

Archaeology Data Infrastructures

Data reuse potentials and limitations to modelling settlement systems (...)

Petr Pajdla

11/12/22

Table of contents

Preface	4
Notes on writing	4
Introduction	6
Context	6
Archaeological heritage management in the Czech Republic	6
Research questions	6
Thesis outline	6
1 Theory	7
1.1 Definitions and terminology	7
1.1.1 Data	7
1.1.2 Data infrastructures	7
1.2 Overview of theoretical concepts	7
1.3 Digital humanities	8
1.4 Digital archaeology	8
1.5 Spatial archaeology	8
1.6 Archaeology as theory- and/or data-driven science	8
1.7 Theorizing data	8
2 Methods	9
2.1 Software	9
2.1.1 Reproducibility	9
3 Data and materials	10
3.1 Data management plan	10
3.1.1 Re-using data	11
3.1.2 Creating and collecting data	11
3.1.3 Processing data	11
3.1.4 Interpreting data	11
3.1.5 Preserving data	11
3.1.6 Giving access to data	11
3.2 Data sources	11
3.2.1 Archaeology information system of the Czech Republic	11
3.2.2 Legacy data sources	11

Preface

Warning

This is a website for the **work-in-progress** PhD thesis of mine. It is **not** intended to be read by anyone except me (*and maybe few other people*) yet. If you do flick through it anyway, consider yourself warned. It might be messy at some places and will definitely undergo serious rewriting.

Note

This work can be read online at <https://petrpajdla.github.io/dataInfrastructures/>. The source repository is on GitHub at <https://github.com/petrpajdla/dataInfrastructures/>.

This document is created in an open-source [Quarto](#) scientific and technical publishing system. You might be asking why is it published and written like this even if it is not intended for any audiences except myself yet. I have no answer to this. One evening I simply decided to give *Quarto* publishing a try and set this whole thing up in less than an hour or so.

Notes on writing

This note is written mostly for a future me, in case I need to set up the working environment again on a different machine and to serve as a memo if I forget how to continue.

As of November 2022, this is written on [Archlabs GNU/Linux](#) machine, mostly in [Visual Studio Code](#) editor and sometimes in [RStudio](#). Changes are tracked with *Git* and a remote repository is on *GitHub* (see the note above), same as the rendered website. The rendered version of the manuscript is in the branch **gh-pages**. See a guide on how to set this up [here](#). The online version is published with this command:

Terminal

```
quarto publish gh-pages
```

In my point of view, there are numerous advantages to scientific writing in this manner over traditional *Office*-based approach. A non-exhaustive list of why to do scientific writing this way is below.

- **Plain text**

Writing in plain text enhanced with a simple *Markdown* syntax and some *Quarto* elements is great because from one source document, a *.pdf*, *.html*, *.docx* (and probably more) document formats can be rendered using [pandoc](#).

- **Version control**

Tracking changes using *git* is easily implemented when writing in a plain text. Keeping track of any changes in the manuscript is obviously crucial for any later revisions etc.

- **Simple citation management**

Bibliography is organized using [Zotero](#) with [Better BibTeX](#) extension which is used to export (and keep updated) necessary collections in a parent folder of the manuscript as *.bib* files. My *Zotero* library is [here](#). To format the citations, a citation style of the *Journal of Computer Applications in Archaeology* is used (.csl file was obtained [here](#)).

- **Embedded code**

Code blocks (and the associated results) can be easily embedded in the text. My language of choice is *R*. For more information on reproducibility see Marwick (2017) and Marwick, Boettiger, and Mullen (2018).

In-text citations

```
@citekey      -> Author (year)
-@citekey     -> (year)
[@citekey]    -> (Author, year)
@citekey [p. X] -> Author (year, p. X)
```

Crossrefs

```
{#sec-label} -> #sec-label
{#fig-label} -> #fig-label
crossref withot numbering: -@sec-label, [Chapter -@sec-label]
```

Introduction

Context

Archaeological heritage management in the Czech Republic

Research questions

Thesis outline

In Chapter 1, *Theory*, the foundation is given by defining basic terms, data, data infrastructures etc. Then, theoretical approaches the work spans from are discussed and the concept of data in archaeology theorized. The dichotomy between archaeology as data- and/or theory-driven science is debated.

In Chapter 2, *Method*, methodological boundaries are set up.

Chapter 3, *Data*, introduces data sources that are used here. Understanding the data models employed in various data sources is vital for any subsequent steps taken in the analytical process. Special attention is thus given to analyzing how reality and facts are represented by the data models of used sources. A data management plan (Section 3.1) details how data is handled in this research.

1 Theory

1.1 Definitions and terminology

1.1.1 Data

The term data is used in a plural form what is the current scientific convention (Kitchin 2022, xvii). As Kitchin (2022, 15) states, “Data are not simply captured or recorded, but are the product of discursively framed and technically mediated processes.”

“The production of data is a social practice, conducted through structured and structuring fields (e.g. methods, concepts, expertise, institutions) that are shaped by and contribute to configurations of power and knowledge.” (Ruppert, Isin, and Bigo 2017)

“(…) databases are designed and build to hold certain kinds of data and enable certain kinds of analysis, and how they are structured has profound consequences as to what queries and analysis can be performed.” (Ruppert 2012)

1.1.2 Data infrastructures

As I was saying in the Section [1.1.1](#).

1.2 Overview of theoretical concepts

Review of current approaches: Spatial and/or Landscape archaeology, Macroarchaeology, Big data archaeology etc. Describe software used!

1.3 Digital humanities

1.4 Digital archaeology

1.5 Spatial archaeology

1.6 Archaeology as theory- and/or data-driven science

i Note

This section is partly based on the *Data-driven Archaeology. Are we there yet?* talk coauthored with Hana Kubelková and Petr Květina. It was presented at the *Central European Theoretical Archaeology Group (CE TAG)* meeting entitled *Theoretical Approaches to Computational Archaeology I* coorganized with Michael Kempf, Jan Kolář and Jiří Macháček in 2021 at the Department of Archaeology and Museology, Faculty of Arts, Masaryk University.

1.7 Theorizing data

Defining archaeological data, micro- to macro-scales;

2 Methods

2.1 Software

Most of the things included here, if not all of them, were achieved using open-source software. Large part of this endeavour is also documented in code. This text was written in plain text with some basic markdown and quarto syntax for formatting, cross references, citations etc. At some places there are R code blocks. The text is processed into three outputs, a [website](#) (HTML document), a [PDF](#) document and a [MS Word](#) document using Quarto. The plain text version, same as the rendered website, is hosted at [GitHub](#). The text was mostly written in the Visual Code Studio, analysis were mostly performed using Rstudio or terminal. Library was organized using Zotero.

Raster graphics were created and edited using GIMP, vector graphics using Inkscape. All the GIS operations that required graphical user interface (GUI), or were more conveniently performed in a GUI, were done in QGIS.

Some data were prepared, extracted or processed using basic GNU/Linux shell or SQL commands or scripts. Data from Wikidata was queried using SPARQL. Any analysis was mostly done in R, a language for statistical computing and graphics (**r**core?). Various packages were used, the most important packages are listed here, the complete list is in an Appendix

2.1.1 Reproducibility

3 Data and materials

i Note

This chapter, especially the Section 3.1: Data management plan, builds up on the project [Data management in Archaeology](#) I cooperated on with Hana Kubelková in 2021 at the Department of Archaeology and Museology, Faculty of Arts, Masaryk University.

Sources of (archaeology) data in the Czech Republic, an overview:

Data models, datafication of past reality, simple vs complex data models; Assessing findability, accessibility, interoperability, and reusability (FAIR) principles; Cultural heritage management data vs research data domains; Archaeological information system of the Czech Republic (AIS CR) as the main data infrastructure.

3.1 Data management plan

Good data stewardship is a crucial element in *Open Science* (Mons 2018, 1–5), an umbrella concept for how scientific research is conducted in a way that knowledge is reusable, modifiable and redistributable. The data management plan (DMP) then stands at the very beginning of every such endeavour. In its essence, a DMP is a stand-alone document detailing how data is handled at each of the steps in its life cycle. This implies that it is not a static, but a living record of how the data was gathered and/or captured, curated, selected, analyzed, interpreted, shared and archived in the course of a project or after its end. A DMP helps in adhering to the FAIR principles, i.e. making data findable, accessible, interoperable and reusable, a set of propositions enabling more effective knowledge discovery, collaboration, and data reuse (Wilkinson et al. 2016; Hollander et al. 2019).

This DMP is partly based on the structure given in the [Data Stewardship Wizard](#) (Pergl et al. 2019), an online tool dedicated to cooperative creation of DMPs, templates created in the [Ariadne project](#) (Doorn and Ronzino 2022) and my own ingenuity.

3.1.1 Re-using data

3.1.2 Creating and collecting data

3.1.3 Processing data

3.1.4 Interpreting data

3.1.5 Preserving data

3.1.6 Giving access to data

3.2 Data sources

3.2.1 Archaeology information system of the Czech Republic

3.2.2 Legacy data sources

What is a legacy data source?

3.2.2.1 Museum databases

3.2.2.2

References

- Doorn, Peter, and Paola Ronzino. 2022. “ARIADNEplus Data Management Plan Tools.” *Ariadne Portal*. <https://vast-lab.org/dmp/>.
- Hollander, Hella, Francesca Morselli, Frank Uiterwaal, Femmy Admiraal, Thorsten Trippel, and Sara Di Giorgio. 2019. “PARTHENOS Guidelines to FAIRify Data Management and Make Data Reusable,” August. doi:[10.5281/zenodo.2668478](https://doi.org/10.5281/zenodo.2668478).
- Kitchin, Rob. 2022. *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. Second. Los Angeles, California: SAGE Publications.
- Marwick, Ben. 2017. “Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation.” *Journal of Archaeological Method and Theory* 24 (2): 424–450. doi:[10.1007/s10816-015-9272-9](https://doi.org/10.1007/s10816-015-9272-9).
- Marwick, Ben, Carl Boettiger, and Lincoln Mullen. 2018. “Packaging Data Analytical Work Reproducibly Using R (and Friends).” *The American Statistician* 72 (1): 80–88. doi:[10.1080/00031305.2017.1375986](https://doi.org/10.1080/00031305.2017.1375986).
- Mons, Barend. 2018. *Data Stewardship For Open Science: Implementing FAIR Principles*. Boca Raton: CRC Press, Taylor & Francis Group.
- Pergl, Robert, Rob Hooft, Marek Suchánek, Vojtěch Knaisl, and Jan Slifka. 2019. “‘Data Stewardship Wizard’: A Tool Bringing Together Researchers, Data Stewards, and Data Experts Around Data Management Planning.” *Data Science Journal* 18 (1). Ubiquity Press: 59. doi:[10.5334/dsj-2019-059](https://doi.org/10.5334/dsj-2019-059).
- Ruppert, Evelyn. 2012. “The Governmental Topologies of Database Devices.” *Theory, Culture & Society* 29 (4-5). SAGE Publications Ltd: 116–136. doi:[10.1177/0263276412439428](https://doi.org/10.1177/0263276412439428).
- Ruppert, Evelyn, Engin Isin, and Didier Bigo. 2017. “Data Politics.” *Big Data & Society* 4 (2). SAGE Publications Ltd: 2053951717717749. doi:[10.1177/2053951717717749](https://doi.org/10.1177/2053951717717749).
- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1). Nature Publishing Group: 160018. doi:[10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).