Archaeology Data Infrastructures

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

Petr Pajdla

2/23/23

Table of contents

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 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 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]
```

Stats

As of February 23, 2023 there are roughly 6.3 pages of text. Length of individual chapters is as follows:

```
w     c                    f
```

```
1  267  1788  chapters/intro.qmd
2  330  2391 chapters/theory.qmd
3  227  1741 chapters/method.qmd
4  802  5424   chapters/data.qmd
5 1626 11344               total
```

Introduction

Overview

The introduction explains:

- What is the general research context of the work.
- How archaeological heritage is managed in the Czech Republic, especially in relation to data.
- What research questions are asked here and why.
- How is the thesis structured into chapters and sections.

Context

Archaeological heritage management in the Czech Republic

How archaeology, its finds, sites and data are managed in various countries is heavily influenced by given legal framework. In this section, issues specific to the case of the Czech Republic are described.

Research questions

Thesis outline

Here is a brief outline of the structure of the thesis. 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 analysing 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.

Summary

1 Theory

Overview

Chapter 1 presents:

- Definitions for fundamental concepts I am building up on further in the text.
- An overview of theoretical approaches the work is determined and shaped by.
- A discussion of archaeological research as theory- and/or data-driven.
- A commentary on data from the theoretical points presented earlier.

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

> **Note**
>
> This section is partly based on the *Data-driven Archaeology. Are we there yet?* talk co-authored 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 organized 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;

Chapter summary

2 Methods

Overview

Chapter 2 describes:

-
- The software that is used in the analysis.

2.1 Software

Most of the things included here, if not all of them, were achieved using open-source software. Large part of this endeavor 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 (**rcore?**). Various packages were used, the most important packages are listed here, the complete list is in an Appendix

2.1.1 Reproduciblity

Chapter summary

3 Data and materials

> **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.

Overview

Chapter 3 details:

- How data is managed in this research. This is described in a data management plan.
- What data sources are available in the Czech Republic.
- What data models are most commonly used in Czech archaeology.
- How is the reality represented in the databases of the *Archaeological information system of the Czech Republic*.

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 captured, created, curated, selected, analysed, interpreted, shared, and archived in 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 ideas on good DMP practice. It is included both as part of the text and as a standalone machine actionable file (link?).

3.1.1 Created, collected and re-used data

3.1.1.1 Data re-use

The work is predominantly based on re-using existing data. The sources of data are listed in Section 3.2. I presume that there are many pre-existing data sets in the Czech archaeology, but most of them are either inaccessible or not findable, i.e. we cannot be sure they even exist. The single most complete source for archaeology data in the Czech Republic is without a doubt the *AIS CR* infrastructure.

There are several well published data sets covering the area of the Czech Republic in the Journal of Open Archaeology Data, the radiocarbon data by Tkáč and Kolář (2021) and the Neolithic settlements data set by Pajdla and Trampota (2021). These have an advantage of well formulated access and re-use policies, explicit licence and other conditions for use.

3.1.1.2 Data creation and collection

3.1.1.3 Vocabularies

I am explicitly using vocabularies that are inherent to data sources from which the data is reused. A principal and authoritative vocabulary for archaeology and related fields is the *Getty Art & Architecture Thesaurus* (AAT). Getty *AAT* subjects are used by the *ARIADNE* infrastructure and many other archaeology infrastructures are mapping their vocabularies to the *AAT* subjects. The emerging *ARIADNE AO_Cat* formal ontology is also taken into account when interacting with *ARIADNE* services. The *AIS CR* vocabularies, although implicit to the data, are yet to be published. A possibly incomplete version can be reverse engineered from the available data sets. If reconciliation between data from different data sources is necessary, the *AAT* is used to map between them.

3.1.1.4 File naming conventions

Persistent identifiers?

3.1.2 Data processing

How will you work with the data? Do you have/need storage? Do you do backups? Are you using object-store? Are you using relational database? Graph database? Triple store? How are changes in data managed?

3.1.3 Interpretation

Specify/list data formats you will be using and their structure. Will different data be integrated?

3.1.4 Data preservation

What data sets are you producing? Is data long-term archived? Will it be usable and accessible after a long period of time?

3.1.5 Access to data

Will the data be as open as possible? What are the reasons your data cannot became open?

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

Chapter summary

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.

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.

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.

Mons, Barend. 2018. *Data Stewardship For Open Science: Implementing FAIR Principles*. Boca Raton: CRC Press, Taylor & Francis Group.

Pajdla, Petr, and František Trampota. 2021. "Neolithic Settlements in Central Europe: Data from the Project 'Lifestyle as an Unintentional Identity in the Neolithic'." *Journal of Open Archaeology Data* 9 (0). Ubiquity Press: 13. doi:10.5334/joad.88.

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.

Ruppert, Evelyn. 2012. "The Governmental Topologies of Database Devices." *Theory, Culture & Society* 29 (4-5). SAGE Publications Ltd: 116–136. doi: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.

Tkáč, Peter, and Jan Kolář. 2021. "Towards New Demography Proxies and Regional Chronologies: Radiocarbon Dates from Archaeological Contexts Located in the Czech Republic Covering the Period Between 10,000 BC and AD 1250." *Journal of Open Archaeology Data* 9 (0). Ubiquity Press: 9. doi:10.5334/joad.85.

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.

A Glossary

Data

Data infrastructure

Data management plan (DMP)

(also data stewardship plan)

Database

Data set

FAIR data principles

Legacy data

Roles

A.0.0.1 Data curator

A.0.0.2 Data steward