# Archaeology Data Infrastructures

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

Petr Pajdla

2022-11-12

# Table of contents

# Preface

> 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.

> 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 then 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 trackeg 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).

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

# Introduction

Briefly describe the structure of the work!

# 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

## 1.3 Archaeology as theory- and/or data-driven science

> ℹ 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.

(based on TAG Brno 2021 talk)

## 1.4 Theorizing data

Defining archaeological data, micro- to macro-scales;

# 2 Methods

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

## 2.1 Reproducibility

# 3 Data

> 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

## 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

Kitchin, R. 2022. *The data revolution: Big data, open data, data infrastructures & their consequences*. Second. Los Angeles, California: SAGE Publications.

Marwick, B. 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: https://doi.org/10.1007/s10816-015-9272-9.

Marwick, B, Boettiger, C and Mullen, L. 2018 Packaging Data Analytical Work Reproducibly Using R (and Friends). *The American Statistician* 72(1): 80–88. DOI: https://doi.org/10.1080/00031305.2017.1375986.

Ruppert, E. 2012 The Governmental Topologies of Database Devices. *Theory, Culture & Society* 29(4-5): 116–136. DOI: https://doi.org/10.1177/0263276412439428.

Ruppert, E, Isin, E and Bigo, D. 2017 Data politics. *Big Data & Society* 4(2): 2053951717717749. DOI: https://doi.org/10.1177/2053951717717749.