1

Introduction

Paradata – Documenting Data Creation, Management and Use

Isto Huvila

1.1 Introduction

The datafication of social life and economy dominates the contemporary experience. It has been claimed, not without reason, to concern 'everything' from the personal self to working life, freetime, scientific and scholarly knowledge-making and civic life (Mascheroni 2020; Millington and Millington 2015). Similarly it has been portrayed both as an opportunity (Mayer-Schönberger et al. 2013) and a potential threat underpinned by a particular ideology that Van Dijck (2014) aptly terms dataism. Data is to an increasing extent used both as a substitute and augmentation of physical things when making decisions and creating new knowledge. Even if repeatedly criticised, more data is routinely equated with better. As a consequence, the quantity of data in circulation has burgeoned and the demands to extract more value out of data through effective reuse have increased. For many, data has become a commodity and resource beyond anything else. Its potential to contribute to economic gain, cementing political power, is difficult to underestimate.

Symptomatic of the surge of data and dataism in contemporary existence, there is no one generally accepted definition of what it entails. The observation that a major obstacle to research data management is the multiplicity of understandings of what counts as 'data' is illustrative of the difficulty in finding a consensus. Data can be many things. In this volume that delves into data and its practical underpinnings, our working definition is purposefully broad. For us, data is all kinds of material – or colloquially, 'the stuff' – used in knowledge production. Different scientific and scholarly disciplines and practical everyday contexts understand it often in more specific, both implicit

and explicit, terms. The purpose of our working definition is to be inclusive of them all.

The rapidly growing reliance on data across society has made it increasingly apparent that just having data or information might not be enough to manage or utilise it. Data can neither be trusted by default nor is it intelligible in and of itself. Even if there are tendencies to frame data as raw, there is nothing raw or neutral about it. Making sense of data, and making it useful and manageable, requires understanding of what the data is all about. Otherwise the vast repositories of data become, as Parisi (2021) fears, a 'purposeless purpose' of the colossal enterprise of contemporary data making. In addition to knowing what data is *about*, it is equally vital to know where it comes from and how it has been processed, manipulated and used. The importance of understanding data-related practices and processes for successful data management and reuse has propelled researchers and professionals in different disciplines across science and scholarship from information and computer sciences to knowledge management and organisational learning to inquire into how to describe, document and communicate processes of how data – and in relation to data, information and knowledge – have come about and been processed, managed and used (e.g., Amairia 2023; Edwards et al. 2017; Faniel et al. 2019). As Barrowman argues, 'the very production of data is . . . always relevant to its interpretation' (Barrowman 2018 p. 133). Understanding how data is produced and reproduced in different phases of its lifespan is crucial for its reuse independent of the purpose of using data to create new knowledge, reproduce and validate earlier knowledge-making efforts, or to create trust. In parallel to developing a comprehensive understanding of data-related practices and processes and practical means to document and describe them for conveying the understanding to others, there is a critical need for theorising the phenomenon, for empirical research, identifying existing and forming of new methods and tools, technologies and infrastructures, standards and practices, and for developing novel concepts on practice and process information.

Paradata, one of the key concepts used to describe information about practices and processes, surfaced independently in multiple disciplinary contexts (Davet et al. 2023; Edwards et al. 2017; Huvila 2022a; Sköld et al. 2022). The aim of this book is to provide a first comprehensive overview of the concept and phenomenon of paradata from data management and knowledge organisation perspectives and an introduction to its practical and theoretical implications for research and practice.

The volume introduces the notion of paradata, how and why it can be a useful concept to consider and use to describe data about data creation, curation and use. It also expands the conceptual discussion to inquire into

practical aspects of how better understanding of the practices and processes creation, management and use of data can contribute to knowledge creation and data, information and knowledge management within and across disciplines to support data sharing and (re)use, and understanding of data practices and their consequences and implications. To support paradata practice, the volume also provides an overview of a selection of methods for capturing and documenting, for managing, and for extracting and employing paradata for data creators and (re)users, and data managers. Finally, the volume develops and synthesises the current state of the art of theory and practice in a paradata reference model to inform understanding of the paradata phenomenon and outlines directions for future research and practice in paradata, and future empirical research into scientific and scholarly information work.

The principal focus of our exposé is on research-related paradata, that is, paradata that is generated and/or used in research. Our understanding of research is broad and goes beyond academic knowledge-making covering corresponding professional and non-academic practices (Börjesson and Huvila 2019). However, even if our main interest is research, many of the insights in what paradata can be and how it works are relevant in a much broader array of contexts that operate with diverse types of data. Especially in a theoretical sense, their ambition and pertinence extends even beyond data to how paradata as a concept and lens can contribute to understanding the workings of knowledge extracted from and incorporated in practices and processes.

To exemplify the importance of understanding data-related practices and processes, consider having a simple spreadsheet filled with several thousands of rows of numbers (Figure 1.1). They are all decimals and a closer look reveals that they vary between 1.3 and 4.43. The filename SmithJ_appendix3 might suggest that they form an appendix to something, and someone called J. Smith had something to do with either the appendix or main product. However, this does not get you very far. Having descriptive metadata that explains that the file contains measurements of the volume of pottery vessels in litres helps a little and knowing what types of vessels and where they were found might help you a lot more. But without paradata including, for instance, explanation of how the volume was measured, whether the vessels were (unlikely) whole or if the measurements were done on the basis of retrieved or preserved sherds, how experienced the measurer was, and how carefully and for what purpose the measurements were taken, the spreadsheet or ‘dataset’, might not be that useful at all. In the worst and not especially hypothetical case, to have a (re)usable dataset, it might be necessary to try to find the original pottery vessels if they still exist, and measure them again using a

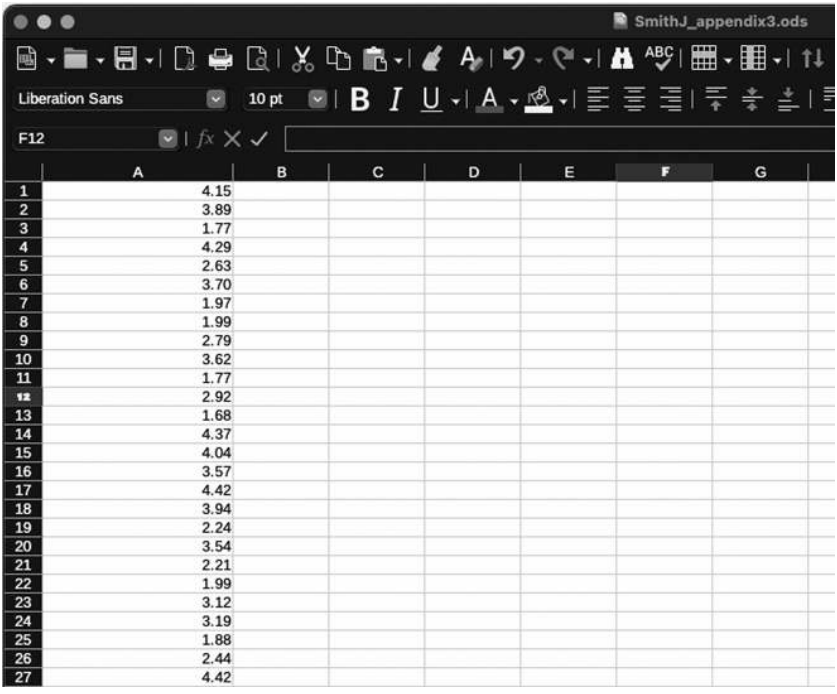

Figure 1.1 A spreadsheet with little explicit paradata.

method and level of accuracy adequate for the purpose. Remaking data takes a lot of time and effort. When this endeavour is multiplied with the escalating number of datasets and other potentially (re)usable assets produced and preserved in the contemporary datafied world, it becomes untenable.

1.2 The Concept of Paradata

As per defining paradata, we start with a working definition of paradata as information about the practices and processes of data making, management and use. It is worth noting that the definition is broader than the first conceptualisation of paradata in survey research that focused on automatically generated metainformation (Couper 1998, 2017) and even broader than such later definitions that have focused on decisions rather than practices and processes relating to data generation as a whole (Rabinowitz 2019). We distinguish information and data from knowledge in an analytical sense as its building blocks, respectively as matters used to inform (information) and typically

construct information (data), while referring to knowledge as being held by human-beings. However, unlike the popular data-information-knowledge-wisdom pyramid, we are not suggesting that information or knowledge would be directly reducible to data (or paradata) (cf. Fricke 2009; Huvila 2022b). By referring to making, management and use this volume aims to capture the full temporal and contextual continuum of data incorporating all types of moves, makings and changes data might end up facing.

Before proceeding to Chapter 2 with a comprehensive overview of paradata and central related concepts, it is useful to point out that neither previous perspectives on paradata nor the one offered in this volume are singular. Rather than trying to force paradata into one form, we have focused on identifying what different things function as paradata, how they do so and when. Paradata from a standards or systems perspective differs from a policy or principle-driven data management perspective. Also, the theorising of data making, management and use has an impact on what makes paradata and when. Paradata on a practice – as conceptualised from the perspective of practice theory – is not necessarily the same as paradata would be for someone approaching it from the perspective of activity theory. To emphasise that paradata is an applicable term across multiple conceptual framings of data-related doings, we write about information relating to data making, management and use the (admittedly somewhat clumsy) notion of 'practices and processes' instead of only one term. We do this in an attempt to capture two of the most prominent conceptualisations used in the literature, namely data practices and processes. Even more importantly, we do this to underline how paradata has relevance across the entire conceptual landscape.

1.3 Perspectives on Paradata

In parallel to approaching paradata as an a priori multifaceted concept, this volume also combines several analytical perspectives. The overall perspective is ours, that is, of the group of authors with background in a spectrum of scholarly disciplines, from information and archival studies to archaeology and cultural research, with an interest in understanding paradata from an information perspective as a form of information relating to informational doings. Paradata is approached from a knowledge organisation perspective as a means for 'describing, representing, filing and organizing documents, document representations, subjects and concepts both by humans and by computer programs' (Hjørland 2016 p. 475). As a form of information on making, managing and using data, paradata unfolds from the perspective of information

behaviour research (Bates 2015) as a referent to (informational) undertakings. This opens up to an inquiry on what paradata is and should be about (Chapter 2), and how paradata is dealt with in different contexts and situations (Chapters 3, 4, 5). We are also interested, from an information retrieval perspective, in how paradata ultimately makes data creation, management and use processes more accessible and intelligible (Chapter 7). From an archival studies viewpoint, we discuss the longevity and preservation of paradata for use and re-use not only in the present but in the near and long-term future (Chapter 6; Edquist 2014). Finally, a fifth complementary perspective draws from professional and personal (research) data management and its interest in storing and managing data for use and re-use as an asset in research and work practices (Chapters 6, 7, 8).

The background of this volume is the research project (funded by the European Research Council) CApturing Paradata for documenTing data creation and Use for the REsearch of the future (CAPTURE). The empirical focus of the project has been archaeology, chosen as an appropriate multi-disciplinary context with a plethora of parallel epistemic perspectives and knowledge interests and where utilising diverse and fragmented data is a commonplace. This focus, even if the curiosity of CAPTURE was not limited to archaeology and archaeological paradata alone, explains why also this volume contains multiple examples from that domain, however, together with others from a wide range of scientific, scholarly and professional contexts from archives, records and knowledge management to survey research and public administration.

We have endeavoured to make this work relevant for both researchers and practitioners across scholarly and professional disciplines: from researchers of paradata to those interested in documenting their data, making sense of others' data, developers of data management systems, data managers, experts and learners alike. However, while venturing in this direction, it is important to acknowledge that the differences in the perspectives of those who work with paradata are not merely a matter of adapting the tone of writing. The baseline we have been confronted with throughout this volume is that the perspectives of data creators, users and managers are not necessarily reconcilable. Borgman (2016) pointed to this direction in her 'Not fade away' paper referring to three perspectives with mutually competing ideals of research data reuse. A social scientist emphasises innovation and novel findings. A data librarian is concerned with optimising the production of well-documented reusable data. At the same time, a policymaker is concerned with timely release of traceable and discoverable datasets for other actors to use for societal benefit. The incompatibility of ideals means in practice that when pondering what is

preferable, good or bad, throughout the following chapters, there is always more than one perspective to consider and a fair chance that, for example, pursuing for manageability or timeliness narrows the usability of data for truly novel discoveries but also that prioritising comprehensive documentation takes time and resources perhaps beyond what is reasonable.

The practical interest in paradata does apparently also vary from the context and field of data practices and processes to another. Paradata has obvious resonance in explicitly data intensive research and work from sciences and technology to medicine and humanities, big data, and even more so in disciplines like data science and digital humanities and social sciences with an episteme that forthrightly builds on 'data'. However, with an intentionally broad take on paradata, the relevance of this book and paradata extends to fields where data is conventionally conceptualised as, for example, material, sources, evidence or documents.

1.4 How to Read This Book

We address different target groups in the different chapters. Some chapters are specifically targeted towards repository managers (Chapter 6), others data creators (Chapter 4), data reusers (Chapter 5), researchers interested in paradata theory (Chapter 2) and practice (Chapter 3). For this reason, the chapters differ in character and style but are hopefully readable both individually or as a part of the book-long narrative.

After this introduction, Chapter 2 provides an overview of the concept of paradata and selected related concepts as discussed in the existing research literature. It delves into the complexity and diversity of the notion to show its broad applicability in both research and practice. Chapter 3 continues to explore the manifestations of paradata by providing an overview of where it can be found. The chapter draws from empirical research conducted in the CAPTURE project focusing on archaeological documentation but also makes ample reference to other disciplinary contexts to exemplify how paradata can be context-specific but also how there can be structures that are technically and epistemically pervasive across domains. Such archetypes function as markers and help to identify commonplace documentation types and 'genres' of how paradata is recorded and embedded in data and data documentation.

In comparison to the beginning of the volume, Chapters 4 to 6 focus on providing practical and conceptual advice for readers considering how to tackle paradata in their work. Chapter 4 showcases a set of methods and

approaches for researchers and other data makers on how to plan for creating and capturing paradata in advance before data making takes place and how to generate and document paradata when datasets are in the making. Chapter 5 presents a similar selection of methods for data reusers, which can be used for finding and identifying implicit and explicit paradata in already existing datasets and data documentation where such information is not necessarily clearly marked as being paradata. In Chapter 6, the focus is on data managers. Rather than presenting a comparable set of methods to Chapters 4 and 5, it sets forth a framework of strategies for managing paradata on a scale from standardised, structured descriptors to material with varying degrees of potential to inform users of data-related practices and processes.

The final two chapters of this volume embark on comprehensive theorisation and reflection of what paradata is and how it works. Chapter 7 presents a paradata reference model with an aim of providing a framework for conceptualising paradata and how it links to practices and processes, and the human working knowledge used to engage with them. Rather than providing an objectivist account of paradata as a thing, the model emphasises the processual nature of paradata as being made and being in the making on a continuum of data being constantly produced and reproduced. At the closing of the volume, Chapter 8 discusses the premise of making paradata matter, paths for future research, and emphasises the need for ethical commitment when pursuing the documentation of data making.

1.5 Conclusion: Paradata Matters

As a final introductory note, it is fair to disclose already at the outset that the general premise of this volume is that paradata matters. Put simply, we posit that without paradata it is questionable if there can be meaningful data reuse. Having said that, we also want to underline at the very beginning that paradata takes many forms that are difficult to find and identify. Also, as the following chapters demonstrate, people still do a lot without knowing what 'needs' to be known, circumventing paradata, and without knowing what they know that they 'need' to know. People also adapt their data reuse to what is possible, changing their behaviour to match what they find doable with the available paradata. Paradata is also quite obviously difficult. Paradata, the work required to make paradata and to make it useful is a part of the invisible work of rendering (Ellingsen and Monteiro 2003) that often remains under the radar and is genuinely demanding to support. What this implies though is not that

paradata does not matter but rather that its complexity makes it both challenging and crucial to understand.

References

Amairia, A. (2023). De la donnée à la connaissance, de l'intelligibilité à l'action [From data to knowledge, from intelligibility to action]. *Communication & Organisation*, **64**(2), 147–160.

Barrowman, N. (2018). Why data is never raw. *The New Atlantis*, (56), 129–135.

Bates, M. J. (2015). Information behavior. In *Encyclopedia of Library and Information Sciences, 3rd ed.*, CRC Press.

Borgman, C. L. (2016). Not Fade Away: Social Science Research Data in the Digital Era, Presented at the Knowledge Rules: Curating Knowledge in the Social SciencesSocial Sciences Research Council Meeting, 2 May 2016, New York Public Library, New York. Retrieved from https://escholarship.org/uc/item/3ps9p9rc

Börjesson, L. and Huvila, I. (2019). Introduction. In Börjesson L. and Huvila, I. (eds.), *Research outside the Academy: Professional Knowledge-Making in the Digital Age*, Cham: Springer International Publishing, 1–19.

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. (2017). Para deta gainen no tanjo to fukyu / Birth and diffusion of the concept of paradata. *Shakai To Chosa (Advances in Social Research)*, **18**, 14–26.

Davet, J., Hamidzadeh, B. and Franks, P. (2023). Archivist in the machine: Paradata for AI-based automation in the archives. *Archival Science*, **23**(2), 275–295.

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

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

Fricke, M. (2009). The knowledge pyramid: A critique of the DIKW hierarchy. *Journal of Information Science*, **35**(2), 131–142.

Hjørland, B. (2016). Knowledge organization (KO). *Knowledge Organization*, **43**(4), 475–484.

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

Huvila, I. (2022b). Making and taking information. *JASIST*, **73**(4), 528–541.

Mascheroni, G. (2020). Datafied childhoods: Contextualising datafication in everyday life. *Current Sociology*, **68**(6), 798–813.

Mayer-Schönberger, V. and Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*, London: John Murray.

Millington, B. and Millington, R. (2015). 'The datafication of everything': Toward a sociology of sport and big data. *Sociology of Sport Journal*, **32**(2), 140–160.

Parisi, L. (2021). Instrumentality. In Thylstrup, N. B. Agostinho, D. Ring, A. D'Ignazio, C. and Veel, K. (eds.), *Uncertain Archives: Critical Keywords for Big Data*, Cambridge, MA: MIT Press, 289–298.

Rabinowitz, A. (2019). Communicating in three dimensions: Questions of audience and reuse in 3D excavation documentation practice. *Studies in Digital Heritage*, **3**(1), 100–116.

Sköld, O., Börjesson, L. and Huvila, I. (2022). Interrogating paradata. *Information Research*, **27**(special issue). https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-490374

Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance & Society*, **12**(2), 197–208.