A Framework for Really Reproducible Research

Scott Jackson

February 18, 2013

Contents

1 Introduction **5**

1.1 Two motivations 5

1.2 Three sides of the coin 6

1.3 How to use this book 6

1.4 Theses and structure of the book 7

2 Principles **9**

2.1 Replication vs. reproduction 9

2.2 A sketchy history of reproducibility 9

2.3 Some core values 10

2.4 Who is this book for? 12

2.5 Goals 13

3 Dimensions of reproducibility **15**

3.1 Domain: the audience 15

3.2 Range: the precision 16

3.3 The minimum 16

3.4 Publication standards 16

3.5 Wide dissemination 16

4 The cycle of research **17**

5 Scholarship **19**

6 Theory **21**

7 Data collection **23**

8 Data analysis **25**

9 Reporting **27**

10 Conclusions **29**

11 Implementation 31
11.1 Guiding principles . . . 32
11.1.1 Free and open source . . . 32
11.1.2 Cross-platform . . . 32
11.1.3 Stable . . . 32
11.1.4 Well-documented . . . 32
11.1.5 Customizable . . . 32
11.2 Summary of tools . . . 32
11.2.1 Emacs . . . 32
11.2.2 Org-mode . . . 32
11.2.3 Git . . . 32
11.2.4 LaTeX . . . 32
11.2.5 Python . . . 32
11.2.6 R . . . 32
11.2.7 (Emacs) Lisp . . . 32
11.3 Scholarship . . . 32
11.4 Data collection . . . 32
11.5 Data analysis . . . 32
11.6 Sharing . . . 32
11.7 Collaboration . . . 32
11.8 Putting it all together . . . 32

Chapter 1

Introduction

1.1 Two motivations

This book is about *reproducible research.* A lot of the book is dedicated to fleshing out exactly what that term means, or perhaps what it *should* mean, but moreover, this book is about *really* reproducible research. In order to tell you what *that's* supposed to mean in a relatively concise way, I'm going to talk about the two different kinds of motivations that drive this book.

One type of motivation is the desire to do Good Science. It's my belief that scientific research should be conducted in a way that is maximally replicable, reproducible, transparent, and trustworthy. I think of this as a kind of ethical obligation for scientists, both to the scientific community and to the people that fund our research (usually taxpayers, in my line of work). I also believe that this is inherent to the success of the scientific method, but I am not an expert on the philosophy of science, so I won't go too deeply into that side of things, beyond expressing some of my personal opinions. From this ethical/philosophical point of view, reproducible research is in line with the value system of science, and *really* reproducible research is a set of real-world methods and techniques that do more than lip service to these ideas; they put them into action.

The second type of motivation is the desire to be a Successful Scientist. This is the more practical, pragmatic side of things. In short, being a successful scientist means doing work that gets published frequently in high-quality journals. Obviously for many academics, there are other sides to professional life like teaching, service, etc. Maybe some of the ideas here can apply some to those domains, too, but I'm going to focus on the research side. A successful researcher, from the most concrete point of view, is one with a good publishing record. From this angle, the motivation of reproducible research is to help the researcher be more efficient, effective, and productive, and to produce higher-quality work that is worthy of the most prestigious journals. And similarly, *really* reproducible research is meant to provide a concrete framework that's more than just recapitulating concepts about effective work habits.

Now hopefully, if you're engaged in research, you're interested in one of these two motivations, maybe both. If you neither care about (a) producing good science worthy of the faith and trust of other scientists and your funding sources, nor (b) being a successful researcher with a great publicaton record, then maybe you're in the wrong field. The central aim of this book is to talk about how reproducibility — *real* reproducibility — can assist you in both of these goals.

1.2 Three sides of the coin

In order to address these two goals, there are three (count 'em!) sides of the coin we need to explore. One side is the conceptual side, trying to lay out a framework of principles for understanding what it means to be reproducible. The other side is a discussion of how these principles play out in the day-to-day activities of carrying out research. But while this second side is more concrete, it's still framed in terms of general principles that could be implemented in many different ways. The third side is an actual, working implementation. You can see the third side (the edge) of the coin as either the part that connects the two faces of theory and practice, or as I prefer to see it, as the edge of a wheel, where (to mix metaphors) the rubber really meets the road. It's fine to talk about good principles of reproducibility, but without a kind of "manual" for exactly how to put things into practice, there will be a high barrier to entry for most people, which means very little change. If we want *really* reproducible research we have to have an actual working system that enables our good intentions to be realized in a practical way.

This book is aimed at the first two, more conceptual sides. The companion to this book is a repository on GitHub that provides a set of tutorials and guides for using a set of (open source, free) software tools to actually implement the ideas in the book. This "companion" set of materials is ultimately what puts the "really" in the title of the book, but I just couldn't see how to effectively squeeze all those tutorials and things into a useful book format. What I will do is include lots of hyperlinks in this book, in order to make it easier to connect from the idea or motivation for a particular kind of habit, workflow, or tool, to a set of tutorials that will help you get up and running with a tool that actually works as quickly as possible. This brings me to the next section.

1.3 How to use this book

Here's how I imagine (and hope) this book could be used. Read through some of the general chapters in the beginning, to get a sense of the big issues. But then feel free to skip around to particular issues or problems that sound interesting. The material in this book will lay out some of my thoughts on the topic, which will hopefully lead you to arrive at your own thoughts and opinions, and if my ideas intrigue you (either because they sound good, or because they sound terrible), you can follow some links to the GitHub repo to see how I personally attempt to implement my ideas. Those links will teach you how to use some piece of software (or non-software solution, if appropriate) to implement things yourself. All the software and techniques I use are extremely customizable, so if you don't like exactly how I do it and want to tweak it for your own use, or if you think you can improve on it more generally, my tutorials will hopefully give you a good start towards being able to do that as well.

At some point, I expect this book to become rather static, in the sense that I will revise things as my thinking changes, and as other revisions/corrections need to be made, but if I ever start to change my mind in a big enough way, I may just need to write a different book. But the repository of tutorials and software is going to be a dynamic thing. It is ultimately a selfish endeavor, because I will base the tutorials around my own usage and needs. As those things change, so will the tutorials. However, since I will be using `git` for version control, you will be able to "roll back" any particular tool or tutorial to an earlier phase, if you thought those worked better for you. At least, you'll be able to do this once you learn how to use `git`, which happens to be the base of what I'm going to recommend for a reproducible workflow! See, this stuff is already sounding useful, and we haven't even really gotten started yet. . .

But the point is that the tutorial stuff will change a lot (assuming I can keep up with it), in an ongoing way. I would be very happy to have feedback on any aspect of this whole enterprise, whether you have suggestions or arguments with things in this book, or whether you have suggestions, problems, or alternatives to the implementation stuff on the GitHub site. Please direct all your love/hate mail to shoestringpsycholing1@gmail.com.

1.4 Theses and structure of the book

This is not the first thing that's ever been written regarding reproducible research. It's a rather hot topic these days, and lots of smart people have been thinking/writing/talking/blogging/tweeting about it. One goal of this book is to push a few new ideas into the discussion. In particular:

1. Whether something is *reproducible* is not an absolute, but is relative to the *range* (i.e., precision, extent) of reproducibility, the *domain* of reproducibility (i.e., what kind of activities are being reproduced?), and the *audience* of reproducibility (i.e., reproducible for whom?).

2. *Reproducible* has to be able to extend beyond just the domain of data analysis; it should apply to all aspects of the cycle of research.

I'll focus on these two theses more or less in turn. I will start with a broad overview in chapter 2. I'll review some of the previous work on the general topic, and further flesh out the broad concepts and motivations I've alluded to so far. This is more or less re-hashing and re-packaging a lot of things that other people have probably said better, though maybe not all in one place. Then in chapter 3 I'll talk about the *dimensions* of reproducibility. I will argue that the notion of *reproducibility* is inherently gradient and scalar, and that any definition or standard will have to make some decisions about the target dimensions that qualify something as "in" or "out" with respect to that local definition of what counts as reproducible. In other words, *reproducible* is a relative goal, and while it may be possible to establish standards for a particular field/context, there is no such thing as a useful universal standard. I will then go on to suggest some possible starting points for where to position a reasonably useful definition along these dimensions. That will conclude all of this nosebleed-level discussion.

I'll then turn to a slightly more concrete discussion of how to make the day-to-day tasks of research more reproducible. In chapter 4 I propose a general schema for the cycle of research, from reading other people's work to producing your own work. Again, my purpose here is not to go too deep into the philosophy of science, so this is just intended as a way of breaking down the research process into some large chunks. I think this is useful for the present purpose, because these different kinds of activities will involve different issues, problems, and standards regarding what it means to be *reproducible*. But the big point is that this broad view of research goes far beyond the realm of what most people talk about when the term *reproducible research* gets thrown around.

In the chapters that follow (5 through 9), I will focus in one one of these domains of research, and discuss some of the special challenges for reproducibility, and at times grapple with the question of whether (and how) activities in this domain could ever be reproducible. In each of these chapters, I will discuss how the other dimensions (*range* and *audience*) play out for that particular domain. Each chapter will also contain a general description of some practices that could lead to better reproducibility, and I will link heavily to my repository of tutorials that will implement specific software (or non-software) solutions to the implementation problem.

The final chapter will conclude with some parting thoughts.

So let's get started!

Chapter 2

Principles

2.1 Replication vs. reproduction

Let's start with a terminological clarification. *Replication* and *reproduction* are terms that often get used to talk about similar yet distinct concepts. The basic idea is that *replication* refers to one of the cornerstones of science, which is that someone can perform an experiment, gets some results, and someone *else* can perform a *similar* experiment, and get a set of results that is consistent with the first set. We can say that these results have been *replicated. Reproducible* usually refers to the ability to get *exactly* the same results starting from the same set of data.

I argue that the meaningful distinction here is that *replication* is about *results* and *reproduction* is about *methods.* This generalizes the notion of reproducible a bit, because it may be possible to reproduce a set of methods almost exactly, but if you're collecting data from a different set of participants, you may not end up replicating some initial result. And conversely, it may be difficult or impossible to reproduce someone's methods exactly, or you may be intentionally running an experiment with a variation on the methods, but you still might get a set of results that replicate some earlier findings.

Throughout this book, I will assume that *replication* is a non-negotiable part of science. If no one can produce results that fit with an earlier set of results, then you assume there was something wrong with the initial results (all else being equal). I propose that *reproducibility* is a methodological approach (or set of approaches) that should make replication easier if a result is true, or should make aberrent or erroneous results more transparent. I'd go so far as to say that reproducibility is the means of facilitating the processes of replication and falsification that are critical to science.

2.2 A sketchy history of reproducibility

I will not go through a detailed, scholarly review of reproducible research, but I will give a brief overview, because it will help orient the reader to how the content of this book fits with other discussion of the general topic.

At its root, the discussion of reproducible research comes from computer science, statistics, and more recently, other disciplines (like genetic biology) that involve increasingly computation-heavy analysis. One of the seminal works is Knuth (1984), which coined the (titular) term "literate programming." the specific idea was that good programs should be written with the ***human*** reader in mind, not (just) the computer that runs the programs. This idea has lead to different ideas and developments in methods of documenting code, methods of writing code to be more legible, and methods of intertwining human language that explains the code with the code language itself. But the broader idea is the foundation of reproducibility in the computer sciences, which is that in order for a program (and its results) to be shared effectively with the broader community, some care needs to be taken in how it is created. A program that is opaque to people other than its creator will be far less useful than a program that can be understood easily, because the latter can be improved, expanded, adapted, modified, and debugged far more easily.

2.3 Some core values

We've already discussed replication, but what else is at stake here? Why should we care about reproducibility?

We live in an exciting but perilous time for science. The rise of the internet into a mature infrastructure, the continuing advances in personal and large-scale (e.g., cloud) computing, great strides in terms of data collection and analysis of types we could barely contemplate 10 or 20 years ago, and so on, provide an exciting new global realm of scientific discovery and collaboration. On the other hand, because of various domains of economic and social change, science is also under attack. Basic scientific education (for both "hard" and "soft" sciences) is in jeopardy in many spheres of public education. Funding for basic science[1] is becoming harder to come by, and more competitive. Several severe cases of academic fraud have received a lot of exposure in the popular press. So while the potential for advances in sciences, including social sciences, has never been greater, issues of risk, accountability, and demonstrating value to society loom heavy over the academic landscape.

An absolutely critical piece for both sides of this picture is the concept of *reproducible research.* On the one hand, as possibilities for new data sources and analyses and collaborations explode, reproducibility is key, in order to maintain trust and order within the scientific community. For example, some recent advances in statistical methods that have recently become much more accessible (e.g., mixed-effects mod-

[1] "Basic" science is typically described as "science for sciences' sake." In other words, science for the sake of increasing understanding. This is contrasted with "applied" science, which is science with an aim of addressing some real-world problem with direct social, economic, or military applications implications. So for example, theoretical particle physics (e.g., the search for the Higgs boson) or theoretical linguistics (e.g., the search for abstract linguistic universals) are basic science, and developing a particular branch of particle physics to explore some new energy weapon, or applying a theory of universal grammar to problems in machine translation, are examples of applied sciene. In reality, there's a large continuum. The point here is that in all sorts of domains, it's harder and harder to do research without some kind of applied angle to justify funding.

els, Bayesian analysis) are still not fully integrated or fully understood by researchers in many fields. This means both that some researchers may be employing methods they do not (yet) fully understand, and that some journal reviewers may be resistant to new methods, even if they do not have good reason to be, simply because they are unfamiliar. More transparent, reproducible methods of employing these and other more novel analyses would greatly facilitate the ability to share, evaluate, and critique these methods.

Reproducible methods promote transparency and trust. When people outside the academic scientific community can pick up and replicate analyses and results, it can help break down the "ivory tower" metaphor. It increases accountability and decreases the possibilities for fraud and scandal. If anyone with a computer can re-run and inspect for themselves some important analysis of (e.g.) climate change, voter fraud, economic disparity, health issues, etc., then there is far more opportunity to bring discourse of such topics into the realm of facts and better decisions, and out of the realm of heresay and fact-twisting partisanship. There is a growing practice of people circulating various plots and graphs through social media like Facebook or Twitter, showing things like debt growth under Democrats vs. Republicans, relationships between gun laws and gun violence, etc., etc. But without an ability to replicate the methods (and directly inspect the data) that went into creating such graphs, there is no real reason to trust any of them. A bar graph can lie just as easily as anything else, especially if you can't see how it was made.

Within academia, there has been a growing recognition and dissatifaction of the problem of replication. The standards for publications in most fields reward studies (by publication and dissemination) for showing effects, while "null results" or failed replications of the same studies may have an extremely difficult time getting published, even though many well-done failed replications should cast significant doubt on the initial published effects. To make matters worse, replication is more difficult and resource-consuming if the original study is not very thoroughly described. By making replication easier, we can save costs and time by reducing the amount of resources wasted on failed replication attempts. There are some interesting current projects trying to address the so-called "file drawer" problem of unreported failed replications, but increased reproducibility is a critical piece of making such efforts successful.

Currently, reproducible research is a fairly hot topic in some circle. The rise in popularity of the open-source statistical software `R` has generated a fair amount of interest, because this software encourages and enables reproducible research in a way that the more popular commercial software packages (SAS, SPSS, etc.) do not. The ideas of *literate programming* (Knuth 1984) have spilled over from the programming world into other academic disciplines, and many of the tools developed for programmers to work with code are turning out to be useful for reproducible research in other fields. There are websites dedicated to the topic, countless blog posts on the virtues of `R` and `Sweave` for reproducible research, and in a new column in the journal *Chance* dedicated to ethics in statistics, prominent statistician Andrew Gelman started off by examining a case of non-reproducible results (in the practical sense, that the authors refused to share the data that would allow for reproduction). A boycott movement against big for-profit publishers like Elsevier and proposed legislation like the Research Works Act have re-invigorated the dialogue about open access to publication of scientific results. So with all this current interest and influx of new tools, what's the purpose of this thing you're reading? And conversely, why should *you*, gentle reader, bother to care? As the first part of the answer, I will discuss who this book is intended for.

2.4 Who is this book for?

Full disclosure: this book is primarily a selfish effort. I am putting this system together to improve my own methods and productivity, and having it all written down in an organized way is a helpful way for me to put my thoughts together into a system that's documented and clear. But this point also gets to the heart of it: I believe that adopting more reproducible methods will lead not only to Better Science, as alluded to by the intro above, but a more efficient and productive workflow. In other words, I believe that Really Reproducible Research (more on what I mean by that in section **??**) is not just The Right Thing for Science, but The Way to Get More Done and Published.

Because of this last point, I intend my primary audience (after myself) to be academic scientists. It also follows from my self-centered goals that the specifics will be geared towards academics working in linguistics, and psycholinguistics. However, I expect that people from other fields could find the discussion and implementation helpful, and easily adapted. To frame it another way, ask yourself the following questions:

- Have you ever picked up an old paper of yours and wished you could remember some detail of how the data was collected/analyzed?
- Have you ever needed to update a figure/table/statistical analysis after a change in the data or analytic procedure?
- Have you ever read someone else's paper and wished you could see exactly how they did their analysis?
- Have you ever been frustrated in how much time you spend chasing down and re-typing/re-formatting the same set of references across multiple papers?
- Have you ever had a request for your data/analysis/other details from a paper and shuddered at the effort needed to share it in an accessible way?
- Have you ever had a problem with inconsistency in a paper, where the stats are from one data set, but the figures (or summary tables, or stats in another section) are taken from a different data set (e.g., after some additional data, or some additional data-cleaning, or something)?
- Have you ever lost track of what kinds of data-cleaning (outlier trimming, transformations, missing data, etc.) have been performed on a data set, and which ones were applied to results in a given paper or presentation?

- Have you ever gone through some laborious data-organization or analysis process (e.g., sorting/labeling/tweaking/cleaning things in Excel by hand), only to have to do it over and over when you discover mistakes or when the data changes in some way?

- Have you ever taken hours to carefully construct some kind of complex figure or diagram by hand (e.g., graph, flowchart, theoretical model, syntactic tree), only to have to re-format it for a journal submission, or a talk handout, or a PowerPoint presentation, or some other formatting issue?

If you are still in the early stages of your career, and are unsure about whether anything like this may happen to you, just do a quick poll of your advisor, more advanced students, etc. If none of these things apply to you, you are likely either (a) not an academic, (b) an academic in a non-scientific field, or (c) already doing a fantastic job doing reproducible research. But if any of these things apply, and you like the idea of doing something about it, then my hope is that this book will help.

Finally, this book does *not* assume you already have facility with programming, etc. Many of the implementation tools I'll discuss in Part 11 involve some level of savvy in programming, using command-line tools, and other things normally associated with steep learning curves. My intention is to present arguments for why these tools are worth the effort to learn and use, but I will start out assuming that the reader is a user of commercial products like the Windows or Mac operating systems, programs like Microsoft Word and maybe a little of Excel, and a graphical stats package like SPSS or JMP, if anything. My hope is that this book could be picked up by people early in their academic careers and applied as they go. My greater hope is that the ideas will be appealing enough and the implementation easy and effective enough that even experienced, established academics could find some utility in improving some of their habits and/or tools. People tend to get entrenched, though, so I'm not holding my breath on the latter group. But one can hope...

2.5 Goals

So what exactly do I hope to accomplish with this book? What exactly should you, the reader, expect to be able to get out of it? To return to my motivations and audience, I would like to enable linguists and psycholinguists (and others, perhaps) to produce Really Reproducible Research, from soup to nuts. Currently, there are bits and pieces of resources and ideas spread around multiple fields and websites and repositories. My purpose here is to collect what I think are the best of the best, and assemble them into a system of principles, tools, and methods that will work well together for a "complete" system of Really Reproducible Research. Additionally, many resources on reproducble research (including the website of that name) are geared primarily towards computational or statistical work, and their principles can be summed up as "share your data, include your code, and make your code legible to others." These principles are certainly relevant, but they don't capture the whole messy system of producing scientific research that is truly reproducible.

Therefore, on the one hand I aim to present a more general discussion and system for carrying out reproducible scientific research beyond "include your code," and on the other hand, I aim to provide a very specific configuration of tools geared towards carrying out reproducible research in linguistics and psycholinguistics. The book is organized with these goals in mind. In the rest of this first part of the book, the discussion will remain tool agnostic in general, although the principles discussed will end up favoring some kinds of tools over others. The goal of this part of the book is to lay out the principles and concepts for what it means to carry out Really Reproducible Research, and what the benefits and drawbacks might be.

The second part of the book makes this more concrete by spelling out a particular implementation. The implementation is partly a set of software recommendations and partly a set of workflows, procedures, and methods for doing typical research tasks in a way to support Really Reproducible Research. There will be plenty of room for customization, because I don't expect that any two researchers will want to do things in exactly the same way, but the goal is to be as specific and concrete as possible, so that you are not left wondering about how to connect the dots. Some suggestions for alternatives will be included, but I will focus on tools that I use and that I think are best for the job, and I will not go through an exhaustive review of tools I'm less familiar with.

Finally, the third part of the book is a set of tutorials designed to enable you to use the tools in the implementation. For example, I discuss Emacs and org-mode as major tools in the implementation. Most people are not Emacs or org-mode users. Both Emacs and org-mode have extensive documentation, including books, tutorials, and tons of articles spread across the web. However, existing documentation is both more and less than what you would need to implement the system I describe in Part 11. They are *more* in the sense that there are *tons* of functions in both Emacs and org-mode that may be great features and very useful, but not relevant or necessary for the system I outline here. Existing tutorials and documentation also provides *less* than what I do here, in the sense that using these tools in the specific way I describe in Part 11 may not be obvious, even if you worked your way through the general manuals or tutorials already available. In other words, my goal is not to teach you Emacs for all general purposes, but rather to teach you how to use Emacs in the system of Really Reproducible Research. Even if you know Emacs, there may be something useful for you in my tutorials, but there will also be lots more to learn about Emacs after you've mastered my tutorials. And as I mentioned in the previous section, I will assume that you have experience with Word, and that's about it, so you should approach the tutorials with minimal anxiety.

With these three parts, my ultimate goals are to (1) describe what I think are the critical elements of reproducible research and convince you that these are worthy and useful goals, (2) describe a concrete system for achieving reproducible research in the real world of working academia, and (3) enable readers with no knowledge of the tools I describe to learn and apply these tools in their own personal approach to reproducible research. This way, I hope that the end result of this book is not just a series of suggestions, but the actual means to implement and improve upon my idea in your own work.

Chapter 3

Dimensions of reproducibility

Thus far, I have only hinted vaguely at what *reproducible* really means. I have frequently used the phrase "Really Reproducible," implying that some values of "reproducible" may be less than desired. In this chapter, I will tackle the definition of "reproducible" in a more systematic way. I argue that "reproducible" is a continuum, and even more so, a two-dimensional continuum. With this understanding, we are in a better position to zero in on appropriate principles and standards in defining a target for what Really Reproducible means.

The first approximation of reproducible comes from the general idea in the scientific method that results should be able to be replicated. That is, I can present some data and an analysis, and in order for it to qualify as "good science," it should be possible for someone else to also collect similar data and perform a similar analysis and get (generally) the same result. Put another way, if no other scientist/lab in the world can get the same results you can, that's a big problem.

But when you start thinking about this seriously, it's apparent that this raises two questions. Reproducible by who? How "similar" must the data, analysis, and results be for it to qualify as "reproducible"? These are what I call the *domain* and *range* of reproducibility.

3.1 Domain: the audience

The first dimension of reproducibility is the *domain* or the *audience.* In short, *who* do you expect to be able to reproduce your work? On one end of the continuum is yourself. If you cannot reproduce your own work, how could you expect anyone else to? On the other end of the continuum is virtually anyone in the world, which is probably almost always impossible. Figure 3.1 illustrates a few important values for the domain.

3.2 Range: the precision

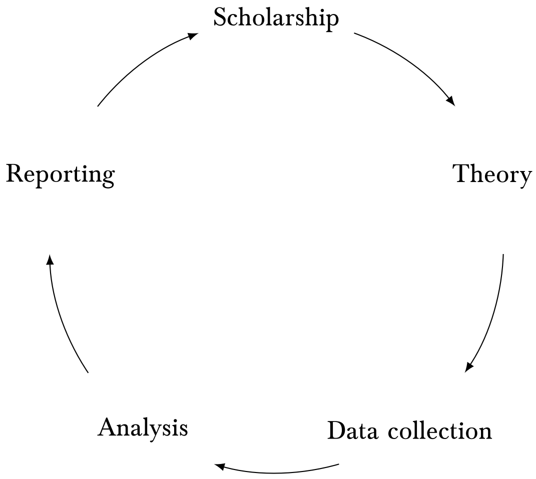
3.3 The minimum

3.4 Publication standards

3.5 Wide dissemination

Chapter 4

The cycle of research



Chapter 5

Scholarship

Chapter 6

Theory

Chapter 7

Data collection

Chapter 8

Data analysis

Chapter 9

Reporting

Chapter 10

Conclusions

Chapter 11

Implementation

11.1 Guiding principles

11.1.1 Free and open source

11.1.2 Cross-platform

11.1.3 Stable

11.1.4 Well-documented

11.1.5 Customizable

11.2 Summary of tools

11.2.1 Emacs

11.2.2 Org-mode

11.2.3 Git

11.2.4 LaTeX

11.2.5 Python

11.2.6 R

11.2.7 (Emacs) Lisp

11.3 Scholarship

11.4 Data collection

11.5 Data analysis

11.6 Sharing

11.7 Collaboration

11.8 Putting it all together

Bibliography

Knuth, D.E. (1984). “Literate programming”. In: *The Computer Journal* 27.2, pp. 97–111.