Documentation of data making, processing and use facilitates future reuse of research data: the CAPTURE project

Isto Huvila[1] and Stefan Ekman,[1]

[1] *Department of ALM, Uppsala University, Thunbergsvägen 3H, Uppsala, Sweden*

Abstract

Reuse of research data requires knowing what the data is about but also of how it was created and previously processed, interpreted and used. The major challenges in capturing enough – but not too much – such process information, termed *paradata*, are to know what to document and how to document it in adequate detail and form. This paper showcases research and findings from the ERC-funded research project CAPTURE, which develops in-depth understanding of how paradata is being created and used today and which elicits and explores methods for capturing paradata. From a research infrastructure perspective, the most challenging question for managing paradata is how to enable and support the creation of paradata that is sufficient, relevant for its future reusers, and not too labour-intensive to produce and maintain. Considering the significant extent to which paradata is coincidental and exists because of the lack of data cleaning and management, a major challenge is also how to strike a balance between too much and too little standardisation.

Keywords

paradata, research data, research process, process documentation, research data management

1. Introduction

Reuse of research data requires not only knowing what the data is about but also a comprehensive understanding of how the data has been created and previously processed, interpreted and used (e.g. [1,2]). Without sufficient documentation of data, we risk ending up in a digital dark age [3] where hard-to-(re)use "dark data" dominates [4]. In the worst case, lack of context around the creation of archaeological research data – both digital and analog – can lead to substandard datasets that are difficult to interpret. Such datasets may not support future research and the creation of new archaeological knowledge to a sufficient extent.

The research project CApturing Paradata for documenTing data creation and Use for the REsearch of the future (CAPTURE) is funded by the European Research Council. It investigates the previously fairly unexplored question of exactly what information about the creation, management, and use of research data is necessary for the data to be reusable in the future. CAPTURE also examines how this information can be captured in a way that is both efficient and comprehensive enough to support data reuse. The empirical focus of the project is archaeology. As a transdisciplinary and paradigmatically diverse field that operates with a broad range of data types, from textual evidence to measurements, visual information, and physical evidence, it provides an outstanding context to delve into the complexities of data documentation.

In the project, data about data creation and manipulation processes is referred to as paradata [5]. The concept, first introduced in survey research to refer to data that describes or concerns processes [6], was introduced in the early 2000s in cultural-heritage visualisation research through the London and Seville principles [7], which stipulate fundamentals for the documentation of visual (re)presentations in archaeological and cultural heritage-related contexts. More recently, paradata has been instituted in information and data management [8] and AI and records management contexts [9].

Huminfra Conference 2024, Gothenburg, 10-11 January 2024.

isto.huvila@abm.uu.se (I. Huvila); stefan.se.ekman@abm.uu.se (S. Ekman)

The aim of this paper is to highlight the significance of capturing, documenting, and preserving research-data-related paradata in research infrastructures and, on the basis of the ongoing research in CAPTURE, to outline key challenges and opportunities relating to the management of paradata.

2. New knowledge on paradata that supports research data management and open science

The purpose of CAPTURE is to create new knowledge and increased understanding of how paradata is created and used today, as well as to develop and test methods for working with paradata. Based on the results, CAPTURE contributes to creating standards and tools for paradata and advances in data-intensive research areas that use heterogeneous research data of different origins. The project does this by creating knowledge that supports the implementation of national, European, and global policies for data management and open data (cf. [10–12]).

CAPTURE also contributes to the effective sharing and reuse of research data in discipline-specific, thematic, and interdisciplinary knowledge ecosystems and repositories (cf. [13]). The project develops a critical understanding of the social contexts and use of infrastructures emphasised in recent research agendas (e.g. ARIADNE [14–16]) and empirical research (e.g. [17]). It creates new knowledge about what data creators and users find important to document regarding data-related processes, what explicit and implicit needs for documentation there are, and how these needs can be satisfied in practice.

3. Document enough, not too much

A major challenge for capturing and preserving paradata is that different data users have different needs in different situations. Literature on data reuse has identified differing needs across various disciplines and how these needs depend on what methods and theoretical perspectives underpin the scholarly enterprise [18]. At the same time, it is both practically impossible to document everything and very hard to decide what should and should not be documented. The variety of needs along with the difficulty of predicting what needs exist make it complicated to document data-related processes.

Determining how to document and preserve just enough therefore becomes a key issue. Like all data about data [19], paradata will be incomplete. As a consequence, it is necessary to focus on striking an acceptable balance between what can be captured automatically and what has to be documented manually (e.g. [20]). It is thus important to examine what information is already embedded in the data itself [21,22] and what can be left to future users to find out for themselves through various types of "archaeological" or "forensic" post-hoc methods [23] for "excavating" existing data. To date, a great deal of research has explored each of these approaches but there has been a lack of research covering the entire paradata phenomenon and how it can be used to support the reuse of research and survey data.

The CAPTURE project uses several methods to investigate the intellectual processes that underlie the creation and use of research data within and outside of archeology and to propose and develop strategies for capturing them. This palette of methods consists of document and documentation studies (e.g. [24–26]), conceptual [27] and citation analysis [28], ethnography, review and testing of previously proposed and newly developed methods for documenting paradata, as well as interviews [29] and focus group discussions with key stakeholders.

4. Much paradata is available in existing documentation

The results from interview and survey studies and the analysis of research publications and data show that a great deal of paradata is already available in the existing scholarly output. In archaeology, survey reports constitute an important source of paradata. They are expected to document both research results and the investigation process. In addition to regular job descriptions, they convey knowledge of work processes, for example in description of results and in information about participating actors. Photographs provide an important paradata source, especially those showing work in progress, environment, and physical conditions at investigation sites [24]. Even a close reading of the dataset can contribute information about underlying processes. Word choices, descriptions, and time stamps are just a few examples of elements in databases that can yield process knowledge. Much of this information is not documented explicitly but is inherent in the messiness of primary research data. Standardised data and metadata formats lack the flexibility to document all possible forms of process information and are unsuited to preserve such implied or inherent paradata. Perhaps somewhat counter-intuitively from a research data management perspective, therefore, extensive standardisation and data cleaning therefore risk resulting in a loss of essential paradata [26].

The fact that much paradata can already be found in existing documentation means that the main challenge with process documentation is not necessarily to expand its quantity or scope. One of the problems is that the paradata is fragmented across different parts of the documentation and that it can prove complicated to get an overall grasp of what paradata is available. Key challenges involve finding the paradata, understanding what is missing, and complementing it with the necessary additional information.

5. Documenting useful information

Another problem in finding paradata is that it is not always available or that available paradata do not correspond with user needs. In particular, information about data management procedures, standards, and structuring of data is rarely documented in detail. Results also demonstrate that data creators and users often have different views on what paradata is needed [26]. When paradata is documented, the data creator would probably focus on those elements that are obvious to them, that accord with their ideas about what is central to data creation, and that are easy to document. Data users, on the other hand, expect and need paradata that help them understand the data based on their particular situation.

The apparent gap between what data creators and data users consider to be important makes it difficult to create and provide paradata that is meaningful to both parties. Data creators have to understand how users think, anticipate what paradata is likely to be helpful, and consider data usage when creating and documenting data. The users similarly need insight into how data creation has taken place and capacity to understand how the data creation process works. An additional complication is that the specific needs depend on the purpose, context and situation of data use. Reproducing research and reanalysing data again for the same purpose and in the same research field as the original study to verify or disprove results require a different set of paradata than if the purpose is to extend the original analysis temporally, spatially, or for example socially, by combining data with other (possibly new) data in the same research field. The same applies if the data is used, possibly in combination with other or new datasets from the same or other research fields, for analysis in another research field, to produce new results using new analytical methods, or to conduct a historical study of a phenomenon related to the dataset or to the research itself.

6. Conclusions

The practical key challenges in providing enough – but not too much – paradata to make research data usable relate to documenting data creation, processing and use: what to document, but also how to document it in adequate detail and form. It is equally crucial to realise what is understood as paradata. The term “paradata” is used with different meanings in different contexts [27]. Therefore, it is necessary to clarify exactly what is meant when the term is used in theory as well as in practice. From a research infrastructure perspective, the most challenging question is how to enable and support the creation of paradata that is sufficient, relevant for its future reusers, and not too labour-intensive to produce and maintain. Considering the significant extent to which paradata is coincidental and exists because of the lack of data cleaning and management, a major challenge is also how to strike a balance between too much and too little standardisation.

7. Acknowledgments

Capturing Paradata to document data creation and use for future research projects (CAPTURE) has received funding from the European Research Council (ERC) Grant number 818210.

8. References

[1] B.L. Voss, Curation as research: A case study in orphaned and underreported archaeological collections, Archaeol. Dialogues. 19 (2012) 145–169. https://doi.org/10.1017/S1380203812000219.

[2] I. Faniel, E. Yakel, Practices do not make perfect: Disciplinary data sharing and reuse practices and their implications for repository data curation, in: L.R. Johnston (Ed.), Curating Res. Data Vol. One Pract. Strateg. Your Data Repos., ACRL, Chicago, IL, 2017: pp. 103–126.

[3] K.D. Bollacker, Avoiding a Digital Dark Age: Data longevity depends on both the storage medium and the ability to decipher the information, Am. Sci. 98 (2010) 106–110. https://doi.org/10.1511/2010.83.106.

[4] G. Geser, F. Niccolucci, D2.4: Final innovation agenda and action plan, ARIADNE. (2016).

[5] L. Börjesson, O. Sköld, I. Huvila, The politics of paradata in documentation standards and recommendations for digital archaeological visualisations, Digit. Cult. Soc. 6 (2020) 191–220. https://doi.org/10.14361/dcs-2020-0210.

[6] M.P. Couper, Usability Evaluation of Computer-Assisted Survey Instruments, Soc. Sci. Comput. Rev. 18 (2000) 384–396. https://doi.org/10.1177/089443930001800402.

[7] J.M. Carrillo Gea, A. Toval, J.L.F. Alemán, J. Nicolás, M. Flores, The London Charter and the Seville Principles as sources of requirements for e-archaeology systems development purposes, Virtual Archaeol. Rev. 4 (2013) 205–211. https://doi.org/10.4995/var.2013.4275.

[8] I. Huvila, Improving the usefulness of research data with better paradata, Open Inf. Sci. 6 (2022) 28–48. https://doi.org/10.1515/opis-2022-0129.

[9] J. Davet, B. Hamidzadeh, P. Franks, Archivist in the machine: paradata for AI-based automation in the archives, Arch. Sci. 23 (2023) 275–295. https://doi.org/10.1007/s10502-023-09408-8.

[10] A. Beck, C. Neylon, A vision for open archaeology, World Archaeol. 44 (2012) 479–497. https://doi.org/10.1080/00438243.2012.737581.

[11] DCC, An analysis of open data and open science policies in Europe, May 2017, SPARC Europe & DCC, Apeldoorn, 2017.

[12] E. Kansa, S. Kansa, Toward a do-it-yourself cyberinfrastructure: Open data, incentives, and reducing costs and complexities of data sharing, in: E. Kansa, S. Kansa, E. Watrall (Eds.), Archaeol. 20 New Approaches Commun. Collab., CA: Cotsen Institute of Archaeology, UC Los Angeles, Los Angeles, CA, 2011: pp. 57–91.

[13] G. Bruseker, N. Carboni, A. Guillem, Cultural heritage data management: The role of formal ontology and CIDOC CRM, in: M.L. Vincent, V.Manuel. López-Menchero Bendicho, Marinos. Ioannides, T.E. Levy (Eds.), Herit. Archaeol. Digit. Acquis. Curation Dissem. Spat. Cult. Herit. Data, Springer, Cham, 2017: pp. 93–131.

[14] N. Aloia, C. Binding, S. Cuy, M. Doerr, B. Fanini, A. Felicetti, J. Fihn, D. Gavrilis, G. Geser, H. Hollander, C. Meghini, F. Niccolucci, F. Nurra, C. Papatheodorou, J. Richards, P. Ronzino, R. Scopigno, M. Theodoridou, D. Tudhope, A. Vlachidis, H. Wright, Enabling european archaeological research: The ARIADNE E-infrastructure, Internet Archaeol. 43 (2017). https://doi.org/10.11141/ia.43.11.

[15] G. Lambourne, L. Stoakes, M. Cassar, K.V. Balen, M. Rhisiart, M. Thomas, R. Miller, L. Burnell, Strategic Research Agenda, JPI Cultural Heritage and Global Change, Rome, 2014. http://www.jpi-culturalheritage.eu/wp-content/uploads/SRA-2014-06.pdf.

[16] G. Geser, Achievements of the ARIADNE Initiative for Archaeological Data Sharing and Research, Internet Archaeol. (2023). https://doi.org/10.11141/ia.64.2.

[17] M.S. Mayernik, D.L. Hart, K.E. Maull, N.M. Weber, Assessing and tracing the outcomes and impact of research infrastructures, JASIST. 68 (2017) 1341–1359.

[18] K. Gregory, L. Koesten, Data Needs, in: Hum.-Centered Data Discov., Springer International Publishing, Cham, 2022: pp. 19–32. https://doi.org/10.1007/978-3-031-18223-5_3.

[19] M.S. Mayernik, A. Acker, Tracing the traces: The critical role of metadata within networked communications, J. Assoc. Inf. Sci. Technol. 69 (2018) 177–180. https://doi.org/10.1002/asi.23927.

[20] M. Stamatogiannakis, P. Groth, H. Bos, Looking inside the black-box: Capturing data provenance using dynamic instrumentation, in: Proven. Annot. Data Process. 5th Int. Proven. Annot.

Workshop IPAW 2014 Cologne Ger. June 9-13 2014 Revis. Sel. Pap., Springer, Cham, 2015: pp. 155–167. https://doi.org/10.1007/978-3-319-16462-5_12.

[21] J. Huggett, Promise and paradox: Accessing open data in archaeology, in: Proc. Digit. Humanit. Congr. 2012 Humanit. Res. Inst. Sheff., 2012.

[22] S. Gant, P. Reilly, Different expressions of the same mode: a recent dialogue between archaeological and contemporary drawing practices, J. Vis. Art Pract. 17 (2017) 100–120. https://doi.org/10.1080/14702029.2017.1384974.

[23] M.G. Kirschenbaum, R. Ovenden, R. Gabriela, Digital forensics and born-digital content in cultural heritage collections, Council on Library and Information Resources, Washington, D.C., 2010.

[24] I. Huvila, O. Sköld, L. Börjesson, Documenting information making in archaeological field reports, J. Doc. 77 (2021) 1107–1127. https://doi.org/10.1108/JD-11-2020-0188.

[25] I. Huvila, L. Börjesson, O. Sköld, Archaeological information-making activities according to field reports, Libr. Inf. Sci. Res. 44 (2022) 101171. https://doi.org/10.1016/j.lisr.2022.101171.

[26] L. Börjesson, O. Sköld, Z. Friberg, D. Löwenborg, G. Pálsson, I. Huvila, Re-purposing Excavation Database Content as Paradata: An Explorative Analysis of Paradata Identification Challenges and Opportunities, KULA Knowl. Creat. Dissem. Preserv. Stud. 6 (2022) 1–18. https://doi.org/10.18357/kula.221.

[27] O. Sköld, L. Börjesson, I. Huvila, Interrogating paradata, Inf. Reseach Proc. 11th Int. Conf. Concept. Libr. Inf. Sci. Oslo Metrop. Univ. May 29 - June 1 2022. 27 (2022) paper colis2206. https://doi.org/10.47989/ircolis2206.

[28] I. Huvila, L. Andersson, O. Sköld, Citing methods literature: citations to field manuals as paradata on archaeological fieldwork, Inf. Res. 27 (2022) paper941. https://doi.org/10.47989/irpaper941.

[29] L. Börjesson, I. Huvila, O. Sköld, Information needs on research data creation, Inf. Res. 27 (2022) isic2208. https://doi.org/10.47989/irisic2208.