The Reproducible Document Stack reinvents the journal publication for a world of computationally reproducible research

Michael Aufreiter(1)* and Naomi Penfold(2)

Affiliations:

(1) *presenter; Substance, Linz, Austria, GmbH; michael@substance.io [ORCID: 0000-0003-0715-1832]

(2) eLife, Cambridge, UK; n.penfold@elifesciences.org [ORCID: 0000-0003-0568-1194]

Source code

- Dar: https://github.com/substance/dar
- Texture: https://github.com/substance/texture
- Stencila: https://github.com/stencila

Background

In science, the manuscript is still considered the primary unit of publishable discovery, a simple artefact composed of mostly a textual narrative supported by images. So, to cater for the publication of every other aspect of the scientific process, there have emerged a constellation of secondary services and tools dedicated to sharing protocols, biological and chemical resources, raw and processed data, and the software that underpins the data analysis and derivation of computational models. The Research Object initiative aims to connect together all the outputs of research to form a package that accurately represents a whole research project. Meanwhile, researchers are documenting the relationships between these objects using tools such as Jupyter notebooks and languages such as R Markdown. Combining text descriptions with code snippets and the computed output facilitates greater reproducibility of computational research outputs. Some early adopters are beginning to share reproducible computational lab books as executable versions using links to services such as Binder (for example, see methods in [1]) and as reproducible packages through services such as CodeOcean (see examples on F1000Research blog [2]).

Several journals have experimented with more interactive and reproducible formats in the past decade. However, the current publishing system only supports the dissemination of these outputs as supplementary files with the article or in external repositories, as secondary outputs to the publication itself. Meanwhile, authors today still have to export their manuscripts in a flattened format (as a Word doc or PDF, for example) or submit in LaTeX. This process means that what may have started as a self-contained, fully reproducible document is fragmented into a loose collection of linked assets, jeopardising both the coherence and ongoing discoverability of the research.

The eLife Reproducible Document Stack

In September, eLife, Substance and Stencila started a project to develop the technology required to treat code and data aspects of research as first-class citizens in the academic publishing process, and create the tools needed to support these new reproducible document formats through authoring, sharing and publication. Specifically, we are developing the eLife Reproducible Document Stack [3], a suite of tools to make it possible for researchers to share their research in a publishable format whereby data (raw or by reference), code and computed output (graph or statistical result) are embedded and executable within the text narrative of a traditional manuscript. We are also enabling the open technologies needed for other journals to support the new submission format.

The creation of an open standard for the exchange, submission and publication of reproducible documents is critical for widespread adoption by academic publishers, and will be beneficial for the discovery and persistence of research reported in this form. We have therefore drafted an open Document Archive Format (Dar), which allows add code and data elements to be embedded in a reproducible document and presented online by publishers as an enhanced version of the published research article. The draft manifest is open for comments.

To support the authoring of structured research documents from the get-go, Substance has developed Texture as an XML-first editor [4]. With Texture, authors can produce an open format that can immediately be processed by publishers without the need for conversion.

Building on Texture, Stencila offers an office suite for reproducible research, combining the WYSIWYG document editor with a spreadsheet editor [5]. We are working closely with Stencila to enable researchers to both author new reproducible documents and interact with and edit published reproducible articles. Where possible, we aim to allow visitors to an online journal the opportunity to not only read and download the research assets for offline use, but also to interact and reproduce individual results in the browser, without the need to independently match the input(s), process(es) and output(s) for such reproduction.

Between now and the RO2018 workshop, we plan to release the first test reproducible manuscript with eLife: we are currently working with a researcher to convert their published article in eLife into the Dar format and to present this enhanced version online in the eLife journal.

Seeking collaborations to extend the Reproducible Document Stack

In this talk, we'll present the latest progress and outputs from the project and explore the next steps. In particular, we are seeking collaborations to make Dar interoperable with the Research Object ecosystem. Doing so would allow users of the reproducible document stack of tools to also benefit from further innovations that build on this ecosystem, such as mechanisms to credit contributions beyond publication authorship, and to facilitate the exchange of metadata information across web platforms. We also welcome the community's suggestions as to the impact they wish to see from this project, and ways in which we could measure this impact together.

Acknowledgements and attributions

We wish to publicly acknowledge all the researchers and technologists who have engaged with us, contributed feedback and tested the tooling throughout the development of this project. Thank you for sharing your experience, critical evaluations and course corrections, and for introducing us to relevant similar and complementary projects along the way. We also thank Nokome Bentley and Aleksandra Pawlik for their continuing contributions with Stencila.

Sections of this abstract are derived from the eLife Labs blogpost describing the project in September 2017 [3], contributed to by Naomi Penfold and Giuliano Maciocci, eLife, and Michael Aufreiter, Substance, and shared for reuse under the Creative Commons Attribution Licence (CC-BY 4.0).

References

[1] Mathelier, Anthony et al. 2016. DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. Cell Systems 3:3, p.278–286.e4. DOI: 10.1016/j.cels.2016.07.001.

[2] Ingram, Thomas. 2017. Reanalyse(a)s: making reproducibility easier with Code Ocean widgets. F1000Research blog. Accessed online July 9, 2018. URL: https://blog.f1000.com/2017/04/20/reanaly-seas-making-reproducibility-easier-with-code-ocean-widgets/.

[3] Penfold, Naomi. 2017. Reproducible Document Stack – supporting the next-generation research article. eLife Labs blog. Accessed online July 9, 2018. URL: https://elifesciences.org/labs/7dbeb390/reproducible-document-stack-supporting-the-next-generation-research-article.

[4] Aufreiter, Michael. 2017. Texture – an open science manuscript editor. eLife Labs blog. Accessed online July 9, 2018. URL: https://elifesciences.org/labs/8de87c33/texture-an-open-science-manuscript-editor.

[5] Aufreiter, Michael, Pawlik, Aleksandra and Bentley, Nokome. 2018. Stencila – an office suite for reproducible research. eLife Labs blog. Accessed online July 9, 2018. URL: https://elifesciences.org/labs/c496b8bb/stencila-an-office-suite-for-reproducible-research.

This abstract was submitted for consideration for the Research Object 2018 workshop within the IEEE eScience 2018 conference.