OPEN, CURATED, PRESERVED: SAFEGUARDING SCIENTIFIC INFORMATION IN THE JOURNALS OF THE AMERICAN ASTRONOMICAL SOCIETY (AAS)

Response to RFI: Implementation and Changes to Science Policy Document (SPD)-41

Authors Kevin B Marvel, AAS Executive Officer, kevin.marvel@aas.org; Ethan Vishniac, ethan.vishniac@aas.org; Frank Timmes, frank.timmes@aasjournals.org; Chris Lintott, chris.lintott@aas.org; Julie Steffen, julie.steffen@aas.org; August Muench, august.muench@aas.org; Greg Schwarz, greg.schwarz@aas.org; Peter Williams, peter.williams@aas.org

Science Division Relevance: Astrophysics | Heliophysics | Planetary Science

AAS Recommendations for NASA

- Guide researchers on publishing small- to medium-sized datasets in journal articles versus larger data products that are best delivered to a NASA-funded repository.
- Ensure open data products are properly curated by data editors, librarians, or repository staff, including the application of granular domain-specific metadata.
- Partner with extant domain-specific repositories to support the ingest and aggregation of NASA datasets, making them truly interoperable and reusable.
- Support improved metadata for data and software by adopting the Unified Astronomy Thesaurus and funding a solution to the many-identifier problem for software citation.

Introduction

The American Astronomical Society (AAS) is a major international organization of professional astronomers, astronomy educators, and amateur astronomers. Its membership of approximately 8,000 also includes physicists, geologists, engineers, and others whose interests lie within the broad spectrum of subjects now comprising the astronomical sciences. The mission of the AAS is to enhance and share humanity's scientific understanding of the universe as a diverse and inclusive astronomical community. The AAS publishes major research journals in the field that are fully open access, include significant accompanying data, link to major external datasets and software, and will be preserved in perpetuity by the AAS.

Since leading the community with early electronic editions of its research journals starting in 1995, the AAS has encouraged and enabled authors to practice good scientific citizenship by including relevant underlying research data in their published articles. Currently, >20% of published AAS content contains some amount of curated data; this fraction increases yearly as we work with authors to expand data publication (Figure 1). In addition, AAS journal articles containing data products receive a greater number of citations (median of 5, mean of 11) than articles without (median of 4, mean of 9). The integration of research data into AAS journal articles addresses the SPD-41 requirement to make it publicly accessible no later than publication. In alignment with NASA SPD-41 III.P, AAS has committed to archive its entire

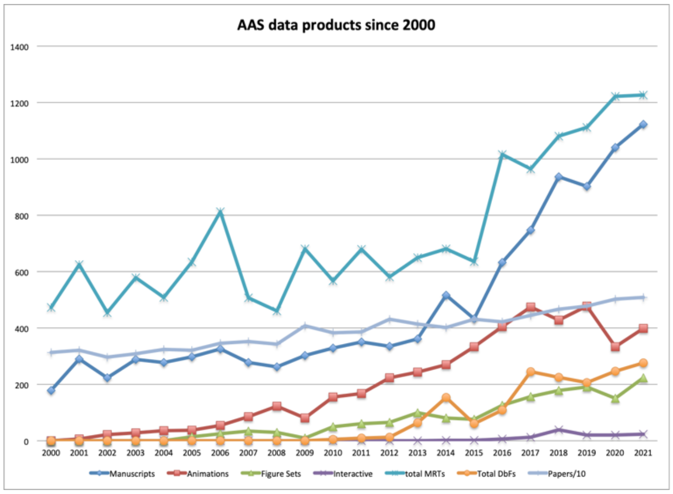

*Figure 1: Datasets published in the AAS Journals per year as a function of type. Plot includes: submitted **manuscripts** with curated data; sum of **papers** published by the Journals (divided by 10); **DbFs** are datasets behind figures curated by Editors; **MRTs** are extended tabular material curated by Editors. Also shown is the growth of visualizations in the Journals including **animations**, **figure sets** or atlases, and **interactive** figures.*

journal content corpus, including data products, in perpetuity through arrangements with third-party archiving services and other means.

Finally, 44% of AAS Journal first authors are based in the United States, which means that AAS is accountable to a wide range of global science organizations and funders. Our responses include (I) the impact of the proposed changes on the authors of journal publications; (II) the guidelines, services, and protocols necessary to support authors for the implementation of SPD-41.

I. Impacts on Authors

Open access articles in the AAS Journals provide a central resource for the discoverability and accessibility of small- to medium-sized datasets. Two PhD Data Editors for the AAS Journals solicit datasets from authors during the manuscript review process and curate these data following best practices established by leading international archives, e.g., NASA NED. Through publication as integral parts of journal articles, datasets are woven into them as extended tables and "data-behind-the-figures."

New requirements forcing authors to mobilize and port data out of journal articles and into repositories will not improve the FAIR (findable, accessible, interoperable, and reusable) qualities of the datasets. Such a mandate would lead authors to face a series of problems: (a) limited choices for repositories; (b) asynchronicity of article publication and data deposit; (c) loss of data review and curation.

Repositories: Astronomy, heliophysics, and planetary sciences currently lack sufficient domain-specific "FAIR" repositories for highly refined data products, for simulation and modeling results, and for null result data. Existing NASA archives focus on supporting large, "Treasury" or high-level science products and are not aligned with the needs of small- to medium-sized datasets published on rapid timescales. Authors that face limited choices for where to deposit their data may be driven to use generalist services that unfortunately provide little or no domain-specific metadata, or interoperability, e.g., the services of the virtual observator(ies).

Asynchronicity An example of best-of-class service for small datasets is Vizier (http://vizier.u-strasbg.fr), a data aggregation and repository created and supported by The Strasbourg astronomical Data Center (CDS), which is the same entity responsible for the NASA-funded Simbad database. Small datasets published with AAS Journal articles are typically ingested and redistributed by Vizier 6-18 months after article publication. We have been advised that the typical peer-review timescale for datasets submitted to the Planetary Data System (PDS) is 6-12 months. High-level science data transferred to archives such as the Mikulski Archive for Space Telescopes (MAST) may not be available for weeks or months as the data are standardized and prepared for release. These timescales are not aligned with article publication timelines unless authors are required to submit data to archives well in advance of peer-review. Authors will face a conundrum when publishing to external repositories: release their data early so that they are synced with the journal publication or publish their article without a formal data release. This asynchronicity does not benefit NASA authors wanting to provide their results to the wider community, nor does it benefit the reader wishing to replicate results or reuse data.

Curation: Faced with limited options for domain-specific repositories or long timescales for data publication, authors may choose to deposit data in generalist repositories. Such repositories lack domain-specific metadata and interoperability, and most institutional or generalist repositories do not have open curation interfaces that would allow journal editors to review and improve these materials in the same ways that AAS Journal data editors review, standardize, and improve data published in the AAS Journals. This loss of curation does not further the cause of open science and in some cases distorts the scientific record.

II. Guidelines, Services, and Protocols to Support SPD-41

II.A Guide Authors for Publishing Small- and Medium-sized Datasets in Articles

AAS research articles contain many thousands of data products, mostly of the small- to medium-sized category (Figure 1). NASA should guide authors by stating that data for tables, charts and graphs do not fall into the policy compliance definition that requires direct deposit of all data in a repository. Collections of raw or uniformly processed data now housed in NASA data centers such as MAST, IRSA, or PDS are subject to a separate set of criteria. AAS journals have been linking to raw data via persistent identifiers (DOIs) created at these data centers and will continue to do so, while expanding this direct linking between datasets and literature to other NASA data archives and continuing to work closely with NASA's Astrophysics Data System (ADS) to expose these links. Successful implementation of SPD-41 requires that NASA prioritize the simplification of approaches to creating dataset identifiers and accompanying researcher support at all NASA data centers.

II.B Require and Support Curation

The AAS strongly endorses the need for curation and quality assurance mechanisms for handling research data and software. We believe that curation of small- to medium- sized data should occur at the point of journal article publication and not be delayed to a subsequent or separate step. Two full-time trained astrophysicists are employed at AAS to curate the research data published in AAS journals. Their curatorial duties include a) checking that author-supplied data adheres to NASA's NED best practices for data publication (Chen et al. 2021); b) encouraging authors to archive data otherwise hosted on personal websites; c) obtaining important data behind figures such as light curves and radial velocities; d) checking that software is cited and highlighted in compliance with AAS software policy and e) verifying author provided DOIs for linking to datasets in NASA data centers. AAS data editors routinely uncover errors in the tabulation of results that would otherwise not be detected without their efforts.

To achieve the FAIR objectives for research data set out in NASA's SPD-41, any utilization of a generalist repository for astronomical, planetary or solar data should include mechanisms to enable the pre-publication review and curation of such data. Generalist repositories endorsed by NASA should include open curation platforms for data scientists or data librarians to participate in the refinement of NASA-funded research data products. NASA support of curation as a base condition for research data will, in turn, engender the trust of the community and foster further data sharing. Unreviewed data do not further the cause of open science and in some cases distort the scientific record.

II.C Partner with Extant Repositories for the Interoperability of Published Datasets

Domain-specific repositories greatly enhance the interoperability and reuse of NASA data. This enhancement is especially powerful compared to data quickly deposited to generalist or institutional repositories sans curation. Many of the existing NASA and international repositories have active relationships with the AAS Journals to help authors to create and the Journals to improve data links in published articles. These existing archives have already implemented many protocols and services provided by the virtual observatories. NASA should partner with extant domain-specific repositories, including international partners, to enhance how its data are deposited, harvested, and made interoperable.

II.D Further Develop Domain-Specific Metadata

Successful reuse and interoperability of astronomy research data depends heavily on existing community-specific metadata schema such as the standards of the International Virtual Observatory Alliance (IVOA). The FAIR principles set out in NASA's SPD-41 are concerned with collection-level metadata only; NASA should apply this collection-level metadata only in conjunction with more granular domain-specific metadata. NASA's ADS provides a community clearinghouse for the interface between the literature and data centers and the metadata required to make it effective.

AAS is the steward of the freely available, open-source, community supported Unified Astronomy Thesaurus (Frey & Accomazzi 2018; https://astrothesaurus.org/) and asks authors to use UAT concepts to tag their AAS journal articles for semantic enrichment of the literature for better search and discovery. NASA should endorse the adoption of this controlled vocabulary for all astrophysics publications and data funded by NASA.

II.E Foster and Improve Software Citation

Authors submitting to AAS journals can have their software reviewed in parallel by The Journal of Open Source Software (JOSS) (Vishniac & Lintott 2018) and are given guidance on where to deposit their software and how best to cite it (Vishniac & Lintott 2016). The AAS has worked with NASA ADS and CERN's data repository, Zenodo.org, on the Arthur P. Sloan Foundation funded project Asclepias (Muench et al. 2020; https://asclepias.aas.org/), which has just released a portal and broker for connecting software tools with scientific results to make research progress in astronomy faster, more open, and more reproducible.

As a result of this we have identified a preeminent challenge in the software citation ecosystem: a piece of software can be identified in many ways. There are journal articles describing the use of the software in science, and other publications, e.g., JOSS, that review and validate the quality of the software. Software can have identifiers assigned by libraries or catalogs while there are also persistent identifiers for each version of the software. These different identifiers are not redundant; each provides unique information necessary to ensure incremental credit for software developers and precise replication of results in journal articles.

The large number of software identifiers leads to a problem for the researcher trying to cite software for attribution and reproducibility: these identifiers are *disparate* and *disconnected*. None of the extant services (e.g., NASA ADS; The Astrophysics Source Code Library, https://ascl.net; the Asclepias broker) links these different identifiers in a way that improves their discovery and citation. NASA should fund an ongoing effort to link and collate these identifiers to improve the discovery, reuse, and citation of these software products.

Summary

Based upon 30 years of trial-and-error in data publishing, the AAS Journals do not think that a "one size fits all" approach to datasets will work for improving open science and data publication. Small- and medium-sized datasets can be curated and securely published in peer-reviewed journal articles, which are the nexus linking research literature, datasets, and software objects. The AAS Journals' open-access articles contain data curated by our data editors and harvested into interoperable repositories such as CDS/Vizier or NASA's NED. Supplemented by indexing and linking provided by NASA's ADS, these articles represent some of the most "FAIR" data published in astronomy. Such datasets are far more valuable and reusable for the community than data deposited to generalist repositories. Policy mandates meant to accelerate data openness must consider the need to document and curate them, which inevitably requires domain-specific infrastructure. In summary we think NASA should:

- Guide researchers on publishing small- to medium-sized datasets in journal articles versus larger data products that are best delivered to a NASA-funded repository.
- Ensure open data products are properly curated by data editors, librarians, or repository staff, including the application of granular domain-specific metadata.
- Partner with extant domain-specific repositories to support the ingest and aggregation of NASA datasets, making them truly interoperable and reusable.
- Support improved metadata for data and software by adopting the Unified Astronomy Thesaurus and funding a solution to the many-identifier problem for software citation.

References

Chen, T. X., Schmitz, M., Mazzarella, J. M., et al. 2021. *Best Practices for Data Publication in the Astronomical Literature*, arXiv:2106.01477

Frey, K., Accomazzi, A. 2018. *The Unified Astronomy Thesaurus: Semantic Metadata for Astronomy and Astrophysics.* The Astrophysical Journal Supplement Series 236, 24. doi:10.3847/1538-4365/aab760 | https://astrothesaurus.org/

Muench, A., Accomazzi, A., Holm Nielsen, L., et al. 2020. *Asclepias: An Infrastructure Project to Improve Software Citation across Astronomy.* Astronomical Data Analysis Software and Systems XXVII, 522, 711. bibcode:2020ASPC..522..711M

Vishniac, E. T., Lintott, C. 2018. *Editorial: A Cooperative Agreement with the Journal of Open Source Software.* The Astrophysical Journal 869, 156. doi:10.3847/1538-4357/aaf876

Vishniac, E. T., Lintott, C. 2016. *Editorial: The AAS Journals Corridor for Instrumentation, Software, Laboratory Astrophysics, and Data*. The Astronomical Journal 151, 21. doi:10.3847/0004-6256/151/2/21 | http://journals.aas.org/news/policy-statement-on-software/

NOI Number: N2-SPD41RFI-0034