The Roles of Data Editors in Astronomy

August Muench

Introduction

In 2000, the Journals of the American Astronomical Society (AAS)[1] completed the adoption of electronic editions for all its journal titles. Writing at the time, Editor-in-Chief Robert Kennicutt asserted that "electronic publishing made it possible to efficiently archive scientific contents that never could be reproduced on a printed page … range[ing] from extended data tables that are useful only in machine-readable form to various forms of on-line graphics …".[2] These data and online graphics were also explicitly considered part of the article's version of record. They were expected to be contextually integrated into the flow of the article rather than relegated to a paragraph after, or footnote under, the text. Kennicutt also recognized that while the journals could house small- to medium-sized datasets permanently, they needed to support larger datasets, which required cross-linking the data in astronomy archives with the articles.

By enveloping its digital assets as part of an article's version of record, the journals established a clear editorial need. Thus in 2000, the AAS employed its first Data Editor: a PhD-trained astrophysicist whose role is to support and assist researchers with submissions of data and graphics, curate and standardize data for improved reuse and long-term storage, and review and establish links to data centers. The role of the AAS Data Editor, which expanded to a journal staff of three full-time PhDs in 2023, has the expressed goals of increasing data sharing and improving the overall published results. How this role is engaged and achieves these goals is an evolving and expanding story that is driven both by its origin and by rapid changes in the treatment of data and software in scholarly journal articles today.

August Muench (https://orcid.org/0000-0003-0666-6367) is Journals Data Editor, American Astronomical Society.

https://doi.org/10.36591/SE-D-4601-04

Categories and Growth of Data Editing

The types of data products curated by AAS Data Editors and the corresponding curves of growth over the past 22 years are graphed in Figure 1. For the purposes of explaining the editing that we do, I have simplified the different article elements into graph categories: "data" and "interactive graphics" and "data links". In practice, data comprise multiple ways of contextualizing data in an article. These elements include machine-readable tables or different types of "data behind a figure." Similarly, readers would experience interactive graphics in online AAS journal articles in various forms: a browsable atlas of related figures, animated figures streamed into the journal article, and HTML5/JavaScript-driven interactive figures. Finally, data links include dataset DOIs from, or data citations to, astronomy archives. This also includes data and software deposited into domain-specific archives or into generalist repositories as part of the review and publication process.

Data Review

For the first 17 yr, the data editors focused on the data products only after a manuscript's acceptance. However, many of the necessary curation tasks are ones only the authors can complete. A good example is enforcing compliance with the AAS Software Policy.[3] As our duties expanded, we asked the question: "Can we change data editing from a serialized process that begins only after a manuscript is accepted to a parallel one?"

In 2017, we implemented a data review of initial manuscript submissions; approximately 90% of all peer-reviewed manuscripts are given this initial data review, which is a considerable effort for the roughly 5200 articles published annually by the AAS. The AAS Data Editor review is included with the first peer-review report for the authors to consider and respond. This review has a large scope. Examples include asking for the data in a particular figure; suggesting data or software citations missing from an article; and informing authors of journal data and graphic requirements, including the need to create accessible captions for animated figures. This is also the prime time to solicit authors to move data that is too large or disconnected from the text, as well as their software artifacts, to persistent repositories. Authors are directed to our helpdesk if they

CONTINUED

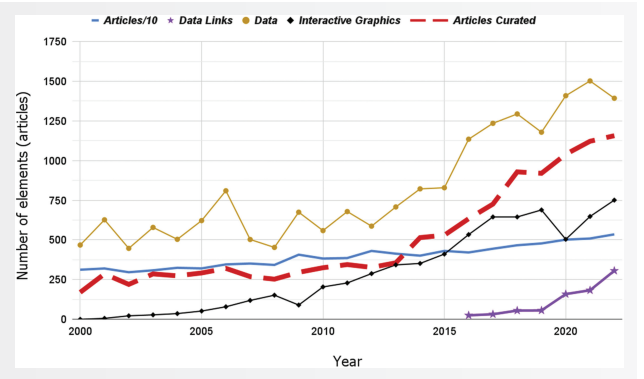

Figure 1. Growth rate of data, online graphics, and data links in American Astronomical Society journal articles. The number of individual elements published annually is compared to the growth rate of total articles (divided by 10, blue solid line) and articles curated by the Data Editors (red dashed line).

have questions or encounter problems implementing our data review requests.

The AAS journals, like many society journals in astronomy, have a very high acceptance rate (more than 85%); our primary editorial mission is to validate, improve, and preserve the astronomy literature. Thus, providing data review at submission leads to little or no wasted effort on rejected manuscripts. Some gains from adding the initial data review were extensive—examples include increasing software citations and improved compliance for descriptive captions of animated figures. Other efforts are rate-limited—we can't ask all authors to submit all data for all figures because we are not capable of curating it all.

Curation

Curation of data occurs after the manuscript has completed peer review and been accepted by the Science Editor. It involves a series of regulated processing steps by the Data Editors that have evolved little except as the domain boundaries of our journals have edged into new fields, such as planetary science. To start, all large tables are standardized, including column labels, units, descriptions, and references, to an ASCII format developed and shared by the Strasbourg Astronomical Data Center (CDS),[4,5] and is used for data in other astronomy journals.[6] Besides providing readers with regularized, curated content, adopting such a format allows information systems such as CDS/VizieR[7] to straightforwardly ingest and serve the calculations using the application programming interfaces (APIs) of the Virtual Observatory.[8] In addition to regularization, we check that author-supplied data adheres to the NASA/IPAC Extragalactic Database's best practices for data publication.[9]

Data-behind-the-figure can be submitted in tabular form (e.g., the data points in a scatter plot, or in several astronomy- or instrument-specific formats, such as FITS)[10] images and spectra. Again, the data are regularized by the editors, especially if the files supplied were headerless. These refined data files are bundled with documentation, which is provided as a template-based ReadMe file, into archive files that are then linked to the related figure's caption in the published article. Between 60%–70% of the manuscripts we work on have curated data in machine-readable tables or as data behind the figures.

Interactive Figures

Data Editors have important roles in the creation of interactive online graphics, beginning with making authors aware that they can provide their readers with something other than the static figures constrained to the PDF copy. The Data Editors build the LaTeX framework that is transformed into online flipbooks or atlas of related figures. We edit authors' animations of observational data, such as the 12-s cadence of multi-spectral images from the Solar Dynamics Observatory or of model and simulation data. Finally, Data Editors guide authors when building HTML5/JavaScript graphics. We give authors direction, edit their figures in Jupyter notebooks, and help to package and document the data that support such figures, which will out-live any technologies used to render the graphics on the Web. Authors have embraced the functionality, creating over 150 JavaScript graphics using technologies such as Bokeh, Plot.ly, and X3D.*

Repositories, Software, and Data Linking

A rapidly growing effort involves authors' use of generalist repositories for hosting data and software related to articles. Surprisingly, there are few domain-specific repositories in astronomy that accept prepublication datasets. This means authors end up using generalist repositories such as Zenodo[12] and Harvard Dataverse[13] if they do not or cannot submit the data to the journals. While these repositories can include rich metadata for deposits, the quality of that metadata depends on guidance and training, which for our authors, come from the Data Editors via data review, curation, and help desk inquiries.

For software, we have evolved from originally soliciting software code to be hosted in the journals in archive files to encouraging authors to provide openly licensed software on collaborative codebases, such as GitHub, and release versions of their code into persistent repositories (e.g., Zenodo). In addition to directing authors to undertake these steps and checking the resulting deposit metadata, the Data Editors also ensure that the software is correctly cited in the final manuscript.

Finally, the AAS has spearheaded efforts to link to important, related data sets in federally funded data repositories, such as the Barbara A. Mikulski Archive for

CONTINUED

Space Telescopes (MAST).[14] Built into our submission system is a prompt for authors asking if they've used data from MAST or IPAC at Caltech.[15] The prompt directs them to find and insert (or cite) the corresponding datasets by their digital object identifiers (DOIs). Data Editors use the authors' responses to these submission prompts to provide additional instruction via initial data review.

Challenges and Future Directions

Two major challenges face the AAS Data Editors. One is how to raise our curation rate to cover a much larger fraction of the literature. In Figure 1, an inflection point in the growth rate of curated articles occurred in 2014, which was when I was hired as the second AAS Data Editor. Even with a third Data Editor starting in 2023, reaching higher will be a challenge. It may be that many of the tasks for data review can be automated using tools for entity extraction or identifying deficiencies in open data and code. This would be akin to the automated peer-review steps in the ScreenIT pipeline for COVID-19 preprints.[16,17] The second major challenge is one of repositories. Data submitted without curation to generalist repositories are of limited value and may be missing critical details, for example, units, which are necessary for either human or machine reuse. Funder mandates to expand the amount of openly accessible data and software could exacerbate this problem unless the Data Editors can keep up.

The future of our data editing almost certainly lies in embracing fully reproducible research. Efforts coming out of astronomy such as "showyourwork!"[18] point directly at articles bound together with the data and software artifacts into complete, compilable results. This future is also apparent in the genesis of Data Editors in fields such as economics,[19,20] ecology, evolution, and behavior research,[21,22] and evolutionary biology,[23] which are focused on the replication of results. Spinning up computing resources to validate such packages may pose a future challenge.

Journals that want to include data editing have many options. Initial data review, which we found to have enormous benefits, could be done by soliciting volunteer data reviewers to perform these tasks quickly at the start of a submission. Developing new author guidance on data and software can follow the recent work of other journals; examples include the guidance developed by the American Geophysical Union[24] or the Data and Code Availability Standard.[25] While our AAS curation tasks seem in-depth, reviewing the metadata and contents of the deposits made by authors to repositories is a simple way to ensure that reusable and good quality data and code are being linked to your final articles. Partnering with repositories, such as Dryad, is another way to add deposit curation at a high-level or even domain-specific depth.

*The full set of ~8000 online graphics can be browsed using the AAS Astronomy Image Explorer.[11]

References and Links

1. https://journals.aas.org/
2. Kennicutt RC Jr. Editorial: electronic supplemental materials for the Astrophysical Journal. Astrophys J. 2001;557:1. https://doi.org/10.1086/323678
3. Vishniac ET, Lintott C. Editorial: the AAS journals corridor for instrumentation, software, laboratory astrophysics, and data. Astronomical J. 2016;151:21. https://doi.org/10.3847/0004-6256/151/2/21
4. https://cds.u-strasbg.fr/
5. Ochsenbein F. Standard for astronomical catalogues and tables, version 2.0. http://vizier.u-strasbg.fr/doc/catstd.htx
6. Brouty M, Woelfel F, Bruneau C, Brunet C, Claude H, Dubois P, Eisele A, Genova F, Lesteven S, Neuville N, et al. Information scientists: between editors and data centers. 2010;433:195. https://ui.adsabs.harvard.edu/abs/2010ASPC..433..195B
7. https://vizier.cds.unistra.fr/
8. https://ivoa.net/
9. Chen TX, Schmitz M, Mazzarella JM, Wu X, van Eyken JC, Accomazzi A, Akeson RL, Allen M, Beaton R, Berriman GB, et al. Best practices for data publication in the astronomical literature. Astrophys J Suppl Ser. 2022;260:5. https://doi.org/10.3847/1538-4365/ac6268
10. https://fits.gsfc.nasa.gov/
11. https://www.astroexplorer.org/
12. https://zenodo.org/communities/aas
13. https://dataverse.harvard.edu/dataverse/aas
14. https://archive.stsci.edu/publishing/doi
15. https://www.ipac.caltech.edu/dois
16. Schulz R, Barnett A, Bernard R, Brown NJL, Byrne JA, Eckmann P, Gazda MA, Kilicoglu H, Prager EM, Salholz-Hillel M, et al. Is the future of peer review automated? BMC Res Notes. 2022;15:203. https://doi.org/10.1186/s13104-022-06080-6
17. Weissgerber T, Riedel N, Kilicoglu H, et al. Lessons learned from automated screening of COVID-19 preprints. Eur J Public Health. 2022;32(Supplement_3). https://doi.org/10.1093/eurpub/ckac129.127
18. https://show-your.work/en/latest/
19. Duflo E, Hoynes H. Report of the search committee to appoint a data editor for the AEA. AEA Papers Proceed. 2018;108:745. https://doi.org/10.1257/pandp.108.745
20. Vilhuber L. Report by the AEA data editor. AEA Papers Proceed. 2022;112:813–823. https://doi.org/10.1257/pandp.112.813
21. http://comments.amnat.org/2021/12/guidelines-for-archiving-code-with-data.html
22. http://comments.amnat.org/2020/10/data-archiving.html
23. https://jevbio.net/data-editor-position-available-at-the-journal-of-evolutionary-biology/
24. Fox P, Erdmann C, Stall S, Griffies SM, Beal LM, Pinardi N, Hanson B, Friedrichs MAM, Feakins S, Bracco A, et al. Data and software sharing guidance for authors submitting to AGU journals. Zenodo. https://doi.org/10.5281/zenodo.5124741
25. Koren M, Connolly M, Lull J, Vilhuber L. Data and code availability standard. Zenodo. https://doi.org/10.5281/zenodo.7436134