Emerging Topics in Life Sciences (2018) **2** 755–758
https://doi.org/10.1042/ETLS20180165

Perspective

Digital publishing isn't enough: the case for 'blueprints' in scientific communication

Laura D. Jennings-Antipov and Timothy S. Gardner

Riffyn, Oakland, CA, U.S.A.

Correspondence: Timothy S. Gardner (tgardner@riffyn.com)

Since the time of Newton and Galileo, the tools for capturing and communicating science have remained conceptually unchanged — in essence, they consist of observations on paper (or electronic variants), followed by a 'letter' to the community to report your findings. These age-old tools are inadequate for the complexity of today's scientific challenges. If modern software engineering worked like science, programmers would not share open source code; they would take notes on their work and then publish long-form articles about their software. Months or years later, their colleagues would attempt to reproduce the software based on the article. It sounds a bit silly, and yet even, this level of prose-based methodological discourse has deteriorated in science communication. Materials and Methods sections of papers are often a vaguely written afterthought, leaving researchers baffled when they try to repeat a published finding. It's time for a fundamental shift in scientific communication and sharing, a shift akin to the advent of computer-aided design and source code versioning. Science needs reusable 'blueprints' for experiments replete with the experiment designs, material flows, reaction parameters, data, and analytical procedures. Such an approach could establish the foundations for truly open source science where these scientific blueprints form the digital 'source code' for a supply chain of high-quality innovations and discoveries.

Reading the fine print leads to frustration, not reproducibility

When I (L. Jennings-Antipov) was working as a biochemistry researcher in graduate school and later during my post-doc, my colleagues and I followed the literature in our field closely. Whenever an interesting paper came out that we thought we could build upon in our laboratory, we would first try to reproduce the findings. This entailed literally reading the fine print — that is, the Materials and Methods section of the paper, which is often published at the end, in smaller print than the 'top-billed' Background, Results, and Discussion sections. However, reading this fine print could only get you so far. Often, the methods simply referenced another paper, which referenced another paper, which may give you some indication of the experimental design and conditions, but in not nearly enough detail to exactly repeat what was done. It would often take us months of trial and error to get the experimental design correct (to the best of our understanding), and even then, often times we could not repeat the exciting result reported in the paper. This left us frustrated because we weren't sure if the experiments were not repeating because of a misinterpretation of findings by the original researcher, or if we simply did not understand what was done and therefore couldn't possibly repeat it.

I wish the above story were an anomaly, but an informal poll of peers tells me my experience is the norm. And study after study has revealed that the above experience, and many more like it, are contributing to a reproducibility crisis resulting in ~$30B wasted every year in the U.S.A. alone on unrepeatable scientific research [1].

Received: 1 November 2018
Revised: 12 November 2018
Accepted: 19 November 2018

Version of Record published: 21 December 2018

Our reporting tools haven't changed since the time of Newton and Galileo

It seems to me that we have it all backwards. The most important part of any scientific publication is the Materials and Methods section of the paper. After all, what good is a result if others cannot repeat it and build off of your work? Sadly, though, our tools for reporting the experimental design and conditions used, and the subsequent results, have not changed since the time of Newton and Galileo. Sure, we've shifted our reporting from pen and paper to electronic format, but we still write out how we performed a set of experiments and the results of these experiments as a long-form 'essay'. And we publish summarized, graphical representations of our results in static pictures. This is akin to a software engineer spending months building a piece of software, taking notes and observations of her work along the way, then sharing it with the world by writing up a long-form article of what she did, complete with cartoon illustrations summarizing the outcome of the code. If you asked a software engineer to share her code in that way, she would look at you like you had two heads. Instead of doing that, software engineers share their open source, annotated code on sites like GitHub, so others can directly use it and build off of it. If Silicon Valley were relying on journal-style publications to advance the state of the art in the high-tech world, companies such as Google and Apple probably would not exist.

Don't just share results, share experiments

Why is our method for sharing data so inefficient? It's simple. We, as researchers, have focused on the wrong goal. Our goal as researchers should not be simply to report our interpretation of the results of an experiment. Our goal should be to share the experiment itself. It is only by sharing experiments (the experimental design, the conditions used, the results obtained, and the analyses performed) that we will contribute something tangible and reusable to the scientific community.

When you stop thinking of the experiment as the means to an end, and start thinking of the experiment as the 'thing' you are contributing to science, your thinking shifts. Instead of asking 'is my hypothesis true?', you ask 'is my measurement a precise, accurate and reproducible assessment?' You begin to treat an experiment as a thing that can be seen, known, and improved upon like bits of code shared among computer programs. When you treat an experiment as a thing, you create a 'supply chain' of scientific methods and experimental data that are reusable and can be built upon by others. When you pass around methods and data like that, open source science becomes a reality, an explosion of computationally enabled discourse and analysis will transpire, and the pace of discovery will leap [2].

New tools are needed for sharing experiments

In recent years, the limitations of traditional publishing approaches in scientific communication have given rise to new digital tools. For example, some journals now accept interactive figures generated by platforms such as Plotly and Code Ocean [3]. Preprints servers, such as bioRxiv (biorxiv.org), allow access to manuscripts ahead of peer review and publication. And several journals, such as *Nature*, *PLoS*, *Science*, and *PNAS*, require authors to deposit datasets into publicly accessible repositories. We applaud all these digital sharing efforts. But while these digitization efforts speed up communication, they do not address the fundamental limitations of document-based and figure-based sharing of science. The datasets in general repositories, such as FigShare (figshare.com) or Dryad (datadryad.org), are in *ad hoc*, unstructured, or semi-structured format and lack a detailed context provided by the methodological information used to generate those data. This makes it very hard for others to interpret the data and extremely time-consuming to reuse the data if they can decode them.

Just as GitHub was developed as the need to share open source code arose, the scientific community is in desperate need of new tools that allow experimental designs and their corresponding results and analyses to be shared publicly, ***and*** in a form ready for statistical analysis, machine learning, and other computation. Existing tools — where sharing experiments means writing a few paragraphs about what you did and turning your results into a static, summarized graph — are no longer adequate. Instead, we need tools that allow researchers to share reusable experimental design templates and complete, annotated datasets that can be interpreted and built upon by others and their data algorithms.

Imagine a world where, when a new study gets published, you read a synopsis of it online, and then click a link to open up the experiment. This experiment tells you exactly what steps were completed, in which order, and which reagents were used. Individual data points (such as the absorbance measured from an ELISA assay) are reported and linked to the procedural data (such as buffer compositions and incubation times) used to generate that data. The analysis that was performed is also attached to the experiment as a script that you can run on your computer, so you can dynamically explore the data rather than looking at a static PDF plot. What's more, the experimental design is a reusable template that you can use to generate data using the exact same protocol, and you can look at your data alongside that of the published paper. Published experiments can be repeated in a week rather than months (if at all). When you are ready to publish your findings using this experimental design template, you simply click a button that says 'share publicly', and now researchers across the globe can access your experiment, the experiment of the original researcher, and they too can repeat the experiment using the reusable design template.

Riffyn is turning fairytale into reality

The above scenario might seem like a fairytale, but companies like Riffyn are turning that fairytale into reality. The Riffyn Scientific Development Environment (SDE) was designed as a new medium for creating and sharing scientific experiments the way programmers share source code. Riffyn SDE delivers all experimental methods, their associated data, and their analysis scripts in a modular, reusable, and instantly computable digital medium. It provides a platform for transforming experimental protocols into visual process flow diagrams. These process flow diagrams become the backbone to which both procedural information and results data get attached. When the researcher is ready to perform analysis on the experiment, a fully annotated, complete dataset (joining procedural information to results data across all the steps of an experiment) can be downloaded and analyzed in any third-party statistical software. Moreover, analysis scripts in any scripting language can be attached to the Riffyn SDE experiment to allow other researchers to dynamically explore the data collected on that experiment, or researchers can collect additional data using that experiment template and analyze these data alongside the original dataset.

In October 2018, the Jack Pronk laboratory, of the Department of Industrial Microbiology at Technical University Delft, published the first article of its kind in *Scientific Data* [4]. The article used Riffyn SDE as the method for designing microbial fermentation experiments, collecting and analyzing physiological and reactor data, and sharing these experimental methods and data. The Pronk group noted that the initial set-up of reusable experimental design templates in Riffyn SDE was more time intensive than writing conventional protocols in prose format. However, upfront costs were offset by greater time savings later. Prior to Riffyn SDE, data analysis for the Pronk laboratory meant manually searching, extracting, cleaning, reshaping, and combining data across multiple notebooks, spreadsheets, and databases. Then, additional time was spent manually examining different methods of data analysis across multiple researchers and 'back correcting' for any errors and differences in calculation methods. Use of Riffyn SDE automated all these activities, allowing the users to generate a fully annotated, cleaned, and structured dataset from multiple experiments in seconds rather than hours or days. Riffyn SDE also eliminated ambiguity in data analysis by allowing the researchers to pipe data from multiple experiments into objective, automated data calculation pipelines.

We note that the Pronk laboratory experiences shed light on a common question: does 'X' software save me time in the laboratory? The answer is yes, but not without a bit of transitional effort to recast their experimental methods into a new form. This effort paid off dramatically by reducing the time-consuming data preparation and analysis phases of experimentation, and by allowing more consistent and repeatable analyses — time and effort that are often not considered in calculations of 'time in the laboratory'. Moreover, the quality and transparency of the resulting methods will save tremendous time and effort for collaborators as well. Notably, in the Pronk *Scientific Data* paper, a link to the Riffyn SDE experiment itself was used in place of the Materials and Methods section since it contained everything researchers needed to repeat the experiment as well as access to the raw datasets and analysis scripts. When contextualized data are so easily accessible, the deepest benefits become apparent: scientists begin asking questions of their data they never before could ask.

'Riffyn SDE's systematic description of experimental methods eliminates many limitations that are inherent to traditional lab protocols and methods sections in journals', explained Dr Jack Pronk, Professor of Industrial Microbiology at the Delft University of Technology and coauthor of the study. 'I expect this type of detailed, web-based description of experimental procedures to become the new standard in scientific publications, thereby increasing the repeatability of scientific research and facilitating meaningful reuse of published data'. [5]

Also in October 2018, Riffyn announced that their SDE platform is now open access for researchers at nonprofit institutions [6]. This news, coupled with the proof-of-concept article in *Scientific Data*, could be the driving force for moving the scientific community out of the Renaissance era and into the 21st century where we belong.

Abbreviations

SDE, Scientific Development Environment.

Competing Interests

The Authors declare that there are no competing interests associated with the manuscript.

References

1 Freedman, L.P., Cockburn, I.M. and Simcoe, T.S. (2015) The economics of reproducibility in preclinical research. *PLoS Biol.* **13**, e1002165 https://doi.org/10.1371/journal.pbio.1002165

2 Gardner, T.S. (2017) Buried in Data and Starving for Information. Retrieved from: https://riffyn.com/riffyn-blog/2017/10/5/buried-in-data-and-starving-for-information

3 Perkel, J.M. (2018) Data visualization tools drive interactivity and reproducibility in online publishing. *Nature* **554**, 133–134 https://doi.org/10.1038/d41586-018-01322-9

4 Juergens, H., Niemeijer, M., Jennings-Antipov, L.D., Mans, R., Morel, J. and van Maris, A.J.A. (2018) Evaluation of a novel cloud-based software platform for structured experiment design and linked data analytics. *Scientific Data* **5**, 180195 https://doi.org/10.1038/sdata.2018.195

5 Gardner, T.S. (2018) Riffyn SDE published in a ground-breaking study in the journal *Scientific Data*. Retrieved from: https://riffyn.com/riffyn-blog/2018/10/3/riffyn-sde-published-in-a-ground-breaking-study-in-the-journal-scientific-data

6 Gardner, T.S. (2018) Riffyn launches Open Access for scientists at non-profits. Retrieved from: https://riffyn.com/riffyn-blog/2018/9/11/riffyn-launches-open-access-for-scientists-at-non-profits