# Archaeology Data Infrastructures

**Data reuse potentials and limitations to modelling settlement systems**

Petr Pajdla

2023-12-16

# Contents

# List of Figures

# List of Tables

# 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.

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 without numbering: -@sec-label, [Chapter -@sec-label]
```

### Stats

As of September 8, 2023 there are roughly 19.3 pages (1800 characters per page) of text. Length of individual chapters is as follows:

- Introduction: 1 pages
- Theory and Method: 15 pages
- Data and materials: 3 pages

```
     w     c ns                  f
1  284  1905  1  chapters/intro.qmd
2 3962 27369 15 chapters/theory.qmd
3  805  5473  3   chapters/data.qmd
4 5051 34747 19              total
```

# Introduction

**Chapter overview:**

- What is the general research context of the work.
- How archaeological heritage is managed in the Czech Republic, especially in relation to data findability, accessibility, interoperability and reusability.
- 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 shortly described to give a basic context for the study.

## 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.

Chapter 2, *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 2.2) details how data is handled in this research.

## Summary

# 1 Theory and method

**Chapter 1 introduces:**

- Definitions of 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 here.

## 1.1 Definitions and terminology

This section presents a discussion of how I (and others) understand **data** and **data infrastructures**. Data are a corner stone of this work and looking closely at how scholars, policy makers, and various other stakeholders define data is crucial for further understanding of what is possible and what is not in the process of knowledge building.

### 1.1.1 Data

In the current *information age*, word *data*[1] is almost omnipresent and it became such a generic term that almost everyone has some notion or idea of what data are. Most of the formal definitions understand data as building blocks of *information* (e.g. Kitchin 2022: 4). These pieces of information are percieved as pre-factual and pre-analytical. That is, in contrast to facts, false data are data nonetheless, disproven facts are no longer facts (Rosenberg 2013: 17–18). What more, data need to be analysed and interpreted to make a meaning. Data are furthermore understood as incontrovertible and non-deconstructible units. These assumptions migh lead to an impression that data exist in the outside world as entities or phenomena independent of the one observing them. And this is indeed one way to comprehend data, as '*things given*' that just need to be discoverd (cf. Huggett 2020). This inductive explanation of data brings in itself a danger that the process of discovering the data is viewed as an atheoretical endeavour.

[1] In this text, I adhere to the current scholarly convention as noted by Kitchin (2022, xvii), i.e. using term *data* in the plural form, with a singular form of *datum*.

On the contrary, if data are understood as *'things made'* they are created at the moment when they are captured by observation, measurement, or derived from other data. This implies theory-laden creative process of data generation, recording, etc. In this case, it is clear that the one creating or recording the data does that based on the experience, objective and knowledge he/she has. The whole process of data recording, creation, or capture is thus discursively framed and technically mediated (Huggett 2020; cf. Kitchin 2022: 4–15). As Kitchin notes, what we nowadays label as data are in fact *capta*, i.e. *'things taken'*.

Ruppert, Isin and Bigo (2017) voice the idea that the practice of data production does not happen through unstructured social (and/or political) practices, but through structured and structuring processes that add to configurations of power and knowledge. Hence, I do not understand *data* as ubiquitous phenomena that are waiting out there to be discovered but as *capta* that are actively and creatively made, recorded, generated etc. by a theory- and agenda-laden researchers. What more, the practice of data production is shaped by and contributing to structure of power and knowledge, because what data are recorded and how they are structured has consequences as to what can be later devised, i.e. what knowlege can be generated.

If the chosen strategies in data recording and generation influence the knowledge generated further along the way, an idea of good and bad data comes to mind. Bad data, like false data, do not exist. Data can be useless, but the value of data is determined by the current goal in mind, data deemed useless by one project can become useful on other occassion, perhaps previously unforeseen and unintended. Nevertheless, there are some signs of good-quality data as cited[2] by Kitchin (2022: 4):

- they are discreet and intelligible, i.e. each datum is individual, separate and separable, and clearly defined;
- they are aggregative, that is they can be built into sets;
- they have associated metadata (data about data);
- they can be linked to other datasets to provide insights not available from a single dataset.

### 1.1.2 Infrastructures

Although *infrastructures* became quite a buzzword in recent years both among the policy makers and researchers, it is rather difficult to define what an infrastructure actually is. Many slightly different variations of more-less the same name and concept are circulating in official documents, various reports, research articles etc. Thus, we encounter terms such as **research infrastructures**, large research infrastructures, open science infrastrucutres, **data infrastructures**, and perhaps more, even though most of their definitions are variations of the very same concept.

[2] Author's note: Kitchin cites Rosenberg (2013), but I was not able to locate this part in Rosenberg's paper, which is primarily focused on the history of the term *data* in English language, not the quality of data.

Hallonsten (2020) maps the field of European (research) infrastructures and identifies the principal problem as the difficulty to come up with a single definition that would fit all of those who are considered to be (or consider themselves to be) an *infrastructure*. Hallonsten (2020: 630) concludes that:

> *(...) "while a politically viable definition seems to be either already in place (...) or unneeded, an analytically workable definition is out of reach unless the scope is limited and the aim of the definition is made more precise."*

In the next paragraphs, we look at several of the *political* definitions of research infrastructures and later we focus on an *analytical* definition of data infrastructures.

**Political definitions**

In the European Union, research infrastructures are currently defined in the Article 2(1) of EU Regulation No 2021/695 establishing Horizon Europe (2021) as facilities providing resources and services for the research communities in their respective fields. These include:

- human resources, major scientific equipment or sets of instruments;
- collections, archives or scientific data, i.e. knowledge-related facilities;
- computing systems and communication networks;
- any other research and innovation infrastructure of a unique nature open to external users.

The *Regulation* also states that infrastructures may be used beyond research, i.e. for education, public services etc. This broad definition given by the European Union covers almost any kind of an infrastructure. In the legal framework of the Czech Republic, given by Act No 130/2002 Coll. on the Support of Research, Experimental Development and Innovation (2002), Article 2(2), *large research infrastructure* is defined as follow (english translation from Roadmap of Large Research Infrastructures 2019):

> *"(...) a facility necessary for conducting comprehensive research and development with high financial and technology demands, approved by the Government and established to be also used by other research organisations."*

This political definition is rather an opportunistic one in demanding that the infrastructure is approved by the Czech government and has high financial and technology demands. Lastly, the UNESCO Open Science Recommendation (2021) adds to the research infrastructures a strong element of open science, addressing them as *open science infrastructures*.

To conclude, the political definitions of research infrastructures are mostly broad enough to fit any kind of a facility, that is deemed appropriate. This point is highlighted especially in the Czech definition, where an approval of the Government is required. In general, there is an emphasis on provision of services (etc.) to various stakeholder communities and cooperation. In the EU Regulation, a subset of a larger field of infrastructures is labeled *knowledge-related facilities*, it is exactly this part that is discussed here as *data infrastructures*.

**Data infrastructures**

Kitchin (2022: 47–48) builds up the definition of data infrastructures by comparing them to *data holdings* and *data archives*. *Data holdings* are any data stored informally, presumably by an individual (scientist), without long-term preservation or sharing for reuse in mind. Such data are inevitably lost when the researcher retires, dies (etc.), because proper metadata descriptions are missing and the data, although they might be organised in some way, lack documentation and it is difficult, if not impossible, to reconstruct the context of the data. In contrast, *data archives* are formal collections that are structured, curated and documented by appropriate metadata with plans for preservation, access and discoverability.

The role of an interconnected digital world is then highlighted in Kitchin's definition of a data infrastructure (Kitchin 2022: 50):

> *"A data infrastructure is a digital means for storing, sharing, connecting and consuming data holdings and archives across the internet."*

Data infrastructures are thus information systems, repositories, archives, databases etc. shared by multiple shareholder groups that are essential in supporting open science and research.

## 1.2 Overview of theoretical concepts

This section aims at briefly introducing theoretical approaches this work is shaped by. At first, digital humanities as an umbrella concept connecting the digital and/or computing world with the humanitites are introduced followed by discussion of digital, quantitative and computational archaeology. And lastly, spatial archaeology as a main concept framing this study is reviewed.

### 1.2.1 Digital humanities

### 1.2.2 Digital, computational and quantitative archaeology

### 1.2.3 Spatial archaeology

## 1.3 Archaeology as theory- and data-driven

> **Note**
>
> This section is partly based on the *Data-driven Archaeology. Are we there yet?* talk co-authored with Hana Kubelková and Petr Květina presented at the *Central European Theoretical Archaeology Group (CE TAG)* meeting focused on *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.

In his vision for the future of archaeology, Kristiansen (2014: 14) sees the opposition between theory and data disappear. Here we examine the dichotomy between data- and theory-driven approaches. The idea that scientific method and/or theory is dead was addressed by many, but I will start this discussion with a trending comment by Anderson (2008).

Anderson uses Boxes famous aphorism *"(...) all models are wrong (...)"* (Box 1976: 792), later extended to *"All models are wrong but some [models] are useful."* (Box 1979: 202) and extends it to *"All models are wrong, and increasingly you can succeed without them."* (Anderson 2008). By this statement, the author means that science as we know it, i.e. building a hypothesis

a (research) question can be tackled by accumulating and analysing large quantities of data (or *big data*[3]) because

## 1.4 Theorizing archaeological data

Defining archaeological data, micro- to macro-scales;

Earlier in this chapter it was explained that here, data is understood as *capta*, i.e. they are actively and creatively made, recorded, generated etc.

## 1.5 Assessing data infrastructures

In this section I define a framework for an assessment of the quality of data infrastructures. As there are some signs of good-quality data (as cited from Kitchin (2022: 4) in the Section 1.1.1), i.e. they are discreet and intelligible, aggregative, have associated metadata and can be linked to other datasets, by extension, good-quality data infrastructures allow data to be discreet and separable from other data, enable data aggregation, contain metadata and allow linking to other data sources. This gives us a general idea about requirements for a

[3] *Big data* is a phrase often heard in archaeology to address large amounts of data. In fact, it is a technical term for a data defined by their large volume, velocity, variety, exhaustivity, resolution, etc. (Kitchin 2022: 61–74). For instance some geophysical, 3D, remote sensing, or archaeogenetic and similar data can fit under the definition of big data, but most of archaeological data, like databases of sites and assamblages etc. are not big data by definition, even though they might cover large spatial and/or temporal regions and consist of many records.

data infrastrucutre, but is difficult to assess in practice. These general ideas are concretized and further developed by the FAIR data principles (Wilkinson et al. 2016), which are aimed at enhancing the reusability of data holdings with an emphasis on the ability of machines to find and use the data. The FAIR data principles are measurable what gives us an opportunity to assess how *FAIR* an infrastructure is.

FAIR data principles, i.e. findability, accessibility, interoperability and reusability, as originally defined by Wilkinson et al. (2016: 4), are as follows. To be *findable*, data and metadata have globally unique and persistent identifiers; are described with rich metadata which include the identifier of the data they describe; and are registered or indexed in a searchable resource. To be *accessible*, (meta)data are retrievable by the identifier using a standardized (open, free and universally implementable) communications protocol that allows for an authentication and authorization procedure; and metadata are accessible, even when the data are no longer available. To be *interoperable*, (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation; contain vocabularies that follow FAIR principles; and include qualified references to other (meta)data. To be *reusable*, meta(data) are richly described with a plurality of accurate and relevant attributes; are released with a clear and accessible data usage license; are associated with detailed provenance; and meet domain-relevant community standards.

With archaeology audiences in mind, the principles are further explained together with tips for implementations etc. by Hollander et al. (2019). Here I build up on Wilkinson et al. (2016) and Hollander et al. (2019) to come up with a set of formal, i.e. measurable and/or determinable criteria for assessment of data infrastructures. This framework is then used to evaluate archaeology data infrastructures in the Czech Republic in Chapter 2.

### 1.5.1 Assessment framework

Assessment criteria are grouped together according to the FAIR principle they relate to. Table 1.1 lists criteria for findability of resources. Data should be easy to find by both humans and machines and well documented by metadata in order to be reusable by other researchers. To be able to use the data object in any way, it must be possible to uniquely identify it, find it, and refer to it (F1). This implies that an identifier of some sort, preferably persistent, i.e. immutable and long-lasting, is assigned to the resource (data and/or metadata). Persistent identifiers typically take form of DOIs, Handles, PURLs and URNs to name just a few examples[4].

Furthermore, it is convenient to be able to locate the resource (preferably on the internet) if the identifier, and possibly prefix of some sort, is known (F2). Since Marwick and Birch (2018)

[4] Handles and DOIs (Digital Object Identifiers) are composed of a `prefix/suffix` and typically resolved at https://doi.org/. URNs (Uniform Resource Names), in form `urn:namespace:name`, are mostly used in the Semantic Web and are not resolvable, i.e. URNs do not have information about the location of the object. PURLs (Persistent Uniform Resource Locators) are an extension of URLs and are resolvable. URNs and (P)URLs are both subsets of URIs (Uniform Resource Identifiers). See e.g. DuCharme (2013: 21–23) for details.

explore the lack of data citation and reuse in archaeology and suggest a standard for data citation, one of the criteria is whether the infrastructure makes it easy to cite the resources it publishes (F3). Another feature that enables findability of data is a rich metadata description (F4–F5).

Table 1.1: Framework for quality assessment of data infrastrastructres, Findability

| ID | Findability | Value |
|---|---|---|
| F1.1 | Are there unique identifiers? | True/False |
| F1.2 | Are the identifiers persistent? | True/False |
| F1.3 | Are the identifiers in any standard form? | True/False |
| F2 | Is it possible to locate the resource by the identifier? | True/False |
| F3 | Is it possible (and made easy) to cite: | - |
| F3.1 | - the data infrastructure, | True/False |
| F3.2 | - its parts and/or | True/False |
| F3.3 | - individual resources? | True/False |
| F4.1 | Is the metadata scheme described, i.e. explicit? | True/False |
| F4.2 | Does the metadata scheme follow a standard? | True/False |
| F5 | Are the metadata searchable? | True/False |

Criteria for accessibility are listed in Table 1.2. Accessible data are retrievable under well-defined conditions using standardised protocols. Certification of a data infrastructure guarantees that its repository is trustworthy, the data are stored safely and will be available over a long period of time. Certifications may include CoreTrustSeal, nestor seal etc. (A1). By a standardised exchange protocol (A2.1) a well-documented technology created and maintained by a recognized authority (e.g. World Wide Web Consortium, W3C) is meant. For example SPARQL is a query language for semantic data created by W3C, OAI-PMH, a Protocol for Metadata Harvesting, is created and maintained by the Open Archives Initiative etc. By a standardised format (A2.2) a machine-readable format is meant, for instance XML, a W3C format for hierarchical data representation, RDF, a W3C standard for semantic data, etc.

For (meta)data to be easily accessible, the policies for the access need to be clearly stated (A3), i.e. definitions of who can access what and when needs to be explicitly communicated, for example existence of differentiated user roles and/or embargo periods. This gives both the users accessing the data clear instructions on how to access the objects they need, and the users depositing the data sets options to protect sensitive data etc. Existence of policies how to handle situations if a data object is no longer available (e.g. deleted, superseded etc.) and presence of metadata tombstones is a good practice how to communicate that a data object existed, but does not anymore (A4).

Table 1.2: Framework for quality assessment of data infrastrastructres, Accessibility

| ID | Accessibility | Value |
|---|---|---|
| A1 | Is the repository trustworthy? | True/False |
| A2.1 | Are the (meta)data retrievable using a standardised protocol? | True/False |
| A2.2 | Are the metadata in a standardised format? | True/False |
| A3 | Is the access policy clearly stated? | True/False |
| A3.1 | Are there embargo periods? | True/False |
| A3.2 | Are the access rights differentiated? | True/False |
| A4 | Is the metadata available even after the data is not? | True/False |

Interoperability is the ability of the (meta)data to be easily combined with other data sets, Table 1.3 lists the interoperability criteria relevant to data infrastructures. Machine interoperability is closely related to the availability of APIs and their quality and human interoperability derives from the existence and extensiveness of documentation.

To enable interoperability, (meta)data model[5] needs to be described clearly and accessibly (I1) and employed controlled vocabularies need to be explained and published, preferably following the FAIR principles (I2). Explanation of the given data model and vocabularies describing exact meanings embedded in the data are a prerequisites for building understanding by other people. Furthermore, well-documented (meta)data models allow creation of mappings between different metadata schemes and data infrastructures. Similarly, the existence of machine actionable APIs (application programming interfaces, I4) that allow harvesting of (meta)data through standardised protocols and return responses in standardised formats (cf. A2) ensure machine interoperability.

Table 1.3: Framework for quality assessment of data infrastrastructres, Interoperability

| ID | Interoperability | Value |
|---|---|---|
| I1 | Is the (meta)data model explained and documented? | True/False |
| I2.1 | Are the vocabularies published and/or well-known? | True/False |
| I2.2 | Are the vocabularies FAIR? | True/False |
| I3 | Are other metadata referenced properly? | True/False |
| I4.1 | Is there a machine-actionable API? | True/False |
| I4.2 | Is the API well documented? | True/False |

By reusability the process of making data ready for future processing and analysis is meant. This is crucial for reproducibility of scientific research. Data, repositories and infrastructures that are systematically documented by manuals, tutorials, guides, codebooks etc. and

[5] The term *data model* is used here in the sense of how phenomena present, observed and/or measured in the real world are encoded in the data, what ratinale is behind the chosen abstraction process, and what is actually meant by the given wording.

transparent about what they do and do not contain foster reuse, because researchers reusing the data have clear notion of what to expect from the data source (R1). Reusability is also enhanced by using widely used and open source file formats (R2). In the long run, long-term preservation (LTP) is a prerequisite for reusability, because if the file format in which the data is saved gets obsolete, it is often difficult to retrieve the original data, see Brin et al. (2013) for recommended file formats, online as *Guides to Good Practice* (n.d.).

Integrity of the (meta)data and existence of multiple versions of the given data objects is also important to consider, because if this information is not properly communicated, different versions of the data objects with identical identifiers might get mixed up (R3). This closely relates to the provenance of the data, i.e. the documentation of the origin of the data object and record of any changes with a rationale behind these processes. Knowing why changes in the (meta)data happened, whether it was a correction of a previous mistake or something else, might be useful for data reuse in the future. Lastly, releasing the (meta)data with proper license information, preferably under a standard data license, for instance a Creative Commons Licence, and any information on a rights holder is neccessary for future reuse because without this information, it is unclear what the terms of (meta)data use are.

Table 1.4: Framework for quality assessment of data infrastrastructres, Reusability

| ID | Reusability | Value |
|---|---|---|
| R1 | Are there documentation, manuals, tutorials etc? | True/False |
| R2.1 | Are common file formats used? | True/False |
| R2.2 | Are file formats suitable for long-term preservation? | True/False |
| R3.1 | Is the (meta)data provenance documented? | True/False |
| R3.2 | Are there any version control mechanisms in place? | True/False |
| R4.1 | Are the rights holders and terms of use clear? | True/False |
| R4.2 | Are the resources released under a standard license? | True/False |

The framework consists predominantly of qualities that are measurable and builds up on the FAIR data principles. CARE data principles, as defined by Carroll et al. (2020), were considered as well, but their goal is to increase the indigenous data sovereignity and self-determination by being people and purpose-oriented, while the FAIR data principles are primarily focused on the characteristics of the data. CARE data principles are put together to address imbalances of power in the knowledge societies and economies and protect indigenous and human rights. Hence the extent to which a data infrastructure adheres to CARE data principles is difficult to determine and/or measure.

The framework for assessment of the quality of data infrastructures is used in Chapter 2 to evaluate the quality of archaeology data infrastructures in the Czech Republic.

## 1.6 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 (R Core Team 2023). Various packages were used, the most important packages are listed here, the complete list is in an Appendix

### 1.6.1 Reproduciblity

## Chapter summary

# 2 Data and materials

**Chapter 2 details:**

- How are data 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*.

**Note**

This chapter, especially the Section 2.2: 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.

## 2.1 Sources of archaeology data in the Czech Republic

### 2.1.1 Archaeology information system of the Czech Republic

### 2.1.2 Legacy data sources

What is a legacy data source?

### 2.1.3 Museum databases

## 2.2 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 (Hollander et al. 2019; Wilkinson et al. 2016).

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?).

### 2.2.1 Created, collected and re-used data

#### Data re-use

The work is predominantly based on re-using existing data. The sources of data are listed in Section 2.1. 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.

#### Data creation and collection

#### 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.

**File naming conventions**

Persistent identifiers?

### 2.2.2 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?

### 2.2.3 Interpretation

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

### 2.2.4 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?

### 2.2.5 Access to data

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

## Chapter summary

# References


Anderson, C. 2008 The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*.

Anon. 2002 *Act No 130/2002 Coll. (The Act on the Support of Research, Experimental Development and Innovation), as amended*.

Anon. 2019. *Roadmap of Large Research Infrastructures of the Czech Republic for the years 2016-2022*. Update 2019. Prague: Ministry of Education, Youth and Sports.

Anon. 2021 Regulation (EU) 2021/695 of the European Parliament and of the Council. *32021R0695*.

Archaeology Data Service and Digital Antiquity. n.d. *Guides to Good Practice*. Available at https://archaeologydataservice.ac.uk/help-guidance/guides-to-good-practice/ [Last accessed 18 August 2023].

Box, GEP. 1976 Science and Statistics. *Journal of the American Statistical Association* 71(356): 791–799. DOI: https://doi.org/10.1080/01621459.1976.10480949.

Box, GEP. 1979 Robustness in the Strategy of Scientific Model Building. In: Launer, RL and Wilkinson, GN (eds.). *Robustness in Statistics*. Academic Press. pp. 201–236. DOI: https://doi.org/10.1016/B978-0-12-438150-6.50018-2.

Brin, A, McManamon, FP, Niven, K, Archaeology Data Service and Digital Antiquity (eds.). 2013. *Caring for digital data in archaeology: A guide to good practice*. Archaeology Data Service and Digital Antiquity. Oxford ; Oakville: Oxbow Books.

Carroll, SR, Garba, I, Figueroa-Rodríguez, OL, Holbrook, J, Lovett, R, Materechera, S, Parsons, M, Raseroka, K, Rodriguez-Lonebear, D, Rowe, R, Sara, R, Walker, JD, Anderson, J and Hudson, M. 2020 The CARE Principles for Indigenous Data Governance. *Data Science Journal* 19: 43. DOI: https://doi.org/10.5334/dsj-2020-043.

Doorn, P and Ronzino, P. 2022. *ARIADNEplus Data Management Plan Tools*. 25 March 2022. Available at https://vast-lab.org/dmp/ [Last accessed 17 November 2022].

DuCharme, B. 2013. *Learning SPARQL: Querying and updating with SPARQL 1.1*. Second edition. Sebastopol, CA: O'Reilly Media.

Hallonsten, O. 2020 Research Infrastructures in Europe: The Hype and the Field. *European Review* 28(4): 617–635. DOI: https://doi.org/10.1017/S1062798720000095.

Hollander, H, Morselli, F, Uiterwaal, F, Admiraal, F, Trippel, T and Giorgio, SD. 2019 *PARTHENOS Guidelines to FAIRify data management and make data reusable* DOI: https://doi.org/10.5281/zenodo.2668478.

Huggett, J. 2020 Is Big Digital Data Different? Towards a New Archaeological Paradigm. *Journal of Field Archaeology* 45: S8–S17. DOI: https://doi.org/10.1080/00934690.2020.1713281.

Kitchin, R. 2022. *The data revolution: A critical analysis of big data, open data & data infrastructures*. 2nd ed. Los Angeles, California: SAGE Publications.

Kristiansen, K. 2014 Towards a New Paradigm? The Third Science Revolution and its Possible Consequences in Archaeology. *Current Swedish Archaeology* 22(1): 11–34. DOI: https://doi.org/10.37718/CSA.2014.01.

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 and Birch, SEP. 2018 A Standard for the Scholarly Citation of Archaeological Data as an Incentive to Data Sharing. *Advances in Archaeological Practice* 6(2): 125–143. DOI: https://doi.org/10.1017/aap.2018.3.

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.

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

Pajdla, P and Trampota, F. 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, 0): 13. DOI: https://doi.org/10.5334/joad.88.

Pergl, R, Hooft, R, Suchánek, M, Knaisl, V and Slifka, J. 2019 'Data Stewardship Wizard': A Tool Bringing Together Researchers, Data Stewards, and Data Experts around Data Management Planning. *Data Science Journal* 18(1, 1): 59. DOI: https://doi.org/10.5334/dsj-2019-059.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rosenberg, D. 2013 Data before the fact. In: Gitelman, L (ed.). *'Raw Data' is an Oxymoron*. Cambridge, MA: MIT Press. pp. 15–40.

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

Tkáč, P and Kolář, J. 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, 0): 9. DOI: https://doi.org/10.5334/joad.85.

UNESCO. 2021 *UNESCO Recommendation on Open Science*.

Wilkinson, MD, Dumontier, M, Aalbersberg, IjJ, Appleton, G, Axton, M, Baak, A, Blomberg, N, Boiten, J-W, da Silva Santos, LB, Bourne, PE, Bouwman, J, Brookes, AJ, Clark, T, Crosas, M, Dillo, I, Dumon, O, Edmunds, S, Evelo, CT, Finkers, R, Gonzalez-Beltran, A, Gray, AJG, Groth, P, Goble, C, Grethe, JS, Heringa, J, 't Hoen, PAC, Hooft, R, Kuhn, T, Kok, R, Kok, J, Lusher, SJ, Martone, ME, Mons, A, Packer, AL, Persson, B, Rocca-Serra, P, Roos, M, van Schaik, R, Sansone, S-A, Schultes, E, Sengstag, T, Slater, T, Strawn, G, Swertz, MA, Thompson, M, van der Lei, J, van Mulligen, E, Velterop, J, Waagmeester, A, Wittenburg, P, Wolstencroft, K, Zhao, J and Mons, B. 2016 The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3(160018): DOI: https://doi.org/10.1038/sdata.2016.18.