

LANDSCAPE ARCHAEOLOGY AND REPRODUCIBLE RESEARCH AT THE 2017 BERLIN SUMMER SCHOOL

July 28, 2017 Ben_Marwick Day of Archaeology 2017, Digital Archaeology, Education, Explore Posts, Germany, Science Archaeology, Berlin, Landscape archaeology, reproducible research, Science

For this year's Day of Archaeology, I'm writing about what I was up to the week before. This is because on the actual Day of Archaeology I am quietly working alone on my computer to prepare a lecture for undergraduates, write some text into a few publication drafts, and send/receive a bunch of emails. Not very exciting to look at and much less fun than what I was doing last week. Last week, I was at the 2017 Summer School on Reproducible Research in Landscape Archaeology at the Freie Universität Berlin (17-21 July), funded and jointly organized by [Exc264 Topoi](#), [CRC1266](#), and [ISAAKiel](#). With a group of 16 archaeologists and geographers from Berlin, Kiel and Cologne, we spent the week learning how to make our research more reproducible, and learning advanced geostatistics.

Summer School
July 17–21, 2017

Reproducible Research in (Landscape) Archaeology

A change is happening at the moment. „Open“ is the new paradigm for software, data access, exchange and publication. One of the greatest benefits of this change is that re-producible research actually becomes possible. However, there are some elementary things a researcher needs to take care of in order to allow others (or themselves) to re-produce and re-analyze their published study. Prof. Ben Marwick, one of the leading practitioners in the context of (landscape) archaeology, will present techniques and tools for reproducible research which are applied and tested within this summer school.

Extracting Sunbeams Out of Cucumbers:
Why Archaeology Isn't a Science, and How It Can Become One
Public Keynote Lecture given by Ben Marwick
Monday, 17 July 2017 at 7:00 pm

www.topoi.org/event/42329/

Einstein Center Chronoi • Freie Universität Berlin • Otto-von-Simson Str. 7 • 14195 Berlin

Ricarda Braun
Excellence Cluster Topoi
Research Group (A-1) Ancient Colonizations of Marginal Habitats
Freie Universität Berlin, Physical Geography
Malteserstr. 74-100, 12249 Berlin
ricarda.braun@fu-berlin.de

Daniel Knitter & Wolfgang Hamer
CRC1266 – Scales of Transformation
Project A2 „Integrative Modeling of Socio-Environmental Dynamics“
Christian-Albrechts-Universität zu Kiel Physical Geography
Ludewig-Meyn-Str. 14, 24118 Kiel
knitter@geographie.uni-kiel.de

Oliver Nakoinz
CRC1266 – Scales of Transformation
Project A2 „Integrative Modeling of Socio-Environmental Dynamics“
Christian-Albrechts-Universität zu Kiel, Pre- and Protohistoric Archaeology
Johanna-Mestorf-Straße 2-6, 24118 Kiel

EXCELLENCE CLUSTER TOPOI Freie Universität Berlin EINSTEIN CENTER Chronoi BERLINER ANTIKE-KOLLEG CRC1266 SCALES OF TRANSFORMATION CAU Christian-Albrechts-Universität zu Kiel

Poster for our Summer School (copyright: TOPOI)

Our Summer School was housed in a beautiful post-war villa (remodeled to be a university building) in the leafy suburb of Dahlem in south-west Berlin. Most of this suburb is filled with buildings for the Freie Universität, one of the most selective and high-ranking universities in Germany. It was founded shortly after WWII in what was then West Berlin, and the name is a deliberate contrast to the older Humboldt University located in the former East Berlin. Both universities are now free and flourishing, and have

active archaeology programs. Many of their current archaeological research activities are linked to the [Excellence Cluster Topoi](#), a unique long-term multi-disciplinary collaboration of researchers from all subjects related to ancient studies and archaeology. The Topoi is part of the [Berliner Antike-Kolleg](#), which partners with the [Max-Planck institute for the History of Science](#), the [Berlin-Brandenburgische Akademie der Wissenschaften](#), the [Stiftung Preussischer Kulturbesitz](#), [Einstein Foundation](#), and the [German Archaeological Institute](#) (an archaeological research institute under the auspices of the federal Foreign Office of Germany), whose headquarters were right around the corner of our venue. So we were in a very stimulating environment for doing archaeology. Our group consists of researchers and PhD students from Topoi project as well as the collaborative research center 1266 of Kiel university. The Summer School, part of a series of summer and winter schools in turns organized between Kiel and Berlin, had a dual focus on methods for reproducible landscape archaeology, and advanced techniques in point pattern analysis and spatial interpolation.

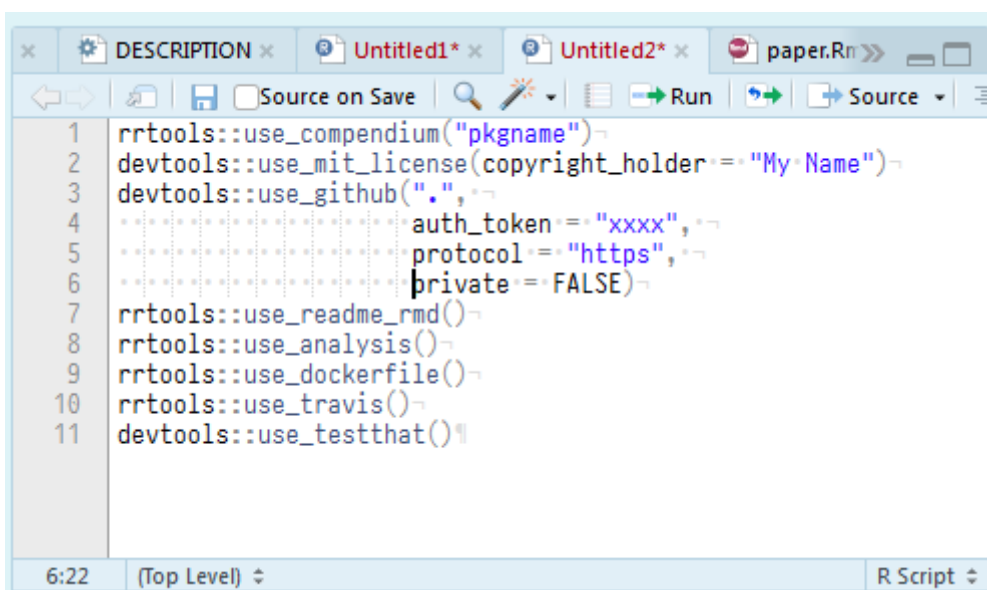
Reproducibility is a relatively [new concept for archaeology](#) that has only recently [received some attention](#), and so we encountered many challenges in teaching and learning this unfamiliar subject. The [general idea](#) is that a piece of research, such as publication, should contain enough information for another researcher to reproduce the results in that paper. In former times, when archaeology was simpler than today, this was relatively straightforward. However, today this is not the case, with computers and complex software and algorithms playing a key role in our data analysis. Thanks to computers, the complexity of modern archaeology means that it is difficult to fit all the details of an analysis into the text of a typical journal article. There just isn't the space. This is a serious problem because the lack of detail in modern publications makes it hard for us to decide if the research is reliable or not, and slows us down in reusing published research in our own projects.

In our Summer School we explored solutions to this problem of reproducibility. Our focus was the [R](#) programming language, which allows us to write instructions for every step of our data analysis. Many archaeologists are now turning away from Excel and SPSS towards [R](#) to write scripts for their data analysis. There are many advantages to this. R works on any computer, and is free to install and use, unlike commercial products. It also has over [10,000 add-on packages](#), so there is tremendous flexibility in the types of analysis it can do. And, of course, because we write scripts to use with R, we have a record of our analysis that we can share with our publications. This means that people reading our publications can easily see all the decisions we made during our data analysis. They can take our code and change it to explore the effects of making different decisions in the analysis.

Our group consisted of already advanced R users, and the challenge we faced was how to organise our projects in a simple, logical way to enhance reproducibility. We wanted to take advantage of existing conventions in R to organise our work in a way that makes it easy for other users to navigate. Our goal was to come up with some simple steps for getting organised with a new or ongoing project that uses R. We were also keen to investigate how we can use modern software engineering tools to automate some of the processes to check the reproducibility of our research. We wanted our final product to embody the [best practices](#) that we've read in several [recent high-profile publications](#).

Our solution was one of the main results of our workshop. This is the new R package `rrtools`, or ‘reproducible research tools’. This free and open source package provides instructions, templates, and functions for making a basic compendium suitable for doing reproducible research and writing a journal article or thesis with R. After much experimentation and trial and error in our group, we boiled the set-up process down to eight quick and easy steps. [These steps](#) include:

- . Creating a new empty R package as the [research compendium](#). This is important, because most R users recognise the file structure of an R package. So they will know where to look for the different parts of the compendium.
- . Adding [licenses to specify how we want the contents of the compendium to be reused](#) by other researchers
- . Connecting our compendium to [GitHub](#), a web site for sharing code and data files. GitHub is also an excellent system for collaborating on writing papers and code, and sharing files.
- . Adding some basic, structured machine-readable and human-readable metadata, instructions on how to cite the compendium, and instructions for potential contributors.
- . Adding a file structure and document templates to organise the typical components of a research project and accompanying journal article or book. For example, separate directories for data and code.
- . Adding a Dockerfile that contains a recipe for making an [isolated computational environment](#) for our analysis. This means we’re not constrained to a specific computer with a specific setup to ensure that our analysis can be done correctly. We can create a Docker container to reproduce the exact computational environment that our original analysis was based on.
- . Connecting our repository to [an online service that automatically tests our code](#) each time we update it. This saves us from having to spend a lot of time testing our code – the Travis service runs the tests for us.
- . Adding [tests for our custom functions](#). In case our works includes complex R functions, we should add tests to ensure that our functions do what we expect.

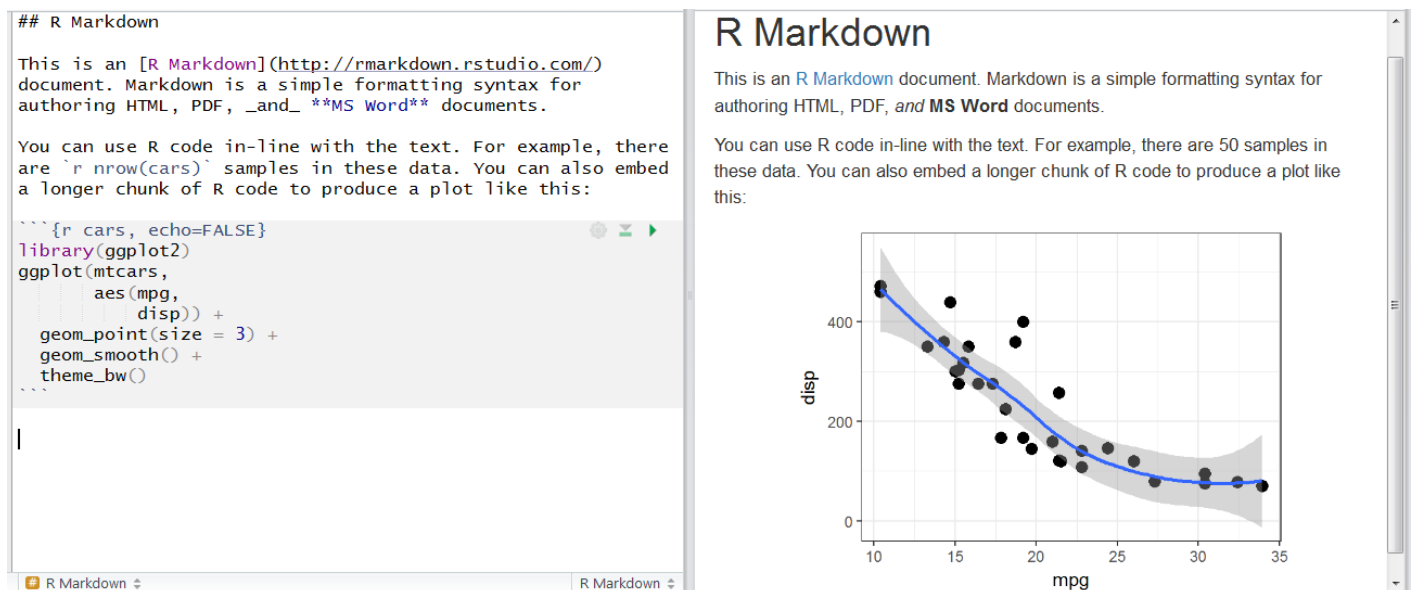


```
1 rrtools::use_compendium("pkgname")
2 devtools::use_mit_license(copyright_holder = "My Name")
3 devtools::use_github(".",
4 ..... auth_token = "xxxx",
5 ..... protocol = "https",
6 ..... private = FALSE)
7 rrtools::use_readme_rmd()
8 rrtools::use_analysis()
9 rrtools::use_dockerfile()
10 rrtools::use_travis()
11 devtools::use_testthat()
```

Screenshot of R code in RStudio, showing the small number of steps needed to create a research compendium that follows current best practices

For each of these steps, we have a one-line function to complete the task. In less than five minutes you can start from nothing, run these functions, and then have a complete research compendium. In our group we mostly used [RStudio](#) to run R, but we also had some [Emacs users](#). The eight functions described above save a lot of intermediate struggles with organising our work to be transparent, open and reproducible. The file structure generated by our package includes several template files. With these templates we can start writing and doing data analysis right away.

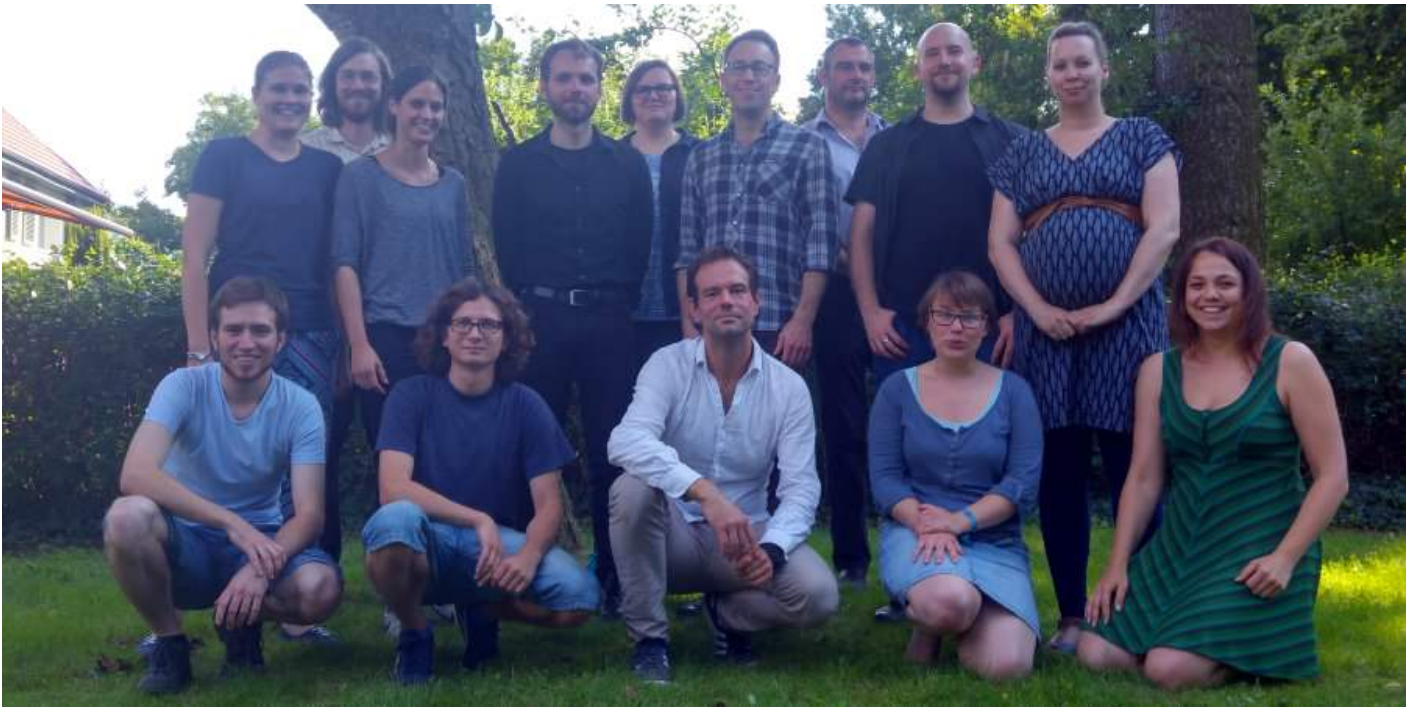
The writing system we use in this approach is called [R Markdown](#). It's different from writing in a word processor, because the document does not look exactly like how it will print out. Instead, we see plain text with some code-like formatting to specify bold, italics, etc. Markdown allows us all the usual scholarly writing tools, such as citing references (we especially recommend [Zotero](#) to help with this), adding tables and figures, and including numbered captions and cross-references. Using R Markdown, we can weave blocks of R code in between the paragraphs of text. These blocks of code can make plots or tables that will appear in the final document. It is possible to write entire journal manuscripts in this R Markdown system, as some of our group have already done.



The image shows a side-by-side comparison of R Markdown source code and its rendered output. On the left, the source code is displayed in a text editor. It starts with a title '## R Markdown', followed by a paragraph of text explaining R Markdown. Then, there is a code block for plotting the 'mtcars' dataset using ggplot2. The code includes: `{r cars, echo=FALSE}`, `library(ggplot2)`, `ggplot(mtcars,`, `aes(mpg,`, `disp)) +`, `geom_point(size = 3) +`, `geom_smooth() +`, and `theme_bw()`. On the right, the rendered output is shown. It has the same title 'R Markdown' and text, but the code block is replaced by a scatter plot. The plot shows 'displacement' (dis) on the y-axis (ranging from 0 to 400) versus 'miles per gallon' (mpg) on the x-axis (ranging from 10 to 35). The data points are black circles, and a blue smoothed line is overlaid, showing a negative correlation. A grey shaded area around the line represents the confidence interval.

This is what R Markdown looks like. We write the text and code on the left, and we get the output on the right (copyright: Ben Marwick)

We're very pleased with this outcome from our Summer School. This [rtools](#) package was born out an intense effort by our group to find simple and easy methods to improve the reproducibility of archaeological research. We're already using it among our own projects. We invite anyone interested to [take a look](#), try it out, and give us your [feedback](#).



Our Summer School group in the lovely garden of our villa, last Friday. We're about to enjoy a BBQ on a pleasant warm evening after the last day of the workshop (copyright: Franziska Faupel)