POLYPUBLIE
Polytechnique Montréal

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

Titre: Beyond advertising: New infrastructures for publishing integrated research objects
Title:

Auteurs: Elizabeth DuPre, Chris Holdgraf, Agah Karakuzu, Loïc Tetrel, Pierre Bellec, Nikola Stikov, & Jean-Baptiste Poline
Authors:

Date: 2022

Type: Article de revue / Article

Référence: DuPre, E., Holdgraf, C., Karakuzu, A., Tetrel, L., Bellec, P., Stikov, N., & Poline, J.-B. (2022). Beyond advertising: New infrastructures for publishing integrated research objects. PLOS Computational Biology, 18(1), e1009651 (7 pages). https://doi.org/10.1371/journal.pcbi.1009651
Citation:

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

URL de PolyPublie: https://publications.polymtl.ca/51280/
PolyPublie URL:

Version: Version officielle de l'éditeur / Published version
Révisé par les pairs / Refereed

Conditions d'utilisation: Creative Commons Attribution 4.0 International (CC BY)
Terms of Use:

Document publié chez l'éditeur officiel
Document issued by the official publisher

Titre de la revue: PLOS Computational Biology (vol. 18, no. 1)
Journal Title:

Maison d'édition: Public Library of Science
Publisher:

URL officiel: https://doi.org/10.1371/journal.pcbi.1009651
Official URL:

Mention légale: © 2022 DuPre, E., Holdgraf, C., Karakuzu, A., Tetrel, L., Bellec, P., Stikov, N., & Poline, J.-B. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Legal notice:

Ce fichier a été téléchargé à partir de PolyPublie, le dépôt institutionnel de Polytechnique Montréal
This file has been downloaded from PolyPublie, the institutional repository of Polytechnique Montréal

https://publications.polymtl.ca

PLOS COMPUTATIONAL BIOLOGY

PERSPECTIVE

Beyond advertising: New infrastructures for publishing integrated research objects

Elizabeth DuPre[1]*, Chris Holdgraf[2,3], Agah Karakuzu[4,5], Loïc Tetrel[6], Pierre Bellec[6,7], Nikola Stikov[4,5], Jean-Baptiste Poline[1]*

1 NeuroDataScience–ORIGAMI Laboratory, McGill University, Montreal, Quebec, Canada, **2** The International Interactive Computing Collaboration (2i2c), Berkeley, California, United States of America, **3** International Computer Science Institute, Berkeley, California, United States of America, **4** NeuroPoly Lab, Polytechnique Montreal, Montreal, Quebec, Canada, **5** Montreal Heart Institute, Montreal, Quebec, Canada, **6** Centre de recherche de l'Institut universitaire de gériatrie de Montréal, Montreal, Quebec, Canada, **7** Department of Psychology, Université de Montréal, Montreal, Quebec, Canada

* elizabeth.dupre@mail.mcgill.ca (ED); jean-baptiste.poline@mcgill.ca (J-BP)

Moving beyond static text and illustrations is a central challenge for scientific publishing in the 21st century. As early as 1995, Donoho and Buckheit paraphrased John Claerbout that "an article about [a] computational result is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result" [1]. Awareness of this problem has only grown over the last 25 years; nonetheless, scientific publishing infrastructures remain remarkably resistant to change [2]. Even as these infrastructures have largely stagnated, the internet has ushered in a transition "from the wet lab to the web lab" [3]. New expectations have emerged in this shift, but these expectations must play against the reality of currently available infrastructures and associated sociological pressures. Here, we compare current scientific publishing norms against those associated with online content more broadly, and we argue that meeting the "Claerbout challenge" of providing the full software environment, code, and data supporting a scientific result will require open infrastructure development to create environments for authoring, reviewing, and accessing interactive research objects.

Check for updates

OPEN ACCESS

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

Editor: Feilim Mac Gabhann, Johns Hopkins University, UNITED STATES

Published: January 6, 2022

Copyright: © 2022 DuPre et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was partially funded by awards to JBP from the National Institutes of Health (NIH) NIH-NIBIB P41 EB019936 (ReproNim), NIH-NIMH R01 MH083320 (CANDIShare), and NIH RF1 MH120021 (NIDM); the National Institute of Mental Health under Award Number R01MH096906 (Neurosynth); as well as the Canada First Research Excellence Fund, awarded to McGill University for the Healthy Brains for Healthy Lives initiative and the Brain Canada Foundation with support from Health Canada (NeuroHub, Canadian Open Neuroscience Platform). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Publishing as curating, promoting, and archiving content

Scientific publishing platforms—traditionally, scientific journals—fulfill a variety of roles in their communities. Three of the most prominent of these are curating, promoting, and archiving research. Although these roles have adapted to online spaces, they have not been fundamentally reshaped. Indeed, contemporary scientific articles are disseminated primarily as portable document formats (PDFs), directly translating paper-based workflows into digital workspaces. Here, we briefly review how publishing fulfills these roles today: curation via peer review, short-term promotion via online dissemination, and long-term access via archiving.

Across many kinds of media, curating online content is challenging both due to its scale and its style of interaction, which often blurs the boundary between creating and consuming information. For scientific publishing, formal and independent peer review is widely considered to be a key demarcation [4] and provides an immediate mechanism to curate research objects. Curation in peer review involves checks on a submission's ethical and scientific rigor, in addition to its relevance to a particular research community. Even as many other forms of curation are possible—including crowd sourced or algorithmically driven [5]—these remain relatively uncommon in neuroscience (cf., arxiv-sanity.com).

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: ED, AK, LT, PB, NS, and JBP were all founding members of Neurolibre. CH is a founding member of 2i2c.

In addition to curating (i.e., reviewing and selecting) research objects, publishing also serves an important role in promoting and archiving content. This occurs in the short term through activities such as website hosting and advertising on social media platforms [6]. Ongoing promotion to an ever-evolving scientific community is enabled through the long-term archiving and the references system. These roles can be fulfilled independently or in an arbitrary order. For example, online interactions have allowed peer review to expand into postpublication peer review on platforms such as PubPeer (https://pubpeer.com) and Sciety (https://sciety.org; [7]).

Even as scientific publishers have successfully moved online, they have not yet embraced the full potential of web-first workflows. We briefly review how 2 norms of online content, connectivity and interactivity, are currently reflected in scientific publishing before arguing for infrastructure that allows for more directly interactive and reusable content.

Rich linking for research objects: Connecting through hybrid content types

Much of the rich, content-driven interactivity of the web depends on access to structured data such as user content on social media platforms. To separate out this content from its presentation, data formats such as XML have been developed to link online content with its supporting resources [8]. Although scientific publishing workflows are largely built around the XML format, the need to output PDF documents means that resources that cannot be directly embedded—such as executable code or supporting data—have been largely excluded from academic publishing. Thus, the scientific narrative has historically been detached from its other associated research objects.

Recently, growing awareness of this problem has led to an increase in publishing what we term "hybrid research objects." Hybrid research objects are distinct from traditional publications in that they make multiple content types available in the same object; that is, they contain narrative text and at least one or more examples of code, data, and computation (e.g., [9]). Multiple paths exist to make these objects available. One path is to include direct links to each resource such as through data and code availability statements [10], without constraining their format or content. Alternatively, some publishers require that linked research objects adhere to specified standards and are explicitly included in the review process. For example, the journal *Scientific Data* from Nature Research publishes descriptors of datasets [11] that include links to dedicated, domain-relevant data hosting infrastructure such as OpenNeuro (https://openneuro.org). Importantly, this raises new questions on how to appropriately handle their peer review, questions for which there is no current consensus [12].

As hybrid research objects have become more prominent, best practices in publishing these objects continue to evolve. We hope to see more hybrid research objects where each linked object is formatted with domain-relevant standards (e.g., neuroimaging data organized according to a domain standard such as the Brain Imaging Data Structure (BIDS); [13]) and bidirectionally linked using persistent identifiers. Nonetheless, because the linked research objects are hosted on unique platforms without clear checks on interoperability across the hybrid object components, it can be difficult to interact with the code, data, or their combination, for example, when trying to perform minimal quality checks on a dataset. It further prevents eventual readers from assessing the reproducibility or generalizability [14] of presented results. Enhancing this experience requires making these research objects interoperable, improving their reusability. Here, we introduce the idea of "integrated" research objects to explicitly test the interaction of included research objects in reproducing a scientific result.

Bridging the gaps: Interactive and integrated research objects

Interactivity is an attractive feature of online content and one that scientists have been especially eager to adopt [15]. This enthusiasm has spurred development of platforms such as Bokeh (https://bokeh.org) and Plotly (https://plotly.com), enabling scientists to provide multiple views of their data through interactive figures and dashboards. Although this work is impressive, it is limited: Researchers remain unable to modify or reexecute the code used to generate these figures when shared through HTML documents. This hinders deep engagement with the presented results.

Achieving deeper interactivity requires interaction between the code, data, and computation supporting a scientific result. One approach to achieve this is to focus on what we call "integrated research objects." Integrated research objects not only make multiple kinds of research objects available and tightly coupled, but they also do so in formats (e.g., computational notebooks) that foreground their interaction by allowing reexecution. In doing so, they offer a clear answer to the Claerbout challenge.

There are limits on the kinds of experiments that can be supported through integrated research objects; for example, experiments relying on cell cultures or other biological samples may only have digital representations of the statistical analyses and end results rather than the experiments themselves. Nonetheless, researchers should be encouraged to provide access to research objects that can be digitized. This is particularly important for computational work, where experiments are carried out in silico and so computation and the resulting narrative are closely linked.

Despite their immediate appeal, the infrastructure required to support integrated research objects is less straightforward. In particular, authoring, curating, and archiving these research objects all introduce significant challenges. Further, requiring that these objects be archivable imposes strong constraints on the kinds of technologies that can be used. Most archival services discourage submitting complex HTML objects with external dependencies as these documents are unlikely to retain their full functionality with evolving versions of HTML, JavaScript, and web browsers [16].

To sidestep this concern, current pilots for publishing integrated research objects consider them as secondary to a traditional, archivable article. For example, *eLife* authors can develop additional, web-first materials to accompany their accepted research articles. Codeveloped with Stencila (https://stenci.la), these executable research articles (ERAs) inherit their structure from the Jupyter notebook [17] format. ERA development has explicitly focussed on improving the authoring experience, and authors are supported in ensuring that all relevant code and data files are included in the ERA environment. While this support reduces the technical barrier in creating integrated research objects, it also means that ERAs are necessarily only developed at the end of the publication process after scientific analyses are finalized. In this way, the traditional, narrative text–based document remains privileged as the primary research object.

Centering integrated research objects will require infrastructure development to both ease the authoring experience as well as represent these objects in an archivable format. Although several standards for integrated research objects could serve as potential starting points, we argue that sustainable development demands open standards with multistakeholder governance and leadership to ensure that resulting specifications are not driven by a single stakeholder.

Authoring integrated research objects with open standards

Perhaps the 2 most broadly adopted standards for integrated research objects are the RMarkdown (https://rmarkdown.rstudio.com) and Jupyter notebook [17] formats. Both technologies

allow researchers to create integrated research objects that include narrative text, code, and computation, although they do so using different internal implementations. Specifically, RMarkdown is based on YAML and markdown formats, while Jupyter notebook is based on the JSON format.

Recent development on Jupyter Book (https://jupyterbook.org) has led to the creation of a MyST markdown format (https://myst-parser.readthedocs.io) that extends Jupyter to build from a combination of YAML and markdown, improving handling for scientific publishing use cases. Thus, RMarkdown and MyST allow researchers to directly describe their scholarship—the code, data, and computation that support a given scientific result—such that it can be easily source controlled and archived. They each also enable generation of user-focused HTML and PDF documents, including PDFs formatted for several major scientific journals (using, e.g., "rticles"; [18]), from user-provided markdown content.

These technologies differ, however, in that RMarkdown development is controlled by a single stakeholder, RStudio. Although its product is openly licensed, developed with community consultation, and freely available, decision-making power rests with RStudio employees. This model is distinct from multistakeholder governance, in which formats are not controlled by individual entities but instead benefit from consensus across organizations. We thus focus on standards developed within the Jupyter ecosystem.

Open standards development within Jupyter has enabled other initiatives such as Stencila and Curvenote (https://curvenote.com) to overlay with additional views and functionality. Integrating these technologies into existing standards (e.g., the Journal Article Tag Suite (JATS) XML format) via translation or conversion processes remains an active area of work. Perhaps their largest departure from existing formats, however, is that they can be reexecuted in an integrated computational environment that includes the supporting data files.

Centering complex objects in scientific publishing with cloud infrastructure

Cloud infrastructure enables browser-based access to computational environments. A major challenge in extending these cloud infrastructures for scientific publishing is the associated cost, both for initial peer review as well as for the long-term preservation of included research objects. User-focused cloud technologies such as Binder (https://mybinder.org; [19]) enable easy access to these environments, but they do not directly address dataset storage. Neuroscience datasets may involve terabytes of data and hundreds of CPU hours of compute time, making cloud computing and data hosting nontrivial. Including multiple versions of a given dataset—from raw data to analysis-ready derivatives—only compounds this problem.

Creating economically viable, noncommercial options will likely involve the coordination of multiple academic and nonprofit groups such as the International Interactive Computing Collaboration (2i2c; https://2i2c.org) as well as explicit funding calls for projects advancing open standards through modular, composable infrastructure. Large field standard datasets, such as those provided by the Allen Institute for Brain Science (https://alleninstitute.org) or the International Brain Laboratory [20], are likely to further benefit from centralized data and computation. This approach has been pioneered in geosciences by the Pangeo project [21], which provides centralized access to and computation on field standard climatology data via JupyterHubs hosted on commercial clouds. Recently, Rokem and colleagues [22] have prototyped this approach in neuroscience through the development of a Pan-neuro initiative, encouraging optimism about future adoption in other scientific communities.

Smaller datasets collected by individual research groups, however, may require alternative approaches; in particular, decentralized data management offers a promising route forward to

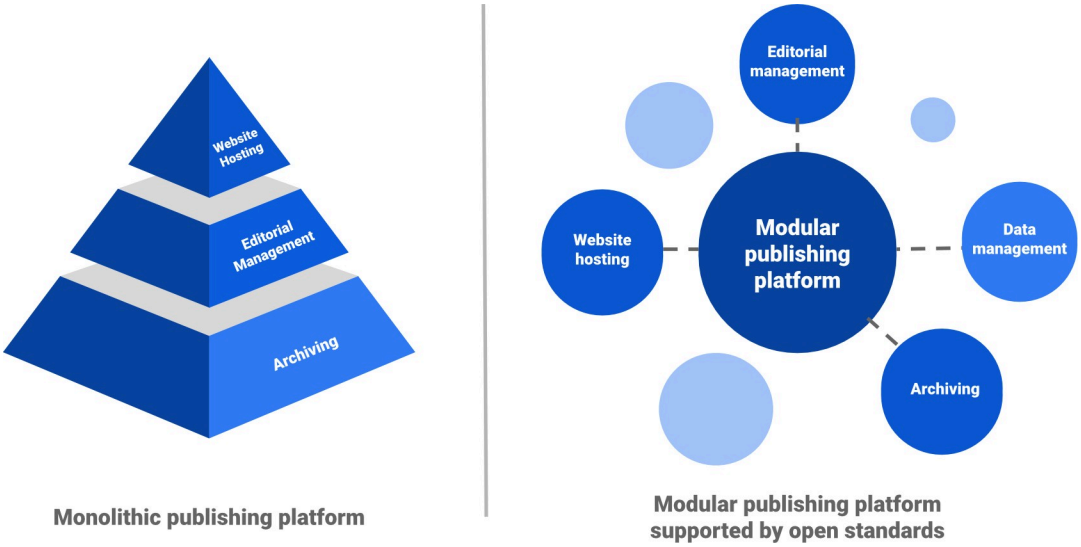

Fig 1. Contrasting monolithic and modular publishing platforms. While monolithic publishing platforms are self-contained, modular publishing platforms rely on open standards across composable infrastructure. In doing so, they create space for additional functionality such as data management that better supports scientific communities.

https://doi.org/10.1371/journal.pcbi.1009651.g001

minimize reliance on a central hosting service in those cases where datasets are small enough to be duplicated [23]. NeuroLibre (https://neurolibre.com) provides one example of this model and relies on nonprofit support to host a curated collection of datasets, each of which support one or more NeuroLibre publications through hosted environments for reexecuting the described analyses.

Although different in scale, we argue that both Pangeo and NeuroLibre share a core approach that should be more broadly adopted. By investing in infrastructure for integrated research objects that heavily relies on open, modular components, we can make strong contributions in individual research domains while still ensuring that these investments can be easily retooled and extended. Fig 1 contrasts this modular, composable infrastructure with more traditional publishing platforms developed on a monolithic technology stack.

NeuroLibre, for example, relies on a combined technology stack from the *Journal of Open Source Software* (JOSS; [24]), Jupyter Book, and BinderHub. Each of these projects independently combines modular technologies to meet existing community needs, and their combination—while currently unique to neuroscience—can easily be repurposed for other research communities, such as the development of Pan-neuro from the Pangeo model.

As scientists increasingly recognize the value in sharing their code and data [25], this approach could facilitate an important transition in scientific publishing. By leveraging MyST as an emerging standard for integrated research objects, alongside modular components for their hosting and reexecution through BinderHub and other open technologies, scientists will be better positioned to author articles that center all the research objects supporting a scientific result, in addition to the underlying narrative.

As science increasingly depends on digital infrastructure, it is clear that scientific publishing is at an inflection point. Reckoning with the Claerbout challenge will require providing access to the research objects supporting the actual scholarship rather than the "advertising" of static scientific articles. Adopting web-based technologies provides the strongest possible path forward, but managing this transition in the face of economic and sociological pressure requires academic communities to advocate for open and sustainable infrastructure development, as

seen in the Pan-neuro and NeuroLibre initiatives. We argue that community-based efforts around open standards, modular and composable infrastructures, and new research object types will underpin the full potential of web-driven publishing.

References

1. Donoho DL. An invitation to reproducible computational research. Biostatistics. 2010; 11(3):385–8. https://doi.org/10.1093/biostatistics/kxq028 PMID: 20538873
2. Piotrowski M. Future Publishing Formats. Proceedings of the 2016 ACM Symposium on Document Engineering DocEng '16. New York, NY, USA: Association for Computing Machinery; 2016. p. 7–8.
3. Keshavan A, Poline JB. From the Wet Lab to the Web Lab: A Paradigm Shift in Brain Imaging Research. Front Neuroinform. 2019; 13:3. https://doi.org/10.3389/fninf.2019.00003 PMID: 30881299
4. Mulligan A, Hall L, Raphael E. Peer review in a changing world: An international study measuring the attitudes of researchers. J Am Soc Inf Sci Technol. 2013; 64(1):132–61.
5. Yarkoni T. Designing next-generation platforms for evaluating scientific output: what scientists can learn from the social web. Front Comput Neurosci. 2012; 6:72. https://doi.org/10.3389/fncom.2012.00072 PMID: 23060783
6. Klar S, Krupnikov Y, Ryan JB, Searles K, Shmargad Y. Using social media to promote academic research: Identifying the benefits of twitter for sharing academic work. PLoS ONE. 2020; 15(4): e0229446. https://doi.org/10.1371/journal.pone.0229446 PMID: 32251463
7. Stern BM, O'Shea EK. A proposal for the future of scientific publishing in the life sciences. PLoS Biol. 2019; 17(2):e3000116. https://doi.org/10.1371/journal.pbio.3000116 PMID: 30753179
8. Guha RV, Brickley D, MacBeth S. Schema.org: Evolution of Structured Data on the Web: Big data makes common schemas even more necessary. Queueing Syst. 2015; 13(9):10–37.
9. Eglen SJ, Marwick B, Halchenko YO, Hanke M, Sufi S, Gleeson P, et al. Toward standard practices for sharing computer code and programs in neuroscience. Nat Neurosci. 2017; 20(6):770–3. https://doi.org/10.1038/nn.4550 PMID: 28542156
10. Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, McGillivray B. The citation advantage of linking publications to research data. PLoS ONE. 2020; 15(4):e0230416. https://doi.org/10.1371/journal.pone.0230416 PMID: 32320428
11. Poline JB. From data sharing to data publishing. MNI Open Res. 2019;2. https://doi.org/10.12688/mniopenres.12772.2 PMID: 31157322
12. Carpenter TA. What Constitutes Peer Review of Data: A survey of published peer review guidelines. arXiv:1704.02236 [Preprint]. 2017.
13. Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Scientific Data. 2016; 3:160044. https://doi.org/10.1038/sdata.2016.44 PMID: 27326542
14. The Turing Way Community. The Turing Way: A handbook for reproducible, ethical and collaborative research. 2021.
15. Perkel JM. Reactive, reproducible, collaborative: computational notebooks evolve. Nature. 2021; 593 (7857):156–7. https://doi.org/10.1038/d41586-021-01174-w PMID: 33941927
16. Davis RC. Five Tips for Designing Preservable Websites. 2011. Available from: https://siarchives.si.edu/blog/five-tips-designing-preservable-websites.
17. Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, et al. Jupyter Notebooks–a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas. Amsterdam, NY: IOS Press; 2016. p. 87–90.
18. RStudio. rticles. 2021. Available from: https://github.com/rstudio/rticles.
19. Jupyter Project, Bussonnier M, Forde J, Freeman J, Granger B, Head T, et al. Binder 2.0—Reproducible, interactive, sharable environments for science at scale. Proceedings of the 17th Python in Science Conference; 2018. p. 113–120.
20. International Brain Laboratory. An International Laboratory for Systems and Computational Neuroscience. Neuron. 2017; 96(6):1213–8. https://doi.org/10.1016/j.neuron.2017.12.013 PMID: 29268092
21. Odaka TE, Banihirwe A, Eynard-Bontemps G, Ponte A, Maze G, Paul K, et al. The Pangeo Ecosystem: Interactive Computing Tools for the Geosciences: Benchmarking on HPC. Tools and Techniques for High Performance Computing. Springer International Publishing; 2020. p. 190–204.
22. Rokem A, Dichter B, Holdgraf C, Ghosh SS. Pan-neuro: interactive computing at scale with BRAIN datasets. 2021.

23. Hanke M, Pestilli F, Wagner AS, Markiewicz CJ, Poline JB, Halchenko YO. In defense of decentralized research data management. Neuroforum. 2021; 27(1):17–25.

24. Katz DS, Niemeyer KE, Smith AM. Publish your software: introducing the journal of open source software (JOSS). Comput Sci Eng. 2018.

25. Boudreau M, Poline JB, Bellec P, Stikov N. On the open-source landscape of PLOS Computational Biology. PLoS Comput Biol. 2021; 17(2):e1008725. https://doi.org/10.1371/journal.pcbi.1008725 PMID: 33571204