4

Methods for Generating and Documenting Paradata

Ying-Hsang Liu and Isto Huvila

4.1 Introduction

Using appropriate methods to capture adequate paradata on data generation practices and processes is an essential step in facilitating reuse of data and understanding the practices and processes of data creation. In the previous chapters of this volume, we have delved into discussing the notion of paradata (Chapter 2) and where it can be found across data documentation (Chapter 3).

An analysis conducted by Juneström and Huvila as part of the CAPTURE project revealed two major categories of methods relevant for data creators to generate paradata: prospective and in-situ ones. Prospective paradata generation takes place before a practice or process is enacted whereas in-situ refers to paradata generation at the time when an activity takes place. Many prospective methods are prescriptive. Specifically, prescriptive methods aim to create structured approaches for directing forthcoming data generation activities. These methods include enforcing the use of formal metadata standards and knowledge representation frameworks (ontology building), registered reports and prescriptive workflows.

In-situ methods involve generating paradata simultaneously with the creation of data. Such approaches include narrative descriptions through note-taking and data storytelling, recordings such as photographs and audio-visual recordings, and the automatic logging of activities.

In this chapter, we will explore how data creators can generate task-appropriate paradata on various practices and processes related to data creation, management and use. The aim of this chapter is to provide insights into methods that can be adopted and adjusted by researchers and data managers to capture paradata across various disciplinary contexts and study scenarios.

Besides offering practicable advice on how to generate and document paradata, these methods serve as examples of diverse approaches that can be employed to enhance methodological transparency in data creation processes, ensuring the understandability, and for instance, accuracy, replicability and credibility of research. The suitability of these approaches depends on the type of transparency sought and the intended purpose of creating and using the paradata. Key references and further reading are provided after each method description.

4.2 Methods Descriptions

The set of methods described in this chapter were chosen based on a scoping review of paradata generation practices in research activities from a wide range of disciplines. Many other methods exist than those covered in this chapter. Some are specific to particular domains, such as survey research (cf. Chapter 2) and meta-analysis, with established approaches on how to document for them relevant aspects of research-related practices and processes (Kreuter, 2013; Schmid et al., 2021). A preliminary framework of paradata generation developed at the beginning of the CAPTURE project formed a baseline for identifying methods of paradata documentation and generation (Huvila, 2022). This was supplemented by reviewing a large number of articles sourced from project team members over the first four years of the project. The goal was to understand how paradata is generated in different settings and how different approaches eventually could be applied to create different types of paradata.

The chapter starts with formal metadata, a section that is also longest, to discuss the approach that is systematically promoted as the principal method for comprehensive documentation of data and a key premise for its findability, accessibility, interoperability and reusability (Wilkinson et al., 2016). Techniques that are specific to certain disciplines and study contexts are mentioned briefly when relevant, to illustrate their potential for wider application (e.g., protocol registration for clinical trials and experimental protocols in life sciences). However, many such methods that are particularly domain specific have generally been excluded from this chapter to maintain its focus on general principles and approaches with potential relevance across a broader range of domains. However, it is important to note that disciplinary specificity does not necessarily imply a lack of broader relevance. Some approaches originating from specific fields, like registered reports in social psychology, clearly have the potential to inform practices far beyond their initial context.

The following sections introduce and discuss five categories of methods for generating and documenting paradata: 1) formal metadata; 2) narrative descriptions; 3) recordings; 4) logging; and 5) planning and workflows.

4.2.1 Formal Metadata

Formal metadata refers to data descriptions produced according to standards that establish consistent criteria, methods, processes and practices for describing resources for particular purposes, including data structures, data contents, data values and means of data exchange (Zeng and Qin, 2022). It is one of the principal approaches used for describing and organising information and data. Formal metadata systems provide context and consistency for data descriptions. They are often called 'authority files'. A formal metadata system may be specific for one organisation, national or international consortia, or associations and cover a subject specialty or a broader discipline. Formal metadata systems are often organised in tree structures in which the metadata terms (concepts) are arranged hierarchically from the broad to the narrow, or by relationships indicating states of connectedness. Relationships can be associative, relational or equivalent (for synonymous terms). Formal metadata systems provide both descriptions of terms and rules for their use. Formal metadata makes data and paradata easier to manage and contributes to its discoverability and retrievability. It also aids in identifying and differentiating between similar and dissimilar resources. Key instruments of formal metadata include data dictionaries, label sets, controlled vocabularies and metadata standards.

Data Dictionaries and Label Sets

A data dictionary is a tool that provides detailed information about the structure, meaning and relationships of variables within a dataset for describing data contents. For survey research it might include, for example, the types of data (e.g., numerical, categorical and free text), wording of survey questions, question types (e.g., multiple choice, Likert scale and rating scale) and the meaning of the data values. Relevant information for other types of research varies but may include comparable documentation of data types, terms, methods and descriptions of data.

A data dictionary helps to capture details of a data collection process more comprehensively, improves consistency and provides means to turn implicit practices explicit. From a paradata perspective, data dictionaries are especially helpful for addressing data users' needs (cf. the need of knowledge organisation and representation paradata in Börjesson et al., 2022). They help users

understand how a dataset is organised, the conventions and considerations applied when it was structured, the process of transforming data into knowledge and how this knowledge is represented.

The primary limitations of data dictionaries are their lack of long-term stability. They are created at one point of time but as people start using new terms, old terms change meaning and are abandoned; the dictionaries become obsolete. Also, while they enable data creators to articulate their assumptions, contextualise datasets and delineate the scope of the data, they often fall short in documenting the social, political and historical contexts of their creation (Poirier, 2022). When creating a data dictionary, it is difficult to know what contextual facets need to be described and to what extent. Further, they often incorporate assumptions their creators might not recognize or deem necessary to document. Consequently, in practice, data dictionaries are less reliable than commonly assumed (Poirier, 2022).

Label sets, either included in data dictionaries or held separately, classify data or information by predefined categories or groups. Labels are specific words, expressions or notations that are assigned to data points to categorise them. For example, whenever survey participants' highest level of education is discussed, a standardised terminology can be used to organise education programmes and related qualifications by levels in a way similar to how biology uses standardised names based on binomial nomenclature to classify species by genus and species names.

For documenting survey paradata, comparable standardised expressions can refer to the names of specific methods of data collection, management and analysis, such as collecting individual data points online, by telephone or by mail. Examples of applicable labels include 'structured face-to-face interview', 'online questionnaire survey', or a 'semi-structured telephone interview'. To clarify their meaning, they need to be accompanied with a detailed procedural description in a label set. Similarly, labels can be used to standardise the names of the procedures of recording variables on a new scale (e.g., Min-Max scaling, standardisation or robust scaling for common procedures), normalised, or when new variables are computed from earlier ones, for instance, by multiplying values of an earlier variable or by calculating the current age of survey participants based on their date of birth.

As a part of the description of categorical variables, label sets can be documented within a data dictionary. Sufficient documentation of data using data dictionaries and label sets is crucial for data harmonisation. This contributes so that the variables described are self-explanatory enough for other researchers to reuse the data, and that the results are replicable across studies. A typical example of a label set in survey research is the description of labels

for various education levels that can be based, for example, on the list of qualifications in the International Standard Classification of Education (ISCED 2011). This set may also specify whether participants were asked to select the highest level of education attained or completed, or for example, to indicate all types of education they have completed. Diagnostic research uses classifications and label sets for similar purposes, such as to provide labels for different stages of cancer that can be used to describe diagnoses.

A clear challenge is determining what constitutes sufficient detail and how to be reasonably certain that the label descriptions are understandable for their intended users. Producing workable label sets is usually possible within a single domain with shared vocabulary and concepts whereas it tends to be difficult in interdisciplinary research where such an understanding is lacking.

Controlled Vocabularies

Controlled vocabularies are useful for consistently describing data contents by unifying the various terms used to describe the same concept (Liu and Wacholder, 2017; Svenonious, 1986; Zeng and Qin, 2022). Controlled vocabularies have been developed for many domains, such as MeSH[1] (Medical Subject Headings) for biomedical domain, INSPEC Thesaurus[2] for the engineering and information technology fields, and European Language Social Science Thesaurus (ELSST)[3] for the social sciences. A major benefit of using controlled vocabularies for documenting paradata is that they enable consistent indexing and retrieval of resources, and to distinguish between documentation of practices and processes (paradata) and the documentation of documents describing them.

For example, in the biomedical domain, the MeSH term 'Clinical Trial Protocol' is a preferred term (i.e. controlled) for other term variants, including 'Clinical Trial Protocols', 'Trial Protocol' and 'Trial Protocols' to document a specific 'Publication Type' that reports a clinical trial protocol. A different term – 'Clinical Trial Protocols as Topic' – is used to describe a clinical study, including the study's objectives, design and methods. In the context of data repositories, the adoption of controlled vocabularies by open repository software can enhance the interoperability across repositories. For example, both 'clinical study' (referring to research reports) and 'clinical trial data' (referring to data from a clinical trial study) are controlled terms of Resource Types in COAR[4] (Confederation of Open Access Repositories) Controlled

[1] https://meshb.nlm.nih.gov/. [2] www.theiet.org/publishing/inspec/guides-and-support.
[3] https://elsst.cessda.eu/. [4] https://vocabularies.coar-repositories.org/.

Vocabularies, making distinctions among different types of resources in data repositories.

Many controlled vocabularies contain elements relevant for expressing paradata, but so far few paradata-specific vocabularies exist. One exception is the Data Practices and Curation Vocabulary (DPCVocab), which can be useful for characterising research data practices like data collection and generation, processing and analysis. This vocabulary helps achieve more consistent data descriptions among curators, data producers, system developers and other stakeholders involved in the data curation process (Chao et al. 2015). Another example of a scheme with direct relevance to the documentation of paradata is a recently developed annotation scheme for characterising research data practices in the disciplines of sociology, economics physics and biology (Lee et al. 2023). This scheme includes research data practices of collecting, processing, analysing, representing and publishing or citing data, along with the dimensions of action, object and instrument. The finding that each type of research data practice varies across disciplines regarding action (e.g., interview, gather or observe), object (e.g., participant, tissue or circuit) and instrument (e.g., questionnaire, centrifuge or tensor) can be further developed for paradata-specific vocabularies applicable to specific domains.

The primary advantage of using a controlled vocabulary is that it can help to enhance the consistency of paradata documentation, facilitate the development and sharing of a common understanding of concepts used to describe practices and processes, and make paradata more transparent and easier to compare and integrate with data from multiple sources. To embed the use of controlled vocabularies within the work routine, the design of controlled vocabularies can be enhanced and guided by investigating people's interactions with information within their information-seeking activities and constraints (Mai 2008). However, the usefulness of a controlled vocabulary is limited by the search interface design and extensive training required for using it effectively for information searching in various domains (Golub et al. 2023; Liu and Wacholder 2017). Further, accommodating new concepts quickly is challenging and the maintenance and update of controlled vocabularies are resource-intensive, requiring inputs from domain experts.

Metadata Standards

Metadata standards for research data documentation exist across disciplines, although not universally, with many domains lacking dedicated specifications. These standards incorporate elements relevant for describing practices and processes, that is, paradata, to varying degrees. Domain- or discipline-specific standards are typically more effective at describing data generation practices

and processes at a higher level of granularity than cross-disciplinary standards, as shown in the specificity of the paradata-related metadata elements in Table 4.1.

For example, Darwin Core and NetCDF Climate and Forecast (CF) Metadata Conventions enable more consistent descriptions of biodiversity data by providing a vocabulary standard for describing practices and processes in great detail. In contrast, many general metadata standards are far less detailed. For example, the WHO Trial Registration Data Set is a standard for describing clinical trials. It stipulates on the inclusion of some contextual information on the data collection process in a trial study, including descriptions of clinical interventions, study type and recruitment of participants. For observational studies in the social, economic and behavioural sciences, the broader methodology and processing details can be documented using the Data Documentation Initiative (DDI) Codebook standard, which describes the variables, files, source material and study level information.

While lists of metadata terms and descriptions of terms are comprehensible for humans, they are difficult to process using computer programmes. In other words, they are not machine-readable. Machine-readability facilitates automatic processing of formal metadata, and effective searching and linking of metadata terms. It is particularly useful for processing large amounts of data and data descriptions.

Metadata schemas are specifications developed to make metadata terms and vocabularies machine-readable, that is, readable and processable by computer programmes. They provide an implementation blueprint for a metadata standard by encoding a metadata element set into a machine-readable format, much like putting together the pieces of a puzzle (Zeng and Qin, 2022). Encoding metadata schemas involves converting metadata elements into a structured, machine-readable format, such as XML (Extensible Markup Language) which is currently the lingua franca for storing and exchanging information across systems. One possible application of using metadata standards to generate and document paradata is that they can be encoded and be made machine-readable, which will facilitate data exchange among the resources.

Examples of machine-readable metadata schemas with potential relevance to documentation of paradata include the Encoded Archival Description (EAD)[5] developed for encoding information on archival records in the XML format, including paradata-relevant information of their provenance, and many others. Schema.org,[6] another widely used metadata schema for structured data

[5] www.loc.gov/ead/. [6] https://schema.org/.

Table 4.1 *A selective list of metadata standards for research data documentation*

Metadata standards	Discipline/ domain	Type of research	Metadata elements useful for describing paradata
Darwin Core (DwC)[1]	Biological diversity	Observational studies	• Occurrence • Organism • MaterialEntity • MaterialSample • Event • Location • GeologicalContext • Identification • Taxon • MeasurementOrFact • HumanObservation • MachineObservation • . . .
NetCDF Climate and Forecast (CF) Metadata Conventions[2]	Climate science	Observational studies	• Description of the Data ○ Units ○ Ancillary Data • Coordinate Types • Coordinate Systems and Domain • Data Representative of Cells • Reduction of Dataset Size • . . .
WHO Trial Registration Data Set[3]	Biomedical research	Observational and interventional studies	• Countries of Recruitment • Health Condition(s) or Problem(s) Studied • Intervention(s) • Key Inclusion and Exclusion Criteria • Study Type (including study design) • Sample Size • Recruitment Status • . . .
Data Documentation Initiative (DDI), DDI-Codebook	Social, behavioural, economic and health sciences	Observational study	• anlysUnit (unit of analysis) • codeBook • cohort • collMode (mode of data collection) • dataProcessing • instrumentDevelopment • . . .

[1] www.tdwg.org/standards/dwc/.
[2] http://cfconventions.org/cf-conventions/cf-conventions.html.
[3] www.who.int/clinical-trials-registry-platform/network/who-data-set.

in the web pages about datasets, can be used to enhance data integration by search engines. Schema.org is not explicitly a paradata schema but includes types and properties useful for representing diverse aspects of practices and processes, including the type Action for describing 'actions' performed directly or indirectly by agents on 'objects', making it practical especially for many less complex documentation needs.

As envisioned by Tim Berners-Lee, the inventor of the World Wide Web, scientific papers published on the Semantic Web would contain machine-readable content (Berners-Lee and Hendler, 2001). The Resource Description Framework (RDF)[7] has emerged as an alternative general framework for representing and linking data based on a simple data model that uses subject-predicate-object expressions to make statements about resources and their qualities. For example, one might state that Amy (person, subject) created (predicate) a dataset A (object). Linked Open Data (LOD) employs these standards, including RDF, to link resources and resource descriptions together and make them openly available to the world (Nurmikko-Fuller, 2023). The RDF-star extension to RDF is a framework to make assertions about RDF statements, providing a mechanism to describe, for example, their origins and intellectual underpinnings (Rupp et al. 2024). Standardised encoding and open linking provide opportunities for both dissemination and utilisation of paradata. They also enable the use of formal metadata elements across individual standards and vocabularies to enrich documentation of practices and processes.

The goal of using controlled vocabularies, metadata standards, and schemas is to enhance data publishing interoperability. For example, Google's Dataset Search[8] can understand Schema.org and the equivalent structures represented in W3C's Data Catalog Vocabulary (DCAT)[9] format, an RDF-based specification for describing and publishing data catalogues on the web. However, there are gaps in the controlled vocabularies that limit their applicability for describing paradata with pre-defined terms in metadata schemas, such as lack of support for incorporating external controlled vocabularies (Wu et al., 2023) and the difficulty of making legacy vocabularies machine-readable to support the findability, accessibility, interoperability and reusability of data (Cox et al., 2021). As the practice of publishing research datasets as part of research outputs becomes increasingly common, adopting appropriate metadata schemas can facilitate the sharing and publishing of crucial information on practices and processes underpinning the creation, management and use of datasets.

[7] www.w3.org/RDF/. [8] https://datasetsearch.research.google.com/.
[9] www.w3.org/TR/vocab-dcat-3/.

Ontologies and Knowledge Graphs

Unlike metadata schemas, which aim to describe resources and their attributes, ontologies provide a structured framework for representing knowledge within and across domains, capturing the semantics of data. They enable the definition, naming and representation of entities (including concepts and data), and the relationships between them. As a standard language for representing ontologies in the Semantic Web, The Web Ontology Language (OWL)[10] is a knowledge representation language with tools to create Semantic Web ontologies that can be utilised in the development knowledge graphs,[11] that is, a network of interlinked knowledge entities.

One of the major knowledge graphs, Google's Knowledge Graph[12] is a critical component of its search engine. It functions as a database of interlinked pieces of information and is used to source facts about people, places, things, events and their relationships shown on the search engine results pages. For example, typing 'Eiffel Tower' into Google search box will display a panel showing a selection of facts about the Eiffel Tower, such as its location, address, height, date of construction started, opening date and architects. The usefulness of ontologies and knowledge graphs for documentation of paradata lies in their ability to provide a formal representation of all types of metainformation in a single framework. This ensures that paradata is not isolated from other documentation but that they are instead complementary to each other and can be queried together.

Many domain-specific ontologies include mechanisms for documenting and developing specific documentation schemes to represent information about complex practices and processes. There are also dedicated process- and practice-oriented ontologies. For example, PROV is a group of specifications developed for the exchange of provenance information across systems (PROV-Overview 2013). Process Specification Language is another process ontology developed and used for documenting industrial manufacturing processes (ISO 18629–1:2004).

The use of ontologies and knowledge graphs can be illustrated with an example from the cultural heritage domain. The CIDOC Conceptual Reference Model (CRM)[13] is a widely used formal ontology for documenting cultural heritage data. It can be used to produce a knowledge graph that makes connections among researchers, data and practices in a network of relations

[10] www.w3.org/OWL/.
[11] There is no firm consensus of a definition knowledge graph and consequently, how they are distinct from, for example, ontologies (cf. Huck 2022).
[12] https://support.google.com/knowledgepanel/answer/9787176?hl=en.
[13] www.cidoc-crm.org/.

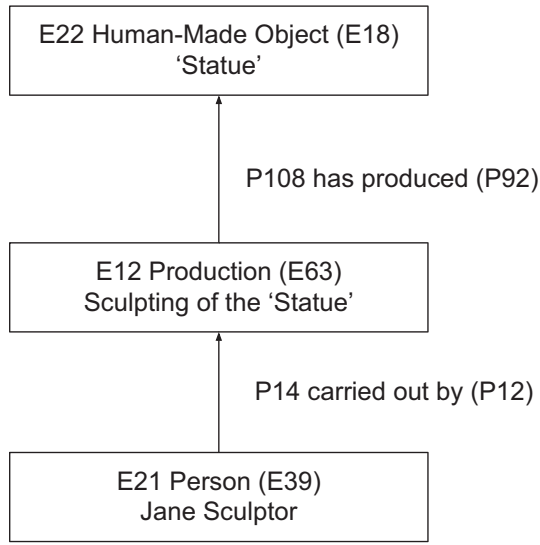

Figure 4.1 Knowledge graph of an artwork titled 'Statue' incorporating simple paradata.

(Oldman and Tanase, 2018). A knowledge graph based on CIDOC CRM can include paradata to incorporate information on the processes involved in creating, collecting and curating cultural heritage data, such as how, when and under what conditions the data was gathered or digitised.

Figure 4.1 shows the knowledge graph of an artwork, entitled 'Statue' representing simple paradata about its production. Further paradata about the sculpture, copies, drawings and photographs of it but also further details such as how the 'Statue' has been referred to in the literature can also be stored in the graph. The uniqueness of multiple originals in the archival ecosystem provides a digital space for studying the history of art (Caraffa et al. 2020). CIDOC CRM is an extensible ontology meaning that it can easily incorporate additional types. A number of CIDOC CRM extensions have been developed including, for instance, CRMdig for documenting provenance information (Doerr et al. 2016; Theodoridou et al. 2010), CRMpe for cross-research-infrastructure information (Bruseker et al., 2017b), CRMsci for information about scientific observation, measurements and processed data in descriptive and empirical sciences (Doerr et al., 2014), and CRMInf on argumentation and inference making in descriptive and empirical sciences (Stead and Doerr, 2015).

A similar example to CIDOC CRM, in the biomedical domain, is the SMART Protocols Ontology that can be used for representing information

about the practices and processes related to experimental protocols (Giraldo et al., 2018). An experimental protocol is essentially a recipe for an experiment, detailing its method and design. The SMART ontology includes data elements relevant to documenting paradata, providing means for describing the execution of protocols and procedures, as well as relevant material artefacts such as laboratory equipment, consumables and software. Figure 4.2 offers an overview of the elements and relations within the SMART Protocols Ontology, illustrating how it can be used to encode experimental protocols consisting of procedures and sub-procedures and their elements. It also shows how individual protocols can be linked to other protocols and the literature through a set of relationships.

Advantages and Limitations of Formal Metadata

Formal metadata is useful for representing knowledge both within specific disciplinary and study contexts and across domains. Its primary advantage in documenting paradata lies in how it contributes to standardising vocabulary, concepts and labels and documentation of practices and processes, which improves the discoverability and interoperability of documentation. Formal metadata-based documentation of paradata allows for making paradata machine-readable and technically compliant to the FAIR (Findability, Accessibility, Interoperability, Reuse) principles for data management (Wilkinson et al., 2016) which are increasingly embraced by funding agencies and research administrative authorities around the world. When a specific term from a documentation standard is consistently used to refer to the same practice or process, all descriptions of that method can be found simultaneously. Similarly, when a standard stipulates what needs to be documented, it is more likely that the particular information will be available for its users, provided data creators follow the standard. When described using formal metadata, research data can be found in digital data repositories, accessible for digital data analysis in open formats, technically and semantically interoperable through metadata standards, and easier to reuse when accompanied by appropriate data licensing information.

One of the major drawbacks of formal metadata is that not all practice and process knowledge is easy to formalise in an ontology. There is a risk of data loss when nuances are not captured by formal terms and relationships. Uncertainties are similarly difficult to formalise. In this sense, formal metadata works best for the representation of a subset of facets of practices and processes that all key stakeholders can agree upon.

Formal data documentation is also hampered by the broader limitations that are inherent to objectivisation of knowledge. All formalisations enact a

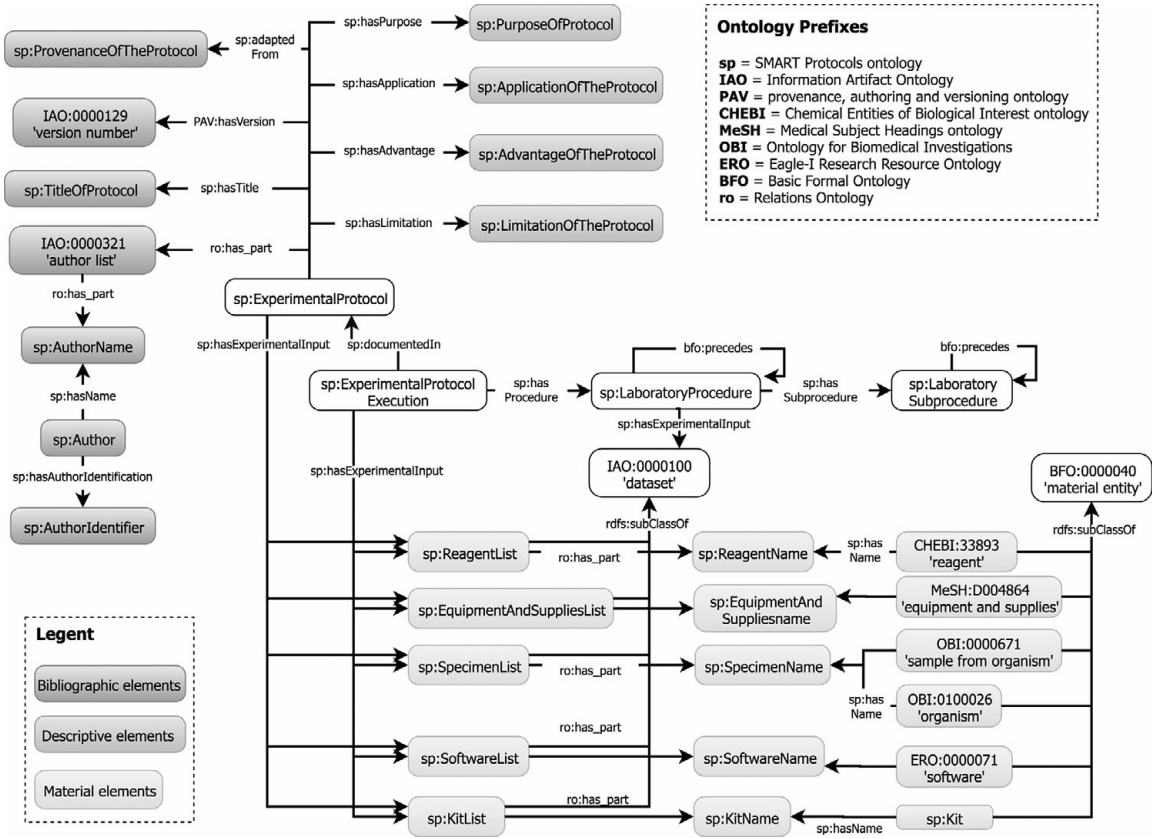

Figure 4.2 Hierarchical organisation of data elements in the SMART Protocols Ontology, sourced from Giraldo et al. 2018 (Image: CC BY 4.0 DEED Attribution 4.0 International).

multitude of cultural and discursive assumptions that are inherent and specific to particular domains and communities. Differences can be found between countries, research disciplines and professional communities but also between diverse sub-specialties within the same discipline. Different communities use different words to describe similar aspects of practices without necessarily being cognisant of their disparities.

For example, archaeologists interviewed in the CAPTURE project highlighted many examples of such differences. A prime example is the naming and temporal bounds of historical periods in different European countries. The term 'Middle Ages' might refer to a period of time that begins several centuries earlier in southern Europe compared to the Nordic countries. There are also parallel categorisation systems that are used independently by particular communities, sometimes without awareness of the existence of multiple classification methods. Many of the assumptions and discursive conventions are also invisible, unconscious and difficult to articulate for community members, leading to gaps in documentation and communication within data standards. A seemingly tidy surface can easily give a false impression of comprehensiveness. The danger of false security is, as Baird remarks, 'in the things the system does not know where to put or how to classify, or in the metadata it does not think to write' (Baird 2023, p. 19).

Another practical complication is that formal data documentation is resource-intensive and it is debatable to what extent all data and associated paradata need to be documented in a minute level of formal detail. Ontology development projects tend to require significant resources and multiyear research funding arrangements, involving a team of researchers with various areas of expertise, such as project managers, domain experts, software engineers and project scientists (e.g., Chao et al., 2015; Hyvönen, 2023; Oldman and Tanase, 2018).

At the same time, however, the development of templates for individual standards and reusable tool sets can make it easier for even a relatively non-technical researcher to utilise ontologies for paradata documentation. The Finnish Sampo-model (Hyvönen, 2023) and the long series of accordingly titled cultural heritage linked open data portals exemplify how an ontology toolkit can be deployed to facilitate work with a broad variety of documentation across multiple contexts. Because of the complications of applying formal metadata to document paradata, especially in cases when established standards are yet to be developed, it is advisable to consult research data and data documentation experts. However, when used mindfully and with care, controlled vocabularies and metadata standards can significantly enhance the transparency of research practices and processes for data publishing.

Key References and Further Reading

- Chao T. C., Cragin M. H. and Palmer C. L. (2015). Data Practices and Curation Vocabulary (DPCVocab): An empirically derived framework of scientific data practices and curatorial processes. *Journal of the Association for Information Science and Technology* 66(3), 616–633. https://doi.org/10.1002/asi.23184. The vocabulary can serve as lingua franca to facilitate communication among the stakeholders in data practices and curatorial processes.
- Nurmikko-Fuller T. (2023). *Linked Data for Digital Humanities,* London: Routledge. https://doi.org/10.4324/9781003197898. A monograph that introduces linked open data in the context of digital humanities, written in non-technical language.
- Zeng M. and Qin J. (2022). *Metadata, 3rd ed.* Chicago: ALA Neal-Schuman. An authoritative and clear introduction to metadata, ranging from fundamental concepts to metadata standards.

4.2.2 Narrative Descriptions

Narrative descriptions encompass written descriptions of connected events for data creation processes, including in the laboratory and field environments, diary keeping, marginalia and fieldnotes. Note taking may go beyond narratives and often involves recording individual data points and pieces of information. We return briefly to this aspect of field notes in the following section when discussing recording as a method of paradata generation.

In field sciences, notebooks and diaries have traditionally been the primary means of documenting both field observations and contextual information, including information on the fieldwork process that can be termed paradata (Canfield, 2011; Rytter et al., 2020). Since the turn of the millennium, digital-, audio- and video-based note taking using mobile phones and in some cases dedicated devices has become increasingly common. In the CAPTURE project, Kaiser has experimented with a prototype of an audio-to-text note-taking application for documenting archaeological fieldwork paradata. Note-taking and narrative descriptions are also prevalent in survey research (Edwards et al., 2017) and laboratory notebooks in natural and life science experiments (Rheinberger, 2023).

Narratives of practices and processes can be documented in various forms and formats, such as in annotations to data, research materials, diaries, research reports and publications, and notebooks. They can also be found in secondary documents and artefacts relating to a practice or process (e.g., Lin, 2021). In many disciplines, the published narratives of research processes tend to be

sparse. Methods sections published in research articles are often brief, focusing on essential content as dictated by author guidelines and disciplinary conventions. However, in disciplines like anthropology, it is conventional to produce extensive methodological reflections as a part of what Geertz (1973) famously termed 'thick description'. While usually only short passages of such descriptions find their way to research publications, book-length studies sometimes also allow researchers to write chapter-length descriptions of methods and the research process (e.g., chapter 4 in Smith, 2020).

Narrative descriptions are often combined with illustrations, diagrams, photographs and other types of content as part of a single document. For example, written notes and sketches recorded in notebooks and experiment sheets can be useful for documenting initial and work-in-progress findings and conceptualisations of data. Tracking how descriptions of findings and concepts evolve over time can help both note-takers and others follow the research process. Paradata recorded in lab notebooks and experiment logs facilitate the transition of data from the lab bench to journal articles, research reports etc., using tools such as lists, tables, curves and graphs for combining and summarising data points, observations and experimental results (Rheinberger, 2023).

In field sciences, note-taking is useful for accurately recording observations since the richness of contextual information captures the various conditions and contexts that shape field research experiences and findings (Emerson et al., 2011). It is recommended to promptly document detailed field notes after fieldwork, preferably within twenty-four hours (Williamson 2018). The intended audience matters as well. Working notes need only to be understandable for their author, while notes for external audiences should be written with their intended readers in mind. The language must be appropriate and understandable for an audience that includes both those who plan to reuse data and those who are merely searching for it. Further, it is important to write standalone descriptions with enough context and a clear structure (Phillips and Smit, 2021) to ensure that their readers do not need to find and access a lot of additional documentation to understand the descriptions.

In addition to the notes documenting field observations (field notes), other note-taking approaches, such as method notes (reflections on techniques used and their descriptions) and theory notes (ideas about the observed phenomena and their connections with the theoretical framework) are useful for capturing the data creation process (Chatman, 1992). Notably, the use of method notes reflects a growing trend towards reflexivity in social research by examining the researcher's influence on both the research process and its results (Goodwin et al., 2017).

Besides documenting the practical steps taken during a research process, paradata captured in fieldnotes offers researchers a means to introspectively reflect on their own biases and preconceptions throughout the research process, making them transparent to others. Ortlipp (2008) describes her use of reflective journals in documenting qualitative research providing both theoretical insights and practical advice. In archaeology, the reflexive diaries recorded during the Çatalhöyük project (in Anatolia) from the 1990s until 2010s in text and video (Sandoval, 2020), exemplify narrativising data creation process for increased reflexivity.

The relevance of narrative descriptions for documenting paradata lies in the deep embeddedness of narratives in how people understand and communicate their experiences. Dourish and Cruz (2018) emphasise that data is never self-explanatory; it needs to be narrated to give it shape and meaning. Various techniques of data storytelling have been developed to narrate data during processes of interpretation and meaning-making (Dykes, 2020; Knaflic, 2019; Matei, 2021). In this sense, narrative descriptions go beyond mere note-taking and function as thinking aids for their creators. They can also give paradata shape and meaning, mobilise it and help to put it to work. This applies both within the domain crafting the narrative and as a meta-story (Holtorf, 2020) of how the domain portrays itself to external audiences.

Key References and Further Reading

- Dourish P. and Cruz E. G. (2018). Datafication and data fiction: Narrating data and narrating with data. *Big Data & Society* 5(2), 1–10. https://doi.org/10.1177/2053951718784083. An article that discusses the use of narratives in data-driven analysis from the perspective of ethnographic practices.
- Emerson R. M., Fretz R. I. and Shaw L. L. (2011). *Writing Ethnographic Fieldnotes, 2nd ed*. Chicago: University of Chicago press. A manual that provides clear and detailed instructions on writing and processing fieldnotes in ethnographic research.
- Rheinberger H.-J. (2023). *A Phenomenology of Experimentation*. Chicago: University of Chicago Press. An insightful research monograph of experimentation and elements of experimental research including traces, models, grafting and note-taking.

4.2.3 Recordings

Besides developing narrative descriptions, practices and processes can also be captured concurrently through various means, such as photographs, audio, video and 3D data capture. The forms of paradata generation share the

common feature of active and purposeful generation of a real-time 'record' of a practice or process. In this sense they can be termed *recordings*. These types of recordings can also be captured using text, either in narrative form or as a series of notes or data points recorded either hand-written on paper or pro forma sheets or digitally in a database.

For instance, during fieldwork documentation, anthropologists take photographs and make films to document their interactions with study participants, cultural artefacts, and the local environment. These photographs and films provide insights into the research process and help interpret the data collected for both data makers and reusers.

When rich enough, they can provide a 'thick depiction', akin to a thick description (Hann, 2021), incorporating multiple levels of interpretations and documentation of the subject of study *and* the study process in one. Such recordings can help to disclose at the best minute details of both data generation practices and processes and their underpinning theories and ideologies. Investigating the colonial legacy in anthropological audiovisual materials, for example, has revealed inherent Western biases and colonial power dynamics by analysing their inception and aesthetics (Giglitto et al. 2023).

As such, recordings provide contextual information and paradata about the research setting and including the environment, interactions between researchers and participants, and non-verbal cues that may not be captured in written notes alone. Whenever audio or video recording does not lead to ethical dilemmas regarding the protection of the privacy and security of involved individuals or groups, they are effective options for generating rich descriptions of practices and processes.

Recordings can document the procedures followed during data collection, including interview techniques, experimental protocols and observational methods. For example, in qualitative studies using interview techniques, audio and video recordings are typically used (Mason, 2018). Researchers can cross-reference interview transcripts and analyses with the recordings to confirm their accuracy. In quantitative studies based on surveys and experiments, recordings are used to ensure that the experiment protocols are followed closely and in interview research, that the interviewer does not deviate from the interview guidelines (Kunz et al., 2024). In observational studies, recordings can capture real-time behaviour, interactions and events in naturalistic settings. The advantage of recording is the ability to collect data that might be impossible or difficult to capture otherwise due to time constraints and the need to engage in concurrent activities. Recordings serving as paradata for documentation of procedures thus contribute both to transparency and replicability in research and in general, of practices and processes.

There is, however, another layer to add. As digital recording devices become increasingly accessible, there is growing recognition of the importance of paradata about recordings. For example, live broadcasting of theatrical events leads to complex documentation processes, in which para-documents (i.e., documents related to a play beyond its core content, such as audience reactions to the primary text), have enriched the impact of a single performance (Abbott and Read, 2017).

One limitation of this approach is the complexity of recordings, which may require the creation and maintenance of paradata specifically for the recording process. This can result in a proliferation of documentation (including documentation of documentation), posing challenges for the management of data and paradata (see Dawson and Reilly, 2019 for reference).

In cultural heritage contexts, 3D scanning methodologies and technologies are increasingly recognised as valuable tools for artefact documentation (Homburg et al. 2021). For example, since 3D scanning accurately captures the exact dimensions and intricate surface details of artefacts, enhanced 3D representations of coins can be achieved by integrating fine photometric details from photographic images with precise geometric data from a 3D laser scanner (MacDonald et al., 2017). The accuracy of such data is high, exceeding the current possibilities of generating comparable information from photographs using photogrammetry, a technique for obtaining information on the physical properties of objects depicted in photographs and video. From paradata perspective, the advantage of both 3D scanning and photogrammetry is not the accurate representation of artefacts per se, but rather the possibility to document a process, for example, the progress of an archaeological excavation or change in natural landscapes, using a series of scans undertaken at different points in time.

In addition to the potential overall quality and accuracy of 3D scanning outputs, capturing paradata in 3D recording practices is important as they provide context and support the reasoning process behind the creation and interpretation of final 3D visualisations (Demetrescu et al., 2023; Opgenhaffen 2022). Further, 3D models have also been proposed as a potentially fruitful approach to knowledge integration, comparable to knowledge graphs, by providing an interface to different forms of knowledge pertaining to both physical and abstract entities.

As interpretative representations, 3D recordings offer a form of knowledge production distinct from those based on text and linear one- or two-dimensional narratives (Derudas, 2021; Sullivan, 2020). Multiple examples of prototypes exist that aim to help archaeologists envision and theorise how different physical elements and multisensory considerations recorded in 3D

models may have influenced the sensory experience of a particular archaeological site. Viewers can interact with the data via metadata accessible through the online 3D browser, as well as through documentary metadata and paradata provided in the model's comprehensive publication (Sullivan, 2020, 2023).

Overall, paradata documentation through recording can enhance the transparency and replicability of practices and processes and support data interpretation. A major disadvantage is the large amounts of data generated, which can be difficult to manage and interpret, as evidenced by the video diaries recorded at the Çatalhöyük excavation (Sandoval, 2020). The information density of recordings is not necessarily high and it can take a lot of time to find relevant evidence. Some of these problems can be alleviated by careful planning of what, how and when to record, and by documenting recordings for searchability and findability. The retrieval and summarisation of information from recordings can also be facilitated by artificial intelligence techniques, though there are likely to be limits in the level of detail and precision that automation can achieve in interpretative tasks.

Ethical considerations must also be taken into account before recording everything, since recording can interfere with both legal and ethical bounds of individual privacy. This includes both those who intentionally recorded and those incidentally present when practices and processes are recorded. Recordings can also be easily misused for surveillance and evaluation of individuals' work even if not originally intended for such purposes.

Despite these challenges, recording – particularly when guided by a careful documentation strategy – offers a powerful method for enhancing the comprehensiveness of paradata. Even if a recording is never completely raw, as it is underpinned by multiple choices of what, how and when to record, it provides opportunities to capture aspects of practices or processes in real-time rather than planning them in advance or narrating them afterwards.

Key References and Further Reading

- Sant, T. (ed.) (2017). *Documenting Performance: The Context and Processes of Digital Curation and Archiving*. London; New York: Bloomsbury Methuen Drama. A broad collection of papers addressing the issues of documenting processes in drama and performance studies highlighting many pertinent issues of the documentation of practices and processes independent of domain and context.
- Opgenhaffen L. (2022). Archives in action: The impact of digital technology on archaeological recording strategies and ensuing open research archives. *Digital Applications in Archaeology and Cultural Heritage* 27, e00231. https://doi.org/10.1016/j.daach.2022.e00231. A journal article featuring a

detailed account of recording research processes that emphasises the need for transparency in the digital recording process, advocating for a thorough documentation of the decisions and techniques used in creating 3D models.

4.2.4 Logging

Logging is closely affiliated to recording as a method for generating paradata. Log files (also known as system logs) are documents created in real time during an on-going practice or process. Unlike recordings, which result from deliberate acts of recording, logging is automatic and generated by registering events and actions within a computer system or software application. As researchers extensively utilise digital devices and applications for data collection, processing and analysis, log files provide a straightforward method for automatically collecting evidence of these processes.

Since the advent of paradata during the data collection process, computer-assisted social survey research has received more attention from survey methodologists (Durrant and Kreuter, 2013). For example, survey researchers using computer-assisted personal interview (CAPI) software programs for data collection can automatically log numerous parameters related to interviews, such as time spent on each question, keystrokes, types of events and the person's role in the study. CAPI has been particularly useful for helping researchers manage fieldwork activities by collecting the paradata of timestamps, GPS (Global Positioning System) coordinates and interviewer characteristics. For example, the collection and analysis of paradata of interviewers' movements in the field using automatically logged GPS (Global Positioning System) coordinates can ensure that sampling protocols are followed correctly (Choumert-Nkolo et al., 2019).

Additionally, the collection and analysis of such paradata as response times in web-based surveys can reveal respondents' difficulties of understanding individual questions asked or the effort they invest in taking the survey (Kunz et al. 2024). Paradata of mouse movements have also been used to predict question difficulty in online surveys by taking into account individual differences in mouse-tracking measures, though there is some room for improvement in accuracy with this technique (Fernández-Fontelo et al., 2023). Incorporating such paradata into survey research not only enhances the integrity of data collection processes but also provides insights into respondent behaviour and fieldwork management.

Besides survey research, logs can generate practice and process data also in various other contexts and domains. Logging is currently being tested for capturing interactions with large language models (e.g., Trippas et al., 2024) and in the field called Robotic Process Automation, to log work processes in

minute detail (Fani Sani et al., 2023). In scientific and scholarly field research, many digital measuring devices from digital cameras to GPS units, log significant amounts of information besides primary photographic or spatial data. For example, photographs shared on social media platforms like Instagram can be used to study the everyday experiences and sensory perceptions of participants in the field by asking them to take pictures and posing them short questions about their feelings and experiences relating to the topics of the photographs (Shortt and Warren, 2020). In education research, the log files from large-scale cognitive assessments of adult compentencies have been analysed to extract process indicators of test-taking, such as total time on task, time to first action, and the number of interactions, to infer the underlying cognitive processes (Goldhammer et al., 2020). Logging extends to data analysis in virtual research environments that help to collect detailed data on minute steps of data management and use (Bentkowska-Kafel et al., 2012; Sant, 2017).

One of the major challenges with logging is to ensure the coherence of logged information. Blockchain technology provides a robust and secure framework for ensuring the integrity of the logged activities by algorithmically guaranteeing the immutability of logged information and the transparency of activities visible to all relevant parties on a decentralised network (Swan 2015). Envisioned as an alternative to having a trusted third-party to guarantee the practical irreversibility of financial transactions – that a payment once made cannot be undone – blockchain allows what Lemieux describes as trustless trust. While blockchain really makes erasing once recorded information of transactions impractical rather than completely impossible, it provides a method to produce a trustworthy record of consequent activities that stands on its own without an institution or individual to guarantee its integrity (Lemieux, 2022). Since every transaction is registered in a block that connects to prior transactions, a sequential, immutable chain is created. A major benefit of blockchain is how it can be utilised for maintaining the integrity of log files, or paradata in general. All steps of a data creation, management or use are registered and cannot be altered afterward.

Moreover, the blockchain itself incorporates paradata through supplementary or metadata associated with blockchain transactions and operations, including contextual information about the transactions registered on the blockchain, such as timestamps, transaction metadata and participant identities. So far, blockchain technology has been used for diverse purposes ranging from safeguarding patient privacy and data security by storing and sharing 3D augmented reality surgical navigation data through peer-to-peer decentralised technology (Batchu et al., 2023) to ensuring the integrity of archival records (Lemieux 2019). However, the effectiveness and trustworthiness of blockchain systems heavily

rely on comprehensive and accurate documentation, as it is often challenging to ascertain the presence, type and location of records within these systems (Lemieux, 2022).

Logging shares many benefits and concerns with recording regarding its usefulness for generating paradata. The primary benefit is its automation, which requires no effort from data creators. This leaves room for directing the conscious human effort to documentation tasks that are difficult or impossible to automatise.

However, similarly to recording, there are ethical issues related to logging practices and processes and keeping logs, especially due to the invisibility of paradata generation. Another challenge with retrieving paradata from log files is that the log files themselves are seldom self-explanatory. Depending on the log, extensive contextual information on both the device or system and its use may be necessary for the logs to make sense to their eventual users. We will return to this final question later in Chapter 5 of this volume when discussing how to use quantitative methods to backtrack past practices and processes.

Key References and Further Reading

- Kunz T., Daikeler J. and Ackermann-Piek D. (2024) Interviewer-observed paradata in mixed-mode and innovative data collection. *International Journal of Market Research* 66(1), 14–26. https://doi.org/10.1177/14707853231184742. A journal article introduces the interviewer-observed paradata in mixed-mode data collection methods.
- Lemieux V. L. (2022) *Searching for Trust: Blockchain Technology in an Age of Disinformation*. Cambridge: Cambridge University Press. A book that discusses the relation of record-keeping, blockchain technologies and trust and emphasises the need for thorough record-keeping and associated documentation and transparency in blockchain systems as a premise to establish and maintain the authenticity of archival records.

4.2.5 Research Plans

In the preceding sections, we have delved into methods for generating paradata either during or directly following practice or process. An alternative approach is to produce documentation in advance. Juneström and Huvila's analysis suggests that this approach can be used either to delineate potential future activities or to prescribe them in advance.

One of the most common approaches to prescribing and prospectively describing practices and processes is by planning and producing corresponding documents, that is, different types of plans. To illustrate plans and their

potential function as sources of paradata, this section takes a closer look at data management plans, registered reports, experimental protocols and clinical trial registries.

Data management plans (DMPs), as a type of research plan, are useful for prospectively generating paradata on data-related practices and processes. They provide a structured framework and promote standardised documentation practices to support data sharing and reuse. DMPs serve as structured descriptions of activities that are expected to be followed to reach a particular outcome in research data management, guiding researchers in planning their work.

The increasing use of DMPs has been influenced by funding agencies seeking to enhance transparency of research and promote the sharing and reuse of research data (Smale et al. 2020). It has been posited that to make research data findable, accessible, interoperable and re-usable (FAIR), a well-constructed DMP should describe research data management procedures planned for an entire research project, with particular attention to the collection, processing and generation of data, applied methodologies and standards, data sharing and open access, and data curation and preservation, with a guideline and template to follow.[14]

The effectiveness of DMPs for researchers can be hampered by stakeholder tensions and the generic nature of templates. If aligned with researchers' paradata needs and discipline-specific norms and data practices, they have the potential to be useful for both their creators and the reusers of documented datasets (Kvale and Pharo, 2020; Smale et al., 2020). In contrast, if reduced to mere administrative paperwork, their value is likely to remain questionable. Overall, if implemented properly and aligned to support the planning of relevant aspects of data creation, management and use, DMPs can function as useful devices for eliciting prospective paradata, which in turn can improve transparency of data creation, management and use practices to promote the sharing and reuse of data.

As an alternative form of research plan, registered reports provide detailed plans for a research study, subjected to peer review and publicly registered before execution. These reports outline research questions, hypotheses, methodology, data collection methods and data analysis techniques (Nosek and Lakens, 2014). Registering a research protocol in advance can assist researchers adhere to a predetermined procedure, thus resulting in more accurate documentation of the process compared to a retrospective description (Huvila and Sinnamon, 2022). To improve the computational reproducibility

[14] Guidelines on FAIR Data Management in Horizon 2020 http://dx.doi.org/10.25607/OBP-774

of registered reports in statistical data analysis, recommended practice is to include a codebook in data files, annotating and structuring the code for clarity, ensuring reproducibility of codes post-revisions, and listing required software packages and versions (Obels et al., 2020). Pre-registered reports are a useful starting point for replicating experimental studies and comparing research findings across studies. Including a codebook in data files and associated paradata in procedures of data analysis can improve computational reproducibility of shared data, thus enhancing methodological transparency and data sharing and reuse.

Experimental protocols are a related approach to pre-registered reports used in experimental research, with the goal of functioning as a recipe for running an experiment. Their specifics and level of detail can vary by laboratories and publications even if they are expected to follow discipline-specific norms (Giraldo et al., 2018). They are expected to provide a description that is sufficiently thorough to give enough information for an external colleague to replicate an experiment. Similarly to registered reports, their prominent aim is to improve transparency and reproducibility of research by prompting data creators to describe their procedures prior to data generation and mobilsing research findings from the laboratory bench to the research publication (Rheinberger, 2023). Their major advantage is in their potential to reduce the number of unplanned ad hoc changes to plans that do not end up being documented. Their principal drawback is that they reduce the flexibility of research work thus making them less suitable for qualitative and exploratory research based on the rationale of adapting data generation methods to the evolving research situation.

As a final example of a plan, clinical trial registries are publicly accessible online databases that provide access to information regarding clinical trials prior to their initiation. As an alternative form of research plans, their focus is on documenting details of clinical trials, including study descriptions, participation criteria and study plans (experimental design and outcome measures).

The purpose of trial registries is to disseminate information concerning clinical trial research, thereby contributing to enhancing the transparency and quality of trials. Registries are typically specific to countries or regions, such as ClinicalTrials.gov, provided by the National Library of Medicine (NLM) in the USA, the Australian New Zealand Clinical Trials Registry (ANZCTR) (www.anzctr.org.au/), and the European Union Clinical Trials Register (currently transitioning to the Clinical Trials Information System, CTIS) (https://euclinicaltrials.eu/search-for-clinical-trials/?lang=en) for trials conducted in the European Union (EU) and European Economic Area (EEA). Additionally, the

World Health Organization (WHO) manages the International Clinical Trials Registry Platform (ICTRP) (https://trialsearch.who.int/), which serves as an aggregator of registries worldwide. Like registered reports, a registration process is in place. However, there is generally no peer-reviewing process and the policies of whether the information is entered by investigators and research sponsors or national authorities with authorisations and ethics reviews included depend on the registry.

The clinical trial registries are useful resources for other researchers conducting meta-analysis studies, developing healthcare guidelines or seeking collaborative research opportunities (Liu et al., 2023). Clinical trial registries offer valuable access to pre-initiation information about trials, including paradata about the various aspects of the trial process, which in turn helps to enhance the methodological transparency on trial studies. As a highly specific and resource-intensive approach to documenting planned research, the approach lacks transferability to domains where a comparable level of regulation and formalisation of data generation is not feasible. At the same time, they show how planning can be a highly effective method of generating detailed paradata to stipulate forthcoming data creation.

Key References and Further Reading

- DeVito N. J., Morley J., Smith J. A., Drysdale H., Goldacre B. and Heneghan C. (2024) Availability of results of clinical trials registered on EU Clinical Trials Register: Cross sectional audit study. *BMJ Medicine* 3(1). https://doi.org/10.1136/bmjmed-2023-000738. A study that describes how the European Union Clinical Trials Register (EUCTR) functions as a repository for accessing unique trial results and can support literature searching for systematic review studies.
- Gajbe S. B., Tiwari A., Gopalji and Singh R. K. (2021) Evaluation and analysis of Data Management Plan tools: A parametric approach. *Information Processing & Management* 58(3), 102480. https://doi.org/10.1016/j.ipm.2020.102480. This article provides a comprehensive review of data management tools and guides the selection of tools that best suit the researchers' needs.
- Nosek B. A. and Lakens D. (2014) Registered reports: A method to increase the credibility of published results. *Social Psychology* 45(3), 137–141. https://doi.org/10.1027/1864-9335/a000192. This article provides an overview of the concept of registered reports and discusses their rationales and use for enhancing the transparency and credibility of experiments in social psychology.

4.2.6 Prospective Workflows

In addition to (research) plans that vary in their degree of formality, prospective workflows represent a future-oriented technique for describing and, often in parallel, prescribing planned practices and processes. Similar to plans, this approach is useful for generating prospective or potential paradata (cf. Chapter 6) – documentation of forthcoming activities that ultimately becomes paradata when the practice or process is enacted.

In the literature on procedural workflows, a workflow refers to the series of tasks or steps involved in completing a particular process or achieving a specific goal. Workflows provide a structured approach to organising and executing work efficiently to accomplish desired outcomes. They are particularly popular in contexts incorporating repetitive tasks that need to be executed repeatedly in the same order, such as consecutive steps in scientific experiments, industry and computational tasks.

Workflow-based approaches systematically outline the steps to accomplish specific tasks and detail how individuals and automated processes and practices should be executed to achieve a specific goal. In IT, computational workflows are usually executable, containing all necessary information to carry out and complete a task as a whole. In contrast, human workflows often document only key steps of a workflow, omitting details deemed unnecessary for a human executing the task.

Figure 4.3 shows a visual representation of a workflow diagram (Activity Diagram) that can be converted to machine-readable code, for example, in Unified Modelling Language (UML). Workflow diagrams are formal diagrammatic models that aim to provide comprehensive documentation of a process. The diagram can be visualised to facilitate the recording, sharing and explanation of protocols used in generating results, selected outcomes and summarising courses of action (Blaise and Dudek, 2023).

Likewise, to make the research results easier to verify, the workflow of managing research data can be semi-automated in a workflow management system to meet specified external and internal requirements of the documentation of data (Miksa et al., 2021). However, to direct the actual workflow – that is, how tasks are executed – the adoption of a workflow management system must be embedded in the daily work practice of its users, requiring both engagement and resources.

Workflows are frequently depicted using machine-readable diagrams to enhance their discoverability by automated systems (Weigel et al., 2020). The surge in digital data processing and analysis has led to an increasing demand for adequate documentation of computational workflows. Supporting

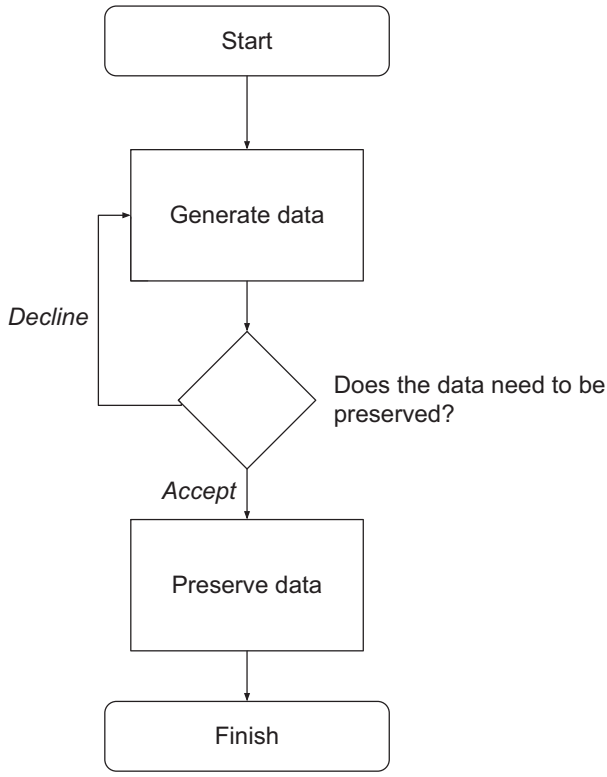

Figure 4.3 A workflow diagram (Activity Diagram) of a simple data generation and preservation workflow.

reproducible scientific workflows in Data Science, a web-based interactive computing platform like Jupyter Notebook (Figure 4.4) enables users to produce annotations by integrating live code, equations, text and media directly in a document that contains both the code and the documentation. It supports the use of various programming languages. The executable workflows of computer codes, with outputs and annotations as interactive documentation, allow data creators to efficiently create reproducible computational workflows. The system incorporates a prompt display of output and the capability to identify necessary documentation updates following alterations to the user interface or algorithms (Beg et al., 2021; Mendez et al., 2019). However, despite the advantages of using dedicated tools, the reproducibility rates sometimes remain low since the documentation and execution of tasks do not necessarily follow the existing guidelines and best practices (Pimentel et al., 2021).

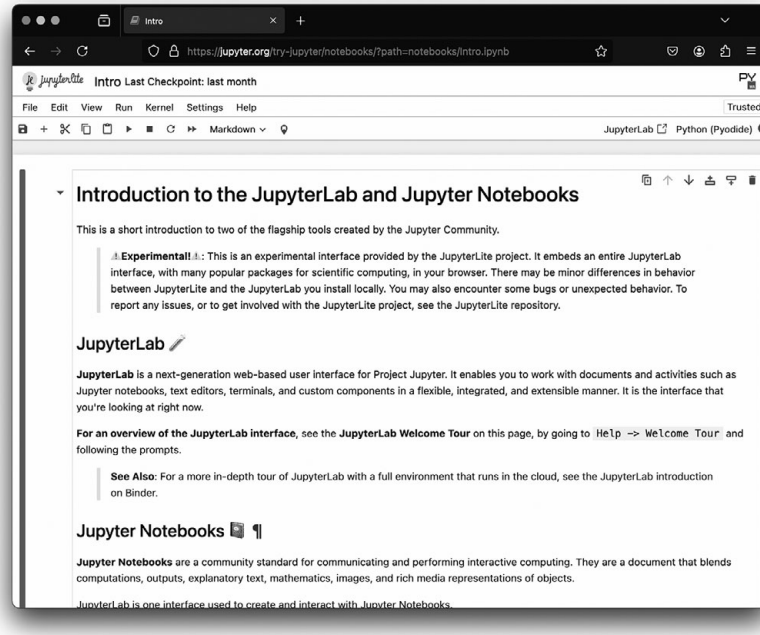

Figure 4.4 Jupyter Notebook (https://jupyter.org/try-jupyter/notebooks/?path=notebooks/Intro.ipynb).

To produce more usable and reproducible research data, Thomer et al. (2018) proposed a method of Research Process Modeling for documenting non-computational data provenance in geobiology fieldwork. This method consists of two inventories and two workflows: 1) an Activity Diagram (see above); 2) an artefact inventory documenting all digital and physical artefacts (including filenames, information on their creating and using processes, related artefacts, class and format); 3) a process inventory with information on involved processes (including title, description, agents, preconditions, inputs and outputs); and 4) a provenance graph modelled according to the PROV specification (PROV-Overview 2013). Research Process Modelling enables researchers to assess their work in relation to their planned activities, providing a means to prescribe and document research activities and artefacts to facilitate future reuse. Related workflow-based approaches to paradata generation have been targeted to specific tasks in other domains as well, including data harmonisation in survey research (Kołczyńska, 2022), life sciences (Cohen-Boulakia et al., 2017), and bioinformatics (Oinn et al., 2004).

Overall, workflow-based approaches offer a practical method for generating potential paradata in advance. This enhances the transparency of data creation processes before data is created, standardises data-related practices and processes, and facilitates data reuse. However, especially when applied to complicated practices and processes, workflows tend to become increasingly complex, and thus difficult to implement and manage. Even if the execution of an existing workflow intuitively might sound like a trivial task, capturing paradata relating to running a workflow for validating its correctness and performance remains a major concern. An illustrative context where this has become increasingly apparent are the GLAM (Galleries, Libraries, Archives and Museums) institutions that struggle with publishing their digital collections (Candela et al., 2023).

A related problem is the integrity of workflows and their associated paradata (Hoopes et al. 2022). For many data generation contexts, especially in social research, the relevance and even possibility of modelling tasks on a strict step-by-step basis remains an open question. However, even if the workflow remains an incomplete simplification, it can still be helpful in directing attention to key steps of practices and processes to document, even if the aim of paradata production is not to generate a comprehensive step-by-step model.

Key References and Further Reading

Kvale L. and Pharo N. (2020) Understanding the data management plan as a boundary object through a multi-stakeholder perspective. *International Journal of Digital Curation* 15(1), 16. https://doi.org/10.2218/ijdc.v15i1.729. A journal article that provides insights into the perceived usefulness of data management plans (DMPs) from the perspectives of their different stakeholders.

Nosek B. A. and Lakens D. (2014) Registered reports: A method to increase the credibility of published results. *Social Psychology* 45(3), 137–141. https://doi.org/10.1027/1864-9335/a000192. An article that discusses rationales and the use of registered reports in experimental research.

International Clinical Trials Registry Platform (ICTRP) (https://www.who.int/clinical-trials-registry-platform) A knowledge base of trial studies across clinical trials registeries.

Thomer A. K., Wickett K. M., Baker K. S., Fouke B. W. and Palmer C. L. (2018) Documenting provenance in noncomputational workflows: Research process models based on geobiology fieldwork in Yellowstone National Park. *Journal of the Association for Information Science and Technology* 69(10), 1234–1245. https://doi.org/10.1002/asi.24039.

A research study that describes an approach to document provenance in non-computational workflows in geobiology fieldwork.

4.3 Discussion

An overview of a selection of potential approaches for generating and documenting paradata – that is, information on practices and processes – sourced from the literature reveals a plethora of usable methods. However, the overview also shows that since the methods introduced in this chapter are primarily designed for data creators, there is a tendency to stipulate the generation of paradata by following specific procedures and standardised formats.

Narrative descriptions represent a more *post hoc* approach that emphasises documenting practices and processes as they are experienced and observed. Recording and logging, on the other hand, generate paradata in the moment, at least in theory. In practice, however, both recording and logging are shaped by the technologies employed to capture paradata and the underlying concept of relevant information that guides what is documented.

A data management plan is designed with guidelines and templates for researchers to follow. Registered reports and computational workflow methods aim to provide comprehensive documentation of data creation processes by adhering to predefined procedures. Since the formal metadata-based methods are concerned with the standardisation of data for data sharing and exchange, their use requires systematicity and compliance to standards. Data creators need to be aware of relevant frameworks for data publishing, such as the 5 Star Linked Open Data (Berners-Lee 2009) model but also be aware of the limitations and consequences of formalising descriptions, the risk of data loss and the impact of their underlying assumptions and perspectives to data documentation.

One of the main drivers for working with paradata is to contribute significantly to the documentation of contextual information (Huvila, 2022). However, effective means to this end are not yet well developed or commonly used. It appears that as a research approach becomes sensitive to situation and context, it also becomes less structured either regarding the specifics of generated paradata, its structure or both.

Apart from narratives and recordings, there are some proposals for capturing contextual information, including paradata, about the data, using a template for data summary. Philips and Smit's guidelines for unstructured descriptions of datasets emphasise the importance of presenting the dataset as a research output and elucidating the context in which the data was generated (Philips

and Smit, 2021). Koesten and colleagues' (2020) dataset summary template is another example of an approach that provides a structured framework for generating meaningful textual representations of data (Koesten et al., 2020).

There are also proposals of reporting standards and guidelines for experimental protocols and data analysis procedures in a checklist format (Considine and Salek, 2019; Giraldo et al., 2018). Such checklists also include the large set of reporting standards developed for enumerating various facets of various types of study designs, data collection and analysis methods in publication. The EQUATOR network[15] (2008-) guidelines in the health domain are a prominent example that are also partially applicable to other domains, although careful consideration must be applied to their suitability when applying them and appropriate tailoring may be necessary.

Collecting best practices across research contexts can be helpful. Proven methods for documenting practices and processes offer guidance to improve the transparency of data creation processes, though they may be difficult or impossible to integrate into the research process if the best practices emerge from far off domains. Another challenge, not to be taken lightly and necessary to solve on a domain to domain basis, requires data creators and users to work together with real data creation and use cases and scenarios to determine precisely what information needs to be documented.

Given the contextual and situational nature of paradata and paradata generation, rather than aiming at using a single approach for generating and documenting paradata, using a combination of the methods introduced in this chapter can be helpful in broadening the scope of generated documentation. However, this should be done thoughtfully, considering what kind of information-specific methods would optimally provide the relevant amount and breadth of documentation.

The diffusion of paradata across data and data documentation suggests that a promising approach to limiting the excess of paradata generation and simultaneously enabling and improving the use and usefulness of existing information is to move towards more integrated interlinking of data and diverse forms of data documentation and secondary information resources. The examples from using 3D documentation as a centrepiece that provides an interface to all relevant information relating to a physical object are taking steps to this direction. Also, Mosconi et al. (2022) who argue that the integration of a narrative layer in data curation can capture the contextual and cultural nuances necessary for qualitative research, are essentially suggesting to package data and method description

[15] www.equator-network.org.

together. This approach, even if developed for science education, has wider applicability, featuring the integration of the metadata standards and promoting ongoing curation activities within daily workflows.

A growing number of authors advocate for shifting paradata generation from retrospective documentation to planning and creating documentation during and, where possible, before practices or processes are enacted. The combined use of registered reports and a narrative description of the data creation processes can assist researchers with following a predetermined procedure, leading to more precise documentation of the processes compared to a retrospective account (Huvila and Sinnamon, 2022). The Research Process Modelling method (Thomer et al., 2018) takes this approach. However, it is important to weigh the benefits of advance planning against the risk of rigidifying practices and procedures, making them less agile and flexible to context and situation.

Broadening the scope of paradata generation also involves embracing multiple forms and formats of paradata. Text and visualisations are not merely different approaches to mapping knowledge of practices and processes but also produce it differently, ideally complementing each other (cf. Schwandt, 2022; Vancisin et al., 2023). Keenan and Walker (2017) provide an example of combining different modes of representation to document data which can be applicable to the processing of paradata. The documentation of a research dataset of seismic data collected during a survey project in 1970 preserved in the University of Montana institutional repository consisted of narrative descriptions, formal Dublin Core metadata, and primary datafiles. The librarians responsible for the work also considered how to organise the datafiles and documentation in the repository and how to make it accessible for its intended users, including how to address possible hindrances caused by individuals' functional variation, for example, lack of eyesight or hearing impairment. There are many benefits of planning ahead and producing narrative descriptions during processes of interpretation and meaning-making for capturing ongoing research activities, in combination with formal metadata-based methods, such as metadata standards and ontology, for more precise documentation of data provenance.

4.4 Conclusions

The methods discussed in this chapter provide guidance for data creators seeking to capture and document paradata effectively during data generation processes. The distinction between prospective and in-situ methods provides a framework

for understanding when and how paradata is generated in relation to the activities being documented. These methods range from formal metadata schemas and structured planning approaches such as data management plans and registered reports, to workflow-based techniques and narrative descriptions.

Formal metadata methods, including metadata standards, label sets and controlled vocabularies, play a critical role in ensuring consistency, discoverability and interoperability of data across diverse domains. However, their implementation may pose challenges for researchers lacking specialised technical expertise and resources, due to their inflexibility and inherent assumptions.

Narrative descriptions serve as rich sources of paradata, capturing contextual details, insights and reflections throughout the data creation process. From field notes in qualitative research to laboratory protocols in life sciences, narratives provide opportunities for conveying a nuanced understanding of data generation practices and interpretations.

Recordings, including images, audio and video, offer tangible documentation of processes, enabling researchers to visualise and analyse activities in detail. Logging methods, such as log files and blockchain technology, automate the documentation of events and actions within computer systems, enhancing transparency and security in data collection and analysis processes. While log files offer detailed records of system activities, techniques like the blockchain can help to secure the immutability and integrity of data transactions, particularly in sensitive domains like medical research.

Rather than documenting on-going or past practices and processes, it is also possible to prospectively generate descriptions to guide forthcoming work. The major advantage of this approach is that a prospective protocol or plan guides the practice in advance, helping to increase the consistency of practices and processes and minimising problematic ad hoc measures that may not be adequately documented.

Another advantage of prospective paradata generation is that in-situ and retrospective documentation might overlook crucial steps and measures. If paradata generation on-the-fly might increase workload by being a secondary undertaking to the documented practice or process, the shortfall of retrospective documentation is the difficulty of remembering what actually happened. However, while such prospective methods as protocols or data management plans offer guidelines and templates for researchers to follow in processing and managing research data, their effectiveness can be limited by varying stakeholder interests and tensions.

As the results of a survey study conducted in the CAPTURE project indicate, both data creators and users find value in documenting closely related aspects of practices and processes. However, their perception of what

constitutes informative data varies (Huvila et al., 2024; cf. Chapters 2 and 3). Similarly, while registered reports aim to improve transparency and credibility by registering detailed study plans before data collection, their practical implementation can be challenging to align with researchers' needs during the research process. A plan should not by default restrict the execution of the planned practice or process. In research, an even more important aspect of planning is that a plan should not constrain the thinking of the data creator, manager or user, leading them to assume that the planned practice or process is the only conceivable option.

This also applies to workflows. Workflow-based approaches, such as diagrams and information visualisation systems, offer systematic ways to document processes and facilitate reproducibility in data analysis tasks. At the same time, there is a risk that these approaches might impose rigid practices and processes that are less desirable in contexts where the goal is understanding rather than reproducibility.

Overall, the selection and application of methods for generating and documenting paradata should be tailored to the specific needs and constraints of individual research contexts. There is no universal method or approach that suits all domains and situations.

However, even if getting the right paradata might still be a wicked problem without apparent solution (cf. Huvila 2022), there are a lot of means to improve the transparency, reproducibility and credibility of the practices and processes of data generation, management and use. Rather than assuming that one approach or type of paradata would be enough, it is necessary to knit together an array of approaches that are contextually and situationally appropriate for the task and together provide enough information. A mindful paradata creator formulates a *paradata finding aid* (as discussed in Chapter 6) that incorporates a map of the methods and generated paradata to facilitate both paradata discovery and its future use.

As we will explore in the next chapter, incomplete documentation before and during a practice or process takes place is not necessarily fatal. A lot of paradata can also be identified and generated retroactively even if it is not explicitly documented as paradata by anyone when data was created, managed or previously used.

References

Abbott D. and Read C. (2017) Paradocumentation and NT Live's 'CumberHamlet'. In Sant, T. (ed.), Documenting Performance. London: Bloomsbury Methuen Drama, 165–187. http://radar.gsa.ac.uk/5068/.

Baird, J. (2023). Unclassified: Structured silences in the archaeological archive. In Raja, R. (ed.), Shaping Archaeological Archives Dialogues between Fieldwork, Museum Collections, and Private Archives, 19–32. Turnhout: Brepols.

Batchu S., Diaz M. J., Ladehoff L., Root K. and Lucke-Wold B. (2023) Utilizing the Ethereum blockchain for retrieving and archiving augmented reality surgical navigation data. *Exploration of Drug Science* 1(1), 55–63. https://doi.org/10.37349/eds.2023.00005.

Beg M., Taka J., Kluyver T., Konovalov A., Ragan-Kelley M., Thiéry N. M. and Fangohr H. (2021) Using Jupyter for reproducible scientific workflows. *Computing in Science & Engineering* 23(2), 36–46. https://doi.org/10.1109/MCSE.2021.3052101.

Bentkowska-Kafel A., Denard H. and Baker D. (2012) Paradata and Transparency in Virtual Heritage Paradata and Transparency in Virtual Heritage. Farnham: Ashgate.

Berners-Lee, T. (2009). Is your linked open data 5 star? Retrieved from www.w3.org/DesignIssues/LinkedData.html.

Berners-Lee T. and Hendler J. (2001) Publishing on the semantic web. *Nature* 410(6832), 1023–1024. https://doi.org/10.1038/35074206.

Blaise, Jean-Yves and Dudek, Iwona (2023) Research workflows, paradata, and information visualisation: feedback on an exploratory integration of issues and practices: MEMORIA IS. *Peer Community in Archaeology*. https://doi.org/10.5281/zenodo.8311129.

Börjesson, L., Huvila, I. and Sköld, O. (2022). Information needs on research data creation. *Information Research*, 27(Special Issue), https://doi.org/10.47989/irisic2208.

Bruseker, G., Carboni, N. and Guillem, A. (2017). Cultural heritage data management: The role of formal ontology and CIDOC CRM. In Vincent, M. L., López-Menchero Bendicho, V. M., Ioannides, M. and Levy, T. E. (eds.), Heritage and Archaeology in the Digital Age: Acquisition, Curation, and Dissemination of Spatial Cultural Heritage Data, 93–131. Cham: Springer.

Canfield M. R (2011) Field Notes on Science and Nature. Harvard University Press.

Chao T. C., Cragin M. H. and Palmer C. L. (2015) Data practices and curation vocabulary (DPCVocab): An empirically derived framework of scientific data practices and curatorial processes. *Journal of the Association for Information Science and Technology* 66(3), 616–633. https://doi.org/10.1002/asi.23184.

Chatman E. A (1992) The Information World of Retired Women. Bloomsbury Academic.

Choumert-Nkolo J., Cust H. and Taylor C. (2019) Using paradata to collect better survey data: Evidence from a household survey in Tanzania. *Review of Development Economics* 23(2), 598–618. https://doi.org/10.1111/rode.12583.

Cohen-Boulakia, S., et al. (2017). Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems*, 75, 284–298.

Cox S. J. D., Gonzalez-Beltran A. N., Magagna B. and Marinescu M.-C. (2021) Ten simple rules for making a vocabulary FAIR. *PLOS Computational Biology* 17(6), 1–15. https://doi.org/10.1371/journal.pcbi.1009041.

Dawson I. and Reilly P. (2019) Messy assemblages, residuality and recursion within a phygital nexus. *Epoiesen*. https://eprints.soton.ac.uk/439599/.

Demetrescu E., Fanini B. and Cocca E. (2023) An online dissemination workflow for the scientific process in CH through semantic 3D: EMtools and EMviq Open Source tools. *Heritage* 6(2), 1264–1276. https://doi.org/10.3390/heritage6020069.

Derudas P. (2021) Archaeological publication systems: Which route to take? A compass for addressing future development. In *The 26th International Conference on 3D Web Technology*, 1–6. Pisa Italy: ACM. https://doi.org/10.1145/3485444.3487648.

DeVito N. J., Morley J., Smith J. A., Drysdale H., Goldacre B. and Heneghan C. (2024) Availability of results of clinical trials registered on EU Clinical Trials Register: Cross sectional audit study. *BMJ Medicine* 3(1). https://doi.org/10.1136/bmjmed-2023-000738.

Doerr, M., Kritsotaki, A., Rousakis, Y., Hiebel, G. and Theodoridou, M. (2014). *CRMsci: The Scientific Observation Model an Extension of CIDOC-CRM to Support Scientific Observation*, Heraklion: FORTH.

Doerr, M., Stead, S. and Theodoridou, M. (2016). *Definition of the CRMdig: An Extension of CIDOC-CRM to Support Provenance Metadata*, Version 3.2.1, Heraklion: FORTH.

Dourish P. and Cruz E. G (2018) Datafication and data fiction: Narrating data and narrating with data. *Big Data & Society* 5(2), 1–10. https://doi.org/10.1177/2053951718784083.

Durrant G. and Kreuter F. (2013) Editorial: The use of paradata in social survey research. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 176(1), 1–3.

Dykes, B. (2019). *Effective Data Storytelling: How to Drive Change with Data, Narrative, and Visuals*, Hoboken, NJ: Wiley.

Edwards R., Goodwin J., O'Connor H. and Phoenix A. (2017) *Working with Paradata, Marginalia and Fieldnotes: The Centrality of By-Products of Social Research*. Edward Elgar Publishing.

Emerson R. M., Fretz R. I. and Shaw L. L. (2011) *Writing Ethnographic Fieldnotes, 2nd ed.* University of Chicago Press.

EQUATOR Network. (2008). Enhancing the QUAlity and Transparency Of health Research. Retrieved from www.equator-network.org.

Fani Sani, M., Sroka, M., and Burattin, A. (2024). LLMs and process mining: Challenges in RPA. In De Smedt, J. and Soffer, P. (eds.), *Process Mining Workshops*, 379–391. Cham: Springer Nature Switzerland.

Fernández-Fontelo A., Kieslich P. J., Henninger F., Kreuter F. and Greven S. (2023) Predicting question difficulty in web surveys: A machine learning approach based on mouse movement features. *Social Science Computer Review* 41(1), 141–162. https://doi.org/10.1177/08944393211032950.

Gajbe S. B., Tiwari A., Gopal ji and Singh R. K. (2021) Evaluation and analysis of Data Management Plan tools: A parametric approach. *Information Processing & Management* 58(3), 102480. https://doi.org/10.1016/j.ipm.2020.102480.

Geertz, C. (1973). *The Interpretation of Cultures : Selected Essays*, New York: Basic Books.

Giglitto D., Ciolfi L., Lockley E. and Kaldeli E. (2023) *Digital Approaches to Inclusion and Participation in Cultural Heritage: Insights from Research and Practice in Europe*, 1st ed. London: Routledge. https://doi.org/10.4324/9781003277606.

Giraldo O., Garcia A. and Corcho O. (2018) A guideline for reporting experimental protocols in life sciences. *Peer Journal* 6, e4795. https://doi.org/10.7717/peerj.4795.

Goldhammer F., Hahnel C. and Kroehne U. (2020) Analysing log file data from PIAAC. In Maehler, D. B. and Rammstedt, B. (eds.), *Large-Scale Cognitive Assessment: Analyzing PIACC Data*. Cham: Springer International Publishing, 239–269. https://doi.org/10.1007/978-3-030-47515-4_10.

Golub K. and Liu Y.-H. (2022) *Information and Knowledge Organisation in Digital Humanities: Global Perspectives*. United Kingdom: Routledge.

Goodwin J., O'Connor H., Phoenix A. and Edwards R. (2017) Introduction: Working with paradata, marginalia and fieldnotes. In Edwards, R., Goodwin, J., O'Connor, H. and Phoenix, A. (eds.), *Working with Paradata, Marginalia and Fieldnotes*. Edward Elgar Publishing. https://doi.org/10.4337/9781784715250.00007.

Hann, R. (2021). Modelling Kiesler's Endless Theatre: Approaches to paradata for heritage visualization. *Theatre and Performance Design*, 7(1–2), 96–115.

Holtorf, C. (2010). Meta-stories of archaeology. *World Archaeology*, 42(3), 381–393.

Hoopes, R., Hardy, H., Long, M., and Dagher, G. G. (2022). SciLedger: A Blockchain-based Scientific Workflow Provenance and Data Sharing Platform. In 2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC), 125–134.

Huck, J. (2022). Knowledge graphs, metadata practices, and Badiou's mathematical ontology. *KULA: Knowledge Creation, Dissemination, and Preservation Studies*, 6(3), 1–17.

Huvila, I. (2022). Improving the usefulness of research data with better paradata. *Open Information Science*, 6(1), 28–48.

Huvila, I., Andersson, L., Sköld, O., and Liu, Y.-H. (2025). Data makers' and users' views on useful paradata: Priorities in documenting data creation, curation, manipulation and use in archaeology. *International Journal of Digital Curation*. 19(1), https://doi.org/10.2218/ijdc.v19i1.892

Huvila I and Sinnamon L (2022) Sharing research design, methods and process information in and out of academia. *Proceedings of the Association for Information Science and Technology* 59(1), 132–144. https://doi.org/10.1002/pra2.611.

Hyvönen E. (2023) Digital humanities on the Semantic Web: Sampo model and portal series. *Semantic Web* 14(4), 729–744. https://doi.org/10.3233/SW-223034.

International Standard Classification of Education (ISCED 2011). (2011), Montreal: UNESCO Institute for Statistics.

ISO 18629-1. (2004) Industrial automation systems and integration: Process specification language Part 1: Overview and basic principles. (2004). (Version 1). Retrieved from https://www.iso.org/standard/35431.html.

Keenan, T., and Walker, W. (2017). Considerations and challenges for describing historical research data: A case study. *Journal of Library Metadata*, 17(3–4), 241–252.

Knaflic, C. N. (2020). Storytelling with Data: Let's Practice!, Hoboken, NJ: Wiley.

Koesten L., Simperl E., Blount T., Kacprzak E. and Tennison J. (2020) Everything you always wanted to know about a dataset: Studies in data summarization.

International Journal of Human-Computer Studies 135, 102367. https://doi.org/10.1016/j.ijhcs.2019.10.004.

Kunz T., Daikeler J. and Ackermann-Piek D. (2024) Interviewer-observed paradata in mixed-mode and innovative data collection. *International Journal of Market Research* 66(1), 14–26. https://doi.org/10.1177/14707853231184742.

Kvale L. and Pharo N. (2020) Understanding the Data Management Plan as a Boundary Object through a Multi-stakeholder perspective. *International Journal of Digital Curation* 15(1), 16. https://doi.org/10.2218/ijdc.v15i1.729.

Lee S., Li W., Zhang P. and Wang J. (2023) Characterizing data practices in research papers across four disciplines. In Sserwanga, I., Goulding, A., Moulaison-Sandy, H., Du, J. T., Soares, A. L., Hessami, V. and Frank, R. D. (eds.), *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity*. Cham: Springer Nature Switzerland, 359–368.

Lemieux, V. L. (2019). Blockchain and public record keeping: Of temples, prisons, and the (Re)Configuration of power. *Frontiers in Blockchain*, 2. https://doi.org/10.3389/fbloc.2019.00005

Lemieux V. L. (2022) Searching for Trust: Blockchain Technology in an Age of Disinformation. Cambridge: Cambridge University Press.

Lin, Y.-T. (2021). Reusing design information: An investigation of the document creation process in service design projects. *Journal of Documentation*, 77(3), 703–721. https://doi.org/10.1108/JD-06-2020-0111

Liu Y.-H. and Wacholder N. (2017) Evaluating the impact of MeSH (Medical Subject Headings) terms on different types of searchers. *Information Processing & Management* 53(4), 851–870. https://doi.org/10.1016/j.ipm.2017.03.004.

Liu Y.-H., Wu M., Power M. and Burton A. (2023) *Elicitation of contexts for discovering clinical trials and related health data: An Interview Study*. Zenodo. Retrieved from https://zenodo.org/records/7839282

MacDonald L., Almeida V. M. de and Hess M. (2017) Three-dimensional reconstruction of Roman coins from photometric image sets. *Journal of Electronic Imaging* 26(1), 011017. https://doi.org/10.1117/1.JEI.26.1.011017.

Mai J.-E. (2008) Actors, domains, and constraints in the design and construction of controlled vocabularies. *Knowledge Organization* 35(1), 16–29.

Mason J. (2018) *Qualitative researching*, 3rd ed. Los Angeles: Sage Publications.

Matei, S. A., and Hunter, L. (2021). Data storytelling is not storytelling with data: A framework for storytelling in science communication and data journalism. *The Information Society*, 37(5). https://doi.org/10.1080/01972243.2021.1951415.

Mendez K. M., Pritchard L., Reinke S. N. and Broadhurst D. I. (2019) Toward collaborative open data science in metabolomics using Jupyter Notebooks and cloud computing. *Metabolomics* 15(10), 125. https://doi.org/10.1007/s11306-019-1588-0.

Miksa T., Oblasser S. and Rauber A. (2021) Automating research data management using machine-actionable data management plans. *ACM Transactions on Management Information Systems* 13(2), 18:1–18:22. https://doi.org/10.1145/3490396.

Nosek B. A. and Lakens D. (2014) Registered reports: A method to increase the credibility of published results. *Social Psychology* 45(3), 137–141. https://doi.org/10.1027/1864-9335/a000192.

Nurmikko-Fuller T. (2023) *Linked Open Data for Digital Humanities*, London: Routledge. https://doi.org/10.4324/9781003197898.

Obels P., Lakens D., Coles N. A., Gottfried J. and Green S. A. (2020) Analysis of open data and computational reproducibility in registered reports in psychology. *Advances in Methods and Practices in Psychological Science* 3(2), 229–237.

Oinn, T., et al. (2004). Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17), 3045–3054.

Oldman D. and Tanase D. (2018) Reshaping the Knowledge Graph by Connecting Researchers, Data and Practices in ResearchSpace. In Vrandečić, D., Bontcheva, K., Suárez-Figueroa, M. C., Presutti, V., Celino, I., Sabou, M., Kaffee, L.-A. and Simperl, E. (eds.), *The Semantic Web – ISWC 2018*. Cham: Springer International Publishing, 325–340. https://doi.org/10.1007/978-3-030-00668-6_20.

Opgenhaffen L. (2022) Archives in action. The impact of digital technology on archaeological recording strategies and ensuing open research archives. *Digital Applications in Archaeology and Cultural Heritage* 27, e00231. https://doi.org/10.1016/j.daach.2022.e00231.

Ortlipp, M. (2008). Keeping and using reflective journals in the qualitative research process. *The Qualitative Report*, 13(4), 695–705.

Phillips, D., and Smit, M. (2021). Toward best practices for unstructured descriptions of research data. *Proceedings of the Association for Information Science and Technology*, 58(1), 303–314. https://doi.org/10.1002/pra2.458

Pimentel J. F., Murta L., Braganholo V. and Freire J. (2021) Understanding and improving the quality and reproducibility of Jupyter notebooks. *Empirical Software Engineering* 26(4), 65. https://doi.org/10.1007/s10664-021-09961-9.

Poirier, L. (2021). Reading datasets: Strategies for interpreting the politics of data signification. *Big Data & Society*, 8(2), https://doi.org/10.1177/20539517211029322.

Poirier, L. (2022). Accountable data: The politics and pragmatics of disclosure datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1446–1456. New York: Association for Computing Machinery.

PROV-Overview: An Overview of the PROV Family of Documents. (2013). Retrieved from www.w3.org/TR/2013/NOTE-prov-overview-20130430/.

Rheinberger H.-J. (2023) *A Phenomenology of Experimentation*. Chicago: University of Chicago Press.

Rupp, F., Schnabel, B., and Eckert, K. (2024). Implementing data workflows and data model extensions with RDF-star. *The Electronic Library*, 42(3), 393–412.

Rytter, M., Andersen, A., Dalsgård, L., Kusk, M. L., Nielsen, M. and Rubow, C. (eds.) (2020) Anthropology Inside Out: Fieldworkers Taking Notes. Sean Kingston Publishing.

Sandoval G. (2020) In pursuit of a reflexive recording: An epistemic analysis of excavation diaries from the Çatalhöyük research project. *Norwegian Archaeological Review* 53(2), 135–153. https://doi.org/10.1080/00293652.2020.1854338.

Sant, T. (ed.) (2017) *Documenting Performance: The Context and Processes of Digital Curation and Archiving*. London ; New York: Bloomsbury Methuen Drama.

Schwandt, S. (2022). Opening the black box of interpretation: Digital history practices as models of knowledge. *History and Theory*, 61(4), 77–85.

Shortt, H., and Warren, S. (2020). Photography: Using Instagram in participant-led field studies. In Ward, J. and Shortt, H. (eds.), *Using Arts-based Research Methods: Creative Approaches for Researching Business*, 237–270. Organisation and Humanities, Cham: Springer International Publishing.

Smale N. A., Unsworth K., Denyer G., Magatova E. and Barr D. (2020) A review of the history, advocacy and efficacy of Data Management Plans. *International Journal of Digital Curation* 15(1), 30. https://doi.org/10.2218/ijdc.v15i1.525.

Smith, L. (2020). *Emotional Heritage: Visitor Engagement at Museums and Heritage Sites*, London: Routledge.

Stead, S., and Doerr, M. (2015). *CRMinf: The Argumentation Model: An Extension of CIDOC-CRM to Support Argumentation, Version 0.7*, Purley: Paveprime.

Sullivan E. A. (2020) *Constructing the Sacred: Visibility and Ritual Landscape at the Egyptian Necropolis of Saqqara*. Stanford University Press.

Sullivan E. A. (2023) The senses & the sacred: A multisensory and digital approach to examining an Ancient Egyptian funerary landscape. In Landeschi, G. and Betts, E. (eds.), *Capturing the Senses*, 37–61. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-23133-9_3.

Svenonius E. (1986) Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science* 37(5), 331–340.

Swan M. (2015) *Blockchain: Blueprint for a New Economy*. Sebastopol, CA: O'Reilly Media.

Theodoridou, M., Tzitzikas, Y., Doerr, M., Marketakis, Y., and Melessanakis, V. (2010). Modeling and querying provenance by extending CIDOC CRM. *Distributed and Parallel Databases*, 27(2), 169–210.

Thomer A. K., Wickett K. M., Baker K. S., Fouke B. W. and Palmer C. L. (2018) Documenting provenance in noncomputational workflows: Research process models based on geobiology fieldwork in Yellowstone National Park. *Journal of the Association for Information Science and Technology* 69(10), 1234–1245. https://doi.org/10.1002/asi.24039.

Trippas, J. R., Al Lawati, S. F. D., Mackenzie, J., and Gallagher, L. (2024). What do users really ask large language models? In *Proceedings of the SIGIR'24, July 14–18, 2024*, New York: ACM. https://doi.org/10.1145/3626772.3657914.

Vancisin, T., Clarke, L., Orr, M., and Hinrichs, U. (2023). Provenance visualization: Tracing people, processes, and practices through a data-driven approach to provenance. *Digital Scholarship in the Humanities*, 38(3), 1322–1339.

Weigel T., Schwardmann U., Klump J., Bendoukha S. and Quick R. (2020) Making data and workflows findable for machines. *Data Intelligence* 2(1–2), 40–46. https://doi.org/10.1162/dint_a_00026.

Wilkinson M. D., et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3(1), 160018. https://doi.org/10.1038/sdata.2016.18.

Williamson K. (2018) *Observation. In Research Methods: Information, Systems, and Contexts*. Elsevier, 405–427. https://doi.org/10.1016/B978-0-08-102220-7.00017-0.

Wu M., Richard S. M., Verhey C., Castro L. J., Cecconi B. and Juty N. (2023) An analysis of crosswalks from research data schemas to schema.org. *Data Intelligence* 5(1), 100–121. https://doi.org/10.1162/dint_a_00186.

Zeng M. and Qin J. (2022) *Metadata, 3rd ed.* Chicago: ALA Neal-Schuman.