

Q&A on dynamic documents

Scott Edmunds 

Published April 16, 2014

Citation

Edmunds, S. (n.d.). In *GigaBlog*. GigaBlog. <https://doi.org/10.59350/q3zb6-3he70>

Keywords

Medicine, GigaDB, Neuroscience, Open Peer Review, Q&A
Feature Image

Copyright

Copyright © Scott Edmunds 2014. Distributed under the terms of the [Creative Commons Attribution 4.0 International License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

At [GigaScience](#) one of our major goals is to take the scientific publishing beyond dead trees and static PDFs to a more dynamic and interactive process, much like science itself has embraced the Internet to become more networked and data driven. One way we have done this is by enabling the histories and analyses from papers to be visualized and executed through our [GigaGalaxy](#) server (see our recent posting on this), but on top of [integrating workflows into our papers through citable DOIs](#), the papers themselves can be generated (and subsequently reproduced) in a similar manner using a number of tools that allow dynamic report generation. With [R](#), the open source software environment for statistical computing continuing to grow in popularity, there are a number of reporting tools being integrated into it such as [Knitr](#) and [Sweave](#). These support reproducible research and automated report generation by supporting execution of R code embedded within various document formats including LaTeX. Our recent [reproducible neurophysiology paper](#) was a great example of this, and following our [interview with lead author Stephen Eglen](#), we thought we would get some further insight into the advantages of dynamic documents by talking to some users about this example.

[c7c935bdc1ddb793b63a7952c98ff1b3](#)Our editorial board member Wolfgang Huber, on top of promoting the archiving of bioinformatics R- packages and workflows through the [bioconductor](#) project is a big advocate of the use of Knitr and Sweave, last year carrying out a [workshop at our BGI hosts](#) that covered this area (see the picture of Wolfgang modeling our GigaPanda t-shirt). Asking him why, Wolfgang summarizes the utility of this approach well: "I do all my projects in Knitr. Having the textual explanation, the associated code and the results all in one place really increases productivity, and helps explaining my analyses to colleagues, or even just to my future self."

Q&A with our reviewers Thomas Wachtler and Christophe Pouzat

As we promote open peer-review and like to credit the work of our reviewers (a process that will be greatly aided by [new moves to credit reviews](#) in ORCID profiles), we also interviewed the referees of the paper on if the review process was improved by the authors providing all of the files required to regenerate the paper. Much of this interview with Thomas Wachtler (group leader at the Ludwig-Maximilians-Universität München, and also on our editorial board) and Christophe Pouzat (Paris Descartes University) has recently been [published in Biome](#), but as with the [Assemblathon2 discussion](#) and some of our [other Q&A's](#) we thought we would focus first on their discussions on dynamic documents, and then provide the "box-set completist" version by publishing the rest of interview in full.

Can you give a little insight about the review process? Did you manage to test and recreate the analyses in the paper in R and how long did it take you?

Thomas Wachtler (TW): Providing code and data with a publication so that it is possible to replicate the analysis is highly valuable.

The paper by Eglen and colleagues is a shining example for such openness in that it enables replicating the results almost as easily as by pressing a button.

To be fair it must be acknowledged that such degree of accessibility may not be practical to achieve for any kind of datasets and studies at this point in time, but this should not be an excuse for not making the best efforts to increase accessibility of any study – to the reviewers as well as to the readers. This paper can be a strong model example for the community.

Christophe Pouzat (CP): It took me a couple of hours to get the data, the few custom developed routines, the "vignette" (that is, in the open-source R statistical software jargon, an executable file mixing description of what the code is doing with the code itself) and to REPRODUCE EXACTLY the analysis presented in the manuscript (using a netbook not a heavy duty desktop computer). With few more hours, I was able to modify the authors' code to change a linear scale for a log scale for their Fig. 4. In addition to making the presented research trustworthy, the reproducible research paradigm definitely makes the reviewer's job much more fun!

Can you say a bit more about how you found the paper?

TW: I was delighted to see this publication, which is the first data publication of electrophysiology data in *Gigascience*, and one of the first ever formal electrophysiology data publications.

Several electrophysiology datasets are available publicly, although the total amount of data is fairly low compared to data collections in other fields. Sites that provide data hosting, like [CARMEN](#), [CRCNS.org](#), or [G-Node](#), play an important role in enabling neurophysiologists to share their data, be it between colleagues or publicly, thus raising awareness for the benefits of data sharing. In some cases datasets have been made public when funding was provided to annotate and document the data. Here the incentive of a publication certainly played a role, which highlights the relevance of journals like *Gigascience* that enable data publications.

Journals that offer data publications enable scientists to immediately gain benefit from sharing their data with the community.

Conventional journals can also help raising awareness of this possibility by explicitly encouraging practices that enhance openness and reproducibility. Whether it is necessary to go so far as to enforce this practice is a decision that a journal should consider carefully. We currently see a growing interest and willingness among neuroscientists to make their data available anyway, so we can expect this to become common practice with time anyway.

Do you think electrophysiological data presents a particular challenge in terms of sharing and reproducing data?

TW: A fundamental requirement for open data to be useful is that it is not only technical accessibility but also practicability, that is, that both data and metadata are provided in standard or at least clearly documented and simple formats. The field of electrophysiology faces a notorious diversity and complexity of data and formats.

To present the data in a unified way, Eglen and colleagues made impressive efforts to read the data from different formats and to convert and annotate them. Ideally, such efforts could be greatly reduced in the future if common standards would be established in the community.

GigaBlog

The INCF's Program on data sharing, where members of [CARMEN](#), [CRCNS.org](#), and [G-Node](#) are also actively participating, are working towards such standards.

How did you find our open peer review process?

TW: Enhancing transparency by double-open peer review as introduced for this journal is in line with the increasing openness in the community and a valuable model alternative to the traditional peer review.

Why is the reproducible research paradigm important to you, and how does this paper address that?

CP: Taking a somewhat "extreme" stance, there is no (natural) science without reproducibility. If there is a long tradition of detailed description of experiments in the literature (making them mostly reproducible), the description standards of analysis / simulations associated to published experimental data have unfortunately been much "weaker". Developing tools making the implementation of the reproducible research paradigm is very important to improve the situation; publishing papers, like the present one, describing simply and beautifully how the paradigm is implemented in a very relevant scientific context is also of paramount importance.

This example went to great effort to make data and code available, and the methods transparent. Do you think it is worth it, and what can we do to encourage others to follow?

CP: Yes, I think it is worth it! I'm sure that by making data (and code) public, researchers will get more citations as well as attract more collaborations. But clearly the funding agencies will have a big role in making scientist switch from the present attitude (or culture) where they consider "their" data and code as private property (even when the work has been entirely funded by public money) toward a situation where they give access to both data and code by default. In order to share data, infrastructures (like the [CARMEN virtual lab](#) or the [g-node](#)) have to be created and maintained and scientists working on their development must get credit for that.

This paper is a would probably not have been possible without the [CARMEN virtual laboratory](#). Do you have anything you would like to say about CARMEN?

CP: A great project!

References

1. Eglen, SJ; Weeks, M; Jessop, M; Simonotto, J; Jackson, T; Sernagor, E. A data repository and analysis framework for spontaneous neural activity recordings in developing retina. *GigaScience* 2014, **3**:3 <http://dx.doi.org/10.1186/2047-217X-3-3>
2. Eglen, SJ; Weeks, M; Jessop, M; Simonotto, J; Jackson, T; Sernagor, E. (2014): Supporting material for "A data repository and analysis framework for spontaneous neural activity recordings in developing retina". GigaScience Database. <http://dx.doi.org/10.5524/100089>

The post [Q&A on dynamic documents](#) appeared first on [GigaBlog](#).