Research Article

Isto Huvila*

Improving the usefulness of research data with better paradata

https://doi.org/10.1515/opis-2022-0129
received November 9, 2021; accepted March 25, 2022.

Abstract: Considerable investments have been made in Europe and worldwide for developing research data infrastructures. Instead of a general lack of data about data, it has become apparent that a pivotal factor that drastically constrains data use is the absence of contextual knowledge about how data was created and how it has been curated and used. This applies especially to many branches of social science and humanities research, where data is highly heterogeneous, both by its kind (e.g. being qualitative, quantitative, naturalistic, purposefully created) and origins (e.g. being historical/contemporary, from different contexts and geographical places). The problem is that there may be enough metadata (data about data) but there is too little paradata (data on the processes of its creation, curation and use). The aim of this position paper is to draw attention 1) to the need for a better and more systematic understanding and documentation of the contexts of creation, curation and use of research data to make it useful and usable for researchers and other potential users in the future, and 2) to specific obstacles that make the capturing of this particular type of metadata, known as *paradata*, especially difficult. Failing to understand what information about the creation, curation and use of research data is needed and how to capture enough of that information risks that the currently collected vast amounts of research data become useless in the future.

Keywords: paradata; data collection; data use; documentation; metadata.

1 Introduction

Considerable efforts and resources have been invested worldwide to develop and establish information and archival infrastructures for managing and making research data available across disciplines. A growing corpus of literature (see e.g. Drude et al., 2016; Gerth et al., 2017) has begun to provide evidence of the value of existing research data infrastructures (e.g. European Strategy Forum on Research Infrastructures, 2011; Moulin et al., 2011; Borgman, 2015a; Borgman et al., 2016) in enabling such, both data-intensive large-scale and qualitative small-scale, research and outreach that has never been possible before (e.g. Benardou et al., 2018; Aloia et al., 2017). In parallel, the potential of data- and collections-based research resting on the utilisation of digital collections of data is becoming increasingly evident (Drude et al., 2016; Gerth et al., 2017). At the same time, however, it has become apparent that there are critical obstacles relating to the description of data and collections that hinder the present infrastructures from properly supporting the use and re-use of information they hold.

Long before the current surge of attention to the complications of data reuse, already from the 1970s onwards, archaeologists have been debating an emerging 'curation crisis' (Flexner, 2016; Kersel, 2015; Voss, 2012; Marquardt et al., 1982). As the critics have warned (e.g. Bauer-Clapp & Kirakosian, 2017; Flexner, 2016; Olivier, 2015; Kintigh & Altschul, 2010; Fagan, 1995), the massive number and ongoing exponential increase of poorly documented and difficult to access data collections are undermining the legitimacy and rationale of the whole discipline and its existence (Voss, 2012). Even if the critique of the archaeological curation crisis refers to general lack of documentation and publishing of collections (Voss, 2012; Friberg & Huvila, 2019), a closer look at its underpinnings has made apparent that a central

***Corresponding author: Isto Huvila,** Uppsala University, Sweden, E-mail: isto.huvila@abm.uu.se

issue, which makes analysing archaeological excavation documentation created by someone who has not excavated at a specific archaeological site a very „difficult endeavour" (Demoule, 2011) is not the lack of information about the data, but that the late-comer *has not excavated there* i.e. does not have enough contextual knowledge to interpret it properly (Voss, 2012; Faniel et al., 2013). While cross-disciplinarity makes archaeology a useful context to explore the problems associated with the lack of knowledge on creation, curation and use of research data, the difficulty of using others' research materials and data is not unique to archaeology but has been observed across disciplines and pertains to all research working – in some meaning – with data (see e.g. Leonelli & Tempini, 2020; Borgman, 2012). In this respect, codicology (Clemens & Graham, 2007) and diplomatics (i.e. the study of ancient documents) have been pioneers both in acknowledging the need for meticulous documentation of research materials and their provenance and developing systematic methods for such work – that have not lost their relevance in the digital era (Duranti, 1998). A critical insight emerging from acknowledging the significance of tracing the history of a particular piece of information is that in addition to the meticulous description of contents, technical characteristics and ownership, authorship, rights and sources of data implemented in the current metadata schemes (cf. Huggett, 2016), a key to ensuring the usability and usefulness of data underlined in the immediate past and on-going research (e.g. Drude et al., 2016; Geser & Selhofer, 2015; Geser & Niccolucci, 2016; Krauwer et al., 2017) is proper documentation of *the processes and practices of its creation*. Further, as curators have sometimes reminded (Voss, 2012), it is equally important to have access to its history of *curation and use*, i.e. where, how and on what premises it was created, and how it has been managed, used and 'remixed' with other data.

Even if a meticulous review of the literature reveals a plethora of remarks that point to the importance of comprehensive information on the creation, curation and use of research data for its future usefulness and usability, it is striking how poorly the issue is understood. Moreover, much of the earlier work has focused on documenting and communicating historical contexts of data and data-related processes even if it would be equally crucial to understand and describe also contemporary and desired future practices and processes. Prompted by the breadth of the problem of the lacking of such information and its implications to research across disciplines, the aim of this paper is to draw attention by reviewing earlier literature to following questions:

- firstly, a systematic understanding of the processes and practices of creating, curating and using research data are crucial across different research disciplines in making the data useful and usable for researchers and other potential users who themselves have not participated earlier in working with the data;
- secondly, obtaining, preserving and passing on such an understanding requires that a certain amount of information – or *paradata* – on its creation, curation and use is available for the future users of the data;
- thirdly, a much better understanding of what information can function as paradata and where it can be found is crucial to the effort, and
- fourthly and finally, there are several different methods researchers and data curators can use to obtain useful paradata and make existing information useful as paradata.

So far paradata has been discussed to some extent in a limited number of fields, including survey research (Edwards et al., 2017) and cultural heritage visualisation (Bentkowska-Kafel et al., 2012) and to varying degrees by referring to such related and partly overlapping concepts as provenance, context information and metadata but its implications to the usability of research data at large have remained largely unexplored. Similarly, there is an apparent lack of comprehensive understanding of the phenomenon of paradata as a whole and what exactly is problematic about it. Many of the examples used to illustrate the argument in this article are drawn from archaeology, a distinctly cross-disciplinary and thus helpful domain to highlight the issues and opportunities relating to paradata and paradata work. In spite of this partial focus on one domain, the discussion meanders across a much wider range of disciplinary contexts as does its relevance that is by no means specific to archaeological data practices.

The structure of this article is as follows. First, we start by exploring the general question of the general importance of contextual information about data, where the issue has been debated and what arguments have been put forward to substantiate it. After setting the stage, the discussion proceeds to explore the notion of paradata as a particular perspective to contextual information about processes. Next, the article reviews major issues relating to the management and use of paradata, specifically the relatively unproblematic question of how to represent paradata and the more tricky questions of what qualifies as paradata and where it can be found. After having outlined key challenges, the discussion turns to interrogating how they could be addressed. The focus is particularly on what underpins useful

paradata and what could eventually help to make paradata useful. Finally, the article turns to outline the practical and theoretical implications and consequences of having better paradata and to conclude by suggesting how future research can contribute to improving documentation of data creation, curation and use processes.

2 Importance of contextual information

In contrast to the specific significance of knowing about creation, curating and using data, the general importance of contextual information about research data and its importance for successful sharing and reusing of research data has been highlighted in the literature across disciplines from information science and research data management (e.g. Faniel et al., 2019; Mostafa, 2018; Wilkinson et al., 2016; Borgman, 2012) to computer-supported cooperative work (e.g. Karasti et al., 2006), different branches of sciences (Kim & Yoon, 2017) including zoology and quantitative social sciences (Faniel & Yakel, 2017), survey research (Edwards et al., 2017), cultural heritage studies (e.g. Bentkowska-Kafel et al., 2012; Kowal et al., 2017; Kolenda & Markiewicz, 2017; Ball, 2011; Niccolucci, 2010) heritage management (e.g. Foster & Rafferty, 2016; Münster et al., 2016; Krauwer et al., 2017; Harrison et al., 2017), and science studies (e.g. Bauman & Raud, 2015). Calls for adequate contextual information have been made in evidence-based studies of research practices (Borgman, 2012; Faniel et al., 2013; Kim & Yoon, 2017) and theoretical and normative literature (e.g. Niven & Richards, 2017; Huggett, 2016; Lake, 2012; Morgan & Eve, 2012) alike. Its significance is acknowledged also in the recent standards and guidelines, including the FAIR principles for scientific data management and stewardship (Wilkinson et al., 2016). At the same time, much of the work so far has remained domain-specific (e.g. Bentkowska-Kafel et al., 2012; Edwards et al., 2017), focused on either highly generic or very specific, not seldom curatorial (as e.g. in Lee, 2011; Beaudoin, 2012) rather reusers', needs (Faniel et al., 2019), and lacked detail how to tackle with the issue in practice.

Unsurprisingly, the importance of contextual information has been underlined especially in data infrastructures and (re)use-oriented research (e.g. Tempini, 2017; Faniel & Yakel, 2017) in such data-intensive fields as, for instance, corpus linguistics (e.g. Pančur et al., 2018), large-scale survey research, astronomy and genomics (Borgman, 2015b). Reuse-oriented research has often been struggling to find and get access to and make data usable (e.g. Löwenborg, 2014; Cooper & Green, 2015; Doran et al., 2019) that raises legitimate concerns of the present, often limited consideration of the usefulness, limits and contextuality of the currently available and (re)used research data (e.g. Bevan, 2015). Outside of the fields with a long tradition of reusing data—for example, the above-mentioned corpus linguistics, large-scale survey research, astronomy and genomics—data reuse is not, however, as widespread as it might be imagined. Many researchers are much less concerned about reuse than collecting and working with new data (Voss, 2012; Borgman, 2015b). The prominence of fieldwork in archaeology (Voss, 2012) and the typical focus of the state-of-the-art methodological innovations on the "interpretation at the trowel's edge" in the field rather than working with legacy collections (Hodder, 2003; Voss, 2012; Dell'Unto et al., 2017) is exemplary of this general tendency, which is not specific to that particular discipline.

While certain contextual information is necessary for using all types of research materials, old and new, in all contexts, in tight disciplinary communities much of such knowledge is readily available and already known by all of its members (Fry, 2006; Prainsack et al., 2014). Meticulous documentation and communication of contextual information are exceedingly critical when (re)users come from diverse disciplinary backgrounds and lack a shared tacit understanding of the priorities and usual practices of obtaining and processing data (Doran et al., 2019). This is typical to cross-disciplinary research topics (incl. geographical information science, data science, and natural sciences oriented work in humanities and social sciences, see Kristiansen 2014; Cooper & Green 2015; Bevan 2015; McCoy 2017) and fields like digital humanities (Poole & Garwood, 2018) and heritage research, and the exceedingly multidisciplinary research, for instance, in historical and linguistic studies and social sciences (Drude et al., 2016).

The fundamental rationale of documenting not only observations but also data collection and manipulation procedures stems from the fact that similarly to the creation and use, also curation (Buchanan, 2016; Voss, 2012), archiving, and data and collections management (Kelleher, 2017; Newman, 2012; Schwartz & Cook, 2002) are non-neutral, generative activities with an impact on data and consequently, on how it can and should be interpreted. In this respect, it is difficult to overemphasise how the contexts of creation, curation and use are not only nice-to-know background information about data but a key to understanding the material itself and its implications to inferences

made when it is used. Independent of discipline, the reasons why contextual information is needed, tend to be similar (see e.g. Faniel & Yakel, 2017; Yoon, 2016; Birnholtz & Bietz, 2003). Without documentation of how data was captured or created, in what types of conditions, what tools were used, who did what and when, what qualifications and experience the individuals had, and for instance, who made decisions and on what premises it is impossible to know the extents of the data, what was left out and why, and in general, how well the data suits for addressing other types of research questions than the original ones, and what needs to be taken into account when using it. The lack of adequate contextual documentation of processes relating to data creation, curation and use – or as Leonelli and colleagues (Leonelli & Tempini, 2020) refer to it, data journey – is a major impediment of open science (Lake, 2012), effective data reuse (e.g. Faniel et al., 2013), and keeping, as Conkey and Gero (1997) put it, human-agency visible in knowledge creation processes. The importance of obtaining and having a proper understanding of the data in context is also of relevance for the quality of research as the proponents of mindfully collected 'slow data' (for the concept, see Kansa, 2016) and 'slow' research (for the concept, see Adams et al., 2014) indirectly underline by emphasising qualitative value over efficiency in measuring outcomes and implications of scientific and scholarly work.

Considering how widely the lack of adequate contextual information has been commented at least in bypassing as a problem, it is conspicuous how slippery concept 'context' is and how difficult it can be to capture. It is easy to list names, places, tools and moments in time whereas really keeping and making 'human-agency visible', to borrow the concept of (Conkey & Gero, 1997), has proven to be much harder. In the following, as we will see, the problem starts with establishing how to conceptualise human agency in terms that make it documentable.

3 What is paradata?

A natural follow-up to the question of whether and why the documentation of contextual information in general and data creation, curation and use, in particular, is important, is what exactly needs to be documented. It is tempting to suggest that the conceptual unclarity surrounding the different concepts referring to various aspects of the context of data and of data creation, curation and use is symptomatic of the issues pertaining to this question. Apart from *paradata* (data that describes processes, Couper, 2000, cf. metadata that describes data, Pomerantz, 2015), also the terms *context* and *context information* (e.g. Faniel et al., 2013, 2019) and *provenance* and *provenance metadata* (e.g. Doerr et al., 2016; Huggett, 2012 cf. the earlier popular notion *provenience* i.e. the origins/birthplace of an object, Buchanan, 2016) have been used in relation to information about the origins, context and processes related to the earlier life of data – each term and proposer with slightly different emphases, premises and objectives underpinning their work. In addition, such concepts as metadata, data documentation or contextual information are sometimes either implicitly or explicitly used in a generic but in practice often either rather broad (e.g. the definition of Brocks et al., 2009) and unspecific, or in a fairly narrow meaning to refer to everything or to only a small subset of all possible information that pertains to the 'context' of data. To exemplify the contrasts, Faniel and colleagues (2019) write of context information literally consisting of in practice all context-related information whereas, for instance, Beaudoin (2012) focuses on information relevant for digital preservation. Figure 1 outlines a diagrammatic illustration of schematic broad and narrow perspectives together with an alternative middle-range perspective proposed and discussed later in this section.

The parallel conceptual vagueness and over-specificity pertains also to the concept of paradata. The term paradata, which refers in a generic sense to process information, has been used in several parallel contexts, however, quite clearly without reaching a well-articulated definition. There is a considerable body of literature on the uses and usefulness of paradata in survey research (Kreuter, 2013; Kreuter & Casas-Cordero, 2010), education (e.g. Milligan et al., 2014), heritage visualisation (e.g. Bentkowska-Kafel et al., 2012) and recently, information science (e.g. Börjesson et al., 2020). Originally, Couper (1998) introduced the concept in the survey research field to denote automatically generated process data of human-system interactions, for example, sequences of keystrokes and other interactions and their timings, but later on, the term has been used to refer to a broader variety of processual meta-level information on processes (Nicolaas, 2011). Further definitions of paradata have been proposed e.g. paradata as "information about how users have interacted with artefacts and sometimes also the purpose, task at hand and other contextual information" (Drachsler et al., 2012) but even they tend to fall within the broad understanding of paradata as data about processes. In survey research, the concept has been found useful in evaluating data quality (Kreuter, 2013), response times (Couper

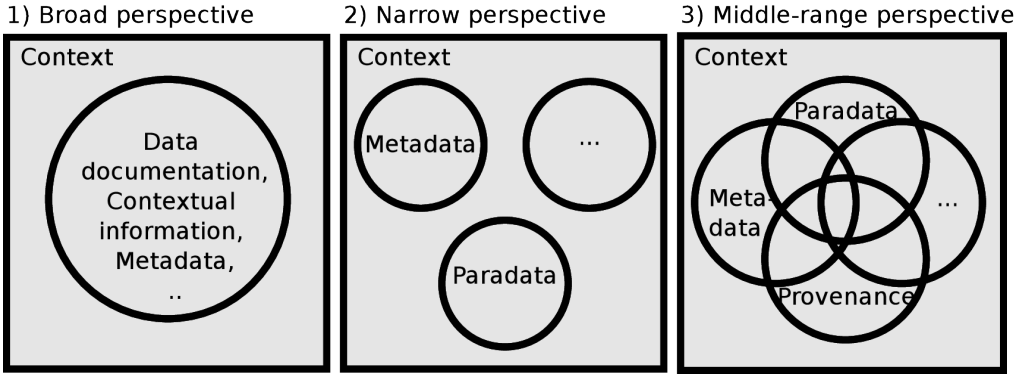

Figure 1: Broad, narrow and middle-range perspectives to contextual information.

& Kreuter, 2013), non-response adjustment (Olson, 2013), and for instance, understanding what respondents do while filling in web questionnaires (Stieger & Reips, 2010). Paradata has also been used for research in data collection practices (Laflamme, 2009; Elliot et al., 2015) and for helping to understand and plan data acquiring procedures in interview research (Durrant et al., 2011).

Outside of survey research, the concept of paradata has gained traction, especially in the context of computer-based visualisation of cultural heritage. A key document in heritage visualisation paradata is the so-called London Charter, which advocates for the need to document interpretative decisions made in the course of 3D visualisation. Even if the London Charter assumes a narrower definition of the term as information about human processes of understanding and interpretation of data objects (The London Charter Organisation, 2009), in practice, having this information means that it is necessary to capture paradata in a much broader sense (as for Couper, 2000, or that Huggett 2012 calls *provenance*) as a premise of these fairly advanced insights.

Even if the diversity of views of what paradata is suggesting the complexity of what it potentially entails in different fields, the proliferation of contexts where it has been debated, largely independently of each other, points to the gravity of the need and utility of process data. Similarly, earlier studies show (e.g. Goodwin et al., 2017; Anderson & Larson, 2013) that collecting and investigating different forms of data from automatically collected technical data to hand-written marginalia can convey a lot of invaluable information on research processes. Still, as a whole, we are far from a concerted understanding of what all paradata in practice might entail and how it relates to parallel concepts and forms of descriptive data. While it is possible to argue for the usefulness of many of the related concepts such as provenance or contextual metadata, a potential advantage of referring to paradata as documentation of the matters of data creation, curation and use is its focus on processes and doings rather than on contextual and historical information in a more general sense. Instead of only looking back, the notion of paradata collects together information that describes historical and ongoing contemporary processes relating to data with planned, anticipated and desired future processes that could be applied to it.

What is suggested here is that similarly to other forms of meta and para information (including metadata and paratext), paradata is a form of data with relation to other data. This means essentially assuming a middle-range perspective (Fig. 1) to contextual information. In contrast to provenance data as data about data provenance, metadata as (usually any type of descriptive) data about data, paradata is data that can help to elucidate past, ongoing and potential processes relating to data. For example, of the categories of context information identified by Faniel et al. (2019), several contain information that is directly characterisable as paradata. These include, for example, data collection, provenance, curation and digitisation information. At the same time, many others are in most cases probably better described as metadata—even those with (in)direct evidential value of, for instance, actors and activities and that can be used as or turned to paradata. Such information could be, for example, data-related locations, identifiers, time periods and access restrictions. As a subset of all contextual information, a complete utopian set of paradata would make it possible to reconstruct and follow all doings and decisions related to a particular data in minute detail. Whether a particular piece of information can function as paradata depends on perspective like the data'ness of things. Like the data that is metadata for someone can be data for someone else, also someone's paradata can be assumed to be

data or metadata for others. One specific piece of data can *function*, for instance, as metadata and paradata but it does not mean that all data would necessarily do that. In this sense, paradata is fundamentally an analytical rather than substantial concept and depends on how certain kinds of things or types of information can be informative in particular situations, tasks and domains and for specific individuals and groups.

4 Problems and non-problems of working with paradata

Even if it is obvious that increasing awareness of the importance of the availability of paradata and documentation of processes and contexts is of critical importance, the mere recognition of the problem is obviously not enough to solve it. The past and ongoing work with direct relevance to the documentation of the intellectual processes (e.g. in the context of archaeology and cultural heritage, in projects such as PARTHENOS, ARIADNE and CARARE) – instead of tracing mere transformations and movement of data – has focused so far to a large extent on scoping the conceptual apparatus (e.g. Bentkowska-Kafel et al., 2012; Huggett, 2012), and on implementing standards for technical representation of the information (e.g. Bruseker et al., 2017b, 2017a; Doerr et al., 2016; D'Andrea & Fernie, 2013) instead of explicating in depth what to document and how to capture it. As a consequence, the understanding of the variety of documentation has remained rather rudimentary (Gant & Reilly, 2017). Further, the interest in the matter has tended to limit to solving confined sets of questions within particular disciplines. In archaeology, much of the past and on-going focus has been directed to the documentation of archaeological 3D data (e.g. Bentkowska-Kafel et al., 2012; D'Andrea & Fernie, 2013; Greengrass, 2008; Kolenda & Markiewicz, 2017; Niccolucci et al., 2013), in social science to survey research (Edwards et al., 2017) and in the information field, there is a considerable corpus of research on automatic tracing of the provenance of digital datasets (e.g. Missier, 2016; Curcin, 2016; De Oliveira et al., 2015; da Cruz et al., 2011; Davidson & Freire, 2008). At the same time, it is evident that the benefits of comprehensive documentation of data creation, curation and use processes are not limited only to these specific disciplines or types of research data.

Regarding paradata, after briefly touching upon the conceptual issue of framing what it entails in the current section, in the following, this text proceeds to discuss the issues of how to represent it, what should be captured and documented, and where this information could be available. Of these questions, especially the critical understanding of *how* to capture the qualitative context and processes of data creation and use (including what actions were taken, why and according to what premises and when data was created or modified) and *what specifically needs to be captured* about the context in order to facilitate data sharing and reuse is still in its infancy.

5 Representation as a relative non-problem

The first question relating to work with paradata, and at least for data curators the most obvious one, is how to represent it. Considering that much of data documentation-related work has focused on developing systems and schemes for representing various forms of data, the primary problem with paradata might not be the lack of means to represent it as Huggett (2012) remarks, or how to manage it (cf. Niven & Richards, 2017). This does not mean, of course, that it would not be necessary to investigate further and determine what approaches function best (e.g. Huvila, 2016, 2011; Vatanen, 2004) and when and how such data could and should be managed in practice (Bentkowska-Kafel et al., 2012).

As a whole, the available general-purpose ontologies such as the CIDOC-CRM in the cultural heritage domain and comparable schemes in other fields often have the potential to accommodate for formal representation of paradata. Moreover, their functionality can be easily expanded using extensions like the CRMDig extension for representing "steps and methods of production ('provenance') of digitization products and synthetic digital representations" (Doerr et al., 2016), or 3D object-specific CARARE 2.0 model (D'Andrea & Fernie, 2013) demonstrate. Similarly, for representing contextual information, there are already several conceptual schemes developed for representing structured descriptions of specific aspects related to the origins and movement of data (for a partial survey, see Romary et al., 2017). Examples of such schemes in the humanities are the provenance information in the OASIS model and CEDARS metadata standard (Day, 2002), the earlier mentioned CRMDig extension of the CIDOC-CRM (Doerr et al., 2016; Doerr & Theodoridou, 2011) for documenting (especially the technical/physical) aspects of the production of digitisation

products and synthetic digital representations. Also, the CRMsci (Doerr et al., 2014b) extension on data about scientific observation, measurements and processed data, and CRMinf (Stead & Doerr, 2015) for argumentation and inference making provide means to represent processes. The datatype-specific STARC metadata model for documenting 3D digital media files (Ronzino et al., 2012), the Music (MEI) and Text (TEI) Encoding Initiatives (Wilson, 2020), and the CARARE 2.0 model for documenting 3D cultural heritage data (D'Andrea & Fernie, 2013) do also incorporate fields for documenting such information similarly to the generic PROV and PREMIS metadata standards for digital preservation (Li & Sugimoto, 2014) and such discipline-specific frameworks as the UK MIDAS heritage data standard (English Heritage, 2012) with elements for describing actors (e.g. persons) and their involvement (e.g. as a surveyor or excavator) in heritage documentation work, and the Extended Matrix and Extended Matrix Framework for formal documentation of scientific processes underpinning archaeological virtual reconstructions (Demetrescu & Fanini, 2017). In addition to general standards and models, multiple technical (e.g. Doerr et al., 2014a for the management of 3D models) and institutional frameworks (e.g. the CARARE model links to Europeana) and tools (e.g. in the context of 3D-COFORM project, Niccolucci, 2010; EM/EMF, Demetrescu & Fanini, 2017) incorporate features that are relevant for managing and maintaining paradata related information. Also, the recent and ongoing data policy work that emphasises the implementation of adequate measures for representing, documentation and management of information on different aspects of data necessary for enabling its reuse, aligns (Hollander et al., 2017) with the requirements of administering paradata.

Finally, on a micro-level, the possibility to accommodate for local and domain-specific vocabularies and attribute data implemented in current data format specifications (e.g. national standards like SIKB0102: Boasson & Visser, 2017 and CSGDM: Metadata Ad Hoc Working Group, 1998; internal formats of individual software packages like in Intrasis, Rosén et al., 2017) opens up for using them in storing paradata. Also, such conventional techniques such as static relational schemes developed for specific needs (e.g. Mi & Pollock, 2017), and unstructured free-text fields (e.g. as in Battini & Sorge, 2017) can be adequate in some cases (e.g. especially for human readers, Huggett, 2016) even if they have apparent shortcomings regarding, respectively, their flexibility and the consistency of descriptions. Lastly, to the long list of potential means to represent paradata could be added also the formal techniques for modelling, representing and sharing stepwise descriptions of specific scientific tasks (e.g. MyExperiment <http://www.myexperiment.org/>; VisTrails, <http://www.vistrails.org/>; Ludaescher et al., 2013; Davidson & Freire, 2008).

An evident conclusion of such close to a daunting list of examples of potential means to represent paradata and paradata-like data is that it appears as a relative non-problem. The proliferation of standards and schemes does not necessarily mean that they help to document all necessary aspects of data creation, curation and use, or that the standards would be easy enough to use. In parallel, the abundance of standards raise questions of the contextuality of descriptions and how much an individual researcher or curator needs to know about specifics of data creation, curation and use processes and their context to create and understand descriptions, what and how to describe. Still, it seems plausible to suggest that the most critical issues in solving the problem of documenting data creation, curation and use processes lie at least partially somewhere else.

6 The problem of capturing what

In contrast to the relative proliferation of at least potential formats and means to represent it (Huggett, 2016), the question of *what* (and on what premises) specific types of information (i.e., content) can be useful as paradata for those who share and reuse data has remained more of a conundrum. There are several reasons that make identifying and preserving relevant paradata difficult. Many of these issues can be traced back to the discipline-specific characteristics of information processes and scholarly information (Huvila, 2019) that complicate their documentation and management. A prominent obstacle is the heterogeneity of data collection and documentation processes in many fields. Even if the major steps undertaken during a research study would follow a fairly linear pattern (Hawthorne, 2021), studies of scientific and scholarly documentation and work practices provide plenty of evidence of the underlying variety and messiness of how the work proceeds even in cases where standardisation has been championed as a major priority (e.g. Bowker, 1994; Law & Lynch, 1988). The uniformity of practices does obviously vary from one case and context to another. For example in archaeology, documentation processes tend to be more uniform in commercial development-

led archaeology than in academic archaeological research (Wallrodt, 2016). Similarly, quantitative survey research is perhaps on average more standardised than qualitative observational studies. Partly, even if some work has been done to develop means to document scholarly processes (e.g. Antonijevic & Cahoy, 2018; Chiquet, 2020), social science and humanities disciplines generally lack equivalent to scientific workflows (Davidson & Freire, 2008) applied in many laboratory sciences that could be used directly as a basis for developing a framework for systematic capturing of paradata. The problem is, however, deeper than developing and implementing a workflow system. Even if workflows and procedural standardisation could work in research that stems from an episteme that is compatible with such forms of regulation, the practical conditions of research in many fields, prevailing epistemic paradigms and their diversity effectively deny the possibility to force all research to be conducted according to a small set of rigorously standardised workflows.

A related practical problem to the variety and complexity of data creation, curation and use practices is that of the granularity and limits of what can be documented. For data creators, it is impossible to document everything and for both data creators and users, it is very difficult to anticipate what prospective users really need to know in the future. There is also a large variation in what researchers consider important to document even within small and tight groups (e.g. Akmon et al., 2011). For example, interview researchers seldom describe in detail the interview situation beyond what was said aloud even if it is not unthinkable that a future researcher might need to know details about the physical setting of interviews and for example, how the interviewer acted physically in the situation to understand interviewees' reactions and comments. Yet requiring complete documentation is naive. A field researcher can impossibly document every step taken and all the conscious and unconscious decisions made when collecting data. It is both it is fair to accept that similarly to all data on data (Mayernik & Acker, 2018), also paradata is bound to be incomplete.

7 Where in the world is paradata? Or the problem of how and where

Similarly to the question of what should be documented and preserved as paradata, the issue of how to do it and where to find that data has remained far more open than the question of means to represent it. Considering the diversity of conceivable traces with a capability to inform about processes and doings, paradata is unquestionably a compound of evidence that needs to be gathered from multiple sources. In rough terms, the parallel sources and approaches to acquire paradata can be divided into three major categories on the basis of where and how it is acquired (Fig. 2). Further, as the paradata'ness of a specific piece of information depends on whether it can *function* as paradata in a specific situation rather than that a specific type of information would be paradata by its nature, the potentially relevant means to find and extract it are obviously largely the same than with all others forms of descriptive data and information, including metadata.

First, a lot of paradata and proto-paradata is already captured or could be captured (1) automatically by existing tools used by researchers, data curators and other data users and producers (see e.g. Kreuter, 2013). Digital cameras collect EXIF (meta)data, including focal length, shutter speed, exposure compensation, metering pattern, and date/time when a photograph was taken (Huggett, 2012) that—if put together—provides indicative evidence of photographing processes. In field research, standalone and integrated GPS devices can be used to collect x, y and z coordinates and the date/time when particular data points were collected. All of this type of automatically collectable data can be informative of decisions and the sequence of actions taken. Many computer applications preserve data on how and when data was manipulated and they can also provide aggregate information on data. With an appropriate set of tools, the manipulation and ingestion of this data can be (semi-)automatised and built-in technical infrastructures (e.g. Felicetti & Lorenzini, 2011). Further, for example, techniques used for automatic capturing of data provenance (Stamatogiannakis et al., 2015) can undoubtedly be applied in capturing 'forensic' evidence to inform of the processes of data creation, curation and use also beyond its mere provenance.

Second, data contains a lot of paradata that can be generated *post hoc* using different (2) 'forensic' approaches. For instance, photographic documentation of study objects in field sciences show often traces of researchers' actions, tools, weather and lighting conditions and more (e.g. Gant & Reilly, 2017). Similarly, diagrams reveal occasionally details of how they were drawn and written notes and the choice of words in documentation contain explicit and tacit references to how they and their underpinning work was done. The same applies to notes, memos and paper drafts that often

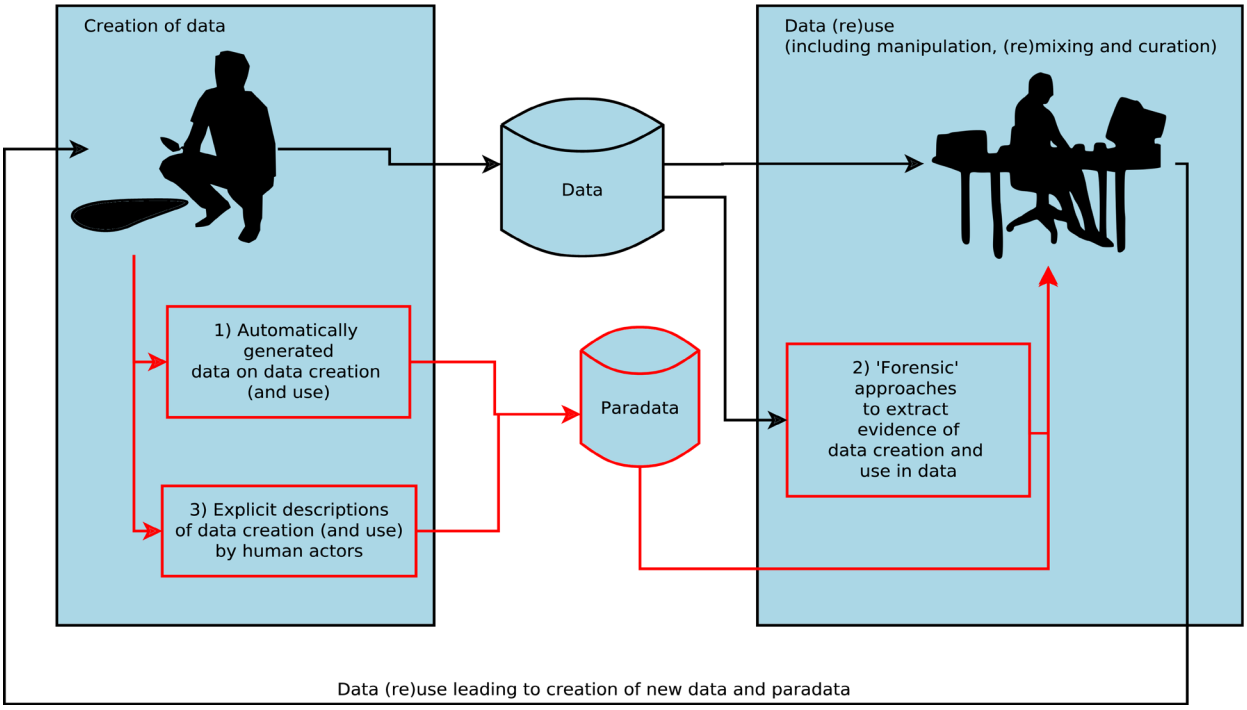

Figure 2: Sources of paradata in the process of data creation and (re)use (data flow in black, paradata flow in red colour).

contain cues of the process that led to final publications, products and outcomes (Latour, 1992; MacLeod & Nersessian, 2016).

Third, the (3) manual and semi-automatic making of paradata (or paradata-like-data) is already a part of certain research and data curation processes. Examples of such practices are, for instance, anthropological and archaeological field documentation, which is generally expected to contain information on both findings and the investigation process (e.g. Gustafsson & Magnusson Staaf, 2001; Huvila, 2011; Rytter et al., 2020; Marsh et al., 2020). Keeping notebooks is typical also in laboratory sciences (Klokmose & Zander, 2010) and increasingly in data-intensive research (Wofford et al., 2020; Graham, 2017). Even if articulating (Carr, 2005; Fook et al., 1997) and sharing (Hinds & Pfeffer, 2003) procedural expertise is difficult, people do not necessarily know why and how they know what they know (Bouwman et al., 1987), and the accuracy and reliability of explicit descriptions of work and research processes and their usefulness in facilitating data reuse have been questioned (e.g. Schloen, 2001; Button & Harper, 1996), it would be too hasty to belittle their value as paradata, especially when used in combination with other information (e.g. Henninger, 2018; Power et al., 2017). Rather than their quality and usefulness as such, their apparent major shortcoming is the time and resources required to produce meticulous documentation.

It is obvious that these brief examples of where paradata can be acquired during and after data creation provide only a small glimpse of all possible sources of such information. To this end, more evidence-based research is needed. In parallel, it is apparent that neither purposefully generated nor coincidental evidence alone provides a comprehensive understanding of data creation, curation and use. In some respects, there is a dire need for more information but especially for a better understanding of what information is already and first then, and in addition to what Huggett (2012) suggested already some time ago, deciding what else needs to be documented, and how it can be done.

8 Working with paradata

In the pursuit of a comprehensive—if not complete—body of paradata through using available sources and complementing them with new information, the two critical questions are what would be applicable means to produce useful paradata,

and what makes paradata useful for its different conceivable users. The mere fact that a lot of information on data creation, curation and use is available and preserved does not mean that the current practices of paradata production are necessarily appropriate for the task or that the resulting paradata would be easy to digest and exploit. The same applies to the practices of curating and using paradata.

9 Making useful paradata

As in general with all documentation work (Huggett, 2012), finding a workable way forward to produce useful paradata calls for recognising its value and incorporating its capturing and preservation as a taken-for-granted part of everyday (work) practices of both data creators, curators and users. Partly, it is important for everyone to improve and develop the current understanding of what paradata is, where it can be found, how it can be best captured and what means are best suited to represent particular types of paradata. Partly, as with provenance metadata in sciences (cf. Davidson & Freire, 2008), it is necessary to connect different types of information that qualifies as and contributes to paradata originating from diverse sources and make them comprehensible and usable for its intended users. As West and Sinibaldi (2013) observe, different sources and mechanisms of processing data and generating paradata have their distinct advances but also their respective sources of errors and omissions. Combining paradata from different sources is especially crucial with digital data, which is intricately embedded with the infrastructure that carries them and makes it difficult to distinguish digital services, data-carrying documents and their contents. In contrast to suggestions that by definition, all paradata can be collected automatically (Drachsler et al., 2012), a comprehensive understanding of the underpinnings of data requires a much more complex body of evidence (Huggett, 2012; Gant & Reilly, 2017) than limiting it to the tracking of the curatorial provenance of datasets within specific digital information infrastructures (cf. Mannheimer et al., 2014). At the same time, however, there is no reason to disparage the significance of the possibility of capturing automatically some of the required information or that some paradata can already be found as a part of other metadata and the data itself.

A parallel obvious possibility to automatic paradata generation is to develop current note-taking and documentation practices in such contexts (e.g. fieldwork, laboratories) where paradata-like-data is already produced. In archaeology, the use of videos, diaries and the introduction of 'daily sketches' at Çatalhöyük (Berggren et al., 2015) serves as an example of some of these possibilities – even if some of them would be unattainable for many less well-resourced research projects. Similar examples of creative use of field notes, diaries and summaries and development of new digital tools and genres can also be found in many other disciplines from laboratory work to industrial research (Schöpfel, 2019; Randles et al., 2017; Bowker, 1994). Data provenance research has suggested automated and semi-automated means of automatic and non-invasive real-time (e.g. Curcin, 2016; Pinheiro et al., 2013) and post-hoc tracking the movements of data implemented by, for instance, dynamic instrumentation (Stamatogiannakis et al., 2015) and citation analysis (Geser & Selhofer, 2015). Recent studies have also shown that it is possible to conduct text-mining on free-form field notes to extract information from them (Power et al., 2017). It is plausible that to a certain extent, similar approaches can be useful in extracting paradata.

Further, as noted earlier, data tends to include pieces of information that contain or can be useful in creating paradata. Even if the limitations of post-hoc extraction of evidence come with multiple limitations all too well-known to everyone who has ever conducted historical research, developing techniques and approaches for close reading of data can help to disclose more implicit evidence of how it came into being. There are many examples (incl. Star & Griesemer, 1989; Bowker, 1994; Pickering, 1995, see also e.g. Giulierini et al., 2020) demonstrating that it is not especially far fetched to suggest that approaches similar to those used by codicologists, palaeographers and art historians can be applied to research documentation as well. Here the key actors are data curators and future users of data as facilitators and performers of this work. Henninger (2018) has demonstrated by combining different datasets documenting the same project (in her case, archaeological excavation finds) that it is possible to find evidence on how researchers are reasoning in the field. Similarly, the automatically captured information on time and location of when and where measurements and photographs have been taken can be useful as a starting point in tracing when and where significant intellectual decisions and choices have been made. The imprints of tools in visual documents (Gant & Reilly, 2017) is another comparable source of information. The traces of, for example, how an ethnologist has moved her pen when

drawing a sketch, an archaeologist has used a trowel to clean a surface, a historian has read a text or an art historian directed her attention to different features of a painting are informative of the intellectual processes of how the studied object was conceptualised.

Similarly to other approaches of acquiring paradata, there are opportunities to advance the explicit documentation of on-going work. A lot of work could be done to incentivise explicit 'articulation work' (Suchman, 1996) i.e. documentation of work processes in data curation and reuse especially for and within research fields whenever there is a reason to consider that their meticulous documentation is relevant. Similarly, because the documentation of work practices, both in research and generally, is seldom considered a priority (Raghu & Vinze, 2007; Stenmark, 2000) and often means to an end (new knowledge) rather than the main task of those who are expected to produce such records, it would be important to decide when such documentation is not worthwhile. Moreover, there is undoubtedly room for developing increasingly rigorous means and approaches to documentation and taking notes. For instance, there is a clear lack of intuitive and unobtrusive tools for collecting structured paradata according to events (e.g. D'Andrea & Fernie, 2013; Niccolucci et al., 2013), workflow (Davidson & Freire, 2008), and argumentation (Vatanen, 2004) based models of documenting activities. A closer consideration of matching documentation approach with the objects being documented (Vatanen, 2003), what Ben-David has recently called a ''structural fit' between medium and method' (Brügger et al., 2019), would be equally important. Finally, there are also undoubtedly a lot of opportunities to develop new and existing types of 'documents' for capturing and representing paradata. For example, digital 3D models have been proposed as a "holistic instrument of describing, recording and disseminating" (Hauck & Kuroczyński, 2015) objects of scholarly interest. This would mean conceiving them as capable of incorporating a comprehensive set of meta and paradata within the object itself.

What a brief walk-through of various thinkable approaches to conceiving, capturing and creating paradata shows is, however, that similarly to how the representation of paradata is probably less a problem than figuring out what it might be, the most crucial question concerning the making of paradata is to advance the understanding of what means are usable and useful in particular contexts of data creation and use. There are many possibilities but it is not evident that they are always feasible or lead to meaningful paradata in every conceivable context and situation.

10 Making paradata useful

In addition to the need to develop means how to capture and document as useful paradata as possible, also the current understanding of its use has limitations to an extent that it is equally crucial to put a serious effort in elaborating the premises of what makes paradata useful. This is a task for (para)data creators but also for data curators and reusers to consider. At the present, as West and Sinibaldi (2013) note in the context of survey research, the quality of existing paradata varies a lot and it is not always very useful. A close reading of earlier literature prompts to suggest at least three different ways forward:

1. Formal paradata is good for computers while the best paradata does not necessarily look like 'data' at all for its human users;
2. Documentation and documented processes need to align with each other not least because they shape each other;
3. Useful documentation needs to consider the scholarship in its complete complexity;
4. Be comprehensive as the usefulness and eventual uses of paradata are likely to be impossible to anticipate.

Considering the first bullet point and what researchers and other users of data do, the customary approach to try to understand the context and use of data tends to be, if available, to turn to textual documentation (Huggett, 2016; Faniel et al., 2013). There is currently little evidence of how different forms of structured paradata would eventually be consumed and how they would work in practice. As Huggett (2016) suggests for metadata, it is not very far-fetched to suggest that the principal value of structured paradata is in its machine-readability by computational tools rather than in being directly usable or useful for human beings. This might suggest that in many cases the best paradata is not 'data' at all in any conventional sense of describing and understanding what qualifies as data. Partly, information that functions as paradata can be features and choice of words that only an experienced insider of a specific research community can half-intuitively interpret. Partly, useful information can be embedded in unstructured narratives, or in

relations and patterns in the data itself that can be accessible only through computational analysis. At the same time, however, studies of data provenance suggest that human readers have found specific types of structured paradata useful in their work (e.g. Asuncion, 2013). Data provenance research demonstrates further the possibility to utilise dedicated computational data analysis and visualisation tools for enquiring into (para)data (e.g. De Oliveira et al., 2015; Curcin, 2016), (re-)generating human-readable accounts of structured paradata (e.g. Gratzl et al., 2016), and in some cases, even reproduce workflows (Ludaescher et al., 2013). However, as a whole, a reasonable premise is that all findable, accessible, interoperable and reusable (cf. FAIR, 2016) (para)data should, can and need not be—at least directly—at the same time both human and machine-readable.

When considering the usefulness of paradata and the second bullet point, an important question is how the documentation supports and fits into the practices and procedures of how researchers conduct their work as a whole. Capturing and documenting paradata is after all, to a significant degree a social and organisational rather than a mere technical issue. In addition to an imminent risk of producing relevant data that is still lacking pertinence, the producing and using—and the mere availability—of paradata is likely to affect scholarly work in many unanticipated and intricate senses beyond providing a better understanding of the processual origins of a specific pile of legacy data. The literature does not lack examples of how the infrastructures of scholarly information (Huggett, 2016) and practices of documenting science (Bowker, 2005) from pro-forma context sheets in field sciences (e.g. Yarrow et al., 2008; Lucas, 2001; Star & Griesemer, 1989) to notebooks (Holmes et al., 2003; Kanza, 2018) in laboratory work have shaped researchers' work for both good and bad.

Considering the third bullet point, while some of the relative (in)compatibilities between paradata and their uses might go unnoticed, it is likely that many of them do not. Researchers tend to be happy when they have a considerable degree of freedom to organise and manage their information as they themselves find relevant. In contrast, an external mandate is often found problematic not only in principle but also in practice (e.g. Kanza, 2018; Huvila, 2012). As Turner puts it when discussing the issue regarding visual cultural heritage information, "[a]t present, these are just technological solutions allowing a method for storing and connecting with the extra data/information that will be required. A true solution to these issues requires an understanding of the complete system, including the human visual system, the researchers' wishes and interpretation, as well as the users' wishes" (Turner, 2012, p. 142). The experienced lack of comprehensive consideration and understanding of the complete sociotechnical and socio-material system composed of not only technologies but also of human actors described by Turner is not specific to visualisations but is common to scientific and scholarly work at large.

Considering the fourth bullet point, the question of producing, preserving and providing useful paradata is not only about avoiding unwanted side effects and catering for explicit wishes but also helping with implicit, unarticulated and unknown needs. As information science research demonstrated already decades ago, users do not (necessarily) know very well what information they really need to succeed in their pursuits (Frické, 2012). As all data on data (Mayernik & Acker, 2018), also paradata is a phenomenon that is embedded in infrastructures as a necessary substrate that makes them work. Therefore, it is possible that the usefulness of paradata is not only and primarily in how it can address the immediate and explicit needs of its individual users but in providing a broad and comprehensive enough understanding of the origins of the data it describes.

Finally, the fifth bullet point directs attention to that in addition to capturing and creating paradata, a crucial part of the process of making paradata useful is the presence of adequate curation and management procedures. While the making and identification of paradata is largely a task for data creators and reusers, the curation and management is primarily a task for data curators and stewards. The management would be in the best case not only administering and keeping it but active facilitation of its reuse and perhaps the development of scholarly data editions comprising of data and contextual information—including paradata—for diverse purposes. As these purposes vary and change, this work does not end but needs to be followed in the continuum of research data (cf. records continuum, McKemmish, 2016) where its relevance and usefulness is pluralised again and again.

11 Why better paradata matters?

Even if the documentation of data creation, curation and use undoubtedly has certain intrinsic value, paradata matters primarily due to its implications across academic and professional practice, and theory. Process information on how research data is created, curated and used helps discipline-specific research but especially work that traverses disciplinary boundaries in multiple senses summarised in Figure 3.

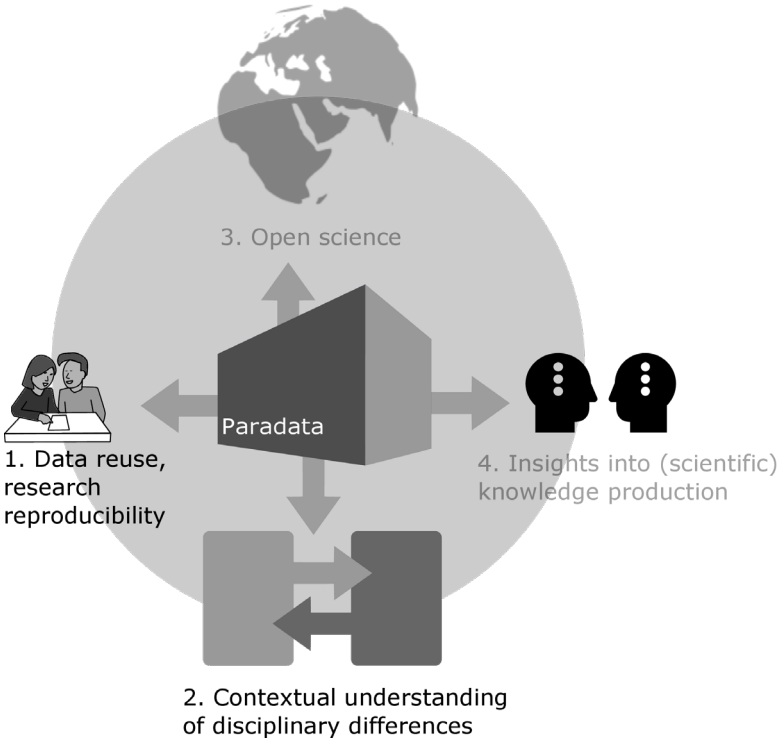

Figure 3: Why paradata matters.

First, a better grasp of data creation, curation and use practices and how they can be captured provides new knowledge and perspectives to how data has been, is and can be (re-)used (Geser & Niccolucci, 2016, see also Harrison et al., 2017). It provides keys to understanding how data (re)use could, and according to its different stakeholders, should be supported (Aloia et al., 2017) and how to facilitate sharing of data by describing how the data has been used earlier by others. When relevant, similarly to provenance information (Missier, 2016), paradata can also contribute to the reproducibility of scientific and scholarly processes by disclosing how earlier work on the data has been performed. The literature on the applications of paradata in survey research provides perhaps the most explicit examples of the practical benefits of identifying, collecting and keeping paradata for understanding data collection practices and evaluating and reusing data (examples e.g. in Edwards et al., 2017).

Second, better paradata holds a promise to open perspectives to the contextualities of research data (cf. Harrison et al., 2017), disciplinary differences, and benefit disciplinary and cross-disciplinary use of research materials. As Geser and Niccolucci (2016) note, while the general principles of providing open access to research data are not specific to individual disciplines, there are disciplinary conditions that have to be taken into account. Closer scrutiny of paradata helps to make explicit what Geser and Niccolucci (2016, p. 54) describe as 'the ways research data are being generated, used and published'. It can also help to foster a common ground they ask for the integration of research infrastructures by making visible and increasing understanding of unique characteristics of scholarly work in different disciplines. As

such it can also help to provide a basis for comparisons and collaborations in a broad interdisciplinary context (urged e.g. by Lambourne et al., 2014; Harrison et al., 2017) within and beyond humanities and social sciences.

Third, a proper understanding and documentation of paradata is also a central premise for the successful implementation of current visions of open science (e.g. Beck & Neylon, 2012; Mendez et al., 2020) and the contemporary open data policies (e.g. DCC, 2017). Moreover, as noted earlier, paradata is a key constituent in the process of opening research data in the broad intra- and cross-disciplinary knowledge ecosystem in the sense of making it not only available but also intelligible (cf. Bruseker et al., 2017b). The proliferation of poorly documented data and inadequate understanding of how to document it properly not only impedes sharing and reusing data but makes it difficult for researchers to see the advantages of open scholarship. Better paradata provides much-needed 'context' (as in Meyers, 1993) to data and can shed light on "dark data" (Geser & Niccolucci, 2016) that is difficult to find and access.

Fourth and finally, especially for science and information studies research, paradata can help to provide new understanding of scholarly knowledge production and in a broader sense, how scholarship is achieved in terms of a 'socio-techno-informational' (Eriksson-Backa et al., 2021) practice. It links to the research agenda advocated by Dallas (2015) to conduct in-depth research on research practices and methods, and to devise and develop curation linked to existing pervasive digital infrastructures from the premises of how research works in specific domains, disciplinary and cross-disciplinary contexts. Even if such a perspective that essentially looks research from outside might appear as somewhat peripheral from the point of view of the research field under scrutiny, it can contribute to the much-wanted evidence-based self-understanding of both established research practices and their implications, and perhaps, especially to the development of new and emerging cross- and sub-disciplinary constellations (cf. e.g. Luhmann & Burghardt, 2022; Malazita et al., 2020) including those emerging from and within digital humanities and social sciences.

12 Conclusions

Without proper documentation of the human processes of creating, understanding and interpreting data objects, there is a risk of creating and archiving large collections of data that are incapable of supporting research and other types of reuse, and even worse, leading researchers and others to work and conduct research on faulty premises and drawing erroneous conclusions on data that has been created under incompatible bases.

Even if the importance of understanding the social and intellectual premises of data infrastructures, including digital archives and libraries, has been acknowledged for some time as a necessary premise of their sustainability and usefulness, this article posits that there is a serious lack of 1) systematic understanding of how researchers work when they are creating information and data, and 2) practicable measures of capturing and documenting this work that would provide enough context for the forthcoming users of the data and not too cumbersome and vague for the data creators, curators and users to understand and use as a part of their daily work. Much work has been done to develop and implement automated methods for tracing movements and transformations of provenance data in digital systems (Missier, 2016; Malik et al., 2010; da Cruz et al., 2011) but what appears to be missing, is a comprehensive understanding of the processes of creation and use in the larger sociotechnical or socio-techno-informational context that goes beyond tracing digital workflows and the transformations and movements of data in digital repositories.

As it seems, a major obstacle – perhaps the most critical one at the moment – to solving the issue of capturing and preserving relevant paradata is, firstly, finding ways to cope with the different needs of different data users have in different contexts and situations. In this respect, an equally critical question to having enough of it is how to avoid having too much paradata and being as creative as possible in identifying and extracting information with the potential to function as paradata in already existing documentation and the data itself rather than engaging in large-scale production of new (para)data. Similarly to research data, also paradata and the understanding of data creation, curation and use are still first and foremost instrumental to the principal rationale of scientific and scholarly work of creating new knowledge. Therefore, secondly, there is a dire need for methods of collecting existing and easily capturable information rather manually creating paradata. Hoarding and keeping paradata for its own sake is as counter-productive as failing to collect and retain enough of it. The gravity of the problem has become especially evident with the proliferation of data-intensive research and big data analysis even if, as a whole, the need to understand data-related processes is equally crucial to the success of small-scale scholarship, conventional interpretative studies and

practice-oriented knowledge transfer to the society. The problem is wicked in the sense that it is difficult to pin down, there are no clear limits to the amount of information to capture, the correctness of solutions is difficult to establish, and situatedness of information needs makes the problem and solutions similarly situated. In practice, different users and uses of data and collections need to be taken properly into account when describing processes rather than assuming that descriptions are stable in time and space. At the same time, even if the risk of not being able to provide a blanket solution to the problem is high, the potential payback of making a fair effort to develop a better understanding of how to capture and document paradata is high in the context of all research relying on the use of secondary data.

Acknowledgements: The author would like to thank Dr. Lisa Börjesson and Dr. Olle Sköld for enlightening discussions on paradata and highly useful comments on the manuscript.

Author's Statements : Funding information: This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant agreement No 818210 as a part of the project CApturing Paradata for documenTing data creation and Use for the REsearch of the future (CAPTURE).

References

Adams, V., Burke, N. J., & Whitmarsh, I. (2014). Slow research: Thoughts for a movement in global health. *Medical Anthropology*, *33*(3), 179–197.

Akmon, D., Zimmerman, A., Daniels, M., & Hedstrom, M. (2011). The application of archival concepts to a data-intensive environment: working with scientists to understand data management and preservation needs. *Archival Science*, *11*(3), 329–348. http://dx.doi.org/10.1007/s10502-011-9151-4

Aloia, N., , Binding, C., Cuy, S., Doerr, M., Fanini, B., Felicetti, A., Fihn, J., Gavrilis, D., Geser, G., Hollander, H., Meghini, C., Niccolucci, F., Nurra, F., Papatheodorou, C., Richards, J., Ronzino, P., Scopigno, R., Theodoridou, M., Tudhope, D., Vlachidis, A., & Wright, H. (2017). Enabling european archaeological research: The ARIADNE e-infrastructure. *Internet Archaeology*, (43).

Anderson, C. M., & Larson, S. (2013). Marginalia, or the knowledge of excess. *Social & Cultural Geography*, *14*(7), 739–743. https://doi.org/10.1080/14649365.2013.811535

Antonijevic, S., & Cahoy, E. S. (2018). Researcher as bricoleur: Contextualizing humanists' digital workflows. *DHQ*, *12*(3). http://www.digitalhumanities.org/dhq/vol/12/3/000399/000399.html

Asuncion, H. U. (2013). Automated data provenance capture in spreadsheets, with case studies. *Future Generation Computer Systems*, *29*(8), 2169–2181.

Ball, A. (2011). International digital curation conference 2010. *Ariadne*, (66). http://www.ariadne.ac.uk/issue66/idcc-2010-rpt/

Battini, C., & Sorge, E. (2017). Amphitheater of Volterra: case study for the representation of the excavation data. *Studies in Digital Heritage*, *1*(2), 269–281.

Bauer-Clapp, H., & Kirakosian, K. (2017). Archaeologists and archives: Revisiting an old challenge. *Advances in Archaeological Practice*, *5*(3), 220–226.

Bauman, Z., & Raud, R. (2015). *Practices of selfhood*. Cambridge: Polity.

Beaudoin, J. E. (2012). A framework for contextual metadata used in the digital preservation of cultural objects. *D-Lib Magazine*, *18*(11-12).

Beck, A., & Neylon, C. (2012). A vision for Open Archaeology. *World Archaeology*, *44*(4), 479–497.

Benardou, A., Champion, E., Dallas, C., & Hughes, L. M. (2018). Introduction: a critique of digital practices and research infrastructures. In A. Benardou, E. Champion, C. Dallas, & L. M. Hughes (Eds.) *Cultural Heritage Infrastructures in Digital Humanities*. London: Routledge.

Bentkowska-Kafel, A., Denard, H., & Baker, D. (Eds.) (2012). *Paradata and transparency in virtual heritage*. Farnham: Ashgate.

Berggren, A., Dell'Unto, N., Forte, M., Haddow, S., Hodder, I., Issavi, J., Lercari, N., Mazzuccato, C., Mickel, A., & Taylor, J. (2015). Revisiting reflexive archaeology at Çatalhöyük: integrating digital and 3D technologies at the trowel´s edge. *Antiquity*, *89*(344), 433–448.

Bevan, A. (2015). The data deluge. *Antiquity*, *89*(348), 1473–1484.

Birnholtz, J. P., & Bietz, M. J. (2003). Data at work: Supporting sharing in science and engineering. In *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work*, GROUP '03, (pp. 339–348). New York: ACM. https://doi.org/10.1145/958160.958215

Boasson, W., & Visser, R. (2017). Sikb0102: Synchronizing excavation data for preservation and re-use. *Studies in Digital Heritage*, *1*(2), 206–224.

Borgman, C. L. (2012). The conundrum of sharing research data. *JASIST*, *63*(6), 1059–1078.

Borgman, C. L. (2015a). *Big Data, Little Data, No Data : Scholarship in the Networked World*. Cambridge, MA: MIT Press.

Borgman, C. L. (2015b). If Data Sharing is the Answer, What is the Question? *ERCIM News*, (100), 15–16. http://ercim-news.ercim.eu/en100/special/if-data-sharing-is-the-answer-what-is-the-question

Borgman, C. L., Darch, P. T., Sands, A. E., & Golshan, M. S. (2016). The durability and fragility of knowledge infrastructures: Lessons learned from astronomy. In A. Grove, D. H. Sonnenwald, L. Harrison, C. Blake, C. Schlögl, I. Peters, B. Endler-Jobst, C. Cool, & Y.-L. Theng (Eds.) *ASIST 2016 Proceedings of the 79th ASIS&T Annual Meeting*. Silver Spring, MD: ASIS&T.

Börjesson, L., Sköld, O., & Huvila, I. (2020). The politics of paradata in documentation standards and recommendations for digital archaeological visualisations. *Digital Culture and Society*, *6*(2), 191–220.

Bouwman, M. J., Frishkoff, P. A., & Frishkoff, P. (1987). How do financial analysts make decisions? A process model of the investment screening decision. *Accounting, Organizations and Society*, *12*(1), 1–29.

Bowker, G. C. (1994). *Science on the run: information management and industrial geophysics at Schlumberger, 1920-1940*. Cambridge, Mass.: MIT Press.

Bowker, G. C. (2005). *Memory practices in the sciences*. Cambridge, MA: MIT Press.

Brocks, H., Kranstedt, A., Jäschke, G., & Hemmje, M. (2009). Modeling context for digital preservation. In E. Szczerbicki, & N. Nguyen (Eds.) *Smart Information and Knowledge Management*, (pp. 197–226). Berlin: Springer.

Brügger, N., Milligan, I., Ben-David, A., Gebeil, S., Nanni, F., Rogers, R., Turkel, W. J., Weber, M. S., & Webster, P. (2019). Internet histories and computational methods: a 'round-doc' discussion. *Internet Histories*, *3*(3-4), 202–222.

Bruseker, G., Carboni, N., & Guillem, A. (2017a). Cultural heritage data management: The role of formal ontology and CIDOC CRM. In M. L. Vincent, V. M. López-Menchero Bendicho, M. Ioannides, & T. E. Levy (Eds.) *Heritage and archaeology in the DigitalAge: acquisition, curation, and dissemination of spatial cultural heritage data*, (pp. 93–131). Cham: Springer.

Bruseker, G., Doerr, M., & Theodoridou, M. (2017b). *D5.1. Report on the Common Semantic Framework*. PARTHENOS.

Buchanan, S. A. (2016). *A Provenance Research Study of Archaeological Curation*. Ph.D. thesis, The University of Texas at Austin, Austin.

Button, G., & Harper, R. (1996). The relevance of work-practice for design. *Computer Supported Cooperative Work*, *4*(4), 263–280.

Carr, S. M. (2005). Knowing nursing - the challenge of articulating knowing in practice. *Nurse Education in Practice*, *5*(6), 333–339.

Chiquet, V. (2020). Supporting sustainable digital data workflows in the art and humanities. *Sharing the Experience: Workflows for the Digital Humanities. Proceedings of the DARIAH-CH Workshop 2019 (Neuchâtel), DARIAH-CAMPUS.*.

Clemens, R., & Graham, T. (2007). *Introduction to Manuscript Studies*.. Ithaca, NY: Cornell University Press.

Conkey, M. W., & Gero, J. M. (1997). Programme to practice: Gender and feminism in archaeology. *Annu. Rev. Anthropol.*, *26*(1), 411–437.

Cooper, A., & Green, C. (2015). Embracing the complexities of 'big data' in archaeology: the case of the English Landscape and Identities project. *J Archaeol Method Theory*, *23*(1), 271–304.

Couper, M. (1998). Measuring survey quality in a CASIC environment. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*.

Couper, M. P. (2000). Usability evaluation of computer-assisted survey instruments. *Social Science Computer Review*, *18*(4), 384–396.

Couper, M. P., & 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.

Curcin, V. (2016). Embedding data provenance into the learning health system to facilitate reproducible research. *Learning Health Systems*, *1*(2), e10019.

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

Dallas, C. (2015). Curating Archaeological Knowledge in the Digital Continuum: from Practice to Infrastructure. *Open Archaeology*, *1*(1), 176–207.

D'Andrea, A., & Fernie, K. (2013). Carare 2.0: A metadata schema for 3d cultural objects. In *2013 Digital Heritage International Congress (DigitalHeritage)*, (pp. 137–143). New York: IEEE.

Davidson, S., & Freire, J. (2008). Provenance and scientific workflows: Challenges and opportunities. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, (pp. 1345–1350). New York: ACM.

Day, M. (2002). *Cedars Guide to Preservation Metadata*. Bath: UKOLN. http://www.ukoln.ac.uk/metadata/cedars/guidance/metadata.html

DCC (2017). *An Analysis of Open Data and Open Science Policies in Europe, May 2017*. Apeldoorn: SPARC Europe & DCC.

De Oliveira, D., Silva, V., & Mattoso, M. (2015). How much domain data should be in provenance databases? In *Workshop on Theory and Practice of Provenance (TaPP)*.

Dell'Unto, N., Landeschi, G., Apel, J., & Poggi, G. (2017). 4D recording at the trowel's edge: Using three-dimensional simulation platforms to support field interpretation. *Journal of Archaeological Science: Reports*, *12*, 632–645.

Demetrescu, E., & Fanini, B. (2017). A white-box framework to oversee archaeological virtual reconstructions in space and time: Methods and tools. *Journal of Archaeological Science: Reports*, *14*(Supplement C), 500–514.

Demoule, J.-P. (2011). We still have to excavate - but not at any price. *Archaeological Dialogues*, *18*(1), 5–10.

Doerr, M., Chrysakis, I., Axaridou, A., Theodoridou, M., Georgis, C., & Maravelakis, E. (2014a). A framework for maintaining provenance information of cultural heritage 3d-models. In *Proceedings of the EVA London 2014 on Electronic Visualisation and the Arts*, EVA London 2014, (pp. 267–274). London: BCS.

Doerr, M., Kritsotaki, A., Rousakis, Y., Hiebel, G., & Theodoridou, M. (2014b). *CRMsci: the Scientific Observation Model An Extension of CIDOC-CRM to support scientific observation*. Heraklion: FORTH.

Doerr, M., Stead, S., & Theodoridou, M. (2016). *Definition of the CRMdigAn Extension of CIDOC-CRM to support provenance metadata*. Heraklion: FORTH., version 3.2.1 ed.

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

Doran, M., Edmond, J., & Nugent-Folan, G. (2019). Reconciling the cultural complexity of research data: Can we make data interdisciplinary without hiding disciplinary knowledge? *Preprint of manuscript submitted to CODATA*. http://hdl.handle.net/2262/83156

Drachsler, H., Greller, W., Fazeli, S., Niemann, K., Sanchez-Alonso, S., Rajabi, E., Palmér, M., Ebner, H., Simon, B., Nösterer, D., et al. (2012). *D8. 1 Review of Social Data Requirements*. Open Discovery Space project.

Drude, S., di Giorgio, S., Ronzino, P., Links, P., van Nispen, A., Verbrugge, K., Degl'Innocenti, E., Stiller, J., Oltersdorf, J., & Spiecker, C. (2016). *D2.1. Report on User Requirements*. PARTHENOS.

Duranti, L. (1998). *Diplomatics : new uses for an old science*. Lanham, MD: Scarecrow Press.

Durrant, G. B., D'Arrigo, J., & 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.

Edwards, R., Goodwin, J., O'Connor, H., & Phoenix, A. (2017). *Working with Paradata, Marginalia and Fieldnotes*. Cheltenham: Edward Elgar.

Elliot, H., Edwards, R., Phoenix, A., & Boddy, J. (2015). *Narrative analysis of paradata from the Poverty in the UK survey: a worked example*. Southampton: NCRM.

English Heritage (2012). *MIDAS Heritage - The UK Historic Environment Data Standard, v1.1*. London.

Eriksson-Backa, K., Hirvonen, N., Enwald, H., & Huvila, I. (2021). Enablers for and barriers to using My Kanta - a focus group study of older adults' perceptions of the national electronic health record in Finland. *Informatics for Health and Social Care*.

European Strategy Forum on Research Infrastructures (2011). *Strategy Report on Research Infrastructures: Roadmap 2010*. Luxembourg: European Union.

Fagan, B. (1995). Timelines: Archaeology's dirty secret. *Archaeology*, *48*(4), 14–17.

FAIR (2016). *Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing*. FORCE11, version b1.0 ed. https://www.force11.org/fairprinciples

Faniel, I., Kansa, E., Whitcher Kansa, S., Barrera-Gomez, J., & Yakel, E. (2013). The challenges of digging data: a study of context in archaeological data reuse. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '13, (pp. 295–304). New York: ACM.

Faniel, I. M., Frank, R. D., & Yakel, E. (2019). Context from the data reuser's point of view. *Journal of Documentation*, *75*(6), 1274–1297.

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

Felicetti, A., & Lorenzini, M. (2011). Metadata and tools for integration and preservation of cultural heritage 3d information. *Geoinformatics FCE CTU*, *6*(0), 118–124.

Flexner, J. L. (2016). Dark and bright futures for museum archaeology. *Museum Worlds*, *4*(1), 1–3.

Fook, J., Ryan, M., & Hawkins, L. (1997). Towards a theory of social work expertise. *British Journal of Social Work*, *27*(3), 399–417.

Foster, A., & Rafferty, P. (Eds.) (2016). *Managing digital cultural objects: Analysis, discovery and retrieval*. London: Facet.

Friberg, Z., & Huvila, I. (2019). Using object biographies to understand the curation crisis: lessons learned from the museum life of an archaeological collection. *Museum Management and Curatorship*, *34*(4), 362–382.

Frické, M. (2012). *Logic and the Organization of Information*. New York: Springer.

Fry, J. (2006). Scholarly research and information practices: a domain analytic approach. *Information Processing & Management*, *42*(1), 299–316.

Gant, S., & Reilly, P. (2017). Different expressions of the same mode: a recent dialogue between archaeological and contemporary drawing practices. *Journal of Visual Art Practice*, *17*(1), 100–120.

Gerth, P., Sieverling, A., & Trognitz, M. (2017). Data curation: How and why. a showcase with re-use scenarios. *Studies in Digital Heritage*, *1*(2), 182–193. https://scholarworks.iu.edu/journals/index.php/sdh/article/view/23235

Geser, G., & Niccolucci, F. (2016). *D2.4: Final Innovation Agenda and Action Plan*. ARIADNE.

Geser, G., & Selhofer, H. (2015). *D2.3: Preliminary innovation agenda and action plan*. ARIADNE.

Giulierini, P., Coralini, A., & Calandra, E. (Eds.) (2020). *Miniere della memoria; scavi in archivi, depositi e biblioteche*. Sesto Fiorentino: All'Insegna del Giglio.

Goodwin, J., O'Connor, H., Phoenix, A., & Edwards, R. (2017). Introduction: working with paradata, marginalia and fieldnotes. In R. Edwards, J. Goodwin, H. O'Connor, & A. Phoenix (Eds.) *Working with Paradata, Marginalia and Fieldnotes*, (pp. 1–19). Cheltenham: Edward Elgar.

Graham, S. (2017). Open notebook research in digital archaeology. *Electricarchaeology.ca*. https://via.hypothes.is/https://electricarchaeology.ca/2017/03/15/open-notebook-research-in-digital-archaeology/

Gratzl, S., Lex, A., Gehlenborg, N., Cosgrove, N., & Streit, M. (2016). From visual exploration to storytelling and back again. *Computer Graphics Forum*, *35*(3), 491–500.

Greengrass, M. (2008). Introduction. In M. Greengrass, & L. M. Hughes (Eds.) *The virtual representation of the past*. Aldershot and Burlington: Ashgate.

Gustafsson, A., & Magnusson Staaf, B. (2001). Rapport om rapporter – en diskussion kring kvalitetsbedömningar av arkeologiska rapporter. Report 2001:3, RAÄ, Stockholm.

Harrison, R., Morel, H., Maricevic, M., & Penrose, S. (2017). *Heritage and Data: Challenges and Opportunities for the Heritage Sector: Report of the Heritage Data Research Workshop Held Friday 23 June 2017 at the British Library, London*. London: Heritage Research, Heritage Futures, Alan Turing Institute, British Library.

Hauck, O., & Kuroczyński, P. (2015). Cultural heritage markup language-designing a domain ontology for digital reconstructions. In *Proceedings of the 2nd International Conference on Virtual Archaeology, Hermitage Museum, St. Petersburg/Russia*, (pp. 250–255).

Hawthorne, S. (2021). An insider's take on data curation: Context, quality, and efficiency. *Journal of eScience Librarianship*, *10*(3).
Henninger, M. (2018). From mud to the museum: Metadata challenges in archaeology. *Journal of Information Science*, *44*(5), 658–670.
Hinds, P. J., & Pfeffer, J. (2003). Why organizations don't "know what they know": Cognitive and motivational factors affecting the transfer of expertise. In M. S. Ackerman, V. Pipek, & V. Wulf (Eds.) *Sharing expertise : beyond knowledge management*, (pp. 3–26). Cambridge, MA: MIT Press.
Hodder, I. (2003). Archaeological Reflexivity and the "Local" Voice. *Anthropological Quarterly*, *76*(1), 55–69. http://www.jstor.org/stable/3318361
Hollander, H., Morselli, F., Admiraal, F., Conrad, A., Trippel, T., Zeldenrust, D., Ronzino, P., Giorgio, S. D., Madonna, A. D., & Hedges, M. (2017). *D3.1. Guidelines for Common Policies Implementation*. PARTHENOS.
Holmes, F. L., Renn, J., & Rheinberger, H.-J. (Eds.) (2003). *Reworking the bench : research notebooks in the history of science*. Dordrecht: Kluwer.
Huggett, J. (2012). Promise and paradox: Accessing open data in archaeology. In C. Mills, M. Pidd, & E. Ward (Eds.) *Proceedings of the Digital Humanities Congress 2012*. Sheffield: Humanities Research Institute.
Huggett, J. (2016). Digital haystacks: Open data and the transformation of archaeological knowledge. In A. T. Wilson, & B. Edwards (Eds.) *Open Source Archaeology, Ethics and Practice*, (pp. 6–29). Berlin: de Gruyter Open.
Huvila, I. (2011). The politics of boundary objects: hegemonic interventions and the making of a document. *JASIST*, *62*(12), 2528–2539.
Huvila, I. (2012). Being Formal and Flexible: Semantic Wiki as an Archaeological e-Science Infrastructure. In M. Zhou, I. Romanowska, Z. Wu, P. Xu, & P. Verhagen (Eds.) *Revive the Past: Proceeding of the 39th Conference on Computer Applications and Quantitative Methods in Archaeology, Beijing, 12-16 April 2011*, (pp. 186–197). Amsterdam: Amsterdam University Press. http://dare.uva.nl/aup/nl/record/412958
Huvila, I. (2016). Awkwardness of becoming a boundary object: Mangle and materialities of reports, documentation data and the archaeological work. *The Information Society*, *32*(4), 280–297.
Huvila, I. (2019). Management of archaeological information and knowledge in digital environment. In M. Handzic (Ed.) *Knowledge Management, Arts and Humanities*, (pp. 147–169). Cham: Springer.
Kansa, E. C. (2016). Click here to save the past. In D. B. Counts, E. W. Averett, & J. M. Gordon (Eds.) *Mobilizing the past for a digital future : the potential of digital archaeology*, (pp. 443–472). Grand Forks, ND: Digital Press at the University of North Dakota. http://dc.uwm.edu/arthist_mobilizingthepast/
Kanza, S. (2018). *What Influence would a Cloud Based Semantic Laboratory Notebook have on the Digitisation and Management of Scientific Research?*. phdthesis, University of Southampton, Southampton.
Karasti, H., Baker, K. S., & Halkola, E. (2006). Enriching the notion of data curation in e-science: Data managing and information infrastructuring in the long term ecological research (LTER) network. *Computer Supported Cooperative Work (CSCW)*, *15*(4), 321–358.
Kelleher, C. (2017). Archives without archives:(re) locating and (re) defining the archive through post-custodial praxis. *Journal of Critical Library and Information Studies*, *1*(2), 1–30.
Kersel, M. M. (2015). Storage wars: Solving the archaeological curation crisis? *Journal of Eastern Mediterranean Archaeology & Heritage Studies*, *3*(1), 42–54.
Kim, Y., & Yoon, A. (2017). Scientists' data reuse behaviors: A multilevel analysis. *JASIST*, *68*(12), 2709–2719.
Kintigh, K. W., & Altschul, J. H. (2010). Sustaining the digital archaeological record. *Heritage Management*, *3*(2), 264–274.
Klokmose, C. N., & Zander, P.-O. (2010). Rethinking laboratory notebooks. In M. Lewkowicz, P. Hassanaly, V. Wulf, & M. Rohde (Eds.) *Proceedings of COOP 2010*, (pp. 119–139). London: Springer.
Kolenda, J., & Markiewicz, M. (2017). A medieval bishop's palace in Milicz: 3d reconstruction as a method of a research hypotheses presentation. *Studies in Digital Heritage*, *1*(2), 428–443.
Kowal, S., Koszewski, K., Słyk, J., & Wrona, S. (2017). Digital method for verifying archaeological hypotheses. medieval gord under Pułtusk castle. *Studies in Digital Heritage*, 1(2), 444–455. https://scholarworks.iu.edu/journals/index.php/sdh/article/view/23412
Krauwer, S., Giorgio, S. D., Verbrugge, K., Spiecker, C., Chatzimichail, C., Degl'Innocenti, E., Gautier, H., Giacomi, R., Henriksen, L., Illmayer, K., Madonna, A. D., & Sanesi, M. (2017). *D2.3 Report on the Assessment of Data Policies and Standardization*. PARTHENOS.
Kreuter, F. (2013). *Improving surveys with paradata analytic uses of process information*. Hoboken, NJ: Wiley.
Kreuter, F., & Casas-Cordero, C. (2010). Paradata. RatSWD Working Paper 136, RatSWD, Berlin.
Kristiansen, K. (2014). Towards a new paradigm? the third science revolution and its possible consequences in archaeology. *Current Swedish Archaeology*, *22*, 11–34.
Laflamme, F. (2009). Data collection research using paradata at Statistics Canada. In *Proceedings of Statistics Canada Symposium 2008: Data Collection: Challenges, Achievements and New Directions*.
Lake, M. (2012). Open archaeology. *World Archaeology*, *44*(4), 471–478.
Lambourne, G., Stoakes, L., Cassar, M., Balen, K. V., Rhisiart, M., Thomas, M., Miller, R., & Burnell, L. (2014). Strategic Research Agenda. Tech. rep., JPI Cultural Heritage and Global Change, Rome. http://www.jpi-culturalheritage.eu/wp-content/uploads/SRA-2014-06.pdf
Latour, B. (1992). 'where are the missing masses? the sociology of a few mundane artifacts. In W. E. Bijker, & J. Law (Eds.) *Shaping Technology/Building Society: Studies in Sociotechnical Change*, (pp. 225–258). Cambridge, MA: MIT Press.
Law, J., & Lynch, M. (1988). Lists, field guides, and the descriptive organization of seeing: Birdwatching as an exemplary observational activity. *Human Studies*, *11*(2/3), 271–303.
Lee, C. A. C. (2011). A framework for contextual information in digital collections. *Journal of Documentation*, *67*(1), 95–143. http://dx.doi.org/10.1108/00220411111105470

Leonelli, S., & Tempini, N. (Eds.) (2020). *Data Journeys in the Sciences*. Cham: Springer.
Li, C., & Sugimoto, S. (2014). Provenance description of metadata using prov with premis for long-term use of metadata. In *Proceedings of the 2014 International Conference on Dublin Core and Metadata Applications*, DCMI'14, (pp. 147–156). Dublin Core Metadata Initiative. http://dl.acm.org/citation.cfm?id=2771234.2771250
Löwenborg, D. (2014). Recuperating GIS data from excavations: On the use, or lack of use, of digital archaeological information. In I. Huvila (Ed.) *Perspectives to Archaeological Information in the Digital Society*, (pp. 11–23). Uppsala: Department of ALM, Uppsala University.
Lucas, G. (2001). *Critical Approaches to Fieldwork Contemporary and Historical Archaeological Practice*. London: Routledge.
Ludaescher, B., Cuevas-Vicenttín, V., Missier, P., Dey, S., Kianmajd, P., Wei, Y., Koop, D., Chirigati, F., Altintas, I., Belhajjame, K., & Bowers, S. (2013). Facilitating Scientific Research through Workflows and Provenance on the DataONE Cyberinfrastructure (Invited). *AGU Fall Meeting Abstracts*. http://adsabs.harvard.edu/abs/2013AGUFMIN53E..07L
Luhmann, J., & Burghardt, M. (2022). Digital humanities–a discipline in its own right? an analysis of the role and position of digital humanities in the academic landscape. *JASIST*, *73*(2), 148–171.
MacLeod, M., & Nersessian, N. J. (2016). Interdisciplinary problem- solving: emerging modes in integrative systems biology. *European Journal for Philosophy of Science*, *6*(3), 401–418.
Malazita, J. W., Teboul, E. J., & Rafeh, H. (2020). Digital humanities as epistemic cultures: How DH labs make knowledge, objects, and subjects. *DHQ*, *14*(3), 000465. http://www.digitalhumanities.org/dhq/vol/14/3/000465/000465.html
Malik, T., Nistor, L., & Gehani, A. (2010). Tracking and sketching distributed data provenance. In *2010 IEEE Sixth International Conference on e-Science*, (pp. 190–197).
Mannheimer, S., Yoon, A., Greenberg, J., Feinstein, E., & Scherle, R. (2014). A balancing act: The ideal and the realistic in developing Dryad's preservation policy. *First Monday*, *19*(8).
Marquardt, W. H., Montet-White, A., & Scholtz, S. C. (1982). Resolving the crisis in archaeological collections curation. *American Antiquity*, *47*(2), 409–418.
Marsh, D. E., Punzalan, R. L., & Johnston, J. A. (2020). Preserving anthropology's digital record: Copar in the age of electronic fieldnotes, data curation, and community sovereignty. *The American Archivist*, *82*(2), 268–302.
Mayernik, M. S., & Acker, A. (2018). Tracing the traces: The critical role of metadata within networked communications. *JASIST*, *69*(1), 177–180.
McCoy, M. D. (2017). Geospatial big data and archaeology: Prospects and problems too great to ignore. *Journal of Archaeological Science*, *84*, 74–94. http://www.sciencedirect.com/science/article/pii/S0305440317300821
McKemmish, S. (2016). Recordkeeping in the continuum: An Australian tradition. In A. J. Gilliland, S. McKemmish, & A. J. Lau (Eds.) *Research in the Archival Multiverse*, (pp. 122–160). Clayton, Victoria, Australia: Monash University Publishing.
Mendez, E., Lawrence, R., MacCallum, C. J., Moa, E., & alii (2020). *Progress on Open Science: Towards a Shared Research Knowledge System: Final Report of the Open Science Policy Platform*. Brussels: European Commission.
Metadata Ad Hoc Working Group (1998). *Content Standard for Digital Geospatial Metadata*, vol. FGDC-STD-001-1998. Washington, D.C.: Federal Geographic Data Committee.
Meyers, T. B. (1993). *Preserving the archaeological record: An archives management program for archaeological documentation*. phdthesis, University of Missouri, Columbia.
Mi, X., & Pollock, B. M. (2017). Metadata schema to facilitate linked data for 3d digital models of cultural heritage collections: A university of south florida libraries case study. *Cataloging & Classification Quarterly*, (pp. 1–14).
Milligan, C., Littlejohn, A., & Margaryan, A. (2014). Workplace learning in informal networks. In A. Littlejohn, & C. Pegler (Eds.) *Reusing Open Resources: Learning in Open Networks for Work, Life and Education*, (pp. 93–103). New York: Routledge.
Missier, P. (2016). The lifecycle of provenance metadata and its associated challenges and opportunities. In V. L. Lemieux (Ed.) *Building Trust in Information*, (pp. 127–137). Cham: Springer.
Morgan, C., & Eve, S. (2012). DIY and digital archaeology: what are you doing to participate? *World Archaeology*, *44*(4), 521–537.
Mostafa, J. (2018). Documents and (as) machines. *JASIST*, *69*(1), 3–5. http://dx.doi.org/10.1002/asi.23993
Moulin, C., Ciula, A., Kelleher, M., Mittler, E., Tadic, M., Ågren, M., Bozzi, A., & Kuutma, K. (2011). Research Infrastructures in the Digital Humanities. Science Policy Briefing 42, European Science Foundation, Strasbourg.
Münster, S., Pfarr-Harfst, M., Kuroczyski, P., & Ioannides, M. (2016). *3D Research Challenges in Cultural Heritage II : How to Manage Data and Knowledge Related to Interpretative Digital 3D Reconstructions of Cultural Heritage*. Cham: Springer.
Newman, J. (2012). Revisiting Archive Collections: Developing Models for Participatory Cataloguing. *Journal of the Society of Archivists*, *33*(1), 57–73.
Niccolucci, F. (2010). Technologies, standards and business models for the formation of virtual collections of 3d replicas of museum objects: The 3d-coform project. In *2010 IST-Africa*, (pp. 1–8).
Niccolucci, F., Felicetti, A., Amico, N., & D'Andrea, A. (2013). Quality control in the production of 3d documentation of monuments. In *Built Heritage 2013 Monitoring Conservation Management*, (pp. 864–873). M: Politecnico di Milano.
Nicolaas, G. (2011). Survey paradata: A review. Discussion paper, NCRM, Southampton.
Niven, K., & Richards, J. D. (2017). The storage and long-term preservation of 3d data. In D. Errickson, & T. Thompson (Eds.) *Human remains: another dimension : the application of imaging to the study of human remains*, (pp. 175–184). London: Academic Press.
Olivier, L. (2015). *The dark abyss of time: archaeology and memory*. Lanham, MD: Rowman & Littlefield.
Olson, K. (2013). Paradata for nonresponse adjustment. *The Annals of the American Academy of Political and Social Science*, *645*(1), 142–170.

Pančur, A., Šorn, M., & Erjavec, T. (2018). SlovParl 2.0: The collection of Slovene parliamentary debates from the period of secession. In D. Fišer, & a. d. J. Maria Eskevich (Eds.) *ParlaCLARIN 2018 Workshop Proceedings May 7, 2018 Phoenix Seagaia Conference Center, Miyazaki, Japan*. Luxembourg: ELRA.

Pickering, A. (1995). *The Mangle of Practice: Time, Agency, and Science*. Chicago: University of Chicago Press.

Pinheiro, R., Holanda, M., Araujo, A. P. F., Walter, M. E., & Lifschitz, S. (2013). Automatic capture of provenance data in genome project workflows. In *2013 IEEE International Conference on Bioinformatics and Biomedicine*. IEEE.

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

Poole, A. H., & Garwood, D. A. (2018). Interdisciplinary scholarly collaboration in data-intensive, public-funded, international digital humanities project work. *Library & Information Science Research*, *40*(3), 184–193. http://www.sciencedirect.com/science/article/pii/S0740818817303511

Power, C., Lewis, A., Petrie, H., Green, K., Richards, J., Eramian, M., Chan, B., Walia, E., Sijaranamual, I., & Rijke, M. D. (2017). Improving archaeologists' online archive experiences through user-centred design. *ACM Journal on Computing and Cultural Heritage*, *10*(1), 3:1–3:20.

Prainsack, B., Schicktanz, S., & Werner-Felmayer, G. (Eds.) (2014). *Genetics as social practice : transdisciplinary views on science and culture*. Farnham: Ashgate.

Raghu, T. S., & Vinze, A. (2007). A business process context for knowledge management. *Decision Support Systems*, *43*(3), 1062–1079.

Randles, B. M., Pasquetto, I. V., Golshan, M. S., & Borgman, C. L. (2017). Using the jupyter notebook as a tool for open science: An empirical study. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, (pp. 1–2).

Romary, L., Banski, P., Bowers, J., Degl'Innocenti, E., Ďurčo, M., Giacomi, R., Illmayer, K., Joffres, A., Khan, F., Khemakhem, M., Larrousse, N., Litke, A., Monachini, M., van Nispen, A., & Ogrodniczuk, M. (2017). *D4.2 Report on Standardization*. PARTHENOS.

Ronzino, P., Hermon, S., & Niccolucci, F. (2012). A metadata schema for cultural heritage documentation. In V. Capellini (Ed.) *Electronic Imaging & the Visual Arts: EVA 2012*, (pp. 36–41). Florence: Firenze University Press.

Rosén, C., Gainsford, S., Toreld, C., Alfsdotter, C., Carlstein, C., & Öbrink, M. (2017). The Nya Lödöse project. strategies and context. *International Journal of Historical Archaeology*.

Rytter, M., Andersen, A. O., Rubow, C., Kusk, M. L., Dalsgård, A. L., & Nielsen, M. (2020). *Anthropology Inside Out: Fieldworkers Taking Notes*. Canon Pyon: Sean Kingston Publishing.

Schloen, J. (2001). Archaeological Data Models and Web Publication Using XML. *Computers and the Humanities*, *35*(2), 123–152.

Schöpfel, J. (2019). Grey literature and professional knowledge making. In L. Börjesson, & I. Huvila (Eds.) *Research Outside The Academy : Professional Knowledge-Making in the Digital Age*, chap. 137-153. Basingstoke: Palgrave MacMillan. https://doi.org/10.1007/978-3-319-94177-6

Schwartz, J. M., & Cook, T. (2002). Archives, Records, and Power: The Making of Modern Memory. *Archival Science*, *2*, 1–19.

Stamatogiannakis, M., Groth, P., & Bos, H. (2015). Looking inside the black-box: Capturing data provenance using dynamic instrumentation. In *Provenance and Annotation of Data and Processes: 5th International Provenance and Annotation Workshop, IPAW 2014, Cologne, Germany, June 9-13, 2014. Revised Selected Papers*, (pp. 155–167). Cham: Springer.

Star, S. L., & 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.

Stead, S., & Doerr, M. (2015). *CRMinf: the Argumentation Model - An Extension of CIDOC-CRM to support argumentation*. Purley: Paveprime, version 0.7 ed.

Stenmark, D. (2000). Leveraging tacit organizational knowledge. *Journal of Management Information Systems*, *17*(3), 9–24.

Stieger, S., & 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.

Suchman, L. (1996). Supporting Articulation Work. In R. Kling (Ed.) *Computerization and Controversy: Value Conflicts and Social Choices*, (pp. 407–423). Academic Press, second ed.

Tempini, N. (2017). Till data do us part: Understanding data-based value creation in data-intensive infrastructures. *Information and Organization*, *27*(4), 191–210. http://www.sciencedirect.com/science/article/pii/S1471772716300653

The London Charter Organisation (2009). The London Charter for the computer-based visualisation of cultural heritage. Tech. Rep. 2.1, London.

Turner, M. J. (2012). Lies, damned lies and visualizations: Will metadata and paradata be a solution or a curse? In A. Bentkowska-Kafel, H. Denard, & D. Baker (Eds.) *Paradata and transparency in virtual heritage*, (pp. 135–143). Farnham: Ashgate.

Vatanen, I. (2003). Deconstructing the (re)constructed: issues on annotation of the archaeological virtual realities. In M. Doerr, & A. Sarris (Eds.) *CAA2002 The Digital Heritage of Archaeology. Computer Applications and Quantitative Methods in Archaeology. Proceedings of the 30th Conference, Heraklion, Crete, April 2002*, (pp. 69–74). Archive of Monuments and Publications, Hellenic Ministry of Culture.

Vatanen, I. (2004). Argumentation paths in Information Infrastructure of the Archaeological virtual realities. In M. der Stadt Wien Referat Kulturelles Erbe Stadtarchäologie Wien (Ed.) *Enter the Past - The E-way into the Four Dimensions of Cultural Heritage. CAA 2003. Computer Applications and Quantitative methods in Archaeology. Proceedings of the 31st Conference, Vienna, Austria, April 2003. (On the accompanying CD-ROM)*, vol. 1227 of *BAR International Series*. Oxford: Archaeopress.

Voss, B. L. (2012). Curation as research. a case study in orphaned and underreported archaeological collections. *Archaeological Dialogues*, *19*(2), 145–169.

Wallrodt, J. (2016). Why paperless: Technology and changes in archaeological practice, 1996–2016. In D. B. Counts, E. W. Averett, & J. M. Gordon (Eds.) *Mobilizing the past for a digital future : the potential of digital archaeology*, (pp. 33–50). Grand Forks, ND: Digital Press at the University of North Dakota. http://dc.uwm.edu/arthist_mobilizingthepast/

West, B. T., & Sinibaldi, J. (2013). The quality of paradata: A literature review. In *Improving Surveys with Paradata: Analytic Uses of Process Information*. Wiley.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018.

Wilson, E. A. (2020). Text encoding with the text encoding initiative (TEI) and the music encoding initiative (MEI). In E. A. Wilson (Ed.) *Digital Humanities for Librarians*, (pp. 87–106). Lanham; Boulder; New York; London: Rowman & Littlefield.

Wofford, M. F., Boscoe, B. M., Borgman, C. L., Pasquetto, I. V., & Golshan, M. S. (2020). Jupyter notebooks as discovery mechanisms for open science: Citation practices in the astronomy community. *Computing in Science & Engineering*, *22*(1), 5–15.

Yarrow, A., Clubb, B., & Draper, J.-L. (2008). Public Libraries, Archives and Museums: Trends in Collaboration and Cooperation. IFLA Professional Reports 108, IFLA, The Hague.

Yoon, A. (2016). Data reusers' trust development. *JASIST*, *68*(4), 946–956. https://doi.org/10.1002/asi.23730