Archaeology Data Infrastructures

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

Petr Pajdla

11/12/22

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.

ℹ **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 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).

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

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. Last but not least, 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. 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

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

> **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.

(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 Digital humanities

2.2 Digital archaeology

2.3 Spatial archaeology

2.4 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, vecor 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. Analysis was mostly performed in an R language for statistical computing and graphics (**rcore?**).

2.4.1 Reproduciblity

3 Data

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.

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.

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.

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 Software

A.1 R Session Information

```r
sessionInfo()
```

```
R version 4.2.2 (2022-10-31)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: ArchLabs Linux

Matrix products: default
BLAS/LAPACK: /usr/lib/libopenblas_haswellp-r0.3.21.so

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

loaded via a namespace (and not attached):
 [1] compiler_4.2.2  magrittr_2.0.3  fastmap_1.1.0   cli_3.4.1
 [5] tools_4.2.2     htmltools_0.5.3 stringi_1.7.8   rmarkdown_2.17
 [9] knitr_1.40      stringr_1.4.1   xfun_0.34       digest_0.6.30
[13] jsonlite_1.8.3  rlang_1.0.6     evaluate_0.17
```

A.2 R Packages

```
installed.packages()[, c(1, 3)] |> knitr::kable()
```

	Package	Version
abind	abind	1.4-5
askpass	askpass	1.1
assertthat	assertthat	0.2.1
backports	backports	1.4.1
base64enc	base64enc	0.1-3
BH	BH	1.78.0-0
bit	bit	4.0.4
bit64	bit64	4.0.5
blob	blob	1.2.3
bookdown	bookdown	0.29
brew	brew	1.0-8
brio	brio	1.1.3
broom	broom	1.0.1
bslib	bslib	0.4.0
cachem	cachem	1.0.6
callr	callr	3.7.2
cellranger	cellranger	1.1.0
classInt	classInt	0.4-8
cli	cli	3.4.1
clipr	clipr	0.8.0
collections	collections	0.3.6
colorspace	colorspace	2.0-3
commonmark	commonmark	1.8.1
cowplot	cowplot	1.1.1
cpp11	cpp11	0.4.3
crayon	crayon	1.5.2
credentials	credentials	1.3.2
crosstalk	crosstalk	1.2.0
crsmeta	crsmeta	0.3.0
curl	curl	4.3.3
cyclocomp	cyclocomp	1.1.0
data.table	data.table	1.14.4
DBI	DBI	1.1.3
dbplyr	dbplyr	2.2.1
deldir	deldir	1.0-6
desc	desc	1.4.2

	Package	Version
DescTools	DescTools	0.99.47
devtools	devtools	2.4.5
diffobj	diffobj	0.3.5
digest	digest	0.6.30
downlit	downlit	0.4.2
dplyr	dplyr	1.0.10
dtplyr	dtplyr	1.2.2
e1071	e1071	1.7-12
ellipsis	ellipsis	0.3.2
evaluate	evaluate	0.17
Exact	Exact	3.2
expm	expm	0.999-6
fansi	fansi	1.0.3
farver	farver	2.1.1
fastmap	fastmap	1.1.0
fontawesome	fontawesome	0.4.0
forcats	forcats	0.5.2
fs	fs	1.5.2
gargle	gargle	1.2.1
generics	generics	0.1.3
gert	gert	1.9.1
ggforce	ggforce	0.4.1
ggplot2	ggplot2	3.3.6
ggrepel	ggrepel	0.9.1
ggspatial	ggspatial	1.1.6
gh	gh	1.3.1
gitcreds	gitcreds	0.1.2
gld	gld	2.6.6
glue	glue	1.6.2
goftest	goftest	1.2-3
googledrive	googledrive	2.0.0
googlesheets4	googlesheets4	1.0.1
gridExtra	gridExtra	2.3
gtable	gtable	0.3.1
gtools	gtools	3.9.3
haven	haven	2.5.1
here	here	1.0.1
highr	highr	0.9
hms	hms	1.1.2
htmltools	htmltools	0.5.3
htmlwidgets	htmlwidgets	1.5.4

	Package	Version
httpuv	httpuv	1.6.6
httr	httr	1.4.4
httr2	httr2	0.2.2
ids	ids	1.0.1
igraph	igraph	1.3.5
infer	infer	1.0.3
ini	ini	0.3.1
isoband	isoband	0.2.6
janitor	janitor	2.1.0
jpeg	jpeg	0.1-9
jquerylib	jquerylib	0.1.4
jsonlite	jsonlite	1.8.3
knitr	knitr	1.40
labeling	labeling	0.4.2
languageserver	languageserver	0.3.14
later	later	1.3.0
lazyeval	lazyeval	0.2.2
leaflet	leaflet	2.1.1
leaflet.providers	leaflet.providers	1.9.0
lemon	lemon	0.4.5
lifecycle	lifecycle	1.0.3
lintr	lintr	3.0.2
lmom	lmom	2.9
lubridate	lubridate	1.8.0
magrittr	magrittr	2.0.3
markdown	markdown	1.3
memoise	memoise	2.0.1
mime	mime	0.12
miniUI	miniUI	0.1.1.1
modelr	modelr	0.1.9
munsell	munsell	0.5.0
mvtnorm	mvtnorm	1.1-3
nabor	nabor	0.5.0
nngeo	nngeo	0.4.6
openssl	openssl	2.0.4
osmdata	osmdata	0.1.10
packrat	packrat	0.8.1
parzer	parzer	0.4.1
patchwork	patchwork	1.1.2
pdist	pdist	1.2.1
pillar	pillar	1.8.1

	Package	Version
pkgbuild	pkgbuild	1.3.1
pkgconfig	pkgconfig	2.0.3
pkgdown	pkgdown	2.0.6
pkgload	pkgload	1.3.1
plyr	plyr	1.8.7
png	png	0.1-7
polyclip	polyclip	1.10-4
praise	praise	1.0.0
prettymapr	prettymapr	0.2.4
prettyunits	prettyunits	1.1.1
processx	processx	3.8.0
profvis	profvis	0.3.7
progress	progress	1.2.2
PROJ	PROJ	0.4.0
proj4	proj4	1.0-11
promises	promises	1.2.0.1
proxy	proxy	0.4-27
ps	ps	1.7.2
purrr	purrr	0.3.5
quarto	quarto	1.2
R.cache	R.cache	0.16.0
R.methodsS3	R.methodsS3	1.8.2
R.oo	R.oo	1.25.0
R.utils	R.utils	2.12.1
R6	R6	2.5.1
ragg	ragg	1.2.4
randomizr	randomizr	0.22.0
rappdirs	rappdirs	0.3.3
raster	raster	3.6-3
rcmdcheck	rcmdcheck	1.4.0
RColorBrewer	RColorBrewer	1.1-3
Rcpp	Rcpp	1.0.9
RcppEigen	RcppEigen	0.3.3.9.2
RCzechia	RCzechia	1.9.4
readr	readr	2.1.3
readxl	readxl	1.4.1
rematch	rematch	1.0.1
rematch2	rematch2	2.1.2
remotes	remotes	2.4.2
reprex	reprex	2.0.2
reproj	reproj	0.4.3

	Package	Version
reshape2	reshape2	1.4.4
rex	rex	1.2.1
rgdal	rgdal	1.5-32
rjson	rjson	0.2.21
rlang	rlang	1.0.6
rmarkdown	rmarkdown	2.17
rootSolve	rootSolve	1.8.2.3
rosm	rosm	0.2.6
roxygen2	roxygen2	7.2.1
rprojroot	rprojroot	2.0.3
rsconnect	rsconnect	0.8.28
rstudioapi	rstudioapi	0.14
Rttf2pt1	Rttf2pt1	1.3.11
rversions	rversions	2.1.2
rvest	rvest	1.0.3
s2	s2	1.1.0
sass	sass	0.4.2
scales	scales	1.2.1
selectr	selectr	0.4-2
sessioninfo	sessioninfo	1.2.2
sf	sf	1.0-8
shiny	shiny	1.7.3
showtext	showtext	0.9-5
showtextdb	showtextdb	3.0
snakecase	snakecase	0.11.0
sourcetools	sourcetools	0.1.7
sp	sp	1.5-0
spatstat	spatstat	2.3-4
spatstat.core	spatstat.core	2.4-4
spatstat.data	spatstat.data	3.0-0
spatstat.geom	spatstat.geom	3.0-3
spatstat.linnet	spatstat.linnet	2.3-2
spatstat.random	spatstat.random	2.2-0
spatstat.sparse	spatstat.sparse	3.0-0
spatstat.utils	spatstat.utils	3.0-1
spData	spData	2.2.0
spdep	spdep	1.2-7
stringi	stringi	1.7.8
stringr	stringr	1.4.1
styler	styler	1.8.0
svglite	svglite	2.1.0

	Package	Version
sys	sys	3.4.1
sysfonts	sysfonts	0.8.8
systemfonts	systemfonts	1.0.4
tensor	tensor	1.5
terra	terra	1.6-17
testthat	testthat	3.1.5
textshaping	textshaping	0.3.6
tibble	tibble	3.1.8
tidyr	tidyr	1.2.1
tidyselect	tidyselect	1.2.0
tidyverse	tidyverse	1.3.2
tinytex	tinytex	0.42
tweenr	tweenr	2.0.2
tzdb	tzdb	0.3.0
units	units	0.8-0
urlchecker	urlchecker	1.0.1
usethis	usethis	2.1.6
utf8	utf8	1.2.2
uuid	uuid	1.1-0
vctrs	vctrs	0.5.0
viridis	viridis	0.6.2
viridisLite	viridisLite	0.4.1
visNetwork	visNetwork	2.1.2
vroom	vroom	1.6.0
waldo	waldo	0.4.0
whisker	whisker	0.4
withr	withr	2.5.0
wk	wk	0.7.0
xfun	xfun	0.34
xml2	xml2	1.3.3
xmlparsedata	xmlparsedata	1.0.5
xopen	xopen	1.0.0
xtable	xtable	1.8-4
yaml	yaml	2.3.6
zip	zip	2.2.2
base	base	4.2.2
boot	boot	1.3-28
class	class	7.3-20
cluster	cluster	2.1.4
codetools	codetools	0.2-18
compiler	compiler	4.2.2

	Package	Version
datasets	datasets	4.2.2
foreign	foreign	0.8-83
graphics	graphics	4.2.2
grDevices	grDevices	4.2.2
grid	grid	4.2.2
KernSmooth	KernSmooth	2.23-20
lattice	lattice	0.20-45
MASS	MASS	7.3-58.1
Matrix	Matrix	1.5-1
methods	methods	4.2.2
mgcv	mgcv	1.8-41
nlme	nlme	3.1-160
nnet	nnet	7.3-18
parallel	parallel	4.2.2
rpart	rpart	4.1.19
spatial	spatial	7.3-15
splines	splines	4.2.2
stats	stats	4.2.2
stats4	stats4	4.2.2
survival	survival	3.4-0
tcltk	tcltk	4.2.2
tools	tools	4.2.2
utils	utils	4.2.2

B Glossary

Data

Data infrastructure

Data management plan (DMP)

(also data stewardship plan)

Database

Data set

FAIR data principles

Legacy data

Roles

B.0.0.1 Data curator

B.0.0.2 Data steward