2

The Concept of Paradata

Olle Sköld

2.1 Introduction

Paradata belongs to a family of concepts – together with, for example, metadata and provenance data – that in different ways express the characteristics of, and the relationships between, forms of information and information about them. Metadata is a term that is often used to denote the informational hierarchy of more complex forms of information (books, journals, datasets) and higher-level or more general descriptions thereof (titles, creators, file formats) (Mayernik, 2020). Paradata, drawing on the provisional definition provided in Chapter 1, is information describing the practices and processes involved in how forms of information are created, curated and used. This very wide characterisation branches in multiple directions and can serve to illustrate some of the complexity appended to the concept of paradata and how it has been put to work in multiple domains of research and in service of a broad range of objectives.

Paradata is at its core information that tells the user about how and possibly why something, or a facet of something, emerges in the way that they do (Andersson et al., 2025; Sköld et al., 2022). This 'something' becoming paradata might offer information about a great many things, but is commonly made by humans, machines or humans and machines together. Paradata can be information about how data and datasets were collected or aggregated (Börjesson et al., 2022a), how a method of analysis was put to use (Huvila et al., 2021b), or the automatically recorded information describing how a piece of equipment carried out its task – like the digital camera embedding camera settings and image metrics in photographs (Kreuter, 2013) or evidence of AI processes (Franks, 2024). In addition to techniques of creation and use, paradata can explicate underpinning intellectual processes (choices, deliberations, interpretations) and horizons (disciplinary perspectives and approaches, theoretical resources) (Bentkowska-Kafel et al., 2012; Huvila and Sköld, 2021). Paradata also, as stressed by InterPARES Trust AI's (forthcoming; cf. Chapter 3 of this volume and Franks, 2024) definition of the term, crucially involves information about the persons involved in the documented data procedures and not only the procedures themselves. There is no specific informational scope associated with paradata, rather paradata can offer either comprehensive and full-range information about the events and processes and shape forms of information, or reflect the same chain of happenings in an incomplete and fragmentary manner (Geiger and Ribes, 2011; Huvila et al., 2023). There is little consensus regarding how to object-agnostically describe paradata. Connecting to the same core theme of paradata being about the factors impacting the creation or re-fashioning of the things being created or re-fashioned, paradata is differently construed to be information about 'processes' (Couper, 2000), 'practices' (Huvila et al., 2021a), 'behaviours' (Stieger and Reips, 2010) 'events' (Sendelbah et al., 2016), 'provenance' (Dahlström and Hansson, 2019), 'traces' (Huvila et al., 2023) or other expressions of chained and to some extent coordinated action (see e.g., Huvila and Sköld, 2021; Rösch, 2021).

The users and uses of the concept of paradata are just as intricate as its empirical characteristics and frameworks. The notion of paradata is present in many fields of research, including extensive use in survey research, archaeology, heritage research, and information studies. In these fields, paradata has garnered both academic and professional interest by parties invested in how paradata can be used, how it can be created and how it can be curated and theorised. In relation to paradata making, it has been suggested that paradata can be captured by structured approaches like workflows and conceptual models (Börjesson et al., 2020; Post and Chassanoff, 2021; Siqueira and Martins, 2022), reflective journals and other narrative means of documentation (Matei and Hunter, 2021; Phillips and Smit, 2021), and video- and photo documentation (Chrysanthi et al., 2016). Paradata use by reading, following or otherwise tracing paradata is seen to facilitate a range of ends all differently connected to paradata being informative about processes that shape scholarly outputs – data, visualisations, conclusions etc. (e.g., Arshia et al., 2021; Fear and Donaldson, 2012). Paradata can, for instance, push algorithmic accountability by facilitating the tracking of operations (Cameron et al., 2023; Davet et al., 2023). It can also be used as an explorative device in grasping how and where provenance is recorded in archival holdings, and how it can be mobilised to underpin archival credibility and trustworthiness (Andresen, 2020). A key paradata use case is cross-stakeholder data reuse, where paradata can be seen to be a part of the core resources required to enable shared research data to be purposefully put to work in support of research tasks other than those originally intended when the data was collected or created by allowing insight into how the data came into being, how it has been processed, and used (Borgman, 2012; Faniel and Yakel, 2017).

This chapter seeks to delve into the concept of paradata and to offer the tools required to link paradata's notable complexity to broad usefulness. As illustrated, paradata can be used as a conceptual explorative device to interrogate how activities of the past intersect with present ends and means via modes of documentation and recording across a wide variety of settings (Reilly et al., 2021; Sköld et al., 2022). Paradata can also be used to refer to specific types of data that can be collected or consulted to find information about processes of becoming (cf. Chapter 3), and to denote corresponding methods of (para)data collection and use. Following Moore's (2004) call to investigate scholarly concepts by approaching them in both theory and academic practice, the present chapter will consider the conceptual dimensions of paradata alongside its prominent empirical examples and scenarios of application. In doing so the chapter will provide the groundwork for subsequent exploration in this volume of how paradata can be be found and harnessed in research documentation (Chapter 3), in methods for documenting and generating (Chapter 4) and identifying (Chapter 5) paradata, and approaches to managing paradata in data repository contexts (Chapter 6). The review of paradata will be done by first investigating the etymology of the concept. Then, paradata definitions in survey research, archaeology and heritage visualisation research will be reviewed. Subsequently, the chapter will move on to discuss metadata and provenance data, two key related terms that will be used to discuss and further interrogate the concept of paradata. The chapter concludes with a discussion of the concept of paradata drawing on the different strands of analysis and exploration previously presented.

2.2 The Etymology of Paradata

The building blocks of the composite term paradata are *para-* and *-data*. Delving into the etymology of these component words specifically as they pertain to scholarship and research settings, it can be observed that *data* is derived from the Latin word with the same spelling (plural of *datum*) signifying '(thing[s]) given' (Online Etymology Dictionary, 2024, no pagination). The classical use of *data* denotes a premise of mathematical calculations

(Online Etymology Dictionary, 2024). Current and colloquial uses of the term include data being a collection, thematisation, or grouping of qualitative or quantitative information (or datum) often in the framework of science, computing and epistemology (Oxford English Dictionary, 2024). While seeing widespread use in scholarship and ancillary fields as a prefix to designate a broad range of activities, skill sets, tools and platforms as having to do with data (data management, data literacy, data repositories), it is difficult to precisely define data because of its epistemic variability (Borgman, 2012; Gitelman, 2013). As shown in Chapter 3, where contextual and practice-led approaches to understanding and identifying data and paradata are discussed and put to work, it can be argued that data is something that ultimately emerges (differently) depending on for what purposes it is being sought and by what methods it is being read, created or curated. Chapter 3 also follows Hjørland (2018) in operationalising recorded data, that is, data manifested in such a way that it can be interacted with, as documents and documentation – for instance: datasets, scholarly papers and monographs, notes and correspondence.

Like 'data-', *para-* is a prefix that sees considerable academic and, to some extent, everyday use. The *para-* prefix has both Latin and Greek roots and its significance varies depending on which origin is being considered. Words like parachute and parasol stem from the Latin prefix meaning 'defense, protection against; that which protects from' (Online Etymology Dictionary, 2023). The *para-* prefix in 'paradata' however relates more closely to the Greek provenance, where it designates 'alongside' and 'beyond' as also seen in the concept of 'paratext' (Genette and Maclean, 1991). The prefix has an array of other meanings in Greek (cf. Online Etymology Dictionary, 2023), which might be a part of the explanation of its wide and varied use in English. The Oxford English Dictionary (2023) notes that the most general application of *para-* is to form terms that are closely related to but distinct from the root word. More specific applications exist in biology and the medical sciences, where the prefix is used to express co-location, proximity, and dysfunction. The use of *para-* to form words and terms in chemistry, which has been said to stem from Berzelius' work in the 1830s (Crosland, 1962 as cited in Oxford English Dictionary, 2023), is notably close to what is manifested in paradata: There, *para-* denotes substances that are varieties or modifications of the root substance and, for paradata most importantly, substances that have been produced or occur alongside the root substance (Oxford English Dictionary, 2023; cf. 'para-archives' in Wall and Hale, 2021). This highlights a discussion continued in Chapters 3, 6, and 7 relating to paradata making: Paradata emerges in related but varied ways – before (prospective paradata), during (in situ paradata) or after (retrospective paradata) – to the enactment or creation of the data, process or practice that the paradata informs about. It is also interesting to note that Efthymiou et al. (2015) stress that the meaning of *para-* in present-day Greek, beyond locational and non-evaluative significances (including resemblances and parallelity), also has a presence in the semantic space of evaluation of divergences and errors. This use of the *para-* prefix corresponds to a certain segment of paradata scholarship use cases, most notably those in survey research (see e.g., Kreuter, 2013, Nicolaas, 2011, and below).

Although the etymology of paradata and its component words is many times more complex than is expressed here, it is possible to note that the *data-* component can be considered to determine its basic significance of the term: Paradata is, crucially, anchored in the semantic space of data and data-related phenomena even though it may vary quite considerably across settings of use what this data is and how it is manifested. The prefix *para-* on the other hand is possibly the main operational force of the two composite parts of the concept of paradata, expressing that paradata is data that has a close relationship to another set of data in the sense of having epistemological utility, but existing beyond and in parallel to it.

2.3 Paradata Definitions

As shown in the beginning of this chapter, paradata is a concept that has seen diverse uses in a broad range of tasks, settings and disciplines; from research to data management and data theorising. Even though paradata might not have received as much attention in theoretical and empirical research as, for example, data and metadata, there have been far-reaching attempts to define paradata in several of the research studies preceding this volume (see e.g., Börjesson et al., 2020; Huvila et al., 2022; Sköld et al., 2022), especially in survey research, archaeology, and heritage visualisation research. These research areas are linked in being interdisciplinary and by their relatively long and established use of the concept of paradata, but they are also distinct in many ways. Survey research and heritage visualisation research can be characterised by being focused on the development, application and interpretation of the range of methods that are the focus of the research area – for instance, photogrammetry and other techniques to build visualisations (3D models, maps) on the basis of heritage and archaeological data (e.g., Niccolucci et al., 2013), and different kinds of survey methodologies (e.g., Edwards et al., 2017; Kreuter, 2013). Archaeology is an academic discipline with a long history that is hallmarked by a diverse array of sub-disciplines where data types, methods of data collection, and modes of work and epistemic horizons, including data curation and use, vary to a notable degree (Huvila, 2014; Khazraee, 2019). Descriptions of data-related practices and processes have been produced in survey research, archaeology, and heritage visualisation research using many other terms (see e.g., Huvila et al., 2021a) than paradata and since before the notion emerged. The review of paradata below, however, is based on literature that employs precisely the 'paradata' concept.

2.3.1 Paradata in Survey Research

There are indications that the earliest academic use of the concept of paradata took place in survey research, although the roots of the concept and particularly the data frame of mind it encapsulates are unclear – see for instance Huvila (2012) referencing Adkins and Adkins (1989) as an early example of paradata-related discussions of how to record past data processes in analogue settings. While Lyberg (2009) has traced the origins of the term paradata even further back to the 1920s, studies that seek to find a core paradata definition commonly reference two papers by Couper from the turn of the millennium (Couper, 1998 and 2000, the former as cited in Kreuter and Casas-Cordero, 2010). Working with the issue of how to improve the usability of computer-assisted survey methods, Couper defines paradata as 'auxiliary data describing the [survey] process' (Couper, 2000, p. 393). Paradata here is described as data that can provide insight into the work of enacting surveys, manifesting for instance in (para)data telling about the survey procedure (the no. of interviewer calls per response, response rates, keystroke logs showing how the interviewer managed the survey system) and the survey results (e.g., interview length per average). Couper (2000) also suggests that useful paradata can trace data from web surveys that would allow evaluation of different parameters that together show how the respondents interacted with the survey website or tool.

Couper's conceptualisation of paradata has seen widespread use both in survey research (e.g., Nicolaas, 2011; Olson, 2013) and in other domains (e.g., Huvila et al., 2021a; Reilly et al., 2021), and it has three key characteristics that are recurrent in a large portion of academic paradata use cases. One such characteristic is that paradata has an 'auxiliary' or supporting role in relation to another dataset that is the principal outcome of the practices or processes described by the paradata, like the survey results in Couper (2000). Another characteristic is that paradata is quite intimately tied to the wide-ranging documentation affordances of technology, where large amounts of data of different granularities can be recorded that describes system interactions almost in parallel to the interactions taking place. A third characteristic is the multimodal temporality of paradata. Paradata is directed towards the past in the sense that it informs about past events and processes. Paradata is, however, also tied to the present and future in its strong utilitarian connotations – paradata is sought because it informs about past events and processes in a way that facilitates present or future actions, evaluations or interpretations in connection to the main product or principal outcome that the paradata describes.

Subsequent survey methodology research has continued to develop the concept of paradata and associated methodologies in several directions, emphasising different facets of the concept and its applications and uses (Nicolaas, 2011). One significant development is that paradata began to shift into something not solely a 'by-product' (Kreuter and Casas-Cordero, 2010, p. 2) of the survey process recorded by the survey software, but also as a type of data collected using multiple means in service of a broader range of goals and intentions. Beyond offering information about the processes underpinning collected data, paradata was seen to be relevant for understanding missing data and data not collected, like survey non-responses, and gauging respondent interest and involvement in the survey instruments (Couper and Kreuter, 2013; Olson, 2013). This more comprehensive definition of paradata further included different types of qualitative data generated both by the data creators themselves and the technologies involved in the research work. Paradata became inclusive of observations by the interviewers about, depending on the present research objective, not only the survey process itself and how the respondent approached answering questions, but also pertaining to the appearance of the respondents and impressions and observations stemming from the broader circumstances of the interview (e.g., state of the respondent's domicile or neighbourhood) (Durrant et al., 2011; Edwards et al., 2017). This type of paradata, manifested commonly in field notes and audio recordings, could include estimations of the respondents' intellectual process, like the extent to which they were certain about their answers and what, and if so how, documentation was consulted by the respondent when providing responses to queries (Nicolaas, 2011).

The concept of paradata was also refined in how it connects to paradata use, driven in part by technological innovation which provided additional and more intricate approaches to creating paradata. For example, eye-tracking methodologies were implemented to record paradata of respondent-computer survey interactions (West, 2011). As available trace paradata in computer-driven survey systems began to increase qualitatively and diversify quantitatively, procedures around how to interpret the resulting data developed. Stieger and Reips (2010) and Sharma (2019) point to the many challenges involved in knowing how log files from systems with different functions in the infrastructures underpinning digital survey systems can be used to capture respondent's often dynamic and idiosyncratic modes of survey-response behaviours. Further expanding the boundaries of the concept of paradata and its use in the survey settings, there have also been calls to delve into existing datasets in search of paradata that can usefully inform researchers about data-related processes. This paradata, what is termed 'retrospective paradata' in Chapter 7, may originally have been recorded for other purposes but can in the current contexts of use provide valuable insights into past data procedures of data creation, curation, and use (Edwards et al., 2017; cf. Börjesson et al., 2022a).

2.3.2 Paradata in Archaeology and Heritage Visualisation Research

The concept of paradata as it emerges in both archaeology and heritage visualisation research is somewhat similar across the two fields, although differences exist. An important point of common reference is The London Charter, a collection of six principles seeking to promote modes of work in computer-driven visualisation efforts that ensures methodological robustness and transparency, and also advance computer-based visualisations as a means to manage and inquire into heritage and archaeological data (The London Charter Organization, 2009b). Principle 4 in the charter outlines how visualisation methods and the resulting visualisations should be documented. Subprinciple 4.6 is titled 'Documentation of process (paradata)' and reads as follows:

> Documentation of the evaluative, analytical, deductive, interpretative and creative decisions made in the course of computer-based visualisation should be disseminated in such a way that the relationship between research sources, implicit knowledge, explicit reasoning, and visualization-based outcomes can be understood. (The London Charter Organization, 2009a)

Writing about the processes and discussions leading up to the establishment of The London Charter, Beacham et al. (2006) characterise the issue of documentation as being among the most difficult challenges addressed in the charter. The difficulties stemmed not principally from attaining a sufficient degree of methodological transparency via documenting the steps and decision-making involved in visualising heritage and archaeological data, but rather to greatly varying methodological expectations rooted in the methodological traditions of the many disciplines and fields of research that engaged in the production and use of computer-based visualisations. Paradata was suggested by Denard (see Denard, 2016) as a way of denoting descriptions of the mental and physical activities involved in creating the visualisations, complemented by articulations of the intellectual context and implicit assumptions and principles, that might serve to bridge methodological frameworks (Beacham et al., 2006). Beacham (Beacham 2011, p. 51 see also Beacham et al., 2006) stresses that paradata are descriptions of 'intellectual capital' being operationalised in the making of computer-based visualisations, further broadening the scope of paradata to possibly include representations ranging from discrete tasks and strategies of action to epistemic horizons.

The London Charter is widely referenced in archaeology and heritage visualisation research discussing paradata (e.g., Huggett, 2012; Morgan and Winters, 2015; Niccolucci et al., 2013), although it has been noted that paradata in the archaeological context is not well defined (Huggett, 2020), and that the principles of the charter require adaptation to the specific circumstances and objectives of archaeological heritage, as done in the Seville Charter (Lopez-Menchero and Grande, 2011). Other and complementary conceptualisations of paradata have been suggested that, for example, also highlight that paradata should include descriptions of the evidence (literary sources, image resources, archival data) involved in creating, curating and using archaeological data and heritage visualisations beyond activities and decisions (Barratt, 2016; D'Andrea and Fernie, 2013). Barratt (2016), referencing Dell'Unto et al. (2013) further expands upon the concept of paradata by categorising different types of evidence (testimonial sources, sources based on current measurements and analyses) and intellectual processes (deduction, comparison, hypotheses) that are useful to include when describing past processes of scholarly work. Exploring paradata from a data management perspective, Kansa et al. (2020; cf. InterPARES Trust AI, forthcoming) determine that information about the data authors and their areas of expertise and training are paradata that can be employed to better understand the data or scholarly product created. Kansa et al. (2020) also underline that it is important to also think about paradata that is more closely associated with the data itself to be of wider scope than the data and include, for example, sampling biases and information about missing data (cf. Ullah, 2015) that can help explain why the data appears in the way it does.

Paradata as it is defined and used in archaeology and heritage visualisation research has several similarities to how paradata is approached in survey research, but there are also other conceptual trajectories emerging. The characterisation of paradata as an auxiliary data type vis-à-vis the 'principal' data or visualisation scholarly outcomes remains, however, with a slight difference in framing. The general understanding of scholarly work as it is represented in this literature resembles a network of interacting research practices, methodologies and varying kinds of data and epistemic circumstances, where the specific relationships between these network components become important due to how they impact each other and ultimately the nature of the scholarly work as a whole (e.g., Denard, 2016; Richards-Rissetto and Landau, 2019). Paradata, in reflecting how research is enacted and data is produced in practice, emerges as a data type that encompasses a wide range of key elements and their interplay (see Reilly et al., 2021), from qualitative paradata describing how understanding and interpretation has happened (D'Andrea and Fernie, 2013) to operational trace data and data detailing machine settings (Reilly et al., 2021). While the concept of paradata remains closely associated with technology and technology's capacity to record representations of how it is used and calibrated (Niccolucci et al., 2010), it is in archaeology and heritage visualisation research to a notably higher degree associated with descriptions of the practices and processes engaged in by the makers of scholarly data and their epistemic attributes. Here, the principal paradata stakeholder group is made up of researchers or other parties seeking to gain insight into data processes by consulting paradata representing the original setting of data creation in a broad-spectrum way, while paradata in survey research commonly reflects the activities of the respondent stakeholder group and how they impact the survey data outputs (Ballin et al., 2006; Nicolaas, 2011).

2.4 Paradata's Conceptual Siblings

There are many concepts that relate to paradata. Among these concepts, metadata and provenance – or provenance data – are arguably most closely connected to paradata in terms of conceptual significance and how they are used in research practice. Metadata, provenance data, and paradata are all concepts that signify 'data about data' in both different and related ways. A working definition of metadata is that it is data about data in the sense that metadata describes data, for example titles, author names and information about standards and tools used in organising and processing the data (Pomerantz, 2015). Provenance data is also data about data, signifying instead data providing information about the actors, activities and resources that have been involved in shaping and maintaining the data (Lemieux, 2016). The close semantic links between metadata and provenance data, and between these terms and paradata, is reflected in their intertwined and diverse uses. As observed by Mudge (2016), metadata and provenance data are frequently used interchangeably in scholarly settings. It is also not uncommon to see the term 'provenance metadata' denoting provenance data (see e.g., Doerr et al., 2016; Reilly et al., 2021) or paradata (Gant and Reilly, 2018). Dahlström and

Hansson (2019, p. 6) write that paradata is 'an interesting form of metadata' that is connected to digital provenance while Beacham (2011, p. 49) likens paradata with 'contextual metadata', noting, however, slight differences in emphasis between the terms. The term 'provenance paradata' is, for instance, used in Chapter 6 of this volume and in Börjesson et al. (2022b, no pagination). The closeness between paradata and provenance data is also repeatedly underlined in the literature (see e.g., D'Andrea and Fernie, 2013; Huggett, 2012; Niccolucci et al., 2013).

Below, metadata and provenance will be explored definitionally and by delving into prominent use cases in a series of domains including e-science, computer science, archival studies, archaeology, heritage visualisation research and the information disciplines. Particular interest is given to how the terms connect to each other and, crucially, how they can also offer a useful window into the concept of paradata.

2.4.1 Metadata

The term 'metadata' is said to have originated in the 1960s within the field of computer science (Furner, 2020) and has since gained widespread recognition, being utilised in numerous and varied contexts. Metadata is firmly established in the LAM (libraries, archives, museums) sector and associated disciplines (information studies, archival studies, museum studies), where it principally refers to the resources used to describe the attributes of physical and digital collections and holdings so their items can be made searchable, findable, and possible to use and manage and preserve (Mayernik, 2020; Ronzino et al., 2012). Gilliland (2008) and Pomerantz (2015) argue, however, that metadata in actuality predates the coining of the term, and that its manifestations have been in existence since the earliest attempts at organising information. The increasing pervasiveness of digital technology in all domains of leisure and labour has been a driver in metadata becoming a colloquial term and a ubiquitous and impactful phenomenon, indispensable in the operation and handling of digital data and digital communications, including social media and other digital platforms and infrastructures (Pomerantz, 2015; Zeng and Qin, 2016).

The literature offers several conceptualisations of metadata that illustrate the purposes that metadata is envisioned to fulfil. Metadata's fundamental function is to in varying ways enrich the form or forms of information it is appended to by facilitating a range of actions. Some of these actions have to do with knowledge organisation and the work of ordering and describing items, data, books and records according to certain principles and systems that support information search (Mayernik, 2020) and required administrative and managerial tasks (Kalová, 2020; Tompkins et al., 2021), including repository storage and long-term preservation (Day, 2002). Metadata also enables both reuse, that is, use by a party external to an item's context of creation) and use of the item described by offering insight into its features and attributes (Gilliland, 2008) and relationships to related items (Force and Smith, 2021). Metadata supports the usability and reusability of forms of information by being a resource that users can employ to validate and contextualise the items, and as evidence in efforts to determine their authenticity, integrity and trustworthiness (Tennis, 2008). Conversely, metadata is also an essential tool in the successful sharing of data, documentation and other forms of information (Chao, 2014). Metadata is, however, no silver bullet in facilitating sharing and reuse. It is necessary that the metadata shared is the metadata needed by the stakeholders set to reuse what is shared (Kim, 2021), and there might be infrastructural and epistemic challenges that impact their ability to apply the metadata (Hansson and Dahlgren, 2021).

Instances and Definitions of Metadata

The literature shows that there are many types of metadata across the diverse domains where it is used. Further, while there are many metadata standards available, metadata is often differently standardised and defined in the contexts of its use (Furner, 2020). In more domain-agnostic writings about metadata, commonly mentioned metadata types are administrative metadata (for managing information), descriptive metadata (for making information findable and identifiable), technical metadata (for managing the systems underpinning information search and repositories), and paradata- and provenance-like use metadata (for describing the use and use-limitations of information) (Gilliland, 2008; Mayernik, 2020; Zeng and Qin, 2016). Metadata types and their naming also vary between applications. For example, projects invested in online interactions can use social metadata describing comments and ratios of positive and negative ratings (Drachsler et al., 2012), while archival metadata applications can include metadata typologies including records creation, record keeping and preservation (Tennis, 2008).

Metadata definitions are similarly diverse and can be found in an array of standards (see Furner, 2020 for an overview of metadata standards), charters, glossaries and scholarly outputs. Zeng and Qin (2016, p. 11) identify the most basic metadata definitions in use as 'information about information' or 'data about data'. Pomerantz (2015, p. 26) remains on a similar level of abstraction, and ties into what is akin to a sociomaterial understanding of information, where information is something that emerges out of human practices rather than being simply encoded into informative objects (see Chapter 3 for a discussion of the sociomateriality of information, documentation and data), by positing that metadata is 'a potentially informative object that describes another potentially informative object'. Gilliland (2008, no pagination) offers an alternate approach to describing metadata by conceptualising it as potentially everything that can be said about different forms of information, while saying that metadata should reflect the content, context and structure of the item it describes. Other definitions point to metadata having more specific characteristics. Greenberg (2003) and Smiraglia (2005) propose that metadata is structured and Zeng and Qin (2016) underline that metadata in addition to being structured also is encoded, that is, in some way involved in a system of description. Tennis (2008) says that metadata is readable by humans and machines, and that it is always artificial and derived on the basis of an analysis of the forms of information being described. Some definitions furthermore tie metadata to certain purposes. Mayernik (2020) understands metadata as something that is always intended to help solve an issue or to be otherwise useful. Greenberg (2003, p. 1876) writes that metadata is supposed to facilitate 'functions' tied to the item in focus of the metadata descriptions. Smiraglia (2005) stresses that metadata should underpin information retrieval, and Zeng and Qin (2016) regard metadata as resources holistically supporting interactions with different forms of information, from identification to evaluation, use, preservation and management.

2.4.2 Provenance and Provenance Data

Like metadata, the concept of provenance is employed in an extensive array of domains and contexts – from computer science to archival science and beyond. Its modes of use and the connotations it carries, including the significance of provenance and the utilities of provenance data, also varies. Provenance has strong connections to the archival field, where it is operationalised in, among other things, the principle of provenance. The principle of provenance is a key tenet of archival theory that describes how archival holdings should be organised by keeping the records from one record creator together and separate from records from other record creators, in this way preserving the meaning and context of the records (Michetti, 2016; Sweeney, 2008). Provenance also sees use in other fields and disciplines concerned with tracking and documenting data trajectories, and using information about past interactions with forms of information (again, data, records, books, manuscripts) to further present analytical or interpretative efforts (see e.g., Lemieux, 2016). Provenance is an important term in the field of computer science, in particular in scholarly efforts to better model and describe data workflows and the management of scientific data (Curcin, 2017; Doerr and Theodoridou, 2011). Additional examples of settings where provenance and provenance data is a concern includes archaeology (Huggett, 2012), information studies (Fear and Donaldson, 2012), museum studies (White, 2017), geology and palaeontology (Reilly et al., 2021), and art history and cinema studies (Bernardi et al., 2021).

While being used in many settings for a range of purposes, provenance data is broadly conceptualised as a resource principally for making visible how forms of information travel between stakeholders and repositories, and to describe the impacts that these travel pathways have had on the forms of information themselves (Gehani et al., 2021; Lemieux, 2016; MacNeil, 2008). The literature shows that the information about the 'histories' of different forms of information provided by provenance data is understood to be useful for a range of more specific reasons. In settings of use more closely affiliated with the data-centric understandings of the concept, provenance data is seen to facilitate examinations of different kinds, including data audits and validations (Davidson and Freire, 2008). Curcin (2017) and Ludäscher (2016) outline how provenance data also can be used to assess results and outcomes by facilitating replicability and reproducibility of data tasks and processes, including both analytic and data management operations. Provenance in the data sciences commonly represent very fine-grained representations of data interactions. Provenance graphs and other provenance accounts can provide detailed insight into the discrete steps involved in data creation, curation and use (Davidson and Freire, 2008; Doerr and Theodoridou, 2011).

The use of provenance in disciplines and contexts more aligned with the archival discourse shows many similarities, but also some differences in emphasis and in how the provenance data is manifested. Similarities include using provenance data to drive data exploration (Davidson and Freire, 2008) and the determination of authenticity and other qualities, including accuracy and relevance (Fear and Donaldson, 2012; cf. Davet et al., 2023). Provenance data is also used to better understand the origins of the data or record at hand, inclusive of both actors involved and the broader contexts in which the materials emerged or were created (Buchanan, 2016; Michetti, 2016), together with impactful epistemic and theoretical circumstances (Huggett, 2012). However, provenance data in the archival setting less often pertains to the objects or records it describes on the item-level, instead existing predominantly at the collection-level. This makes it difficult, often impossible, to base estimations of authenticity and trustworthiness on detailed documentation of data interactions. Instead, an important function of provenance data is to show chains of custody and ownership that can be used to assess relevance and task-related value (Lemieux, 2016). It is also often highlighted that provenance data underpins all basic categories of archival work and functionality (arrangement, description, acquisition, retrieval, appraisal, Lemieux, 2016; Michetti, 2016; Sweeney, 2008), including preservation (Li and Sugimoto, 2014).

Instances and Definitions of Provenance Data

Instances of provenance data discussed in the literature mirror the diverse applications of the concept across its domains of use. There is, for example, data provenance (Asuncion, 2013), archival provenance (Bearman and Lytle, 1985), network provenance (Gehani et al., 2021), digital provenance (Dahlström and Hansson, 2019), and many other types of provenance used to describe the custody trajectories and modes of creation and use of the forms of information they relate to, including provenance detailing field-agnostic occurrences like processes (Cuevas-Vincenttin et al., 2016) and events (Doerr and Theodoridou, 2011). The W3C Working Group (2013) identifies three types of provenance in their PROV Data Model: provenance tied to the agents (people, organisations) involved in creating or manipulating forms of information; provenance tied to the forms of information themselves, describing their origins and connections to other items; and provenance tied to the processes of creating the forms of information. This categorisation connects to the where-provenance, why-provenance, and how-provenance discussed by Huggett (2012), which contains provenance reporting modes of creation (how) and origins of the data or items (where), but to greater extent emphasises the reasons for creating the data and the underpinning choices and deliberations (why). CRMdig, a data model for scientific observations, additionally contains what-provenance and when-provenance, offering information about resources and tools involved in the events described and temporal details respectively (Doerr and Theodoridou, 2011). There are also provenance typologies that are organised wholly on a temporal basis (see Cuevas-Vincenttin et al., 2016; Davidson and Freire, 2008; Ludäscher, 2016). Prospective provenance is used in both the archival and computer science domains. Common examples include 'records management plans' in the former field and 'workflows' in the latter. Prospective provenance is documentation that covers the workflows and procedures involved in creating information. Retrospective provenance, on the other hand, describes how forms of information have been created, comprising – like prospective provenance – of data inputs and other resources and tools employed in their creation.

Definitions of provenance data abound, but share many of the same elements. Sweeney (2008, p. 193) observes that a recurring common denominator is that provenance is information that describes 'the origins of an information-bearing entity or artifact'. Beyond this point, definitions diverge depending on the field. In archival settings, many definitions stress that provenance data ties the record to the individuals or organisations who during the course of their activities created or modified it (see e.g, the archival ISAD(G) standard, International Council on Archives, 2000), sometimes covering also the specific organisational functions or processes involved in shaping or making the record (Lemieux, 2016, see also Sundberg, 2013). Some definitions limit the chronology of relevance for provenance data to the activities prior to inclusion into the holdings of an archive or records centre (International Council on Archives, 2004), while others include custodial and post-custodial provenance (Cook, 1993). Information about the chain of custody and ownership is, however, ubiquitous in the archival understanding of provenance (Lemieux, 2016; Michetti, 2016; Sweeney, 2008).

Definitions of provenance in more data-centred domains commonly stress that the data interactions detailed in the provenance data can be carried out by both human and technological actors (Missier, 2017) – as do several archival definitions, see for instance MacNeil, 2008 – and can have a notably comprehensive scope. The PROV Data Model defines provenance as ‘descriptions of the entities and activities involved in producing and delivering or otherwise influencing a given object’ (W3C Working Group, 2013, no pagination). Provenance definitions in the computer science domain often include accounts of data inputs or other resources employed (see e.g., Davidson and Freire, 2008), and very detailed descriptions of how the object has been modified and transformed (see e.g., Gehani et al., 2021 and W3C Working Group, 2013) intending to allow for the tracing connections between ‘raw’ data and products based on analysis of this data (Davidson and Freire, 2008; Pinheiro et al., 2013).

It can be noted that discussions of how to involve AI applications in archives management and the use of archives for scholarly or other purposes seem to indicate a possible intermingling of different archives- and computer science-based understandings of provenance and provenance data. At times, however, ‘paradata’ is used to describe the information that can be used to audit AI functions and interactions within the archival framework (see e.g., Davet et al., 2023; Franks, 2024; InterPARES Trust AI, forthcoming), possibly to distinguish this particular instance of process information from conventional archival provenance – a further indication of the (sometimes) closeness of the terms.

2.5 Discussion

The goal of this chapter was to examine the concept of paradata. The examination was done on the basis of a review of the etymology of the concept, and analyses of how paradata is defined and put to use in research practice in survey research and heritage visualisation research, two domains where paradata is a comparably established notion. Metadata and provenance data, two of paradata's closest conceptual siblings, were also reviewed alongside examples of relevant use cases so as to provide additional perspectives on the concept of paradata and how it can be understood.

It remains difficult, however, to arrive at an exhaustive description of the concept of paradata. This is not due to a lack of definitional options. Numerous studies have observed that there are many definitions of paradata discussed in multiple fields of study (Reilly et al., 2021), although none of them are in widespread use (e.g., Kreuter and Casas-Cordero, 2010; Nicolaas, 2011; Olson, 2013). Instead, the chapter indicates that there are several complexities at play that make it challenging to explain the concept. One major complexity is that while concepts always can be approached and analysed as such, they are also to some extent inextricably tied to practice (Moore, 2004). This means that any discussion of conceptual significance or definitions are tied to modes of doing and thinking in a particular professional arena or scholarly discipline, regulated by both specific practical circumstances and epistemic frameworks. As will be discussed further on, it is not certain if this heterogeneous array of paradata applications and meanings is a hindrance for grasping and putting the concept to use, or a resource that can be drawn on for the same purposes.

The second major complexity that makes it difficult to arrive at a precise rendering of the concept of paradata is that it co-exists with metadata and provenance in a complicated conceptual space, where the connections and boundaries between the concepts are blurred and shifting according to perspective. The complexities underpinning the difficulties of arriving at an encompassing and clear description of paradata also interconnect. This is visible in the many instances in the literature, where paradata is discussed as an instance of provenance data (Niccolucci et al., 2013) or metadata (Dahlström and Hansson, 2019), and many other intersections of the concepts are present (see e.g., Beacham, 2011; Huggett, 2012; Reilly et al., 2021).

2.5.1 Encompassing Paradata, Provenance Data and Metadata

The challenges of encompassing paradata notwithstanding, this chapter has shown that the concept has several core characteristics and also key correspondences to metadata and provenance data. General features of all three concepts are that they – like data and documentation, see Chapter 3 – exist through and in relation to conditions that are both social and material in nature, consisting of, to different extents, field-specific and field-overarching elements of scholarly practice, tools and systems, and epistemic frameworks. Paradata, provenance data and metadata are thus used and produced by both human and non-human actors by automatic or manual means, and they are, to a large extent, system-dependent in the sense that access, management and storage extensively relies on the purposeful functioning of supporting information systems (Force and Smith, 2021; Hansson and Dahlgren, 2021; Kim, 2021).

The literature also shows that all three data types can be represented 'in' data files and other forms of information, or exist in resources that are distinct from them – like metadata schemas or method descriptions in associated publications (De Oliveira et al., 2015; Pomerantz, 2015). Also, the discussed data types emerge in different forms and in different places: They can be highly structured (found in standards, data models) but also unstructured (present in datasets, work logs or notes) and comprise narrative or trace data. In terms of their temporal orientation, the concepts can be prescriptive and determine what paradata, provenance data or metadata should be documented when carrying out a certain task or process. They can also be retrospective and be used to discern and describe data interactions that are in the past or prospective and focused on speculating or stipulating future activities (Chapter 7; Cuevas-Vincenttin et al., 2016; Davidson and Freire, 2008; Ludäscher, 2016). Finally, paradata, provenance data and metadata are 'meta' concepts (cf. Schenk et al., 2009) in that they express the relationships and attributes of forms of information and other data or descriptive resources related to these items.

The relationship between paradata, provenance data and metadata on one hand and the forms of information that they describe on the other, is arguably – at least from a conceptual perspective – both their main designator and possibly their lowest common denominator. This chapter shows that the differences between the concepts are notable. The use cases of paradata, provenance data and metadata reviewed above support some observations about the relationship between the concepts as they manifest in research practice. A general dissimilarity between metadata and paradata and provenance data is that the former term describes the attributes of forms of information, while the latter two terms describe their histories of creation, curation, use and custody. In situations where the concepts are not considered to be distinct, metadata is most commonly attributed the highest order of abstraction. In these instances, provenance data (da Cruz et al., 2011; Curcin, 2017; Malik et al., 2010) and paradata (Dahlström and Hansson, 2019; Niccolucci et al., 2013; Pomerantz, 2015) are organised as metadata sub-types.

This is the only observable hierarchical relationship between the concepts. Paradata and provenance data are at times posited as alternatives to each other (Huggett, 2012), but several authors consider the concepts to be distinct, albeit closely related. The principal difference between paradata and provenance in these cases is that provenance has a stronger emphasis on the technical circumstances of data interactions. Paradata is instead seen to highlight the activities, processes and frameworks involved in interpreting and understanding data (D'Andrea and Fernie, 2013; Reilly et al., 2021).

The conceptual relationships most frequently discussed in the reviewed literature are those between metadata and paradata, and metadata and provenance data. Li and Sugimoto (2014) also suggest that metadata creation and description should be considered from the viewpoint of provenance, and that it would be valuable to record and keep. Reilly et al. (2021, p. 458) contrasts metadata, described as 'generally uncontentious static properties of data', to the more heterogeneous collection of data inputs and byproducts, scholarly activities and interpretative resources captured as paradata.

Studies in survey research show fairly uniform distinctions between paradata and metadata. Metadata is exemplified as providing information about survey-supporting elements, for example, questionnaires, sample designs, settings and configurations, and the web browsers and operating systems used (Barratt, 2016; Olson, 2013; Stieger and Reips, 2010). This is in contrast to paradata, which describes both how the surveys have been interacted with (scrolls, clicks, typing, call records) and the intellectual work involved in enacting the survey study (Barratt, 2016; Stieger and Reips, 2010). It is interesting to note that the relatively homogenous characterisation of paradata in survey research stands somewhat in contrast to approaches in other fields, where paradata is understood to encompass also the machinery, infrastructures, and tools involved in data creation, curation and use alongside information about their configurations (Huvila, 2022; Kreuter, 2013).

2.5.2 The Concept of Paradata

In conclusion, the concept of paradata as discussed in the literature is diverse. This chapter has shown that paradata encompasses a range of meanings and definitions. It is also conceptualised as being useful in relation to a various set of purposes and intents. Paradata's empirical referents – that is, the standards, data models and data or process descriptions that exemplify and instantiate it – are plentiful and assorted. Paradata is connected in close and complex ways to metadata and provenance data, and while some general observations can be made about how the concepts and their modes of use differ and are alike, it is difficult to draw any definitive boundaries between them. This does not mean, however, that there are not any purposeful approaches available that can be used to grasp the concept of paradata and to discern its possible implementations.

One approach is to consider paradata in a more abstract way. From this perspective, it is possible to determine that paradata is a concept that can be used to examine any information or data phenomenon in any work or leisure setting, although it has been most prominently used to examine scholarly settings and scholarly data in survey research, heritage visualisation research and information studies. Paradata is data about the full range of activities, resources and epistemic frameworks involved in creating, managing and using scholarly data. It can emerge in any form or format, and be created intentionally or as a secondary outcome of scholarly data interactions and processes. Paradata is useful for gaining insight into past data trajectories, processes and events. Purposeful paradata can facilitate a wide range of data management and reuse scenarios, and is an asset in examining data when assessing its relevance, reliability, accuracy, scope and content. From a more abstracted perspective, paradata differs from metadata in that it describes data processes, and metadata describes data. Paradata and provenance data are more alike in that both concepts are broadly concerned with the origins and processing histories of data, but provenance data is more strongly situated in the settings of archives, archaeology, art history and computer science which, as we have seen above, impacts what the data looks like, and how it is created and used.

Another approach is to consider paradata in a more tangible way, which also introduces significantly more changeability to the concept. In this sense, paradata emerges as a concept that is differently defined and applied across research settings and disciplines. In survey research, paradata showcases well-regimented expressions and modelings of past data trajectories and interactions that are akin to the computer science workflow-provenance research and standard-driven metadata descriptions. In other domains, paradata is a notably more fluid concept used as an exploratory and interpretative device (see e.g., Börjesson et al., 2022a; Reilly et al., 2021). Similarly to commonplace framings of metadata (see e.g., Mayernik, 2020; Tennis, 2008), paradata is in some settings data that is always created to function precisely as paradata. In other settings, it – like provenance data – is a 'by-product' of the activities of individuals, groups or institutions. Paradata also stands in other, more complex relationships to provenance data and metadata, and the intersections and similarities between the concepts are many. The significance of paradata and how it relates to metadata or provenance data has to be determined on a case-by-case basis, although there are some general patterns to how the concepts are employed together, as previously discussed in this chapter.

The way of thinking about paradata in a more tangible sense also points to paradata challenges that are important to discuss. In this view, both the definitional and the use case facets of paradata are understood to be bound to the practices and regimens of their setting of application. That being said, data like paradata always travel across several settings of application, where practices may contrast or clash (Brown and Duguid, 1996; Star and Griesemer, 1989). In relation to paradata, this might mean that there are gaps between the understandings and conceptualisations of paradata that risk causing paradata use becoming difficult. Such gaps can manifest also in varying levels of compatibility in information system design and data ontologies with particular paradata-related tasks, purposes and understandings. Examples include research communities with more or less matched comprehensions of paradata, and data models that render paradata in ways that do not match what is required for success in a particular use case (cf. Hansson and Dahlgren, 2021; Kim, 2021; Mayernik, 2011).

Related to this point are a range of often-present insufficiencies connected to the concept of paradata. One insufficiency is that paradata never is a 'reality-checking' implementation (Beacham, 2011, p. 51). No amount of paradata or combination of paradata can give complete access to data-interaction histories, and the work of, for example, reusing data is surely facilitated by paradata, but never fully driven and accomplished by paradata. Tasks involving the use of paradata to achieve some aim or result must always be supplemented also by other non-paradata resources and efforts.

Another insufficiency is that paradata itself is a data type that requires interpretation and sufficient interpretative conditions. This could be taken to mean that, theoretically, there would be reason to to supplement the paradata with additional paradata describing how and why it was created and documented (cf. Gant and Reilly, 2018; Huggett, 2020; Li and Sugimoto, 2014). This 'paradox . . . of potentially infinite regression' (Huggett, 2020, p. 3) must, again, be considered in relation to practice where case-specific objectives and requirements of use, together with available resources to harness or create paradata, will be the actual factors limiting and shaping what interactions with paradata are relevant and possible.

The many definitions of and approaches to paradata are, in some respects, an obstacle for grasping and employing the concept. Equally, from a different viewpoint, it might not be considered desirable to lessen the complexities associated with paradata. As discussed later in this volume in Chapters 4–6, standards and other meta-framing attempts can be useful, but so also can the diversity of meanings and approaches to using paradata for various ends and means throughout paradata's many fields and scenarios of use. As observed by Furner (2020) in his discussion of the wealth of available metadata standards, it is important that connectivities exist between communities of concept use. Such connectivities can be built on the basis of clear and well-documented applications of the concept that make visible connections and disconnects across different scenarios of its application, as has been attempted in this chapter and which this volume pursues further in the following chapters.

References

Adkins L. and Adkins R. A. (1989). *Archaeological Illustration*. Cambridge: Cambridge University Press.

Andresen H. (2020). A discussion frame for explaining records that are based on algorithmic output. *Records Management Journal* 30(2), 129–141. https://doi.org/10.1108/RMJ-04-2019-0019.

Arshia A. H., Rasekh A. H., Moosavi M. R., Fakhrahmad S. M. and Sadreddini M. H. (2021). Traceability mining between unit test and source code based on textual analysis applied to software systems. Digital Scholarship *Humanities* 36(2), 268–285. https://doi.org/10.1093/llc/fqaa017.

Asuncion H. U. (2013). Automated data provenance capture in spreadsheets, with case studies. *Future Generation Computer Systems* 29(8), 2169–2181. https://doi.org/10.1016/j.future.2013.04.009.

Ballin M., Scanu M. and Vicard P. (2006). *Paradata and bayesian networks: A tool for monitoring and troubleshooting the data production process*. Rome: Università degli studi Roma Tre.

Barratt R. (2016). 3D reconstruction in archaeology: Using the future to understand the past in the present. Part 3: Sources and paradata. Available at https://archphotogrammetry.com/2016/06/25/part-3-sources-and-paradata (accessed 29 May 2024).

Beacham R. (2011). Concerning the paradox of paradata. Or, 'I don't want realism; I want magic!' *Virtual Archaeology Review* 2(4), 49–52.

Beacham R., Denard H. and Niccolucci F. (2006). An introduction to the London Charter. In *Papers from the Joint Event CIPA/VAST/EG/EuroMed Event*.

Bearman D. A. and Lytle R. H. (1985). The power of the principle of provenance. *Archivaria*, 21, 14–27.

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

Bernardi J., Usai P. C., Williams T. and Yumibe J. (eds.) (2021). *Provenance and Early Cinema*. Indianapolis: Indiana University Press.

Borgman C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology* 63(6), 1059–1078. https://doi.org/10.1002/asi.22634.

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

Börjesson L., Sköld O., Friberg Z., Löwenborg D., Palsson G. and Huvila I. (2022a). Re-purposing excavation database content as paradata: An explorative analysis of paradata identification challenges and opportunities. *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 62(3), 1–18. https://doi.org/10.18357/kula.221.

Börjesson L., Sköld O. and Huvila I. (2020). Paradata in documentation standards and recommendations for digital archaeological visualisations. *Digital Culture & Society* 6(2), 191–220. https://doi.org/10.14361/dcs-2020-0210.

Brown J. S. and Duguid P. (1996). The social life of documents. *First Monday* 1(1).

Buchanan S. A. (2016). *A Provenance Research Study of Archaeological Curation*. PhD dissertation. Austin: University of Texas at Austin.

Cameron S., Franks P. and Hamidzadeh B. (2023). Positioning paradata: A conceptual frame for AI processual documentation in archives and recordkeeping contexts. *Journal on Computing and Cultural Heritage* 16(4), 1–19. https://doi.org/10.1145/3594728.

Chao T. C. (2014). Enhancing metadatafor research methods in data curation. *Proceedings of the American Society for Information Science and Technology* 51(1), 1–4. https://doi.org/10.1002/meet.2014.14505101103.

Chrysanthi A., Berggren Å., Davies R., Earl G. P. and Knibbe, J. (2016). The camera 'at the trowel's edge': Personal video recording in archaeological research. *Journal of Archaeological Method and Theory* 23(1), 238–270. https://doi.org/10.1007/s10816-015-9239-x-015-9239-x.

Cook T. (1993). The concept of the archival fonds in the post-custodial era: Theory, problems and solutions. *Archivaria* 35, 24–37.

Couper M. P. (1998). Measuring survey quality in a CASIC environment. In *Proceedings of the Survey Research Methods Section,* 41–49.

Couper M. P. (2000). Usability evaluation of computer-assisted survey instruments. *Social Science Computer Review* 18(4), 384–396. https://doi.org/10.1177/089443930001800402.

Couper M. P. and Kreuter F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176(1), 271–286. https://doi.org/10.1111/j.1467-985X.2012.01041.x.

Crosland M. P. (1962). *Historical Studies in the Language of Chemistry*. London: Heinemann.

da Cruz S. M. S., Paulino C. E., de Oliveira D., Campos M. L. M. and Mattoso M. (2011). Capturing distributed provenance metadata from cloud-based scientific workflows. *Journal of Information and Data Management* 2(1), 43–50.

Cuevas-Vincenttin V. et al. (2016). *ProvONE: A PROV Extension Data Model for Scientific Workflow Provenance*. Available at http://jenkins-1.dataone.org/jenkins/view/Documentation%20Projects/job/ProvONE-Documentation-trunk/ws/provenance/ProvONE/v1/provone.html (accessed 29 May 2024).

Curcin V. (2017). Embedding data provenance into the Learning Health System to facilitate reproducible research. *Learning Health Systems* 1(2). https://doi.org/10.1002/lrh2.10019.

Dahlström M. and Hansson J. (2019). Documentary provenance and digitized collections: Concepts and problems. *Proceedings from the Document Academy* 6(1), 1–11. https://doi.org/10.35492/docam/6/1/8.

D'Andrea A. and Fernie K. (2013). CARARE 2.0: A metadata schema for 3D cultural objects. In *2013 Digital Heritage International Congress (DigitalHeritage)*. New York: IEEE, 137–143. https://doi.org/10.1109/DigitalHeritage.2013.6744745.

Davet J., Hamidzadeh B. and Franks P. (2023). Archivist in the machine: Paradata for AI-based automation in the archives. *Archival Science* 23, 275–295. https://doi.org/10.1007/s10502-023-09408-8.

Davidson S. B. and Freire J. (2008). Provenance and scientific workflows: Challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. New York: ACM. https://doi.org/10.1145/1376616.1376772.

Day M. (2002). *Cedars Guide to Preservation Metadata*. Bath: University of Bath.

De Oliveira D., Silva V. and Mattoso M. (2015). How much domain data should be in provenance databases? In *Proceedings of the 7th USENIX Conference on Theory and Practice of Provenance*. Berkeley, CA: USENIX Association.

Dell'Unto N., Leander A. M., Dellepiane M., Callieri M., Ferdani D. and Lindgren S. (2013). Digital reconstruction and visualization in archaeology: Case-study drawn from the work of the Swedish Pompeii Project. In *2013 Digital Heritage International Congress (DigitalHeritage)*. New York: IEEE, 621–628. https://doi.org/10.1109/DigitalHeritage.2013.6743804.

Denard H. (2016). A new introduction to the London Charter. In Bentkowska-Kafel A, Denard H. and Baker D. (eds.), *Paradata and Transparency in Virtual Heritage*. Farnham: Ashgate.

Doerr M., Stead S. and Theodoridou M. (2016). *Definition of the CRMdig: An Extension of CIDOC-CRM to support provenance metadata.* Available at https://cidoc-crm.org/lrmoo/sites/default/files/CRMdig_v3.2.1.pdf (accessed 29 May 2024).

Doerr M. and Theodoridou M. (2011). CRMdig: A generic digital provenance model for scientific observation. In *TAPP11: 3rd USENIX workshop on the Theory and Practice of Provenance*.

Drachsler H. etal. (2012). *D8. 1 Review of Social Data Requirements*. Report written in the Open Discovery Space project.

Durrant G. B., D'Arrigo J. and Steele F. (2011). Using paradata to predict best times of contact: Conditioning on household and interviewer influences. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174(4), 1029–1049. https://doi.org/10.1111/j.1467-985X.2011.00715.x.

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*. Cheltenham: Edward Elgar Publishing.

Efthymiou A., Fragaki G. and Markos A. (2015). Exploring the meaning and productivity of a polysemous prefix: The case of the Modern Greek prepositional prefix para-. *Acta Linguistica Hungarica Acta Linguistica Hungarica* 62(4), 447–476. http://dx.doi.org/10.1556/064.2015.62.4.4.

Faniel I. and Yakel E. (2017). Practices do not make perfect: Disciplinary data sharing and reuse practices and their implications for repository data curation. In Johnston L. R. (ed.), *Curating Research Data: Practical Strategies for Your Data Repository*. Chicago, IL: ACRL, 103–126.

Fear K. and Donaldson D. R. (2012). Provenance and credibility in scientific data repositories. *Archival Science* 12(3), 319–339. https://doi.org/10.1007/s10502-012-9172-7.

Franks, P. (2024). The crucial role of paradata in AI governance. In Duranti L and Rogers C (eds.), *Artificial Intelligence and Documentary Heritage*. Vancouver: InterPARES Trust AI.

Force D. C. and Smith R. (2021). Context lost: Digital surrogates, their physical counterparts, and the metadata that is keeping them apart. *The American Archivist*, 84(1), 91–118. https://doi.org/10.17723/0360-9081-84.1.91.

Furner J. (2020). Definitions of ‘metadata’: A brief survey of international standards. *Journal of the Association for Information Science and Technology* 71(6), E33–E42. https://doi.org/10.1002/asi.24295.

Gant S. and Reilly P. (2018). Different expressions of the same mode: A recent dialogue between archaeological and contemporary drawing practices. *Journal of Visual Art Practice* 17(1), 100–120. https://doi.org/10.1080/14702029.2017.1384974.

Gehani A., Ahmad R., Irshad H., Zhu J. and Patel J. (2021). Digging into big provenance (with SPADE). *Communications of the ACM* 64(12), 48–56.

Geiger R. S. and Ribes D. (2011). Trace ethnography: Following coordination through documentary practices. In *Proceedings from the 44th Hawaii International Conference on System Sciences*. New York. IEEE, 1–10. https://doi.org/10.1109/HICSS.2011.455.

Genette G. and Maclean M. (1991). Introduction to the paratext. *New Literary History* 22(2), 261–272. https://doi.org/10.2307/469037.

Gilliland A. J. (2008). Setting the stage. In Baca M. (ed.), *Introduction to Metadata*. Los Angeles, CA: Getty Research Institute, 1–19.

Gitelman L. (ed.) (2013). ‘*Raw Data’ is an Oxymoron*. Cambridge, MA: MIT Press.

Greenberg J. (2003). Metadata and the world wide web. In Drake M. A. (ed.), *Encyclopedia of Library and Information Science*, 2nd ed. New York: Dekker, 1876–1888.

Hansson K. and Dahlgren A. (2021). Open research data repositories: Practices, norms, and metadata for sharing images. *Journal of the Association for Information Science and Technology* 73(2), 303–316. https://doi.org/10.1002/asi.24571.

Hjørland B. (2018). Data (with big data and database semantics). *Knowledge Organization* 45(8), 685–708. https://doi.org/10.5771/0943-7444-2018-8-685

Huggett J. (2012). Promise and paradox: Accessing open data in archaeology. In Mills C., Pidd M. and Ward W. (eds.), *Proceedings of the Digital Humanities Congress 2012*. Sheffield: Humanities Research Institute. http://dx.doi.org/10.17613/qc5d-9x46.

Huggett J. (2020). Capturing the silences in digital archaeological knowledge. *Information* 11(5), 278. https://doi.org/10.3390/info11050278.

Huvila I. (2012). The unbearable complexity of documenting intellectual processes: Paradata and virtual cultural heritage visualisation. *Human IT* 12(1), 97–110.

Huvila I. (2014). Archaeologists and their information sources. In Huvila I. (ed.), *Perspectives to Archaeological Information in the Digital Society*. Uppsala: Uppsala universitet, Institutionen för ABM, 25–54.

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

Huvila I., Andersson L. and Sköld O. (2022). Citing methods literature: Citations to field manuals as paradata on archaeological fieldwork. *Information Research* 27(3). https://doi.org/10.47989/irpaper941.

Huvila I., Andersson L. and Sköld O. (eds.). (2024). *Perspectives on Paradata: Research and Practice of Documenting Data Processes*. Cham: Springer.

Huvila I., Greenberg J., Sköld O., Thomer A., Trace C. and Zhao X. (2021a). Documenting information processes and practices: Paradata, provenance metadata, life-cycles and pipelines. *Proceedings of the Association for Information Science and Technology* 58(1), 604–609. https://doi.org/10.1002/pra2.509.

Huvila I. and Sköld O. (2021). Choreographies of making archaeological data. *Open Archaeology* 7(1), 1602–1617. https://doi.org/10.1515/opar-2020-0212.

Huvila I., Sköld O. and Andersson L. (2023). Knowing-in-practice, its traces and ingredients. In Cozza M. and Gherardi S. (eds.), *The Posthumanist Epistemology of Practice Theory: Re-imagining Method in Organization Studies and Beyond*. Cham: Palgrave Macmillan, 37–69. https://doi.org/10.1007/978-3-031-42276-8_2.

Huvila I., Sköld O. and Börjesson L. (2021b). Documenting information making in archaeological field reports. *Journal of Documentation* 77(5), 1107–1127. https://doi.org/10.1108/JD-11-2020-0188.

International Council on Archives (2000). *ISAD(G): General International Standard Archival Description*. Paris: International Council on Archives.

International Council on Archives (2004). *Multilingual Archival Terminology: Provenance*. Paris: International Council on Archives.

InterPARES Trust AI (forthcoming). *Terminology Database: Paradata*. Vancouver: InterPARES Trust AI.

Kalová T. (2020). Creating needs-based metadata and research data management services: Exploring the requirements of scientists. *Young Information Scientist* 5, 31–46. https://doi.org/10.25365/yis-2020-5-3.

Kansa S. W., Atici L., Kansa E. C. and Meadow R. H. (2020). Archaeological analysis in the information age: Guidelines for maximizing the reach, comprehensiveness, and longevity of data. *Advances in Archaeological Practice* 8(1), 40–52. doi:10.1017/aap.2019.36.

Khazraee E. (2019). Assembling narratives: Tensions in collaborative construction of knowledge. *Journal of the Association for Information Science and Technology* 70(4), 325–337. https://doi.org/10.1002/asi.24133.

Kim Y. (2021). A study of the roles of metadata standard and data repository in science, technology, engineering and mathematics researchers' data reuse. *Online Information Review* 45(7), 1306–1321. https://doi.org/10.1108/OIR-09-2020-0431.

Kreuter F. (ed.) (2013). *Improving Surveys with Paradata: Analytic Uses of Process Information*. Chichester: John Wiley & Sons.

Kreuter F. and Casas-Cordero C. (2010). *Paradata. Rat für Sozial- und Wirtschaftsdaten (RatSWD) working paper*. Berlin: RatSWD.

Lemieux V. L. (2016). Provenance: Past, present and future in interdisciplinary and multidisciplinary perspective. In Lemieux V. (ed.), *Building Trust in Information*. Cham: Springer, 3–45. https://doi.org/10.1007/978-3-319-40226-0_1.

Li C. and Sugimoto S. (2014). Provenance description of metadata using PROV with PREMIS for long-term use of metadata. In *International Conference on Dublin Core and Metadata Applications*, 147–156. https://doi.org/10.23106/dcmi.952136486.

Lopez-Menchero V. M. and Grande A. (2011). The principles of the Seville Charter. In *CIPA Symposium Proceedings*, 2–6.

Ludäscher B. (2016). A brief tour through provenance in scientific workflows and databases. In Lemieux V (ed.), *Building Trust in Information*. Cham: Springer, 103–126. https://doi.org/10.1007/978-3-319-40226-0_7.

Lyberg L. (2009). *The paradata concept in survey research.* Presentation. Available at https://csdiworkshop.org/wp-content/uploads/2020/03/Lybert2011CSDI.pdf (accessed 29 May 2024).

MacNeil H. (2008). Archivalterity: Rethinking original order. *Archivaria* 66(0 SE - Articles):1–24.

Malik T., Nistor L. and Gehani A. (2010). Tracking and sketching distributed data provenance. *In 2010 IEEE Sixth International Conference on e-Science*. New York: IEEE, 190–197. https://doi.org/10.1109/eScience.2010.51.

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), 312–322. https://doi.org/10.1080/01972243.2021.1951415.

Mayernik M. S. (2011). Metadata tensions: A case study of library principles vs. everyday scientific data practices. *Proceedings of the Association for Information Science and Technology* 47(1), 1–2. https://doi.org/10.1002/meet.14504701337.

Mayernik M. S. (2020). Metadata. *Knowledge Organization* 47(8), 696–713. https://doi.org/10.5771/0943-7444-2020-8-696.

Michetti G. (2016). Provenance: An archival perspective. In Lemieux V. (ed.), *Building Trust in Information*. Cham: Springer, 59–68. https://doi.org/10.1007/978-3-319-40226-0_3.

Missier P. (2017). Provenance standards. In Liu L. and Özsu M. T. (eds.), *Encyclopedia of Database Systems*. New York: Springer, 1–8.

Moore H. L. (2004). Global anxieties: Concept-metaphors and pre-theoretical commitments in anthropology. *Anthropological Theory* 4(1), 71–88. https://doi.org/10.1177/1463499604040848.

Morgan C. and Winters J. (2015). Introduction: Critical blogging in archaeology. Internet *Archaeology* 39. https://doi.org/10.11141/ia.39.11.

Mudge M. (2016). Transparency for empirical data. In Bentkowska-Kafel A, Denard H. and Baker D. (eds.), *Paradata and Transparency in Virtual Heritage*. Farnham: Ashgate, 252–263.

Niccolucci F., Beacham D., Hermon S. and Denard, H. (2010). Five years after: The London Charter revisited. In Artusi A., Joly M., Lucet G., Pitzalis D. and Ribes A. (eds.), *The 11th International Symposium on Virtual Reality, Archaeology and Cultural Heritage*. Graz: The Eurographics Association, 101–104.

Niccolucci F., Felicetti A., Amico N. and D'Andrea, A. (2013). Quality control in the production of 3D documentation of monuments. In *Built Heritage 2013 Monitoring Conservation Management*. Milano: Centro per la Conservazion, 864–873.

Nicolaas G. (2011). *Survey paradata: A review*. Review paper. Southampton: NCRM.

Olson K. (2013). Paradata for nonresponse adjustment. *The Annals of the American Academy of Political and Social Science* 645(1), 142–170. https://doi.org/10.1177/0002716212459475.

Online Etymology Dictionary. (2023). -para (prefix). Available at www.etymonline.com/word/para-#etymonline_v_7162 (accessed 29 May 2024).

Online Etymology Dictionary. (2024). data (noun). Available at www.etymonline.com/word/data#etymonline_v_782 (accessed 29 May 2024).

Oxford English Dictionary. (2023). para- (prefix). https://doi.org/10.1093/OED/3136200135 (accessed 29 May 2024).

Oxford English Dictionary (2024). data (noun). https://doi.org/10.1093/OED/7999740343 (accessed 29 May 2024).

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.

Pinheiro R., Holanda M., Araujo A. P. F., Walter, M. E. and Lifschitz, S. (2013). Automatic capture of provenance data in genome project workflows. In *2013 IEEE International Conference on Bioinformatics and Biomedicine*, 15–20. https://doi.org/10.1109/BIBM.2013.6732621.

Pomerantz J. (2015). *Metadata*. Cambridge, MA: MIT Press.

Post C. and Chassanoff A. (2021). Beyond the workflow: Archivists' aspirations for digital curation practices. *Archival Science* 21(4), 413–432. https://doi.org/10.1007/s10502-021-09365-0.

Reilly P., Callery S., Dawson I. and Gant S. (2021). Provenance illusions and elusive paradata: When archaeology and art/archaeological practice meets the phygital. *Open Archaeology* 7(1), 454–481. https://doi.org/10.1515/opar-2020-0143.

Richards-Rissetto H. and Landau K. (2019). Digitally-mediated practices of geospatial archaeological data: Transformation, integration, & interpretation. *Journal of Computer Applications in Archaeology* 2(1), 120–135. http://doi.org/10.5334/jcaa.30.

Ronzino P., Hermon S., and Niccolucci F. (2012). A metadata schema for cultural heritage documentation. *Electronic Imaging & the Visual Arts*. Florence: Firenze University Press, 36–41. http://doi.org/10.1400/187333.

Rösch F. (2021). From drawing into digital: On the transformation of knowledge production in postexcavation processing. *Open Archaeology* 7(1), 1506–1528. https://doi.org/10.1515/opar-2020-0211.

Schenk S., Dividino R. and Staab S. (2009). Reasoning with provenance, trust and all that other meta knowledge in OWL. In Freire J., Missier P. and Sahoo S. S. (eds.), *Proceedings of the First International Workshop on the role of Semantic Web in Provenance Management (SWPM 2009), collocated with the 8th International Semantic Web Conference (ISWC- 2009).*

Sendelbah A., Vehovar V., Slavec A. and Petrovčič A. (2016). Investigating respondent multitasking in web surveys using paradata. *Computers in Human Behavior* 55, 777–787. https://doi.org/10.1016/j.chb.2015.10.028.

Sharma S. (2019). *Paradata, Interviewing Quality, and Interviewer Effects*. PhD dissertation. Ann Arbor: University of Michigan.

Siqueira J. and Martins D. L. (2022). Workflow models for aggregating cultural heritage data on the web: A systematic literature review. *Journal of the Association for Information Science and Technology* 73(2), 204–224. https://doi.org/10.1002/asi.24498.

Sköld O., Börjesson L. and Huvila, I. (2022). Interrogating paradata. *Information Research* 27. https://doi.org/10.47989/colis2206.

Smiraglia R. P. (2005). Introducing metadata. *Cataloging & Classification Quarterly*, 40(3–4), 1–15. https://doi.org/10.1300/J104v40n03_01.

Star S. L. and Griesemer J. R. (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science* 19(3), 387–420. https://doi.org/10.1177/030631289019003001.

Stieger S. and Reips U.-D. (2010). What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior* 26(6), 1488–1495. https://doi.org/10.1016/j.chb.2010.05.013.

Sundberg H. (2013). Process based archival descriptions: Organizational and process challenges. *Business Process Management Journal* 19(5), 783–798. https://doi.org/10.1108/BPMJ-Jan-2012-0002.

Sweeney S. (2008). The ambiguous origins of the archival principle of 'provenance'. *Libraries & the Cultural Record* 43(2), 193–213. www.jstor.org/stable/25549475.

Tennis J.T. (2008). *Metadata in the Chain of Preservation Model: Draft Metadata Specification Model*. InterPARES 2 working paper.

The London Charter Organization (2009a). *The London Charter for the computer-based visualisation of cultural heritage*. Available at https://londoncharter.org (accessed 29 May 2024).

The London Charter Organization (2009b). *The London Charter for the computer-based visualisation of cultural heritage: Principle 4 - Documentation*. Available at https://londoncharter.org/principles/documentation.html (accessed 29 May 2024).

Tompkins V. T., Honick B. J., Polley K. L. and Qin J. (2021). MetaFAIR: A metadata application profile for managing research data. *Proceedings of the Association for Information Science and Technology* 58(1), 337–345. https://doi.org/10.1002/pra2.461.

Ullah I. I. T. (2015). Integrating older survey data into modern research paradigms: Identifying and correcting spatial error in 'Legacy' datasets. *Advances in Archaeological Practice*, 3(4), 331–350. https://doi.org/10.7183/2326-3768.3.4.331.

W3C Working Group (2013). *PROV Model Primer*. Available at www.w3.org/TR/prov-primer (accessed 29 May 2024).

Wall G. and Hale A. (2021). Art & archaeology: Uncomfortable archival landscapes. *The International Journal of Art & Design Education* 39(4), 770–787. https://doi.org/10.1111/jade.12316.

West B. T. (2011). Paradata in survey research. *Survey Practice* 4(4). https://doi.org/10.29115/SP-2011-0018.

White L. (2017). Provenance of museum objects. In McDonald J. D. and Michael L.-C. (eds.), *Encyclopedia of Library and Information Sciences*, 4th ed. Boca Raton: CRC Press, 3756–3765.

Zeng M. L. and Qin J. (2016). *Metadata*. Chicago, IL: American Library Association.