

6

Methods for Managing Paradata

Zanna Friberg and Isto Huvila

6.1 Introduction

This chapter changes the perspective on paradata offered so far in this volume. It builds on the insights into the creation, use and reuse of paradata from the studies addressed in previous chapters and places the focus on management of paradata in repositories. The chapter draws on the studies conducted in the CAPTURE project, on paradata in the context of creation and use within the research process, to inform practices in a repository setting. As such, this chapter can be read as an exploratory outline of methods for paradata management on that basis, or as possible directions for further exploration of paradata curation.

So far, this volume has covered paradata in the research process both as it is created and used, but between creation and use may very well be a repository housing the dataset in question. This chapter turns the attention to this crucial context. Efforts of researchers to be more aware of paradata as a means of making research data more reusable will be greatly diminished if it is not compatible with the processes in repositories.

Data managers need understanding of paradata in theory and practice, how it can be approached and thought about in management settings, and what is its place as a part of the broader data landscape. If repositories are not equipped to manage paradata when these are recorded or have not established practices to preserve potential sources of paradata, there may be considerable paradata loss when data is deposited. Furthermore, for a holistic approach to paradata in research it would be beneficial if repositories were able to provide data creators with instructions on how best to record paradata in preparation for repository management.

Similarly, repositories can play an important role in directing data users to potential sources of paradata in datasets. In this chapter, we primarily report on an exploration of the types of paradata researchers need (Börjesson et al., 2022) and where paradata might be found (see Chapter 3 in this volume) and frame the insights from those studies in terms that can help increase paradata literacy for curators and researchers. The chapter also utilises some of the lessons learned from other studies in CAPTURE (e.g. Börjesson et al., 2021; Börjesson et al., 2022; Huvila, 2020; Huvila et al., 2025). In doing this, this chapter seeks to provide insights about paradata that can inform management of paradata in repositories.

A few key terms need further explanation. Paradata management in this context is understood as the identification, preservation and dissemination (i.e. management) of information about research processes and practices (i.e. paradata). As discussed earlier in Chapters 2 and 3 of this volume, paradata can exist in many forms and in many places and what is paradata for one person may not be paradata for another. While paradata management goes on throughout the research process, in this chapter the scope is on management in repositories. The term repository is used broadly in this chapter, encompassing data repositories but also other management collections, including museum and archival repositories.

An important caveat needs to be asserted here. It is worth bearing in mind that the broad scope and general character of this chapter comes at a cost. While the aim certainly is to provide actionable advice on methods for the management of paradata, we focus on general strategies rather than on detailed instructions. These general strategies may be more or less applicable depending on research domain, repository and national context. Furthermore, the methods outlined here are purposely not entirely newly developed or experimental approaches, since the chapter aims to build on and support existing infrastructures and practices. What this chapter does is to position paradata (as is understood in this volume) on the map of repository management for those interested in paradata and data curation. It also proposes a framework that can function as a point of reference for dialogue between researchers and data curators on the documentation of research processes and practices.

Proposed best practices for research data management have frequently addressed issues that are relevant in the paradata context even if their focus has not been specifically on managing practice and process information. There are also many texts that provide relevant guidance for both researchers and data managers working in general data repositories (e.g., Austin et al., 2016) and in specific fields, for example, in archaeology (Kansa and Kansa,

2021) and plant science (Leonelli et al., 2017), and with particular types of resources (Thomer et al., 2022; VandenBosch et al., 2023). Surveys of specific (e.g., quality assurance in Kindling and Strecker, 2022) and general data management policies and practices do also provide insights into applicable management methods (e.g., Cofield et al., 2024; Geser et al., 2022; Schöpfel and Rebouillat, 2022). A more in-depth exploration of the definition of paradata can be found in Chapter 2 of this volume and the insights about where paradata is found is expanded on in Chapter 3. Paradata management in creation and use scenarios outside of repositories are to an extent covered in Chapters 4 and 5.

Finally, a note on the structure of the chapter. This chapter comes in three parts. It begins by introducing a way to categorise paradata from a management perspective that is useful for developing appropriate management strategies. The two broad categories being *core paradata* and *potential paradata*. After providing that lens to conceptualise paradata in a management context, the following section of the chapter expands on selected examples of paradata needs that may require management and sources of paradata to be mindful of in managing for paradata preservation. The last part of the chapter brings the two previous parts together to discuss two broad strategies for paradata management, standardisation and embracing messiness, before rounding off the chapter with a discussion of paradata literacy as a framework within which to conceptualise and situate paradata management and use.

6.2 Core Paradata and Potential Paradata

The management of paradata, like its creation and use, is situational, and different paradata types and potential needs require different management approaches. The strategies proposed for paradata management in this chapter focus on identifying ways in which existing structures and methods for research documentation can be adjusted or informed to account for paradata, in its many forms. In this chapter, the methods explored can be boiled down to two complementary approaches to managing paradata, based on categorisations of paradata that need management into two types: core paradata and potential paradata.

Core paradata is documentation of processes where there is a general consensus of the documented information being paradata. Core paradata can be structured or narrative descriptions of ‘how it was done’ or versioning data generated by digital systems. Two fields where the concept of paradata has become more established is survey research and 3D visualisation in cultural

heritage. In these fields core paradata is therefore easier to exemplify. In survey research, records of a survey or interview process, timing, locations, secondary observations and data processing is commonly documented as paradata (Schenk and Reuß, 2024). An example from 3D visualisation would be formal encoding of the decisions preceding the generation of digital 3D objects (e.g., Havemann, 2012; Rabinowitz, 2019).

Potential paradata refers to information that was not purposely created as paradata but may emerge or is extracted in a specific context for a specific use. It encompasses the vastness of auxiliary documentation that could be paradata depending on the user, the use and the context of that use. Work logs, emails and personal notes will be further discussed as such auxiliary documentation on the basis of the discussion in Chapter 3 of paradata in research documentation, drawing on interviews conducted in the CAPTURE project (Börjesson and Sköld, 2021). Other previously identified sources of potential paradata are different types of marginalia, marginal notes and remarks providing additional context to the main document or artefact. Marginalia have been identified as potentially rich sources of paradata about research processes (see e.g. Edwards et al., 2017). A source from which a researcher can forensically extract paradata (that is, they can analyse and interpret its content to find useful process information), that was not purposely documented as such. What paradata they seek and extract would differ depending on what their particular research required and thus be very hard to define beforehand by either the data creator or curator.

Separating paradata into these categories is not clear cut, since there can conceivably be secondary process information within core paradata that would potentially be useful as paradata for some other research endeavour down the line. However, the categorisation helps us see how metadata curation could be adjusted to include paradata and also, more importantly, that paradata curation would in many cases need to be differentiated from metadata curation. If we agree that paradata is useful and needed for data reuse and also that much paradata is situational, we need to think of ways to manage the potential for paradata accordingly.

For core paradata, existing structures (i.e. traditional research documentation such as reports and metadata schemas) are generally useful starting points, and only minor changes of perspective or granularity may be needed to better manage for paradata. When dealing with potential paradata, however, where the potential is at heart context-dependent, standardisation is not only very difficult but also possibly counterproductive. When managing the potential for paradata extraction in research documentation we are instead exploring other strategies to prevent paradata loss, strategies to guide wayfinding and paradata

literacy. All these strategies for core paradata as well as potential paradata, are listed below:

Suggested strategies for paradata management

- Expanding on the standardising work that is already being done
- Adjusting for a level of documentation granularity that is more conducive to paradata
- Preventing loss of paradata potential in datasets due to excessive data cleaning
- Facilitating paradata discovery and wayfinding
- Improving mutual paradata literacy among data creators and data curators

These strategies are expanded on at the end of the chapter as broadly being methods based in standardisation work or methods based in ‘embracing messiness’ as a way to preserve potential paradata. The literacy strategy is thought of as a holistic and overarching perspective. A further discussion of paradata literacy as an overarching strategy makes up the final part of this chapter. That deeper discussion will build on some specific insights gained from the CAPTURE project’s work on paradata, which is the subject of the following part of this chapter.

6.3 Paradata Needs and Sources

There are several ways to paint an overview of the insights on paradata provided through the findings of the CAPTURE project that would facilitate a discussion of paradata management. The focus here will be on some of the needs for paradata among researchers and examples of where and how they seek to meet this need. This section outlines examples of the paradata needed by researchers and maps them to repository management. It also includes examples of how researchers try to satisfy these needs for information on processes and practices. Together, these insights provide a framing for the proposed management strategies explored in this chapter.

The discussion in Chapter 3 of where paradata may be found in research documentation shows the range of sources data reusers actually consult to meet the need for comprehensive information about data and ultimately to better understand what was done and how. To assemble as clear a picture of the research process as possible, the initial source researchers consult is often method descriptions available across research documentation. That includes method descriptions in both published and unpublished scholarly outputs such

as journal articles, monographs, reports, manuals and guides. Though such descriptions can range from highly standardised to more free-form, there is a degree to which they are genre-bound and narratively accessible descriptions of such things as research questions, framing, methods, material, analyses and results.

However, it is well established that datasets and the narratives of how they were created and processed are cleaned and structured to make the often complicated and perhaps messy research process neater (see e.g. Börjesson et al., 2022; Huvila et al., 2021; Kim et al., 2019; Leighton, 2015). These practices and how they are documented do not always correspond with each other (Button and Harper, 1996). In attempting to reuse someone else's data, such formal, tidy descriptions are rarely enough to get a sense of the reliability or suitability of the data (for a further discussion of important considerations in reuse scenarios see Börjesson, Sköld, et al., 2022). At the same time, there can be an expectation of the tidiness of descriptions as an often false measure of the quality of the described practices and processes (Wylie, 2019).

Börjesson et al. (2022) explored four paradata needs that exemplify how data reusers go beyond report documentation to try and gain better understanding of the dataset at hand. In the study, the authors build on and confirm the well-studied general information needs of researchers, while exploring paradata needs as a subset of such information needs. The types of information generally shown to be needed by researchers are, as noted, contextual and varied, but include for example, information for different phases of a research project from idea generation to analysis and dissemination (Toms and O'Brien, 2008).

The needs vary between disciplines and especially different research approaches (Huvila et al., 2021). Concerning needs, paradata is not a mere attachment to research data but rather a parallel output of study results and an integral part of the process of how research conveys knowledge and understanding, as discussed further in Chapter 7.

In their study, Börjesson et al. (2022) suggested four types of interconnected process-related information needs and these will serve as illustrations of needs to take into account in paradata management. The list is perhaps neither exhaustive nor equally applicable in all fields but is useful as a starting point to structure the exploration of methods for managing paradata in this chapter.

Four categories of paradata needs

- **Methods paradata:** process information related to how the data was generated

- **Scope paradata:** process information related to what the data cover
- **Provenance paradata:** process information related to where the data come from
- **Knowledge organisation and knowledge representation paradata:** process information related to how the data are structured and communicated.

Methods paradata is information about research methods, techniques and decision-making related to processes and practices of generating data. This kind of information is to an extent included in more traditional scholarly documentation and metadata schemas, but paradata needs are a need for more detailed process information than can usually be extracted from those. There is often a need for information about the methods used to generate the data on a different level of granularity or detail than traditionally offered. Particularly elusive among the methods paradata is decision-making related paradata, which is more likely to be informally documented and/or cleaned away when reporting on the research. An example of important decision-making paradata could be information about how data outliers or sample contamination were to be treated in the particular data set.

Another example of this type of paradata relevant in archaeology as well as several other field sciences is process information about coordinates. Method descriptions that include coordinates can fail to specify if coordinates were generated through a GPS device, measured using a total station or theodolite, or measured from a map. In a reuse scenario where a researcher wants to either find a specific site or perhaps aggregate field data, this piece of process information about data creation (i.e. methods paradata) may be crucial. The implication of coordinates being generated based on a map could, for example, be a higher likelihood of human error, or potential need to identify and look more closely into what specific map was used (see e.g. Ullah, 2015). In contrast, GPS measurements are susceptible to poor satellite coverage leading to lower accuracy and differences between specific types of GPS devices.

Coordinates are also good examples of how paradata needs are interconnected. Provenance paradata about when or by which actor particular parts of a data set were generated could help establish whether the coordinates at hand were likely to have been created using a GPS device or not, based on when they started to be widely used in a particular setting. Information about, for example, the existence of additional data related to the same coordinates, whether located in the same repository or elsewhere, could constitute scope paradata – a third type of paradata identified by Börjesson et al. (2022).

Scope paradata expands on the type of scope-related metadata that provides information about the extent and field of reference of the data. This could

typically include such information as the time period (temporal scope) and geographical area (spatial scope) covered by the data but also information on social groups or communities covered. An example of scope paradata would be information about where other data related to the previously discussed coordinates exists, regardless of whether it would be stored in the same or another repository, or remain in the hands of private individuals or research teams. Scope paradata could also inform of such process-related information as changes in policies or practices in the course of the research, such changes that may have led to differences in scope of the units of data within the dataset. In archaeology for example, policy changes on reporting practices, like whether it is mandatory or voluntary to report certain types of archaeological finds, will be important to know to be able to understand the scope of a data set. This is equally true on a smaller scale, where changes of the sampling strategies at a field site may result in different scope across a data set (see e.g., Faniel et al., 2021). Both are examples of paradata that is crucial for judging the reliability and usability of the data in a reuse scenario. Information about the sampling processes may also qualify as potential scope paradata and potential decision-making related methods paradata. This again illustrates both the interrelatedness and the contextual nature of different paradata types. Moreover, the same information could also constitute the next type of paradata need, provenance paradata.

Provenance paradata goes beyond structured metadata about the whys, whens and whos of the data. The reuse needs discussed in Börjesson et al. (2022) as needs for provenance paradata concern primarily information about the disciplinary and timebound provenance of data, the epistemological underpinnings of data generation, and the rationale behind the data creation (cf. the discussion of paradata and provenance in Chapter 2). Just as knowing about the effects of a new reporting policy or specific sampling policies is important for the understanding and evaluation of a dataset, the temporal and disciplinary context may significantly condition data. Methods for selecting, classifying and collecting data change over time, just as typologies, theories and tools develop (see e.g. Montoya and Morrison, 2019). Data generation and documentation also differ between disciplines (e.g. between archaeology and biology), between different research environments (e.g. two different departments within the same discipline) and with different aims (e.g. aiming to create a reference database or a database of primary data). The disciplinary and timebound provenance of data could also be very informative in a secondary capacity when trying to meet the fourth paradata need identified by Börjesson, Huvila and Sköld, the need for paradata on knowledge organisation and knowledge representation.

Paradata on knowledge representation concerns how data are structured and communicated, something that can be highly specific to specific disciplines or times. This paradata need entails an expressed requirement to know how gaps, subsets, relationships and data points are represented in the dataset. It also encompasses the need to know not merely which standards, if any, have been used to structure the data but also how strictly a standard has been applied. For example, the format of coordinate information discussed previously forms an example of the type of knowledge representation paradata that can make a significant difference when trying to reuse or aggregate field data collected by someone else. The accuracy of coordinates and which one of the dozens of possible coordinate systems used to represent the information is critical for being able to use coordinate data (Hansen and Fernie, 2010; Verhagen, 2023). Furthermore, knowledge representation paradata can be information about the meaning of non-standardised elements including descriptors and keywords, and other information central to understanding the structure of the data set. Empty cells are a typical recurring element in a data table with multiple possible meanings. Börjesson and colleagues point out the importance of knowing whether the absence signifies, for example, that ‘no data was gathered’, ‘no data was available’, ‘data was gathered but did not meet the database creator’s quality criteria’ or any other of the possible alternatives (Börjesson et al. 2022).

Regardless of whether information on methods, scope, provenance and knowledge representation exists or not, an additional layer of difficulty is that finding paradata on decisions made in relation to data creation can require a great deal of work from researchers. This particular issue led to the formulation of the London Charter document instituting the notion of paradata in the field of 3D heritage visualisation after the turn of the millennium (Denard, 2012).

The discussion in Chapter 3 of paradata types shows that researchers engage in quite impressive detective work to leverage sources of process information to find this kind of decision-making paradata. In gathering paradata to help in their investigation, a conventional strategy used by researchers is to consult other, auxiliary, documentation originating from the research process. Such auxiliary documentation can take the form of, work logs where tasks and procedures are tracked, emails where research activities are discussed within the team of researchers, or personal notes made by individual researchers. These three examples represent a range of auxiliary documentation that can be interpersonal and written with an audience in mind or written with no expectation of them ever being seen by anyone else.

Whether or not the documentation is intended for others to read will also impact its accessibility as a source of paradata, with the prevalence of

abbreviations and shorthand increasing the more personal the documentation is. Wylie (2019) notes, the tidied descriptions written for external audiences have a tendency to be very different from the personal working notes that can be a much more crucial source of information in making datasets reusable. Correspondingly, as observed in Chapter 3, while work logs may require more domain-specific knowledge to interpret than a scholarly publication, a personal note may be mostly unintelligible for someone other than the person who wrote it.

In response to these needs, and the types of paradata sources they are associated with, different methods could be applied in order to manage for paradata. The paradata needs may be met through already existing core paradata or through potential paradata that is suitable for being turned into core paradata through structuring and revising it according to relevant data documentation standards. Some needs, in turn, may be hard to meet for a repository, too difficult or time-consuming to demand of data producers and documenters, or are from the outset best met through keeping the potential paradata as such, that is, by managing the potential rather than the (para)data. In the following section, we turn back to the strategies outlined in the chapter introduction and discuss them in more depth with these insights, on what paradata researchers seek, in mind.

6.4 Paradata Management

At the beginning of this chapter, we pointed to a set of possible strategies for managing paradata that we can now start to map to the insights from the studies of paradata needs and practices in order to create a framework for informing paradata management. The framework (Figure 6.1) builds on categorising practice and process information generated during research as core paradata or potential paradata combined with insights drawn from identified categories of paradata needs and findings on where paradata is likely to be found.

We proceed on the basis of this framework, to outline two broad categories of methods for managing paradata through standardisation work and adopting strategies for coping with and navigating the messiness of practice and process information. The strategies presented at the outset are best imagined on a sliding scale between these two broader extremes of core paradata and entirely contextually dependent potential paradata. The strategies have been placed along this scale according to their alignment with managing either core paradata or potential paradata, as delineated in the Figure 6.1.

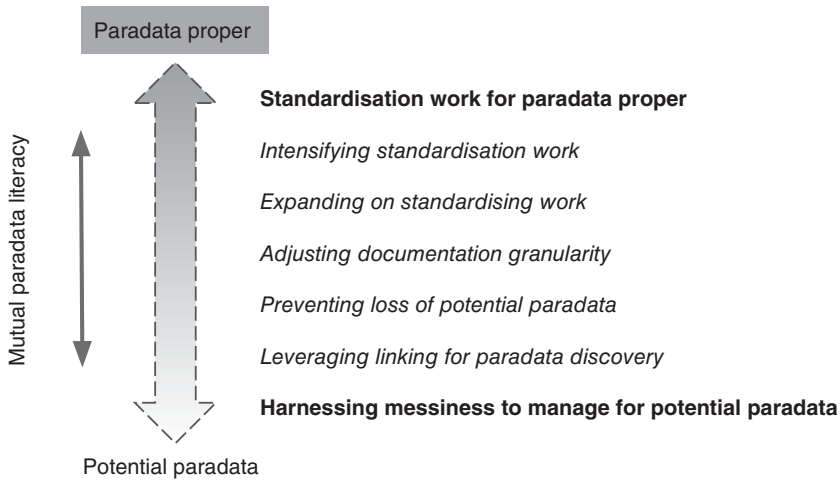


Figure 6.1 A framework for paradata management.

As suggested at the upper end of the spectrum (Figure 6.1), in most cases when a dataset ends up at a repository, there is a subset of information about processes that is already being documented as core paradata in conventional data documentation whenever even a very rudimentary description of a dataset is available. This includes available metadata elements and report documentation, whether the information is termed paradata or not. For this core of core paradata, the very elementary strategies for managing meta-information, that is, adjusting documentation guidelines and standards to include information about research processes and practices could to a larger extent help make the core paradata more visible and easily usable.

Standardisation work for core paradata (Figure 6.1), is a fundamental exercise towards this end. This would be an exercise to map the existing formal descriptors to the four previously described categories of paradata needs and to identify what information relating to the methods, scope, provenance and knowledge representation already is present, how it can be made searchable, and in specific contexts of research, complemented with critical missing elements.

On the far end of the spectrum, *intensifying standardisation work* (Figure 6.1) is the main suggested strategy. Methods paradata is represented to a varying degree in current data documentation standards and, even if the development of methods ontologies and taxonomies (e.g., Borek et al., 2016; Galliers, 1991; Sicilia, 2010; Xian and Fu, 2022) has started in certain fields, there is much variation in how methods are described and what terms and concepts are used. For some methods and data types, for example, measuring

spatial coordinates, standards exist but they are not necessarily applied everywhere. For scope paradata, spatial and temporal scope are covered in standards to a greater extent than, for instance, epistemic and practical issues, and for provenance, many standards put a lot of emphasis on curatorial provenance whereas other processes might be represented in a more cursory manner.

Beyond the core paradata, towards the middle of the spectrum, there is potential paradata that is likely to correspond to common enough paradata needs that it could be reasonable to consider standardising how it is documented to make it core paradata. This kind of potential paradata might include various kinds of report documentation, metadata or reasonably genre-bound auxiliary documentation that has the potential to meet common and identifiable paradata needs. For this potential paradata, *expanding on standardisation work* and *adjusting documentation granularity* (Figure 6.1) are suggested.

Passing the middle of the spectrum going towards the opposite end (towards the bottom of the Figure 6.1), the potential of the paradata is imagined as more and more situational and thus less conducive to any standardisation. At this end of the spectrum of working with core paradata, there is immediate potential to selectively expand standardisation to cover practice and process types according to appropriate taxonomies and ontologies, and standardise vocabulary use for knowledge representation paradata.

Harnessing messiness to manage potential paradata (Figure 6.1) is our suggested method for information that is less likely to be potential paradata in several use cases. This end of the spectrum recognises that in certain isolated cases, it is possible that even very unlikely information can convey crucial knowledge of practices and processes for knowledgeable data users. For this type of paradata, that is too context-dependent to be standardised with any foreseeable success, we are inclined to recommend focusing on the work of preserving messiness and providing wayfinding information. This is represented by the strategies *preventing loss of potential paradata* and *facilitating paradata discovery* in Figure 6.1. Such paradata encompasses most qualitative and narrative descriptions of practice and processes and their elements relating to scope, provenance and how particular methods were applied and how their use deviated from standard practice. It is similarly pertinent when retaining the information inherent in the outputs of non-standard data practices and processes. Analyses of archaeological data (Börjesson et al., 2022) and research outputs (Huvila et al., 2023) provide illustrative examples of how the messiness of working documentation contains a lot of invaluable information on how it was conceived.

Paradata literacy (positioned to the left in Figure 6.1) is an overarching and necessary method for paradata management, valid along the whole scale.

There is evidently a pressing need for greater mutual paradata literacy between researchers creating and using data and curators caring for and disseminating data regarding all process information, regardless of it being deemed core paradata or potential paradata here. In the framework illustrated in Figure 6.1 the strategy presented at the outset as ‘improving mutual paradata literacy among data creators and data curators’ has thus been made such an overarching strategy. As a strategy, paradata literacy is crucial to making paradata work in stitching together measures covered by the other strategies. A working level of paradata literacy is necessary to identify what eventually is informative about diverse aspects of practices and processes and understand the vocabulary used to describe practices, processes and methods. In parallel, paradata literacy plays a key role in understanding how different forms of paradata link to each other, for example, how temporal and spatial scope can be informative about provenance and vice versa.

In the following final part of the chapter, we expand upon these in three sections devoted to discussions on standardisation work, ways to embrace messiness and paradata literacy respectively.

6.4.1 Standardisation

Standardisation is the accepted approach in the literature for facilitating data management and use. Better and more rigorous standardisation of data improves the efficiency (Ibanez et al., 2014; Santos et al., 2016), reliability and quality of research (Herres-Pawlis et al., 2022; Pétavy et al., 2019) and in general of data-based knowledge production (Gal and Rubinfeld, 2019), it helps to make data more manageable (Palaiologk et al., 2012; Pétavy et al., 2019), reusable (Faniel and Jacobsen, 2010; Herres-Pawlis et al., 2022), portable and interoperable (Blobel et al., 2022; Gal and Rubinfeld, 2019; de Mello et al., 2022; Ribes and Lee, 2010), and more easily integrated with other datasets (Ibanez et al., 2014). Standardisation improves the reproducibility of research (Ibanez et al., 2014), facilitates data discovery (Contaxis et al., 2022) and meta-analysis, that is, synthesising results from multiple studies (Laird et al., 2011). Data is also easier and more cost-efficient to preserve if it is well-standardised (Palaiologk et al., 2012). Finally, an often recited benefit is that standardised data is helpful in interdisciplinary collaborations where the different stakeholders lack common knowledge of each others’ data practices (Ribes and Lee, 2010).

However, while standardisation has many evident benefits, it also has several drawbacks. Archaeological information management literature has for a long time debated the relative benefits of standardisation and allowing

researchers to document their data as they find it relevant in their specific contexts of work (Huvila, 2012a). A major drawback of standardisation is that standards are always approximations and never entirely in line with local practices or understanding of the processes or knowledge that is being standardised (Ellingsen, 2004). Standards are not always easily scalable for larger or smaller use cases than was originally envisioned (Hanke et al., 2021). Also important to note here, as we indirectly discuss data reuse, is that even if retroactive standardisation and data wrangling is possible, standardisation works best when done upfront. Retroactive standardisation tends to be problematic not only because there are seldom resources for comprehensive retroactive working through large datasets but also because of the imminent risks of unintentional data loss, misinterpretations and misrepresentation of original data (Mandell, 2022).

Flexibility is often a practical necessity in highly context specific situations where standards are seldom comprehensive and nuanced enough to accommodate local variation, divergent epistemic perspectives and observations that are difficult to classify into predetermined categories. A study of archaeological heritage administrators' information work in Sweden provides evidence of how maintaining a fine balance between standardisation of both work processes and documentation, and what is described in the study as ambiguity, is crucial for the success of work (Huvila, 2021). Besides being cumbersome and costly, over-standardisation can become directly counterproductive and an obstacle to both data creation and use. This is an important perspective to hold on to in discussing something as contextual as paradata and one of the reasons for framing paradata management on a spectrum between standardisation and messiness. Keeping this in mind, the three strategies discussed in this section (i.e. intensifying, extending and adjusting granularity in standardisation from Figure 6.1) should be understood as strategies to complement and be complemented by strategies geared towards embracing messiness.

Intensifying standardisation work (Figure 6.1) is suggested in light of the fact that the standardisation of what we term in this chapter core paradata is still very much work in progress. Returning to survey studies and 3D visualisations (previously mentioned as a context where an emergence of core paradata as a documentation category is identifiable), we can say something about the state of paradata standardisation. Survey research has developed a robust corpus of metadata standards but standards for paradata are close to non-existent (Schenk and Reuß, 2024). A review of heritage visualisation standards shows that while many standards provide means to represent practice and process information, they have varying emphases and cover the full spectrum of practice and process information only partially. One of the most

highly cited documents with direct relevance to paradata in that particular field are charters (including the London Charter, see Denard, 2013; and Seville Principles, or the International Guidelines for Virtual Archaeology, International Forum of Virtual Archaeology, 2011; Bendicho, 2013). The charters provide generic recommendations to document paradata rather than detailed advice on how to do it in practice. Existing metadata schemes and standards cover core paradata, but typically do not use the term paradata (exceptions e.g. CARARE metadata schema, see D'Andrea and Fernie, 2013). When they cover such paradata-like information, it is to varying degrees and often with a focus on historical processes and data curation.

A survey of major data documentation schemes conducted as a part of the CAPTURE project suggests that the situation is similar in many generic standards. At the same time, as Börjesson et al. (2020) suggest, even if there are few or no dedicated paradata standards, such generic conceptual models as the International Council of Museums (ICOM) International Committee for Documentation (CIDOC) Conceptual Reference Model (CIDOC CRM) provide means to represent practice and process information in detail to an extent that dedicated standards are perhaps not always necessary. Combining complete schemas and individual elements can help to achieve the required coverage of details. First and foremost, there is an urgent need to develop new and to adopt existing data documentation standards to cover paradata and devise guidelines for core paradata across data-producing and data-using communities. All of this work does not need to be exhaustive. For example, many fields have informal standards and vocabularies to describe phenomena interesting to them. Many of them have never been written down and the terms have not been explicitly described. As Huvila et al. (2022) remark, while such terms might be generally comprehensible, a systematic documentation of terminology and an effort to invite data producers to commit themselves to a better standardised use of concepts could provide significant benefits with relatively little overhead.

Expanding on standardisation work (Figure 6.1), not just intensifying the current work in progress, is also sorely needed. We particularly find it advisable to expand standardisation work on two fronts: generally to temporarily cover a larger part of the continuum (or life-cycle) of data; and, more specifically, to formalise the documentation of process information that corresponds to established paradata needs.

Along with the temporal extension, we are suggesting standardising critical elements of core paradata already as a part of routine research documentation. This would require, in practice, more structured advice in the methods literature (cf. Huvila et al., 2022b; Huvila and Sköld, 2023) about how to document

key aspects of specific types of data creation and processing activities. In addition, there is a need for greater availability of hands-on advice on what specific paradata should be added to datasets that are produced for sharing and reuse at the planning stage of data collection or generation. This should cover not only what type of information should be added but also more specifically what it is expected to communicate.

As with all data documentation requirements, it should be adjusted to the plausible extent and types of uses relevant to particular datasets. For a lot of data, minimal paradata might be enough whereas datasets with high expectations of meticulous process documentation and high standardised reuse potential should be documented in more detail. With small-scale, highly contextual qualitative data collection with low likelihood of direct reuse of the generated data, a sufficient level of standardised documentation could be a brief narrative of data collection procedures and their underpinning rationale, especially if reusing the data can be expected to require a lot of interpretative work from secondary users. This type of data set and reuse scenario would, however, benefit from such strategies as presented further below, that would prevent the loss of potential paradata that could facilitate that interpretative reuse work. In contrast, there are datasets that require more formal and technical paradata to a much higher degree. This is true for cases such as longitudinal and cross-comparable survey datasets that are produced to be used broadly in both academic and societal knowledge production.

The parallel approach suggested here is equally necessary to starting the generation of core paradata earlier in the data creation process, that is, to formalise the documentation of practice and process information that corresponds to established and common paradata needs. In simple terms, this could mean verifying that the major categories of paradata needs (including those discussed previously in this chapter) are covered to a satisfactory degree in the existing and anticipated core paradata. While doing this it is crucial to consult data producing and using communities to establish what particular information corresponds with categories like methods paradata, scope paradata, provenance paradata and knowledge representation paradata in specific contexts. In such consultations, it is important to both include individuals with a relatively short working experience in the particular field, and those who are more reliable on formal explicit information. Huvila's study of archaeological heritage administrators suggested the experienced administrators could often compensate for missing explicit information due to their experience of working with the data on a daily basis. In contrast, other stakeholders could not be expected to be able to do that but must rely on the formal standardised documentation (Huvila, 2021). The categorisation presented in this chapter

can, in consulting relevant communities, provide a starting point for discussion and a checklist for assessing whether conceivable types of needs have been covered in the documentation requirements.

Adjusting the granularity of documentation (Figure 6.1) is the last of the strategies that are conceptualised as standardisation work in this framework for managing paradata. However, as we shall see, it is placed in the middle of the spectrum for a reason. While determining the necessary level of granularity is certainly difficult, we suggest that it is possible to use the presence and expressed needs of auxiliary documentation as a working heuristic to assess the need of more or less core paradata. If a particular data-creating community is generating a lot of auxiliary documentation that qualifies as paradata, it is probably worth investigating whether some of it might be either replaced by or supplemented with core paradata. In some cases it might be possible if the additional documentation is kept due to the absence of adequate formal documentation standards, making it a question of standardisation work.

In other cases, as in ethnographic research, it is not feasible to replace a notebook with anything that qualifies as core paradata, rather the notebook might be kept or referenced as an identified source of potential paradata (more on this in Section 6.4.2). Correspondingly, repeated demand in data-using communities for either existing or missing auxiliary documentation might signal the need for additional core paradata. This could apply, for example, to the inconsistency identifying data creators or specific measurement instruments in the data documentation. The earlier discussed case of emails and personal information exchange being consulted for augmenting report documentation provides another example of a possible need for augmenting the scope of core paradata. It is seldom feasible to ask for a full email correspondence between researchers to be submitted alongside a dataset to a repository even if it would contain useful information for future reusers of the data. Nor is it reasonable to ask for testimonies or contact information for every single person involved in a research project to be included in report documentation.

However, leaning on our increasing knowledge of key actors and their correspondence being potential sources of paradata, the structures for documenting authors and contributors could be adjusted to increase the likelihood that a reuser can find the right person to talk to in order to better understand, for example, decision-making and contextual premises of making a specific dataset. Here, there are also important opportunities for synergies with the strategy of facilitating discovery and wayfinding, discussed Section 6.4.2 on embracing messiness.

6.4.2 Embracing Messiness

For potential paradata, the usefulness and, indeed, presence of anything deemed paradata will be particularly situational and the only course of action may be to embrace and cater to some level of messiness. It is apparent that cleaning a dataset too much obscures the research process behind it. Literal cleaning is unlikely to happen at a data repository due to the lack of time. Even researchers engage in cleaning proper only selectively. Instead, both data managers and producers alike make numerous decisions of requiring or depositing (only) structured data, leaving out parts of data, datasets or auxiliary information, converting and consolidating that all contribute to the resulting data being a little cleaner than before. It is a result of forcing systematisation of measurements, observations and interpretations, which might not be conclusive enough to be systematised to such an extent. Besides obscuring underlying processes, cleaning also challenges both reliability assessments and several reuse scenarios as discussed above. Of course, it is not possible to keep everything. Rather, as with documentation in general, the question is how to document (or keep) enough, and identify what is most likely to be useful. It is important to try to find a position comparable to what York (2022) terms ‘reuse equilibrium’ where there is sufficient information but not too much. Drawing on findings on where potential paradata is more or less likely to occur in a dataset in different contexts can facilitate the choice of what to leave messy, what to clean or not to keep. In this part of the chapter we explore how the resulting messiness can be managed and navigated. We propose starting by looking to traditional archiving practices for possible avenues for managing potential paradata. Guided by some of the fundamental principles of Western archiving practices we propose ways of embracing messiness that include a strategy to keep messes messy (i.e. preventing loss of potential paradata) and a strategy to navigate the messiness (i.e. leveraging linking for paradata discovery).

Preventing loss of potential paradata (Figure 6.1) is rather vague and not particularly actionable without further explicating by what means loss should be prevented. Unsurprisingly, the realisation that it is near impossible to predict what someone in the future will need to know in order to understand the context of some piece of information is far from new – nor is it born from paying attention to paradata. To some extent, a solution to this was proposed within Western archival practice more than a century ago. Having previously applied subject-based classification akin to classification in libraries, archivists in post-revolutionary France developed another principle for arranging and describing records that would be more efficient than the forbiddingly labour-intensive reorganisation of records.

The resulting system, which would be adopted gradually in the Western archival practice throughout the first part of the 1900s, is commonly referred to as *respect des fonds*. The general idea is that records should be maintained as they were originally accumulated, in their original *fonds*. A ‘fonds’ is all the records of, for example, an administrative authority, or a family, or a person (Trace, 2015, p. 21–24). Two important principles inform how records should be organised within fonds, that is, inform archival arrangement and description; these are the *principle of provenance* and the *principle of original order*. As discussed in Chapter 2, the principle of provenance puts emphasis on the importance of relationships between records and the context, functions and agents (e.g. organisations or individuals) involved in their creation, accumulation and use. The principle of original order promotes maintaining the ordering structure established by the records’ creator in order to preserve important contextual information.

Inherent in the principle of original order is a view of the archival records as evidence of the activity of a particular organisation. The idea of *respect des fonds*, with its underlying principles, becomes rather more complicated when applied in practice, since the application of any standardised model of these principles, like Oliver Wendell Holmes’s (1988) influential nested hierarchical model, will impose structure on the original order. However, the foundational idea of *respect des fonds* offers valuable insights that can be applied to the management of paradata, particularly for endeavours to preserve potential paradata.

The implication of *respect des fonds* and its parallel principles of original order and provenance for managing potential paradata is to keep the fonds of such information intact to an extent that is possible. From a repository perspective, potential paradata is essentially unmanageable when it falls outside the formal structure of how a data repository deals with datasets and metainformation, describes and preserves them, and works towards making and keep them formally FAIR or findable, accessible, interoperable and reusable (Wilkinson et al., 2016). Focusing efforts on keeping such information technically readable, and documenting it only superficially to maintain a rudimentary degree of findability is from a repository perspective obviously a question of managing resources. If keeping the information is useful enough, it is possible to avoid costs of turning it into core paradata.

However, as the previous discussion in this volume especially in Chapter 3 shows, it also has qualitative benefits. Turning potential paradata into core paradata is not only costly but it carries also a risk of losing information potentially available in the organic, unrefined messiness of such resources. For example, with methods paradata, looking into the terminology used to describe

methods as a whole is informative not only of specific methods applied but also of how research practice and process is conceptualised as a whole. Moreover, as the work of Börjesson et al. (2022) shows, comparing earlier and later versions of interpretations in datasets can help to trace interpretative processes otherwise undocumented in the available research material.

Having established that some messiness may be beneficial for the preservation of potential paradata, the next question becomes how we might manage such a mess in a relatively systematic way. Again, we turn to the basic principles of archiving and, more specifically, how the idea of archival finding aids may be adapted to be useful tools for paradata management.

Facilitating paradata discovery (Figure 6.1) is a strategy that takes as its starting point archival finding aids. McNeil offers a useful definition of an archival finding aid as ‘any tool that aims to provide users with intellectual and/or physical access to the holdings of archival institutions’ (MacNeil, 2012, 486). This definition can easily be reframed in a data and paradata context, as a tool that aims to provide users intellectual access to the practices and processes underpinning a particular assemblage of data.

Taking directions from traditional archival practice in this way we can introduce some structure to how we can navigate the messes with paradata potential. There are obviously different means to implement a functioning paradata finding aid. Similarly to paradata itself, finding aids are also contextual. However, in contrast to paradata itself, considering the frame within which they are used, finding aids can be expected to be contextual first and foremost in relation to future data use rather than data creation. In parallel, rather than focusing on what such a finding aid should be, it is more fruitful to plan for what it should *do* (cf. MacNeil, 2012) for data creators, managers and users. It is important to consider earlier evidence of the advantages and affordances of traditional paper-based and digital finding aids when considering their potential adoption to paradata. A finding aid can facilitate paradata discovery through providing links between information resources, keeping traces of previous use of the data and the finding aid itself.

Perhaps the most significant function of a paradata finding aid would be to provide directions. Because all types of potential paradata require a lot of interpretative effort to use, it is superfluous to attempt to describe it beyond what is necessary to use the tool itself. Explicating the aboutness of datasets is crucial for understanding their processual context similarly to providing directions where to find information that helps to interpret it, including vocabularies and descriptions of colloquialisms. A finding aid should also provide an overview of the principles of how data was organised and made, and what standards were used and how un-FAIR as in unfindable, unaccessible,

uninteroperable and unreusable the data is (Börjesson et al., 2022). In parallel, a good finding aid would provide directions across contextual and disciplinary boundaries. In many cases relevant data and paradata is not only disconnected in different logical units within one repository but literally scattered around the world. An illustrative example of this is historical and archaeological data that in many cases has been scattered to museums and research institutions located on different continents (e.g., Börjesson et al., 2022; Sobotkova 2018; Stilborg 2021). Remediating or at least bringing partial alleviation to the greatest obstacle in mediating practice and process information, namely its contextuality, is crucial for any tool proposed for the task. While a universalist ideal *a lieu*x Otlet's Mundaneum or Bush's Memex would be a complete meta-level inter-institutional directory to (all) paradata, much more modest and context-specific directories could if not solve the paradata problem as a whole, at least alleviate it considerably.

However, rather than going entirely contextual, there would also be possibilities to develop standards for formats and contents of finding aids for potential paradata. Solving the problem of diverging expectations of data managers, creators and data reusers and what is feasible to achieve for paradata finding aids requires reaching a compromise between what (all) would be regarded as desirable. Such standardisation would be useful for potential reusers seeking potential paradata specific to their needs, but would also be helpful for data creators in providing rudimentary structure to data documentation efforts. Moreover, as Battley (2013) reminds us, archival findings aids are also records, or with paradata, another layer of paradata describing data management practices and processes (cf. Gant and Reilly, 2017). Generally, the standardisation of finding aids for paradata would also support paradata literacy, which is the topic we will now turn to in the final part of this chapter.

6.4.3 Paradata Literacy

There can be no doubt that solving the conundrum of generating or identifying paradata that is useful for its users, feasible to generate and keep for data creators, and manageable from a repository perspective is not merely an issue of selecting suitable technologies and particular types of resources to solve the problem. Paradata as a type of information, about processes and practices, is a moving target. The most comprehensive method for managing paradata may be simply to acknowledge that there is a need for process information not currently taken into account in documentation strategies and that information plays a much more critical role in the chain of knowledge-making and use than a mere footnote or an auxiliary attachment to data proper.

In contrast to finding aids that provide a map to existing paradata and guidance how to find and navigate it, we suggest looking to paradata literacies as a crucial step in making paradata (proper and potential) actionable in context and especially in working out the conundrums of identifying and making diverse forms of potential paradata. This discussion is both complementary to, and more overarching than, the discussion of paradata usefulness thresholds in Chapter 3. Finding aids and the notion of paradata literacy approach the paradata challenges from different angles. Finding aids can help make paradata and the practices and processes it documents better findable and accessible. Paradata literacies are needed to make data and data practices and processes work both on their own and in relation to each other. In practice, both have a much broader scope and in an ideal situation they would complement each other in a pursuit of increasingly relevant, resonant, ethical and sustainable paradata practices, practices that support corresponding, reflective and epistemically inclusive and sensitive data practices.

The argument for increased mutual paradata literacy is based on the insight that the discrepancy between available and desirable paradata is not simply an information problem. It is also, at least to an equal extent, a problem of disparity in paradata practices and competences, as observed in Chapter 3 where the issue is termed the ‘epistemic usefulness threshold’. Researchers have been criticised in many studies for a lack of adequate competence in data management (e.g., Pálsdóttir, 2021). The same critique could be undoubtedly extended to paradata competences, although as our findings from a survey and interviews on researchers’ paradata practices suggest, data creators, managers and users alike show a high degree of often tacit competence in sharing and acquiring practice and process knowledge. The trouble is that these competences, or literacies, do not necessarily converge.

Contemporary data repositories have structures in place to cater to many of the researchers’ paradata needs, and data managers have skills and capacity to facilitate paradata practices but researchers are not always knowledgeable of what they are, how the existing structures and tools work, and how the available resources could be utilised for generating and preserving better paradata. Data managers are similarly struggling with securing a sufficient understanding of how data is created and used to be able to help researchers and other data users and producers. Therefore we want to underline the importance of improving *mutual paradata literacy* (Figure 6.1) instead of focusing on individual stakeholder groups.

Paradata literacy is a natural extension to what has been discussed in terms of data literacy or literacies (Huvila et al. 2024; 2025). It can unfold as something comparable to the competence and associated abilities that have

been termed metadata literacy (Mitchell, 2013) and metadata literacy skills (Çakmak and Kurbanoglu, 2015). It can also be seen as a complement to data literacy, essentially framing paradata literacy as a key facet of established data literacies. Paradata literacy, in the sense relevant here, would cover the competences of identifying, generating and keeping relevant and adequate core paradata and understanding what it does for data and data use.

The major implication of approaching the conundrum of managing paradata as a literacy question is that of acknowledging that making paradata work requires competence and effort and qualifies as a specific type of undertaking of its own. Data managers, creators and users all have specific roles and need to develop systematic competence to contribute to enacting paradata in practice. Acquiring and developing data management skills identified in the literature helps to develop that competence (e.g., Koltay, 2015; Schneider, 2013; Vilar and Zabukovec, 2019). However, considering the often indefinite boundaries of what paradata is and where to find it, paradata literacy extends beyond being a simple skill set. Paradata literacy is first and foremost a matter of understanding the complexities and challenges of working with paradata where theoretical, ethical and interpretative facets of practice cannot be detached from technical skills and practical doing (cf. the ‘technical usefulness threshold’ in Chapter 3 and Kansa and Kansa, 2020). It links to the broader field of data literacies (Verdi and Deuff, 2021) sharing many of the identified challenges but also opportunities and ways forward (cf. Koltay, 2021).

Like data literacies, paradata literacies are also never value-free. In their complexity and embeddedness in data practice, paradata literacies unfold rather as assemblages of ongoing processes and practices that need to evolve in time and across diverse contexts of data creation, management and use. Paradata literacies are also necessarily reciprocal in the sense that the making of paradata needs to be aligned with its management and use, and vice versa. being a competent paradata creator, manager or user, requires adeptness in the whole lifespan of paradata from its inception to management, use, reshaping and beyond.

Similarly to other literacies, paradata literacy is also about finding a balance between empowering users and providing them relevant services (cf. Huvila, 2012b). The analysis of Weber et al. (2023) points to a similar gap between archival users and archivists that is evident also in the broader sphere of data management. It is likely that much of future paradata use will not be mediated by data managers but rather by data creators and users themselves who are depositing and accessing datasets directly in online repositories.

However, even if this would be the case, there is still need for professional support, probably to a greater extent than is acknowledged in the contemporary

management culture, which often downplays the significance and both qualitative and economic advantages of qualified professional service work. We are still to see a well-grounded division of labour between what users should be able to achieve independently and what types of services are necessary and appropriate to offer and when. Repositories and data managers occupy a position with a broad vista across diverse context-specific data practices. In doing so, they hold a unique position and responsibility to not only act as service-providers but to actively raise questions and act in dialogue with data creators and users to find virtuous paths forward.

References

- Austin C. C., Brown S., Fong N., Humphrey C., Leahey A. and Webster P. (2016). Research data repositories: Review of current features, gap analysis, and recommendations for minimum requirements. *IASSIST Quarterly*, 39(4), Article 4. <https://doi.org/10.29173/iq904>.
- Battley B. (2013). Finding aids in context: Using Records Continuum and Diffusion of Innovations models to interpret descriptive choices. *Archives and Manuscripts*, 41(2), 129–145. <https://doi.org/10.1080/01576895.2013.793164>
- Bendicho V. M. L.-M. (2013). International guidelines for virtual archaeology: The seville principles. In *Good Practice in Archaeological Diagnostics. Natural Science in Archaeology*, 269–283. Springer. https://doi.org/10.1007/978-3-319-01784-6_16.
- Blobel B., Ruotsalainen P. and Giacomini M. (2022). Standards and principles to enable interoperability and integration of 5P medicine ecosystems. In Blobel B., Yang B. and Giacomini M. (eds.), *Studies in Health Technology and Informatics*. IOS Press. <https://doi.org/10.3233/SHTI220958>.
- Borek L., Dombrowski Q., Perkins J. and Schöch C. (2016). TaDiRAH: A case study in pragmatic classification. *DHQ*, 10(1), a235.
- Börjesson L., Huvila I. and Sköld O. (2022). Information needs on research data creation. *Information Research*, 27(Special Issue), isic2208. <https://doi.org/10.47989/irisic2208>.
- Börjesson L. and Sköld O. (2021). *The making and use of paradata: An interview study [dataset]*. <https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-455730>
- Börjesson L., Sköld O., Friberg Z., Löwenborg D., Pålsson G. and Huvila I. (2022). Repurposing excavation database content as paradata: An explorative analysis of paradata identification challenges and opportunities. *KULA: Knowledge Creation, Dissemination, and Preservation Studies*, 6(3), Article 3. <https://doi.org/10.18357/kula.221>.
- Börjesson L., Sköld O. and Huvila I. (2021). The politics of paradata in documentation standards and recommendations for digital archaeological visualisations. *Digital Culture and Society*, 6(2), 1. <https://doi.org/10.14361/dcs-2020-0210>.
- Button G. and Harper R. (1996). The relevance of work-practice for design. *Computer Supported Cooperative Work*, 4(4), 263–280. <http://dx.doi.org/10.1007/BF00749172>.

- Çakmak T. and Kurbanoğlu S. (2015). Metadata literacy skills: An analysis of LIS students. In Kurbanoğlu S., Boustany J., Špiranec S., Grassian E., Mizrahi D. and Roy L. (eds.), *Information Literacy: Moving toward Sustainability* (262–269). Springer International Publishing. https://doi.org/10.1007/978-3-319-28197-1_27
- Cofield S. R., Childs S. T. and Majewski T. (2024). A survey of how archaeological repositories are managing digital associated records and data: A byte of the reality sandwich. *Advances in Archaeological Practice*, 12(1), 20–33. <https://doi.org/10.1017/aap.2023.29>.
- Contaxis N. et al. (2022). Ten simple rules for improving research data discovery. *PLOS Computational Biology*, 18(2), e1009768. <https://doi.org/10.1371/journal.pcbi.1009768>
- D’Andrea A. and Fernie K. (2013). CARARE 2.0: A metadata schema for 3D cultural objects. 2013 *Digital Heritage International Congress (DigitalHeritage)*, 137–143. <https://doi.org/10.1109/DigitalHeritage.2013.6744745>
- de Mello B. H. et al. (2022). Semantic interoperability in health records standards: A systematic literature review. *Health and Technology*. <https://doi.org/10.1007/s12553-022-00639-w>
- Denard H. (2012). A new introduction to the London Charter. In Bentkowska-Kafel A., Denard H. and Baker D. (eds.), *Paradata and Transparency in Virtual Heritage* (57–71). Ashgate.
- Denard H. (2013). Implementing best practice in cultural heritage visualisation: The London Charter. In *Good Practice in Archaeological Diagnostics: Non-invasive Survey of Complex Archaeological sites* (255–268). Springer. https://doi.org/10.1007/978-3-319-01784-6_15.
- Edwards R., Goodwin J., O’Connor H. and Phoenix A. (2017). *Working with Paradata, Marginalia and Fieldnotes*. Edward Elgar Publishing.
- Ellingsen G. (2004). Tightrope walking: Standardisation meets local work-practice in a hospital. *International Journal of IT Standards & Standardization Research*, 2(1), 1–22.
- Faniel I., Austin A., Kansa S. W., Kansa E., Jacobs J. and France P. (2021). *Identifying Opportunities for Collective Curation during Archaeological Excavations*. OCLC. www.oclc.org/research/publications/2021/identifying-opportunities-collective-curation-during-archaeological-excavations.html.
- Faniel I. M. and Jacobsen T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues’ data. *Computer Supported Cooperative Work (CSCW)*, 19(3), 355–375. <https://doi.org/10.1007/s10606-010-9117-8>.
- Gal M. S. and Rubinfeld D. L. (2019). Data standardization. *New York University Law Review*, 94(4), 737–770.
- Galliers R. D. (1991). Choosing appropriate information systems research methodologies; A revised taxonomy. In Nissen H.-E., Klein H. K. and Hirscheim R. A. (eds.), *Information Systems Research: Contemporary Approaches and Emergent Traditions* (327–345). Elsevier.
- Gant S. and Reilly P. (2017). Different expressions of the same mode: A recent dialogue between archaeological and contemporary drawing practices. *Journal of Visual Art Practice*, 17(1), 100–120. <https://doi.org/10.1080/14702029.2017.1384974>.

- Geser G., Richards J. D., Massara F. and Wright H. (2022). Data management policies and practices of digital archaeological repositories. *Internet Archaeology*, 59, Article 2. <https://doi.org/10.11141/ia.59.2>.
- Hanke M., Pestilli F., Wagner A. S., Markiewicz C. J., Poline J.-B. and Halchenko Y. O. (2021). In defense of decentralized research data management. *Neuroforum*, 27(1), 17–25. <https://doi.org/10.1515/nf-2020-0037>.
- Hansen H. J. and Fernie K. (2010). CARARE: Connecting archaeology and architecture in Europeana. *EuroMed 2010: Digital Heritage*, 450–462. https://doi.org/10.1007/978-3-642-16873-4_36.
- Havemann S. (2012). Intricacies and potentials of gathering paradata in the 3D modelling workflow. In Bentkowska-Kafel A., Denard H. and Baker D. (eds.), *Paradata and Transparency in Virtual Heritage* (145–160). Ashgate.
- Herres-Pawlis S. et al. (2022). Minimum information standards in chemistry: A call for better research data management practices. *Angewandte Chemie International Edition*, 61(51), e202203038. <https://doi.org/10.1002/anie.202203038>.
- Holmes O. W. (1964). Archival arrangement: Five different operations at five different levels. *The American Archivist*, 27(1), 21–42.
- Huvila I. (2012a). Being formal and flexible: Semantic Wiki as an archaeological e-science infrastructure. In Zhou M., Romanowska I., Wu Z., Xu P. and Verhagen P. (eds.), *Revive the Past: Proceeding of the 39th Conference on Computer Applications and Quantitative Methods in Archaeology, Beijing, 12-16 April 2011* (186–197). Amsterdam University Press. https://proceedings.caaconference.org/paper/21_huvila_caa2011/.
- Huvila I. (2012b). *Information Services and Digital Literacy: In Search of the Boundaries of Knowing*. Chandos.
- Huvila I. (2020). Information-making-related information needs and the credibility of information. *Information Research*, 25(4), paper isic2002. <https://doi.org/10.47989/irisic2002>.
- Huvila I. (2021). Ambiguity, standards and contextual distance: Archaeological heritage administrators and their information work. *Open Information Science*, 5(1), 190–214. <https://doi.org/10.1515/opis-2020-0121>
- Huvila I., Sköld O. and Börjesson L. (2021). Documenting information making in archaeological field reports. *Journal of Documentation*, 77(5), 1107–1127. <https://doi.org/10.1108/JD-11-2020-0188>.
- Huvila I., Börjesson L. and Sköld O. (2022a). Archaeological information-making activities according to field reports. *Library & Information Science Research*, 44(3), 101171. <https://doi.org/10.1016/j.lisr.2022.101171>
- Huvila I., Andersson L. and Sköld O. (2022b). Citing methods literature: Citations to field manuals as paradata on archaeological fieldwork. *Information Research: An International Electronic Journal*, 27(3). <https://doi.org/10.47989/irpaper941>
- Huvila I., Kaiser J., Sköld O. and Andersson L. (2024). Paradata literacy and the challenges of research data management. *Informaatiotutkimus*, 43(3–4), 109–113. <https://doi.org/10.23978/inf.148594>
- Huvila I., Olsson M., Sköld O., Kaiser J. and Andersson L. (2025). Being literate on data or practices: How paradata functions in the context of literacy. *Proceedings of the 2025 Conceptions of Library and Information Science Conference*.

- Huvila I. and Sköld O. (2023). A fieldwork manual as a regulatory device: Instructing, prescribing and describing documentation work. *Journal of Information Science*. <https://doi.org/10.1177/01655515231203506>
- Ibanez L., Schroeder W. J. and Hanwell M. D. (2014). Practicing open science. In *Implementing Reproducible Research*. Chapman and Hall/CRC.
- International Forum of Virtual Archaeology. (2011). *The Seville Principles: International Principles of Virtual Archaeology*. <http://smarterheritage.com/seville-principles/seville-principles>.
- Kansa E. and Kansa S. W. (2021). Digital data and data literacy in archaeology now and in the new decade. *Advances in Archaeological Practice*, 9(1), 81–85. <https://doi.org/10.1017/aap.2020.55>
- Kim J., Yakel E. and Faniel I. (2019). Exposing standardization and consistency issues in repository metadata requirements for data deposition. *College & Research Libraries*, 80(6), 843–875. <https://doi.org/10.5860/crl.80.6.843>.
- Kindling M., and Strecker, D. (2022). Data quality assurance at research data repositories. *Data Science Journal*, 21, 18–18. <https://doi.org/10.5334/dsj-2022-018>.
- Koltay T. (2015). Data literacy: In search of a name and identity. *Journal of Documentation*, 71(2), 401–415. <https://doi.org/10.1108/JD-02-2014-0026>.
- Laird A. R. et al. (2011). The BrainMap strategy for standardization, sharing, and meta-analysis of neuroimaging data. *BMC Research Notes*, 4(1), 349. <https://doi.org/10.1186/1756-0500-4-349>.
- Leighton M. (2015). Excavation methodologies and labour as epistemic concerns in the practice of archaeology: Comparing examples from British and Andean archaeology. *Archaeological Dialogues*, 22(1), 65–88. <https://doi.org/10.1017/s1380203815000100>.
- Leonelli S., Davey R. P., Arnaud E., Parry G. and Bastow R. (2017). Data management and best practice for plant science. *Nature Plants*, 3, 17086. <https://doi.org/10.1038/nplants.2017.86>.
- MacNeil H. (2012). What finding aids do: Archival description as rhetorical genre in traditional and web-based environments. *Archival Science*, 12(4), 485–500. <https://doi.org/10.1007/s10502-012-9175-4>.
- Mandell R. (2022). Applying new standards to old data: Wrangling metadata in a sports archive. *Journal of Digital Media Management*, 11(1), 18–24.
- Mitchell E. (2013). Trending tech services: Metadata use in everyday situations: What does it mean for libraries? *Technical Services Quarterly*, 30(4), 402–413. <https://doi.org/10.1080/07317131.2013.819750>.
- Montoya R. D. and Morrison K. (2019). Document and data continuity at the Glenn A. Black Laboratory of Archaeology. *Journal of Documentation*, 75(5), 1035–1055. <http://doi.org/10.1108/JD-12-2018-0216>.
- Palaiologk A. S., Economides A. A., Tjalsma H. D. and Sesink L. B. (2012). An activity-based costing model for long-term preservation and dissemination of digital research data: The case of DANS. *International Journal on Digital Libraries*, 12(4), 195–214.
- Pálsdóttir Á. (2021). Data literacy and management of research data: A prerequisite for the sharing of research data. *Aslib Journal of Information Management*, 73(2), 322–341. <https://doi.org/10.1108/AJIM-04-2020-0110>.

- Pétavy F. et al. (2019). *Global Standardization of Clinical Research Data*. 28. www.appliedclinicaltrialsonline.com/view/global-standardization-clinical-research-data
- Rabinowitz A. (2019). Communicating in three dimensions: Questions of audience and reuse in 3D excavation documentation practice. *Studies in Digital Heritage*, 3(1), Article 1. <https://doi.org/10.14434/sdh.v3i1.25386>.
- Ribes D. and Lee C. P. (2010). Sociotechnical studies of cyberinfrastructure and e-research: Current themes and future trajectories. *Computer Supported Cooperative Work (CSCW)*, 19(3), 231–244. <https://doi.org/10.1007/s10606-010-9120-0>.
- Santos L. O. B. et al. (2016). FAIR Data Points Supporting Big Data Interoperability. *Enterprise Interoperability in the Digitized and Networked Factory of the Future*. 8th International Conference on Interoperability for Enterprise Systems and Applications, I-ESA 2016. <https://research.utwente.nl/en/publications/fair-data-points-supporting-big-data-interoperability>.
- Schenk P. O. and Reuß S. (2024). Paradata in surveys. In Huvila I. Andersson L. and Sköld O. (eds.), *Perspectives on Paradata: Research and Practice of Documenting Data Processes*. Springer.
- Schneider R. (2013). Research data literacy. In Kurbanoglu S., Grassian E., Mizrachi D., Catts R. and Špiranec S. (eds.), *Worldwide Commonalities and Challenges in Information Literacy Research and Practice* (134–140). Springer. https://doi.org/10.1007/978-3-319-03919-0_16.
- Schöpfel J. and Rebouillat V. (eds.). (2022). *Research Data Sharing and Valorization: Developments, Tendencies, Models* (1st ed.). Wiley. <https://doi.org/10.1002/9781394163410>.
- Sicilia M.-Á. (2010). On modeling eesearch work for describing and filtering scientific information. In Sánchez-Alonso S. and Athanasiadis I. N. (eds.), *Metadata and Semantic Research* (247–254). Springer. https://doi.org/10.1007/978-3-642-16552-8_23.
- Sobotkova A. (2018). Sociotechnical obstacles to archaeological data reuse. *Advances in Archaeological Practice*, 6(2), 117–124. <https://doi.org/10.1017/aap.2017.37>.
- Stilborg Ole. (2021). A study of the representativity of the Swedish ceramics analyses published in The Strategic Environmental Archaeology Database (SEAD). *Fornvännen*, 116(2), 89–100.
- Thomer A. K., Starks J. R., Rayburn A. and Lenard M. C. (2022). Maintaining repositories, databases, and digital collections in memory institutions: An integrative review. *Proceedings of the Association for Information Science and Technology*, 59(1), 310–323. <https://doi.org/10.1002/pr2.755>.
- Toms E. G. and O'Brien H. L. (2008). Understanding the information and communication technology needs of the e-humanist. *Journal of Documentation*, 64(1), 102–130. <https://doi.org/10.1108/00220410810844178>.
- Trace Ciaran (2015) 'Archival arrangement' in Duranti L. and Franks P. C. (eds.) *Encyclopedia of Archival Science*, Rowman & Littlefield Publishers, Incorporated, Blue Ridge Summit.
- Ullah I. I. T. (2015). Integrating older survey data into modern research paradigms. *Advances in Archaeological Practice*, 3(4), 331–350. <https://doi.org/10.7183/2326-3768.3.4.331>.

- VandenBosch A., Maull K. E. and Mayernik M. (2023). Jupyter Notebooks and institutional repositories: A landscape analysis of realities, opportunities and paths forward. *The Code4Lib Journal*, 58. <https://journal.code4lib.org/articles/17751>.
- Verdi U. and Deuff O. L. (2021). La data literacy distribuée. Périmètres définitionnels, origines documentaire, perspectives réticulaires [Distributed data literacy. Definitional scope, documentary sources and reticular perspectives]. *Les Cahiers Du Numérique*, 16(2–4), 137–173.
- Verhagen P. (2023). Spatial information in archaeology. In Pollard A. M., Armitage R. A. and Makarewicz C. A. (eds.), *Handbook of Archaeological Sciences* 1st ed., (1163–1181). Wiley. <https://doi.org/10.1002/9781119592112.ch57>.
- Vilar P., and Zabukovec V. (2019). Research data management and research data literacy in Slovenian science. *Journal of Documentation*, 75(1), 24–43. <https://doi.org/10.1108/jd-03-2018-0042>.
- Weber C. S., et al. (2023). *Summary of Research: Findings from the Building a National Finding Aid Network Project*. OCLC Research. <https://doi.org/10.25333/7A4C-0R03>.
- Wilkinson M. D., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>.
- Wylie C. D. (2019). Overcoming the underdetermination of specimens. *Biology & Philosophy*, 34(2), 24. <https://doi.org/10.1007/s10539-019-9674-2>.
- York J. (2022). *Seeking Equilibrium in Data Reuse: A Study of Knowledge Satisficing* [Thesis]. <https://doi.org/10.7302/6170>.
- Xian C. and Fu M. (2022). *Towards a Taxonomy of Human-Computer Interaction (HCI) Methods Based on a Survey of Recent HCI Researches*. 2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA), 1192–1196. <https://doi.org/10.1109/ICPECA53709.2022.9718950>.