Proposal

Notebooks Now! Elevating notebooks into scholarly publishing.

Proposal by the American Geophysical Union (AGU) and the Notebooks Community

Problem Statement

The scientific community is increasingly using computational notebooks for executing, managing, and sharing their workflows and analyses, often in cloud-based computing clusters. Popular notebook tools include Jupyter[1] and R Markdown[2] among other resources, and tools that integrate and execute these together, such as Binder. These environments allow for project work to be combined including the data manipulation leading to the final results, the linked software involved in that analysis and manipulation, and the narratives that in effect form the methodology, protocols, and results of a study. Currently, these are then extracted, with loss of functionality and integration, to form a research publication, which is primarily still text-based.

Thus providing notebooks as available and curated research outputs would greatly enhance the transparency and reproducibility of research, integrating into computational workflows. The notebooks allow deeper investigations into studies and display of results because they link data and software together dynamically with what are often final figures and plots. Unfortunately, the current peer-review and publications workflows across the sciences do not readily support notebooks as research outputs or encourage their use and curation. Currently, few publishers allow these as linked supplements. AGU recently developed author instructions (Erdmann, 2021) for depositing them in repositories. Notebooks are not included in the paper peer-review workflow, inhibiting a deeper evaluation by reviewers into the data processing and thus results.

[1] https://jupyter.org/
[2] https://rmarkdown.rstudio.com/

We propose to develop a better approach: an end-to-end scholarly publishing workflow that would treat notebooks, both Jupyter and R Markdown, as a primary element of the scientific record. **This would include an approach where the notebook is the submitted product and is available natively for peer-review.** We intend to transform the publication process in a way that elevates transparent and reproducible work by authors, where data and software, together with narrative, are efficiently documented and shared, where access to computation is more equitable, and where new forms of credit can be extended to the wider research community, including research software engineers or RSEs. Metadata will be extracted to provide well-established publication and discovery services. We envision that certain standards around notebooks would be needed to enable such an end-to-end workflow, for example, around copy-editing, production, platforms, and configuration. The goal would be to maximize functionality and simplify author requirements. We expect that new publication platforms and methods may be needed or current platforms will need to evolve in significant ways.

The signatories to this proposal represent a Steering Committee across the key stakeholders to envision and guide the development of this end-to-end workflow. We propose to engage a larger set of stakeholders (~70 in-person, ~20 virtual) to develop this model in a series of three workshops (one in-person/hybrid, two virtual) with several workstreams in-between focused on steps in the process (e.g., (pre-)submission, review, publication). These workshops will help to align the collective guidance, requirements, and common solutions from the group to support the full-workflow vision and to get buy-in. This proposal is aimed at visioning and designing. It will prepare us for the important steps of implementation. The planned deliverable is a complete design for an end-to-end workflow that includes all stages of the publication process needed for an interactive notebook.

All of the deliverables, documentation, and methods, including the pilot project work, will be provided openly and designed around open standards for broad adoption. Having a standard model for publishing notebooks allows for all publishers to support the growing community of notebook development across all scientific disciplines.

The proposed budget is designed to support an 18-month effort with one hybrid workshop and two virtual meetings, dedicated project management for the full time of the grant, pilot projects to demonstrate solutions, and honorarium for the Steering Committee to incentivize work on this effort. AGU will provide in-kind meeting space for the in-person/hybrid workshop.

Related Work

Notebooks offer the ability to blend narrative, code, analysis, and results in a more seamless, interactive experience, where research can be both read and replicated (Kluyver, 2016). However, much of that experience is lost, i.e., that native interaction, either through providing the notebook as a supplement to a research article or as the final published work. In fact, as open science becomes the norm, authors are asking for a more streamlined experience (Dorigatti, 2021), where code and data are shared together with the narrative. An end-to-end solution that makes it easier for authors to publish notebooks natively via journals is missing where stakeholders from publishing to the researcher/developer community for notebooks can benefit from working together to create richer connections.

According to a 25 March 2022 Lens.org search, there is broader adoption and use of notebooks to communicate research across the list of major academic fields. There is greater use of Jupyter in the Computer Science, Physics, Astronomy, Biology, and Data Science domains

while RStudio[3] is more prevalent in the Medicine, Biology, Psychology, and Computer Science domains. This is also reflected in the prevalence of Jupyter use in domain-centric journals/platforms such as IEEE versus BioMed Central for RStudio.

There is even greater use of notebooks when you look at GitHub[4], whereas as of March 29, 2022, there were over 8.1M search hits for the Jupyter ".ipynb" file extension[5] versus the 1.3M notebooks reported in July 2017 (Rule *et al.*, 2018). According to Rule *et al.*, notebooks are often used for exploration purposes. This is also evidenced through their use at computational clusters from Iowa State University[6] to Caltech[7] and cloud environments such as Azure[8] and Pangeo[9]. But there is a tension between exploration and presentation, as it can take time for researchers to clean and prepare their notebooks for greater sharing.

However, the growth in the use of notebooks has not necessarily translated to stable and citable FAIR digital research objects in the scholarly ecosystem (Wofford, 2019). Likely this is a result of the slow adoption of standard approaches, streamlined integrations for authors, and easy-to-follow guidance, but also the lack of overall recognition for this potential new form of publishing (DuPre, 2022). On the other hand, there are signs that this is changing, for instance, in the case of the Earth and space sciences, where initiatives like Pangeo and NASA TOPS[10] are driving the adoption of notebooks, societies like AGU are exploring how to support and recognize them via their meetings and publications, and authors are looking to publish them.

[3] https://www.rstudio.com/
[4] https://github.com/
[5] https://nbviewer.org/github/parente/nbestimate/blob/master/estimate.ipynb
[6] https://www.hpc.iastate.edu/guides/jupyterhub
[7] https://www.hpc.caltech.edu/documentation/software-and-modules/jupyter-notebook
[8] https://visualstudio.microsoft.com/vs/features/notebooks-at-microsoft/
[9] https://pangeo.io/
[10] https://github.com/nasa/Transform-to-Open-Science

The current state of notebook sharing and "publishing" is disjointed and not well standardized. Rule *et al.* offered a set of rules, tips, tools, and examples that authors should be mindful of when developing and sharing notebooks (Rule *et al.*, 2019). The Rule *et al.* paper is a helpful place to start, but this guidance is not a scalable publishing solution, and additional steps are needed to make notebooks available, published, and to provide credit. This led AGU to develop guidance for both the Jupyter (Python) and R communities (Erdmann, 2021) and EarthCube[11] to develop notebook templates[12] for submitters as well as a peer review rubric (Giglio, 2022) for their yearly meeting calls for notebooks. Later rOpenSci posted guidance as well.[13] While these resources answered some more immediate questions from authors, they still left some unanswered, for instance, a streamlined approach to publishing notebooks as an article, and the supplemental approach still created a disjointed experience where the notebook experience was not fully integrated into the research article. Still, the resources above offer a framework for standardizing and templatizing notebooks. Best practices from them can be used to structure narrative for publication from the process of exploratory analysis.

AGU's current recommendations support the more popular end-to-end workflow that involves hosting, collaborating, and developing Jupyter Notebooks using GitHub (which, can also be used to render/display Notebooks), and creating/linking to a runnable version via Binder[14] while preserving and citing them via a DOI using Zenodo[15]. This creates a somewhat disjointed experience but does speak to current open and FAIR (Findable, Accessible, Interoperable, Reusable) scholarly practices. Alternatively, authors can attach notebooks as supplementary files, for certain publication venues (e.g., ESSOAr[16]). This removes the notebook from the

[11] https://www.earthcube.org/
[12] https://github.com/earthcube/NotebookTemplates
[13] https://ropensci.org/blog/2021/11/16/how-to-cite-r-and-r-packages/
[14] https://mybinder.org/
[15] https://zenodo.org/
[16] https://www.essoar.org/

runtime environment, potentially does not preserve the file, and ultimately does not result in a FAIR digital object. Yet one more approach, exemplified via a service like Hydroshare[17], CUAHSI's online collaboration environment for sharing data, models, and code, allows authors to deposit notebooks alongside their data, code, and other additional files, providing a runtime environment for the notebook and other objects, but also allows the author to create a package or composite with the notebook, and cite with the research paper. This object provides certain benefits per FAIR but still creates a separate experience and some of the granularity of referencing solely the notebook is lost.

Jupyter Book[18] is an emerging option for publishing notebooks natively to the web. It supports the MyST Markdown which allows authors to include citations and cross-references with also some more complex functionality like adding content to the margins. The Canadian Open Neuroscience Platform (CONP) is one example of a group that is leveraging Jupyter Book to create a repository of neuroscience notebooks called NeuroLibre[19]. Computation of the notebooks is supported locally via their compute infrastructure versus leveraging Binder. EarthCube used a similar setup for their 2021 call for notebooks[20], in this case leveraging Binder for execution and Zenodo for the registration of DOIs. While Jupyter and Jupyter Book offer export capabilities to formats accepted by publishers, still work needs to be done to structure, mark up, and copy edit notebooks to make them suitable for publication (in JATS XML[21]).

[17] https://www.hydroshare.org/
[18] https://jupyterbook.org/
[19] https://www.neurolibre.com/
[20] https://earthcube2021.github.io/ec21_book/docs/
[21] https://jats.nlm.nih.gov/

Regarding web authoring tools, there are a number of examples. Authorea[22] (focuses on native web experience) and Overleaf[23] (has a LaTeX[24] focus) are popular tools for authoring manuscripts collaboratively online and submitting them to journals. Authorea offers one approach of embedding notebooks in the manuscripts, either through its own native environment or via Binder (for newer versions) and can leverage Crossref[25] DOIs for the container/forward-facing manuscript for the notebooks. Overleaf offers a hook to allow for manuscripts in its interface to be updated by the analysis in a connected notebook. Curvenote[26], a relatively new collaborative authoring tool, integrates collaborative authoring with the notebook so that an author can comment back and forth more natively in the web environment, while also creating hooks where the analysis in the manuscript can be automatically updated from the notebook. Additional authoring solutions include CoCalc[27] and Google Colab[28]. What is interesting about these tools is that they can serve both as exploratory/learning resources while also speaking to the online collaborative authoring scenario as well.

When it comes to accepting collaboratively developed notebooks, editorial management and manuscript submission systems are primarily structured towards journal articles and notebooks are seen as supplemental material. Some of the most widely used systems include ScholarOne[29], Editorial Manager[30], Open Journal Systems[31], and eJournal Press[32] which AGU uses. These systems leverage templates often in the form of Microsoft Word documents or

[22] https://www.authorea.com/
[23] https://www.overleaf.com/
[24] https://www.latex-project.org/
[25] https://www.crossref.org/
[26] https://curvenote.com/
[27] https://cocalc.com/
[28] https://colab.research.google.com/
[29] https://clarivate.com/webofsciencegroup/solutions/scholarone/
[30] https://www.ariessys.com/software/editorial-manager/
[31] https://pkp.sfu.ca/ojs/
[32] https://www.ejournalpress.com/

LaTeX which ultimately get structured into the Journal Article Tag Suite (JATS), an XML format used to publish scientific literature, which is only applied to journal articles. A comparative analysis of editorial management and manuscript submission systems identified the common functionality that exists between the systems but concluded that they also needed to evolve to meet the needs of a rapidly changing scholarly ecosystem (Kim, 2018). Further integration with the research lifecycle, collaboration, visualization, linking to resources, rethinking contributor roles, these were all seen as aspects that needed to be advanced in these systems. By reframing the primary form of communication around notebooks, all these potential areas of development can be explored and reimagined in these systems. JOSS[33] and JOSE[34] are journals that offer a window into ho how editorial management and manuscript submission systems can be reimagined and repositioned to tap into researcher workflows, for instance leveraging a GitHub workflow approach (e.g., review via pull request).

The concept of a research compendium[35] came up often while AGU was working with community members to develop R-related guidance for publishing notebooks. Simply defined, a "research compendium" accompanies, enhances, or is a scientific publication providing data, code, and documentation for reproducing a scientific workflow. The compendia approach, and the flexibility of R and R Markdown, might account for why the concept of publishing notebooks was new to R users in the AGU community. R users in the AGU community also referenced the Open Science Framework (OSF)[36] as a helpful platform for collaborating, managing, and sharing the compendium of resources. eLife is one publisher that has tapped into the compendia concept. They feature an online collaborative authoring experience that also supports a runtime environment for semantically structuring/styling your project for export in

[33] https://joss.theoj.org/
[34] https://jose.theoj.org/
[35] https://research-compendium.science/
[36] https://www.cos.io/products/osf

multiple, machine-readable formats. eLife calls their approach executable research articles, or ERA, through the use of Stencila, which supports executable document pipelines (Tsang & Maciocci, 2020).

eLife has adopted an open-source end-to-end approach. Meanwhile, a number of AGU authors still leverage commercial/closed solutions such as Matlab[37] and IDL[38]. There have been community efforts to port work using these tools to more open solutions, especially in the Python and R communities (Perkel, 2018).

As identified by Pimentel *et al.*, there is an opportunity to reduce the rate of bad practices in notebooks while raising the rate of good practices and overall reproducibility (Pimentel *et al.*, 2019). This can range from the use of literate programming practices to declaring dependencies. Through the further development of notebooks as a primary element of the scientific record, we can also address big challenges such as citing and crediting dynamic, multiple datasets[39], and software dependencies (Druskat, 2019).

[37] https://www.mathworks.com/products/matlab.html
[38] https://www.l3harrisgeospatial.com/Software-Technology/IDL
[39] https://data.agu.org/DataCitationCoP/

List of Citations

Erdmann, Christopher, Stall, Shelley, Gentemann, Chelle, Holdgraf, Chris, Fernandes, Filipe P. A., Gehlen, Karsten Peters-von, & Corvellec, Marianne. (2021). Guidance for AGU Authors - Jupyter Notebooks. Zenodo. https://doi.org/10.5281/zenodo.5651648

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B.E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J.B., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., & Team, J.D. (2016). Jupyter Notebooks - a publishing format for reproducible computational workflows. ELPUB. https://doi.org/10.3233/978-1-61499-649-1-87

Dorigatti, E. (Guest). (2021, June). Open Science Stories: Sharing source code [Audio podcast]. Season 1, Episode 9. Retrieved from https://anchor.fm/opensciencestories/episodes/S1E9-Emilio-Dorigatti---Sharing-source-code-e107goi.

Rule, A., Tabard, A., & Hollan, J. D. (2018, April). Exploration and explanation in computational notebooks. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-12). https://dx.doi.org/10.1145/3173574.3173606

Wofford, M. F, Boscoe, B. M, Borgman, C. L, Pasquetto, I. V, & Golshan, M. S. (2019). Jupyter notebooks as discovery mechanisms for open science: Citation practices in the astronomy community. UCLA: Center for Knowledge Infrastructures. http://dx.doi.org/10.1109/MCSE.2019.2932067

DuPre E, Holdgraf C, Karakuzu A, Tetrel L, Bellec P, et al. (2022) Beyond advertising: New infrastructures for publishing integrated research objects. PLOS Computational Biology 18(1): e1009651. https://doi.org/10.1371/journal.pcbi.1009651

Rule A, Birmingham A, Zuniga C, Altintas I, Huang SC, et al. (2019) Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. PLOS Computational Biology 15(7): e1007007. https://doi.org/10.1371/journal.pcbi.1007007

Erdmann, Christopher, Meyer, Michael F., Little, John R., Hondula, Kelly, Stachelek, Jemma, Oleksy, Isabella, Brousil, Matthew R., Claborn, Kelly, Mesman, Jorrit, & Dennis, Tim. (2021). Guidance for AGU Authors: R Script(s)/Markdown. Zenodo. https://doi.org/10.5281/zenodo.5647998

Giglio, Donata; Carter, Shay, R.; Chastang, Julien, C; McHenry, Kenton; Schreiber, Lynne; Zaslavsky, Ilya (2022). Evaluation Rubric for Reviewing Standardized Notebooks. In EarthCube Organization Materials. UC San Diego Library Digital Collections. https://doi.org/10.6075/J0V40VCV

Kim, S., Choi, H., Kim, N., Chung, E., & Lee, J. Y. (2018). Comparative analysis of manuscript management systems for scholarly publishing. In Science Editing (Vol. 5, Issue 2, pp. 124–134). Korean Council of Science Editors. https://doi.org/10.6087/kcse.137

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A.,

… Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Tsang, E., & Maciocci, G. (2020). Welcome to a new ERA of reproducible publishing. eLIFE Labs. https://elifesciences.org/labs/dc5acbde/welcome-to-a-new-era-of-reproducible-publishing

Perkel, J. M. (2018). Why Jupyter is data scientists' computational notebook of choice. Nature, 563(7732), 145-147. https://doi.org/10.1038/d41586-018-07196-1

Pimentel, J. F., Murta, L., Braganholo, V., & Freire, J. (2019, May). A large-scale study about quality and reproducibility of jupyter notebooks. In 2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR) (pp. 507-517). IEEE. http://www.ic.uff.br/~leomurta/papers/pimentel2019a.pdf

Druskat, S. (2019). Software and dependencies in research citation graphs. Computing in Science & Engineering, 22(2), 8-21. https://elib.dlr.de/133021/1/druskat-cise-2019.pdf