Archaeology Data Infrastructures

Data reuse potentials and limitations to modelling settlement systems

Petr Pajdla

2024-02-13

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.

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 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).

- **Reproduciblity** Using `renv` package makes reproducibility of the whole project less tedious than I would normally expect. Hints:

  - use `renv::status()` to check the status of the project,
  - use `renv::install()` (`install.packages()` is aliased to it as well) to install new packages,
  - use `renv::snapshot()` to update the lockfile,
  - use `renv::restore()` to restore the state of the project/dependencies recorded in the lock file
  - use `renv::update()` to updated the project.

See https://rstudio.github.io/renv/index.html for more details, as well as this figure:

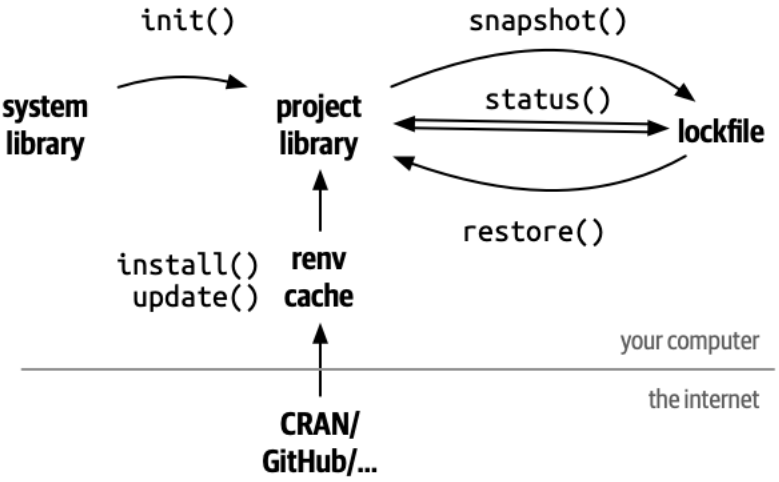

Figure 1.: Basic usage of `renv`

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 February 9, 2024 there are roughly 21.7 pages (1800 characters per page) of text. Length of individual chapters is as follows:

- Introduction: 2 pages
- Theory and Method: 15 pages
- Data and materials: 4 pages

```
     w     c ns                  f
1  508  3551  2  chapters/intro.qmd
2 3963 27385 15 chapters/theory.qmd
3 1222  8089  4   chapters/data.qmd
4 5693 39025 22               total
```

Introduction

ℹ **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

ℹ **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, research and in our case archaeological heritage management.

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.

Digital humanities is now a fully developed interdisciplinary field that utilizes computational methods and digital tools to conduct research, analyze, and interpret humanities data and artefacts. Digital humanities does not create qualitatively *new science*, it enhances and extends the ways in which scholars in the humanities approach their research questions with extensive use of computational methods, statistics, geographic information systems, network analyses, text mining etc.

Archaeology, with its strong interdisciplinary focus and early adoption of many innovations, positions itself somewhat aside from the digital humanities. The difference of archaeology might be given by its specific sources, ie. the material culture objects, which is distinct from many of digital humanities disciplines, which are predominantly based on textual sources. Fieldwork, physical excavations, handling of material culture and focus on physical preservation are processes that lead to archaelogy having apparently closer connections to some of natural sciences, like geology, than to other humanities text-centric disciplines, like history. On the other hand, there are many intersections between digital humanities and archaeology that speak for stronger inclusion of archaeology in this field.

TODO: Describe why DH is relevant, brief intro on Digital, computational and quantitative archaeology, 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 I examine the dichotomy between data- and theory-driven approaches in archaeology, because this work utilises large amounts of data and addressing its data-driven nature is important for methodological transparency and reproducibility, as well as understanding what does it mean in the first place. Abundance of data led to an idea that scientifc method and/or theory is obsolete. This idea was addressed by many, but I will start the 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 scientific method as we know it, i.e. positing a research question based on previous knowledge and observation, formulating a testable and falsifiable hypothesis to address this question, and gathering data or conducting experiments in order to test the hypothesis, is superseded by accumulating large quantities of data (or *big data*[3]) and analyzing it for correlations is good enough, even better, than burdening yourself with building models first. Anderson simply proclaims that *"correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all"*.

Both theory-driven and data-driven aspects are fundamentally rooted in archaeology. The theory-driven aspect of archaeology comes in a plurality of conceptual frameworks or

[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 and variables.

paradigms, e.g. cultural history, cultural ecology, processual and post-processual archaeology etc. Theories in archaeology shape interpretive models that are created to make sense of the past embedded in material objects. They influence the decisions about the design of archaeological projects, like what aspects of archaeological record and data are deemed significant and worthy preserving or what sampling strategy is chosen.

The data-driven aspect of archaeology spans from the focus on material remains of past human activities. The collection, documentation and analysis of the data on artefacts, ecofacts, and other evidence etc. are central part of archaeological practice. The urge to meticulously document each and every detail of the excavated situation is driven by the non-repeatable nature of archaeological research. Together with the use of scientific techniques, such as aDNA analysis, radiocarbon dating etc. and various quantitative methods, GIS, and statistical techniques, an impression or fallacy of objectivity and practice devoid of presumptions or theory arises.

To summarise, theory-driven approach is a top-down process, where knowlege is created by testing models against reality, the data-driven approach is a bottom-up process, where knowledge is created by identifying patterns in large data sets (Maass et al. 2018). Theory and data driven approaches coexist in archaeology in relationship that does I do not see as a strictly binary opposition. Anderson's comment challenges scientific method, emphasizing possibilities of vast data sets for uncovering patterns, yet the vast data sets, or big data, would not exist in archaeology without decisions driven by theory and models how to create them in one way or another in the first place.

In summary, the duality of approaches in archaeology creates an iterative dialogue, where theories guide the formulation of research questions and interpretation of phenomena, while data derived from fieldwork and analysis continually refine theories and models. In this sense, the opposition between theory and data is indeed disappearing, as Kristiansen suggests, and theory- and data-driven approachech permeate and intermingle one with the other, creating a dynamic and evolving discipline.

1.4. Theorizing archaeological data

In the previous section I discussed the interwining nature of archaeology as theory-driven and data-driven practice. Here I delve into the discussion of the heterogeneous nature of archaeological data. Even earlier in this chapter it was explained that here, data is understood as *capta*, i.e. it is actively and creatively made, recorded, generated etc. In this sense, also archaeological data is in fact archaeological *capta*, meaning that in my point of view, it is created the moment it is recorded, it does not exist on its own, an agent's incentive is then a neccessary condition for the *capta*/data to originate.

I will start this section by citing three authors dealing with the nature of archaeological data and use their critical comments on the problematic nature of archaeology data as a starting point for explaining how it is understood throughout this work:

> *"(...) however dense it becomes, archaeological evidence will always remain patchy, with levels of uncertainty and variable expert opinions that are hugely challenging."* (Bevan 2015: 1477)

> *"Archaeological data are shadowy in a number of senses. They are notoriously incomplete and fragmentary, and the sedimented layers of interpretive scaffolding on which archaeologists rely to constitute these data as evidence carry the risk that they will recognize only those data that conform to expectation."* (Wylie 2017: 1477)

> *"Archaeological data is always incomplete, frequently unreliable, often replete with unknown unknowns (...)"* (Huggett 2020: S8)

The citation from Bevan's article could suggests that although abundance of archaeological evidence is accumulated over time, and especially in recent decades in vast amounts (*data deluge*), it is difficult, sic impossible, to draw conclusions about the past. On the contrary, the author acknowledges that *"(...) at least in certain parts of the world, we cannot in all good conscience claim 'we don't yet have enough evidence' or that we should 'wait till the evidence is in'."* In a similar manner, Kristiansen (2014: 18) assumes that *"After 40 years of contract archaeology, real historical knowledge about settlements and landscapes is possible."* However optimistic these claims sound, they are not without problems. The first one is expressed in Bevan's citation, ie. the patchy nature of archaeological evidence and associated uncertainty levels. Although there is most certainly a finite amount of archaeological evidence buried below the surface, in some way our knowledge is never complete, because some (many) patches of land remain unexplored. This makes the evidence, *sensu* data, close to infinite in the sense that we will never be sure whether we are missing any pieces or not.

Wylie addresses these aspects of archaeological data as *notorious incompleteness and fragmentarity*. What more, she stresses that archaeological data is always burdened by the *interpretive scaffoldings*, that is, the theories and models individual archaeologists, the active agents in the data creation process, carry with themselves. This poses a conundrum. If we accept the *theory-driven* aspect of archaeology described in the previous section we approach data recording with certain models in mind, and in turn we fail to recognize and record evidence that does not conform to our expectation, to paraphrase Wylie. If we stick to the *data-driven* aspect of archaeology, we fail to recognize that the data we are re-using was created by someone with certain models and theories in mind. These examples are obviously great simplifications, but illustrate the point that no archaeological data exist without theory, models or expectations in their provenance.

Lastly, Hugget's assertion adds *unknown unknowns* to the mix of archaeological data.

TODO: Defining archaeological data, micro- to macro-scales;

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 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) 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.

[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.

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

[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.

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

ID	Interoperability	Value
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 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

ℹ **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*.

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. Archaeological information system of the Czech Republic

⚠ **Conflict of interest disclosure**

Since 2018 I am part of the AIS CR team at the *Institute of Archaeology, Czech Academy of Sciences, Brno* and since 2023 I am a member of the executive committee of this project. This text is thus influenced by my inside view of how the infrastructure works.

2.1.2. Information system about archaeological data

2.1.3. Legacy data sources

What is a legacy data source?

2.1.4. Museum databases

2.1.5. Other pre-existing data

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.

2.2. Data management

> **Note**
>
> This section 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.

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

The DMP included as Appendix B is generated by Data Stewardship Wizard (DSW, Pergl et al. 2019), an online tool dedicated to cooperative creation of data management plans. It is included both as part of the text and as a standalone machine actionable file. The notes below take into account templates for data management planning created in the Ariadne project (Doorn and Ronzino 2022) as well as the DSW template.

2.2.1. Data re-use

The work is predominantly based on re-using existing data. The sources of data are listed and described in detail in the beginning of this chapter (ie. Section 2.1).

The single most complete source for archaeology data in the Czech Republic is without a doubt the AIS CR infrastructure. The (meta)data of AIS CR infrastrucutre are accessed through a public OAI-PMH API. This data is published under the CC BY-NC 4.0 International license, what makes the conditions for its use clear and transparent.

AIS CR data is aggregated in the ARIADNE infrastructure is accessed through SPARQL endpoint of the dedicated triple store.

The spatial data from ISAD is queried from an ArcGIS rest API. The attributes for this data, ie. SAS database, is webscraped, because it is not published in any other machine-readable form. The ISAD is licensed under CC BY-SA 4.0 International license and SAS metadata, as an integral part of the ISAD ecosystem falls under this license as well, although this fact is not explicitly stated on the website.

The data is accessed through custom scripts developed in the R language.

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.

2.2.1.1. Data creation and collection

No new data is collected or created per se, but re-used data sources are processed, remixed, interlinked etc. into new, different, data. Vector spatial data is stored as geodatabases in OGC GeoPackage, most of other data is stored as comma separated values (CSV) files. Both formats are open standards, CSV is well suited for logn term preservation (LTP) of text-based data. Preservation of code and data is planned in a sense that whole repository of this project will be deposited at Zenodo trustworthy repository and independently archived by the university.

2.2.1.2. Controlled vocabularies and ontologies

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.

2.2.2. Data processing

Data is processed predominantly in the R programming language and environment using varius packages and custom written scripts (see Section 1.6 for detailed overview of software used and how reproducibility is handled). Changes in data, code and text are recorded using a version control system git and preserved online in a repository at GitHub.

2.2.3. 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?

2.2.4. Access to data and code

The data used in this work will be as open as possible, as most of the data is licensed under one of the CC BY-NC or CC BY-SA licenses. The code developed here is published under a very permissive and simple MIT license, where appropriate. As already mentioned, the whole repository will be deposited at Zenodo data repository, as well as archived by the university and already is availbale throuh GitHub interface, including the record of the changes.

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].

Bevan, A. 2015 The data deluge. *Antiquity* 89(348): 1473–1484. DOI: https://doi.org/10.15184/aqy.2015.102.

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.

Maass, W, Parsons, J, Purao, S, Storey, VC and Woo, C. 2018 Data-Driven Meets Theory-Driven Research in the Era of Big Data: Opportunities and Challenges for Information Systems Research. *Journal of the Association for Information Systems* 1253–1273. DOI: https://doi.org/10.17705/1jais.00526.

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.

Wylie, A. 2017 How Archaeological Evidence Bites Back: Strategies for Putting Old Data to Work in New Ways. *Science, Technology, & Human Values* 42(2): 203–225. DOI: https://doi.org/10.1177/0162243916671200.

A. Glossary and abbreviations

A.1. Institutions, infrastructures, services etc.

AAT, Getty AAT Getty Art & Architecture Thesaurus

AIS CR Archaeological information system of the Czech republic (Archeologický informační systém České republiky, https://aiscr.cz/). Central data infrastructure in the Czech republic operated jointly by Institutes of Archaeology of the Czech Academy of Sciences in Prague and Brno. Its main component is the AMČR.

AMČR Archaeological map of the Czech republic (Archeologická mapa České republiky, https://amcr-info.aiscr.cz/). A backbone ecosystem of AIS CR gathering data on archaeological fieldworks in the Czech republic.

ARIADNE ARIADNE (https://ariadne-infrastructure.eu/) is an international infrastructure devoted to archaeological data preservation and sharing. ARIADNE is operating ARIADNE Portal, a central point of access to archaeological resources offered by different partner institutions.

ARUB Institute of Archaeology, Czech Academy of Sciences, Brno (https://www.arub.cz/).

ARUP Institute of Archaeology, Czech Academy of Sciences, Prague (https://www.arup.cas.cz)

ISAD Information system about archaeological data (Informační systém o archeologických datech, https://isad.npu.cz/), an information system operated by the NHI. ISAD consists in principle of spatial data and its metadata, ie. SAS.

NHI (NPÚ) National Heritage Institute (Národní památkový ústav, https://npu.cz/).

NM Natinal Museum (Národní muzeum, https://www.nm.cz/)

SAS State archaeological list (Státní archeologický seznam), part of ISAD operated by the NHI.

A.2. Other abbreviations

DMP Data management plan, also data stewardship plan.

DSW Data stewardship wizard (https://ds-wizard.org/), an online tool for DMP creation. A FAIR Wizard, an instance of DSW operated by the Czech Academy of Sciences, was used to create a DMP enclosed as Appendix B.

FAIR data Findable, interoperable, accessible and reusable data.

B. Data management plan

Machine actionable version (JSON) https://github.com/petrpajdla/dataInfrastructures/blob/main/appendices/dmp.json
Based on Common DSW Knowledge Model, 2.6.3 (dsw:root:2.6.3)
Project phase Before Finishing the Project
Created by Petr Pajdla (pajdla@arub.cz)
Archeologický ústav AV ČR, Brno
Generated on 11 Feb 2024

Data Management Plan created in Data Stewardship Wizard ds-wizard.org

History of changes

Version 1.0

Publication date 2024-02-11

Changes

There are no named versions

B.1. Section A: Data Collection

B.1.1. 1. What data will you collect or create?

B.1.1.1. Re-used datasets

We have found the following non-reference datasets that we have considered for re-use:

- **Archaeological Map of the Czech Republic** (AMCR)

  The Archaeological Map of the Czech Republic (AMCR) is a repository designed for information on archaeological investigations, sites and finds, operated by the Archaeological Institutes of the CAS in Prague and Brno. The archives of these institutions contain documentation of archaeological fieldwork on the territory of the Czech Republic from 1919 to the present day, and they continue to enrich their collections. The AMCR database and related documents form the largest collection of archaeological data concerning the Czech Republic and are therefore an important part of our cultural heritage.

The AMCR digital archive contains various types of records - individual archaeological documents (texts, field photographs, aerial photographs, maps and plans, digital data), projects, fieldwork events, archaeological sites, records of individual finds and a library of 3D models. Data and descriptive information are continuously taken from the AMCR and presented in the the AMCR Digital Archive interface.

B.1.1.2. Data formats and types

We will be using the following data formats and types:

- **Comma-separated Values** (CSV)

  A comma-separated values (CSV) file is a delimited text file that uses a comma to separate values. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format. A CSV file typically stores tabular data (numbers and text) in plain text, in which case each line will have the same number of fields.

  It is a standardized format. This is a suitable format for long-term archiving. We will have only a small amount of data stored in this format.

- OGC Geoackage

  It is a standardized format. This is a suitable format for long-term archiving. We will have only a small amount of data stored in this format.

B.1.2. 2. How will the data be collected or created?

There will be no instrument dataset in this project.

B.1.2.1. Data storage and file conventions

The project will require so little storage space for all data and software (including temporary storage) that it is not a problem.

We will use a filesystem with files and folders. We document how we manage file versioning for files and folders.

We will not be storing data in an ”object/document store” system.

We will not use a database system to store project data.

B.2. Section B: Documentation and Meta-data

B.2.1. 3. What documentation and meta-data will accompany the data?

List of data to be published is given in Section E, Question 9. This also includes information about catalogs where the data can be found. Information about data types used is given in Section A, Question 1.

We will include keywords and relevant ontology references to optimise the possibility for discovery and potential reuse.

Metadata will be openly available. Metadata will available in a form that can be harvested and indexed (managed by the used repository / repositories).

B.3. Section C: Ethics and Legal Compliance

B.3.1. 4. How will you manage any ethical issues?

B.3.1.1. Data we collect

We will not collect any data connected to a person, i.e. ”personal data”.

The data collection is not subject to ethical legislation.

B.3.2. 5. How will you manage copyright and Intellectual Property Rights (IPR) issues?

We will be working with the philosophy *as open as possible* for our data.

All of our data can become completely open over time.

Limited embargo will not be used as all data will be opened.

All data will be owned by the Principal Investigator.

For the reference and non-reference data sets that we reuse, conditions are as follows:

- **Archaeological Map of the Czech Republic** (AMCR)

  The Archaeological Map of the Czech Republic (AMCR) is a repository designed for information on archaeological investigations, sites and finds, operated by the Archaeological Institutes of the CAS in Prague and Brno. The archives of these institutions contain documentation of archaeological fieldwork on the territory of the Czech Republic from 1919 to the present day, and they continue to enrich their collections. The AMCR database and related documents form the largest collection of archaeological data concerning the Czech Republic and are therefore an important part of our cultural heritage.

> The AMCR digital archive contains various types of records - individual archaeological documents (texts, field photographs, aerial photographs, maps and plans, digital data), projects, fieldwork events, archaeological sites, records of individual finds and a library of 3D models. Data and descriptive information are continuously taken from the AMCR and presented in the the AMCR Digital Archive interface.
>
> It is freely available with obligation to quote the source (e.g. CC-BY).

B.4. Section D: Storage and Backup

B.4.1. 6. How will the data be stored and backed up during the research?

Data that project members themselves store adequately backed up and traceable. Therefore data are protected against both equipment failure and human error.

B.4.2. 7. How will you manage access and security?

Project members will not store data or software on computers in the lab or external hard drives connected to those computers.They can carry data with them on password-protected laptops. All data centers where project data is stored carry sufficient certifications. All project web services are addressed via secure HTTP (https://...). Project members have been instructed about both generic and specific risks to the project.

The possible impact to the project or organization if information is lost is small. The possible impact to the project or organization if information is leaked is small. The possible impact to the project or organization if information is vandalised is small.

We are not using any personal information.

B.5. Section E: Selection and Preservation

B.5.1. 8. Which data are of long-term value and should be retained, shared, and/or preserved?

B.5.2. 9. What is the longterm preservation plan for the dataset?

None of the used repositories charge for their services.

We have a reserved budget for the time and effort it will take to prepare the data for publication.

B.6. Section F: Data Sharing

B.6.1. 10. How will you share the data?

Information about used repositories (i.e. where will potential users find out about the data) is provided in Section E, Question 9.

Embargo on the data is described in Section C, Question 5, and Section F, Question 11.

B.6.2. 11. Are any restrictions on data sharing required?

Ethical and legal restrictions are documented under Section C. We have used the Data Stewardship Wizard, which made us aware of options to minimize the restrictions.

No data sharing agreement will be required.

We are not running the project in a collaboration between different groups nor institutes. Therefore, no collaboration agreement related to data access is needed.

B.7. Section G: Responsibilities and Resources

B.7.1. 12. Who will be responsible for data management?

Petr Pajdla is responsible for implementing the DMP, and ensuring it is reviewed and revised.

Petr Pajdla is responsible for reviewing, enhancing, cleaning, or standardizing metadata and the associated data submitted for storage, use and maintenance within a data centre or repository.

Petr Pajdla is responsible for maintaining the finished resource.

B.7.2. 13. What resources will you require to deliver your plan?

To execute the DMP, no additional specialist expertise is required.

We do not require any hardware or software in addition to what is usually available in the institute.

Charges applied by data repositories (if any) are mentioned already in Section E, Question 9.