
REVIEWS:

Reviewer #1:

The authors describe a workflow for “evolving data”. In general I find this contribution to be very strong. It provides a practical solution to a problem that is of pressing need. It does it in a way that enhances reproducibility and builds on existing tools. The long-term impact will depend on training in the use of these tools, so the discussion of software and data carpentries is well placed. I would be excited to see this workflow in a data carpentry module. I have only minor concerns, which I expect the authors could address easily. I do have one discretionary change related to the terminology used in the work, that I think the authors should consider.

We would be excited to see a Data Carpentry module for teaching others to implement the workflow as well! Putting together workshops or online tutorials has been something we've discussed as future directions for this work.

Minor:

“Makes it easier to understand the provenance of the data” -> suggest “it possible to understand...”

Done. We modified this language slightly to read “makes it possible to track the provenance…” to make it more explicit that this information can be used in a more concrete fashion.

After pull requests are introduced but before testing is discussed, it might be nice to have a short reminder of what continuous integration systems do. This would help to remind the reader that they can run test code.

As part of our reorganization to shorten, we have integrated the discussion of what continuous integration systems do with our first discussion of how we use it.

Describing the QA/QC system first as running “unit tests” before then defining unit tests flips the order needed by a reader who is not familiar with the term “unit test”.

We have integrated these into the same sentence so the reader is not left wondering what we're talking about.

How authorship was handled is unclear. Does the PR acceptance denote authorship? Is that recorded somewhere? Would it be worth mentioning the manubot authoring system here (which is also CI-based)?

Authors on this project are the graduate RAs, postdocs, and other personnel who are tasked with responsibility for data collection or significant data management. This may or may not involve being involved with the data management workflow at the PR stage. As a result, while we find the conversation regarding authorship on large open source collaborative projectives to be interesting and important, this project doesn't provide the correct exemplar for discussing it in detail. However, given the importance of this topic, we have mentioned it in our new concluding paragraph as one of the issues our field needs to resolve.

Discretionary:

The terminology "evolving data" is confusing. Does it mean that the data change over time via some sort of mutation and selection process? The "dynamic data" term that is also discussed seems more meaningful. Because "evolving data" also appears to have been used in the past I am making this a discretionary suggestion, but I think the change would enhance clarity.

What to call this type of data has been a source of consternation for us. For scientists with advanced computational training, dynamic data communicates effectively. However, it does not communicate as effectively to less-computationally inclined data generators (e.g., field ecologists). Feedback from twitter (not a phrase we thought we'd ever use in a review) suggested that many of these groups have created their own ad hoc terms for describing this type of data, highlighting their lack of familiarity with the term 'dynamic data'. As a result, we're worried that many data generators may see the term 'dynamic data' and not realize that this paper is actually relevant to them. To communicate to the widest array of scientists, we have opted to simply describe it for what it is: 'regularly updated data'. We have added a box that describes the type of data we are discussing and the various terms that have been used to describe it. We know this is not ideal but we think this is the best compromise for reaching the broadest audience across science.

Reviewer #2 (Melissa Haendel)

Overall, I think this manuscript makes an important point and provides some good guidance for those collecting, archiving, and distributing "evolving" data. Many data sources are "evolving," and so is the science that is dependent upon such data. The manuscript lives somewhere between a community project paper and a 10-rules style paper, largely focusing on trying to genericize lessons learned from the project at hand rather than focusing on it specifically. As someone who has similarly struggled to identify the best location for data science best practices and community standards sorts of papers, I think a PloS community page is a very reasonable location for this work. The manuscript is quite thoughtful, but could be made more impactful and more useful to the community at large if the following constructive comments were addressed:

1. The manuscript is too long such that the important takeaways are somewhat buried. I would try to call these out better. One idea might be to have shorter text sections with a corresponding graphic for each step that highlights using your own platform, an example of how that works for you, and then in the text alternative choices and suggestions on how to make the best choice. You could also adapt the overall figure as a "roadmap" for the whole set of recommendations. It would get reused often I expect.

We have worked hard to follow Dr. Haendel's suggestions for shortening the manuscript. We have shortened the text for each section and moved a lot of details into a supplementary section where readers can find additional information, if they wish. The total length of the main text is reduced from 5671 words to 3980. We also made a more comprehensive workflow figure that shows how the operations in different sections fit together. We think these changes make the manuscript communicate more effectively and appreciate Dr. Haendel encouraging us to do this.

2. Some of the basic descriptions, such as justifying version control, don't seem novel or unfamiliar to most potential readers. Maybe some of this can be reduced, instead provide readers more advice on how to choose between the different options that currently exist, for example for setting up version control/PR/CI.

Our experience working with both more traditional data-collection focused labs and computational groups is that while version control is indeed widely used in the computational groups it is still little known in the more empirically focused groups that are often struggling with this type of data. Our goal in this paper is two-fold: to communicate with more computational groups about how to structure a data workflow for regularly updating data and to provide less computationally advanced groups with the information they need to find suitable training and/or to discuss their data management needs with more computational groups. Hopefully, as a consequence of our shortening the manuscript, the basic descriptions will seem less tedious for the more computationally oriented readers while also still serving as a good entry point for others.

3. The licensing section is too underspecified to be very useful. You may be interested in making recommendations regarding licensing for projects who do not generate all their own data; e.g. not all licenses are compatible and you cannot release as CC0 if some of the integrated content is not under CC0. See http://reusabledata.org/ and https://www.biorxiv.org/content/early/2018/03/16/282830. This is a lot more complexity to data licensing than most people think, and there is much not-quite-legal redistribution going on! (Happy to advise on this part).

We acknowledge that this is underspecified, in large part because the complicated data licensing landscape is beyond the purview of our data management workflow to resolve. Rather than give the impression that our data management workflow helps with figuring out data

licensing issues, we have moved this section to the expanded supplementary section and discuss in that section that our situation allowed us to use a specific open license but that other scenarios, particularly data compilations, may be more complicated. We have added some references to help point readers to more knowledgeable sources. To make sure this is also mentioned in the main text we have added a brief discussion in the new future directions paragraph on the need for the field to resolve best practices for data licensing.

4. I think some people might be confused by the “automated updating of supplemental data”. What you really mean, I think, is that a paper/analysis pointed at a specific dataset, but that that data set is continually evolving. The question is - is the manuscript or other analytics that are built on top of those supplemental data also released accordingly? I think how to help people navigate this socio-technical decision making process better would be useful. This would be another good place for graphics.

We can completely see now how the terminology could be confusing. We do not mean that analyses in papers point to a specific data set but that additional information about data collection efforts are kept in separate tables (not the main data table) for data normalization purposes. We have changed all discussion of these to “supporting tables”. A general example of what we’re talking about would be a database where records from new locations are regularly added. The researchers building that database may want a separate table that keeps information on which locations are in the database and perhaps additional information about that location that can be extracted programmatically from other sources (e.g., average temperature, population size of the town the data was collected from). This supporting table is not an analysis per se, but it is useful for conducting analyses on the main data being collected. Everytime a new location is added to the database, the researchers must either update this by hand or they can use the data management workflow to automatically update those tables whenever new data is added to the main data file. Hopefully this is clearer in our revision.

5. Generally the figures are insufficient in that there are not enough of them and that the ones that are there are not very sophisticated or inclusive of the content of the paper. I think you would have more messaging impact with better figures and less text.

We attempted to address this comment by redesigning our main roadmap figure, which we then use to reference specific steps in each section. We think this new roadmap will be more informative and is definitely better integrated with the main text than the previous version.

6. There doesn’t seem to be much in the way of predicting the future or a conclusions/impact section. How will continually evolving data be used in the data ecosystem of the future? How what you propose enable tracking of provenance/attribution in mashed up data? How can external 3rd parties best participate in fixing data upstream? Etc.

Our interpretation of this comment is that we have not ended the manuscript on a big enough scope. We have revised the final paragraph to provide a more forward looking context for our

summary statements. Many of the specific suggestions here were beyond the scope of our paper, but our workflow does allow 3rd party participation in error fixing and we have a section in the supplement that discusses this.