action Plan released in 2015[2] outlines the need for improved measurements, data assimilation and multidisciplinary, international collaboration to help mitigate risk posed from space weather events. Key points identified to achieve the successful outcomes include developing benchmarks that are well understood and effectively communicated across all stakeholders; creating required infrastructure for sharing measurements for analysis and data assimilation and archival for long term predictions; and improving coordination between all stakeholders to achieve the required preparedness.

Geospace science is largely a grassroots driven effort, which brings together diverse approaches towards solving critical space weather problems. An unintended consequence is the lack of common standards, methods, or tools that help translate the science for effective multidisciplinary collaboration. The grand challenge questions as outlined in the National Research Council Decadal Survey on Solar and Space Physics[3] require a common framework that will enable scientists from various disciplines to work together effectively.

It is widely recognized that the data need to be FAIR (Findable, Accessible, Interoperable, and Reusable)[4], a concept initially developed by Force11.org. There is much work done in various disciplines to make this happen. As one example, we draw your attention to the fact that massive volumes of solar data are now discoverable from the Virtual Solar Observatory (VSO)[5], established in 2005, compiling data from the Solar Dynamics Observatory (SDO, 2.5 petabytes of data from the primary mission between 2010-2015) and other space- and ground-based observatories. The VSO also provides an open and available applied programming interface (API) that can be freely used, extended, and improved by a range of users from experts to the general public. VSO and the solar community's embrace of open data and software sharing policies have enabled broadened participation in solar research and significant advances in

---

1 http://sworm.gov/publications/2015/swap_final__20151028.pdf and http://sworm.gov/publications/2015/nsws_final_20151028.pdf
2 https://news.agu.org/press-release/agu-coalition-receives-grant-to-advance-open-and-fair-data-standards/
3 http://sites.nationalacademies.org/ssb/currentprojects/ssb_056864
4 https://www.force11.org/group/fairgroup/fairprinciples
5 https://umbra.nascom.nasa.gov/vso/

understanding the physics of the sun [*Sola and Tian*, (2009)[6]; *McIntosh et al.*, (2017)[7]] and predicting solar behavior [*Bobra et al.*, (2016)[8]; *Jonas et al.*, (2017)[9]].

Use cases that have benefitted from open data and software sharing practices are increasingly numerous across science and engineering (see, for example, case studies from the Software Sustainability Institute[10], and Figshare's annual *State of Open Data* report[11]). The question we raise in this white paper is how the geospace community can follow these examples to take advantage of digital scholarship to produce discovery science in the near Earth space environment. Many of these issues were discussed at the 2017 CEDAR Workshop during the Digital Geospace session and follow-on discussions during the meeting.

The need for better digital practices that can lead to utilizing large data effectively is also identified by the federal funding agencies, e.g. 'Harnessing the Data Revolution' is one of the top priorities for NSF in FY2019. The NSF's EarthCube[12] initiative is specifically focused on the data and software needs of geoscientists. The lessons from EarthCube and other programs can easily be adopted by the wider geospace community. NASA recently invited the community input on their 'Open Code Policy' for space sciences.

In this white paper, we outline some of the practices used by various science and other communities around the world that can serve as concrete examples for the geospace community to organize discussions around. We are not proposing that these practices be adopted by the geospace community as described, or be made mandatory by the funding agencies. These are meant to stimulate the discussions in the wider geospace community on democratizing digital resources, which in turn would enable creation of an effective multidisciplinary, collaborative culture to tackle the grand challenge questions. We propose that a working group be created to adopt and implement required systemic changes and develop standards for the geospace community.

1. Outline

This document outlines example best practices with respect to reproducible research, open science, and digital scholarship. Specific practices related to data stewardship are in Section 2, and for models and software in Section 3.

2. Best practices for Data stewardship

PROPOSED RECOMMENDATION #D1: All the data produced and used by geospace scientists should be:

1. *available in a shared community repository*, so anyone can access the data

---

[6] Sola, Suarez and K . Wampler Tian. "CME Population Distributions: Science Facilitated by the VSO." (2009).

[7] McIntosh Scott W. and Leamon Robert J. "Deciphering Solar Magnetic Activity: Spotting Solar Cycle 25." Frontiers in Astronomy and Space Sciences. Vol. 4 (2017). https://www.frontiersin.org/article/10.3389/fspas.2017.00004. doi: 10.3389/fspas.2017.00004

[8] http://iopscience.iop.org/article/10.3847/0004-637X/821/2/127/meta

[9] https://arxiv.org/abs/1708.01323

[10] https://www.software.ac.uk/

[11] https://doi.org/10.6084/m9.figshare.5481187.v1

[12] https://www.earthcube.org/

2. *have a license*, so anyone can understand the conditions for reuse of the data
3. *have an associated digital object identifier (DOI) or persistent URL (PURL)*, so that the data is available permanently and URL rot in papers is avoided
4. *cited properly in the paper in the references section*, so readers can identify the datasets unequivocally and data creators can get credit for their work

PROPOSED RECOMMENDATION #D2: When requesting a dataset from a data repository, a user should be given not just a dataset, but also a license and a DOI or PURL as well as a preferred citation that they can use in their papers. Data repositories that serve the Geospace community should consider serving data in this manner.

Permanent Unique Identifiers for Data
DOIs are managed by data repositories and given either to individual datasets or to collections. In order to do so, the repository must forge an agreement with a DOI authority. PURLS can be assigned by anyone to any Web resource, including a dataset that has a URL on the Web. Individual researchers can assign and manage PURLs, using a trusted service such as the W3C's w3id.org. Data repositories also have the option of using PURLS.

Community Data Repositories
Many general data repositories can be used by scientists in any domain, and as such they are available to the geospace community. These repositories will automatically assign a DOI to any uploaded data, and will accept also software, figures, movies, and slide presentations. They will also inquire about choosing a license, and specifying a descriptive name and authors for the dataset. These repositories include Zenodo, Figshare, and Pangea to name a few. The COPDESS organization, which includes dozens of publishers in geosciences, offers a list of recommended data repositories. Universities also offer general repositories, whether developed in-house or as installations of general infrastructure such as Dataverse. These university repositories are typically maintained by library departments and always offer DOIs, licenses, and citations.

Geospace Community Data Repositories
Data repositories that serve the geospace community should adopt mechanisms for assigning DOIs or PURLs to datasets that they serve to users. The management of PURLs or DOIs can be complex, and organizations such as FORCE11, the Research Data Alliance, and ESIP have working groups with extensive and detailed recommendations in this respect.

Licenses for Data
Recommended licenses for data are Creative Commons licenses, preferably CC-BY (unlimited reuse as long as there is attribution) or CC0 (unlimited reuse without conditions).

3. Recommendations for Software used to produce Data

We recognize that funding to implement the following recommendations is currently unavailable to individual investigators in the geospace community. We are not recommending to make open software a mandate by funding agencies and journals at this point. The recommendations below may not apply to all numerical models and are made to motivate further discussion among stakeholders.

PROPOSED RECOMMENDATION #S1: Software used in a paper, particularly models, should be:

1. *available in a shared community repository*, so anyone can access it
2. *have a license*, so anyone can understand the conditions for use and extension of the software
3. *have an associated digital object identifier (DOI) or persistent URL (PURL)* for the version used in the paper, so that the software is available permanently and URL rot in papers is avoided
4. *cited properly in the paper in the references section*, so readers can identify the software version unequivocally and software creators can get credit for their work

When this recommendation cannot be met, a brief explanation should be included in the paper.

PROPOSED RECOMMENDATION #S2: When using or downloading a model or software from a repository, a user should be given the license and a DOI or PURL for the version used as well as a preferred citation that they can use in their papers. Model repositories that serve the CEDAR community should consider serving models in this manner.

PROPOSED RECOMMENDATION #S3: A software registry for geospace science software developed by various stakeholders should be created. Such software registries create a common place for the domain scientists to search for existing software solution and make their own software searchable. The software registry isn't a replacement for software hosting service.

PROPOSED RECOMMENDATION #S4: For open source software solutions using non-proprietary languages, a web-based platform should be implemented that can access data repositories and software to reproduce results, such as from a paper or a proposal with appropriate data and software IDs. This solution is recommended in the light of the recent push for reproducible science.

Permanent Unique Identifiers for Software

A separate DOI should be assigned to meaningful versions of the software, such as a version used for a paper. GitHub offers an option to obtain a DOI for a software version, which is done by storing that version permanently in the Zenodo data repository. Any software can be uploaded manually to community data repositories such as Zenodo, figshare, and Dataverse. PURLS can be assigned by anyone to any software version that has a URL on the Web, using a trusted service such as w3id.org.

Community Software Repositories

Many general software repositories can be used by scientists in any domain, and as such they are available to the geospace community. These repositories will often inquire about choosing a license, and specifying a descriptive name and authors for the software. These repositories include GitHub and BitBucket, to name a few. General data repositories accept software as an entry, and as with any dataset they always offer DOIs, licenses, and citations.

Geospace Community Model and Software Repositories

Model repositories that serve the geospace community should adopt mechanisms for assigning DOIs or PURLs to software versions that they run for users. The management of PURLs or DOIs can be complex, and organizations such as FORCE11, the Research Data Alliance, and ESIP have working groups with extensive and detailed recommendations in this respect. A software registry (e.g., OntoSoft) that includes metadata about models and other software can significantly facilitate software discovery and reuse.

Ancillary software used by the geospace community is currently maintained in personal computers and in some cases in general software repositories. A software registry for geospace community would help make these software available to the larger community subject to the developer's interest.

Licenses for Software
Recommended licenses for software are the standard licenses from the Open Source Initiative, preferably Apache v2, GPL v3 or MIT (unlimited reuse as long as there is attribution) but other more restrictive licenses are available.

4. Developing Common Standards

PROPOSED RECOMMENDATION #C1: Create a working group of data producers (e.g., instrument operators), software producers (e.g., modelers) and users of these in the geospace community to develop standards whereby data formats, types, nomenclature and units have a common framework. Similar effort can be undertaken purely among the modeling community to create the ease of interoperability.

PROPOSED RECOMMENDATION #C2: The development and use of common standards should be supported by the federal funding agency through providing resources and through adopting recommendations from the working group.

5. Summary

This white paper has resulted from a collaboration between the NSF EarthCube Integrative Activity award: Integrated Geoscience Observatory (InGeO) and the NSF EarthCube OntoSoft award for open software sharing in geosciences. EarthCube was initiated by NSF in 2011 to transform geoscience research by developing cyberinfrastructure to improve access, sharing, visualization, and analysis of all forms of geosciences data and related resources. As part of the InGeO project, we have organized several CEDAR sessions on the topic of improving data and software access in geospace community. In those sessions we built on best practices and recommendations developed as part of the OntoSoft GeoScience Papers of the Future initiative. This white paper contains the recommendations that were arrived at through discussions with community members at and outside the CEDAR meetings. It addresses critical pain points for accessing data and software in the geospace community and makes several recommendations for the geospace community to adopt.

One of the key recommendations in this paper is the creation of a working group to develop data and software standards for the geospace community. We believe the geospace research community would rally behind an effort by the agencies to help achieve reproducible science and to remove barriers to accessing and processing data leading to innovative solutions. As cyberinfrastructure tools become more widely available, such an approach will enable the geospace community to use them in achieving long-awaited closure on critical science questions.

The recommendations put forward in this white paper are supported by the following scientists in their individual capacity and not as representatives of their institutions:

Russell Cosgrove, SRI International
Ashton Reimer, SRI International
J. Michael Ruohoniemi, Virginia Tech
Joseph Baker, Virginia Tech
Anthony Mannucci, NASA JPL
Ennio Sanchez, SRI International
Leslie Lamarche, SRI International