# **Pragmatic Reproducible Research** for Analysis, Dissemination and Publication

## TEAM

| QUESTIONS TO ASK YOURSELF | THINKING ABOUT REPRODUCIBLE RESEARCH |
|---|---|
| Who am I working with on this project? | Identify all of the people involved, since everything they work on is subject to change somewhere along the way. |
| How technical are they (and me)? | Reproducible research doesn't have to use technical solutions. Keep thinking of steps you can perform in a reproducible manner without technology, or using technology that may be more approachable. |
| How open are they to approaching reproducible research? | Keep in mind that reproducible research is an up-front investment. It is extra work, and not everyone on the team may be willing to embrace it. If that's the case, you can always focus on your part of the project. |
| Is it clear how the research team is organized (roles, responsibilities, tasks)? | If you are looking to do reproducible research across the whole research project, you need to have effective communication and handoffs.<br>If you are making just your work reproducible, make sure the rest of the team is aware and willing to follow your process.<br>Things won't go according to plan if there is no plan. |
| How are we going to organize our project? | Try to establish a consistent convention for folder naming and structure. You will need to make sure the team (if more than just you) are aware of the plan if it will include files across the whole project. Similar approaches can and should be used for online repositories (e.g. Box, Google Drive)<br>Don't forget to have documentation around your convention, and share it with the whole team (even if they may not be performing reproducible research). This will help in the future when trying to navigate the folders. |

## DATA

| QUESTIONS TO ASK YOURSELF | THINKING ABOUT REPRODUCIBLE RESEARCH |
|---|---|
| Where will (or do) my data live?<br>– Excel document(s)<br>– Cloud-based system (Google Docs, Survey Monkey, REDCap)<br>– Institutional repository / database | Always consider regulations about where you are allowed to store your study data!<br>Start thinking ahead about Dissemination, especially if you are looking to do open science. This may drive how you collect, store, or archive data.<br>Where the data live determines how you can access it (e.g. through special web portals, code, database queries). If you are looking to automate the whole process, you will need to make sure the computer can get to the data. |
| Are my data changing? | Most often you will want to grab a snapshot of your data and work off of that. You can take a new snapshot of the data in the future, if you know that new data are going to be added.<br>By organizing and naming your files properly, you will be able to go back at any point and see the data you were working with.<br>If you think your data aren't changing, you need to be absolutely sure. There are many factors that cause data to change if they are not entirely in your control. |
| How can I make sure my data haven't changed? | Think of how you can prove your data set is unchanged. If you are using files, calculating a checksum (e.g., MD5 hash) can provide more assurances that a file as a whole is unchanged. If you are accessing the data from another system, you could perform some analysis/summary that is guaranteed to produce different results if something changes. |
| Are my data immediately ready for analysis, or will I need to make changes? | Start planning ahead for Analysis. Think of how you will document the data cleaning steps that you took, so that you can apply the same steps again if you get a data refresh.<br>Automation at this point can really help streamline data cleaning in a reproducible way. |

DOI:10.18131/g3-eqv8-gr68

| | |
|---|---|
| How often am I accessing the data? | Especially when you are automating the process, be aware of how often you are actually pulling new data and the impact. For example, if you are accessing data from an API, and you perform that every time you run your code, is that putting a strain on the server hosting the data source/API?<br>Consider future data needs, and how often you would like to refresh your data set (if you know it is changing). If the data are coming from a hosted source, think about downtime – what is the impact to your project if the site is down for a day? A week? Forever? |

## ANALYSIS

| QUESTIONS TO ASK YOURSELF | THINKING ABOUT REPRODUCIBLE RESEARCH |
|---|---|
| Where is my analysis going to run?<br>– Locally on my machine<br>– On a local server<br>– Cloud-based server that I control<br>– Cloud-based analytics platform (Code Ocean) | Are you in control of which components and related versions are installed/available in this environment?<br>If you are considering open science, this will also help you think through where your code can be run now and in the future. |
| How can I know the exact environment in which my analysis ran? | You can start documenting specific versions. Understand if your particular software can give you a printout of the versions of all installed (or used) modules/packages for your analysis.<br>Make note as well about the platform that you are running your analysis on, as well as the operating system (32-bit vs. 64-bit). Keep in mind that a 64-bit machine doesn't guarantee you are running 64-bit software. |
| Have I checked all of the licenses for my software to identify restrictions on their use? | Some environments can help you automate this process.<br>Keep in mind the impact of "viral" licenses, and if you are simply calling a package or incorporating the code into yours.<br>Think ahead to Dissemination – the licenses you choose can affect if and how you share your code.<br>Remember that proprietary (licensed) software can still be used as part of a reproducible research project. |
| How am I going to track what I did, and how that changed over time? | Many solutions are possible (technical and non-technical), but it is important to plan upfront how you will track all of these changes.<br>Possible options are keeping a regular notebook, an electronic lab notebook, and/or using source control.<br>Remember that you can use more than one solution. |

DOI:10.18131/g3-eqv8-gr68

# Pragmatic Reproducible Research for Analysis, Dissemination and Publication

## MANUSCRIPT

| QUESTIONS TO ASK YOURSELF | THINKING ABOUT REPRODUCIBLE RESEARCH |
|---|---|
| What journal(s) am I planning to submit to? | Different journals have different requirements for manuscript submissions. This may require you to submit a manuscript in Microsoft Word, even if you prefer to work in LaTeX. Similarly, it may affect how you render tables and figures (inline or as separate files). |
| Who is going to be involved in the manuscript authoring? | Think about who will need to contribute to authoring, editing, and reviewing the manuscript. If you are choosing a particular technology, like R Markdown, are all of the authors willing to use it? |
| How am I integrating the results of my analysis into the manuscript? | Try to avoid copying and pasting results into your document at all costs.<br>If you are weaving your results directly into the manuscript (e.g., R Markdown, StatTag), think about how to structure long-running analyses and statistical analyses tuned for the manuscript. You don't want a new version of your paper to take 6 hours to compile. |
| Am I able to perform a "dry run" of the methods? | If possible, a colleague not involved with the project could check to make sure that your methods are complete and unambiguous. If you are providing code and/or data along with the manuscript, try running it in a clean environment to make sure you've captured all of the dependencies. |

## DISSEMINATION

| QUESTIONS TO ASK YOURSELF | THINKING ABOUT REPRODUCIBLE RESEARCH |
|---|---|
| Will I have access to my research (notes, data, computation environments, code) in the future? | "Disseminating to myself" - think through how you will retain and access all of the pieces of your research pipeline in the future. |
| Which pieces of my research workflow am I looking to share? | Based on considerations regarding licensing, IP, data access, you may only be able to share parts of your research. Think about how to supplement the pieces that can't be shared with good documentation (in the manuscript and/or the code). |
| How public am I willing to make this? | Reproducible research is not the same as open science, so you don't necessarily have to share things publicly (especially as you are getting started).<br>If you have concerns about sharing parts of your research publicly, think about sharing others. It's not an all-or-nothing proposition. For example, if you're not able to make your data public, think about still sharing your analysis code. |
| How turn-key do I want it to be to reproduce my findings? | Although automation is not a required part of conducting reproducible research, the time spent automating your research pipeline can save a future user (including yourself!) time to set up the same environment again. Even additional time spent documenting can pay off.<br>Remember that this is an up-front investment of your time. If you are willing to put in the time initially, it most often pays off in the long run. |
| What rights do I have to consider to share third-party artifacts (data, libraries, code)? | When planning to share things publicly, seek input from your research team and the institution (especially if you are not the PI). Always make sure you are aware of institutional, state and federal rules on sharing research data.<br>If you need third-party data sources, libraries or other software code to fully reproduce your findings, make sure that you understand what you do and do not have rights to redistribute yourself.<br>Many data sources have additional licenses or agreements that must be acknowledged. If you are not able to share these directly, and are working towards automating your workflow for others, ensure you have substantial documentation to describe what should be used. |

DOI:10.18131/g3-eqv8-gr68