Documenting Information Processes and Practices: Paradata, Provenance Metadata, Life-Cycles and Pipelines

Isto Huvila
Department of ALM,
Uppsala University, Sweden
isto.huvila@abm.uu.se

Jane Greenberg
Metadata Research Center, College of Computing & Informatics
Drexel University
jg3243@drexel.edu

Olle Sköld
Department of ALM
Uppsala University, Sweden
olle.skold@abm.uu.se

Andrea Thomer
School of Information
University of Michigan
athomer@umich.edu

Ciaran Trace
School of Information
The University of Texas at Austin
cbtrace@austin.utexas.edu

Xintong Zhao
Metadata Research Center, College of Computing & Informatics
Drexel University
xz485@drexel.edu

ABSTRACT

Processes and practices—and in general, informational doings and their diverse constellations—are pertinent elements of the information landscape. This panel presents research on documentation and description of processes and practices in the information field addressing: 1) how different conceptualisations of processes and practices influence how they emerge as describable entities; 2) what different approaches to document and describe processes and practices exist and have been proposed in information science and technology research; 3) what aspects of processes and practices different documentation approaches capture, make visible and invisible; and 4) what novel insights from the current state-of-the-art research can be drawn to support practitioners in different areas of the information field, including knowledge organisation, information management, information literacy instruction, and development of information systems and services.

KEYWORDS

pipelines; processes; practices; paradata; provenance metadata.

INTRODUCTION

Processes and practices are pertinent elements of the information landscape and thus of interest across the broad information science and technology field. Besides functioning as sites where information interactions occur, information work takes place and the information itself—understood as broadly and diversely as it is conventionally approached in the information field—happens, they are undertakings that need to be documented in order to understand the nature of their related information and how information unfolds in practice. Besides, a comprehensive understanding of information and processes is a key to enhancing their equity, diversity, inclusion and relevance. Descriptions of such informational doings have been conceptualised, for instance, in terms of workflows, process and life-cycle models, provenance and process metadata, and recently, also as paradata in different areas of the information science and technology field. Even if there is obvious overlap between process- and practice-oriented perspectives and approaches and a comprehensive understanding of informational undertakings has been acknowledged as a key premise of useful and manageable information, so far there has been surprisingly little exchange between the research traditions.

Panel members will present research conceptualizing, documenting, and describing processes and practices in the information field specifically addressing: 1) how different conceptualisations of processes and practices influence how they emerge as describable entities; 2) what different approaches to document and describe processes and practices exist and have been proposed in information science and technology research; 3) what aspects of processes and practices different documentation approaches capture, make visible and invisible; and 4) what novel insights from the current state-of-the-art research can be drawn to support practitioners in different areas of the information field, including knowledge organisation, information management, information literacy instruction, and development of information systems and services. The panelists are information science researchers who have conducted empirical research and concept development relating to different aspects of documenting processes and practices in diverse contexts ranging from archival research and game studies to archaeological and scientific information.

84th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – Nov. 3, 2021 | Salt Lake City, UT. Author(s) retain copyright, but ASIS&T receives an exclusive publication license.

DOCUMENTING PROCESSES AND PRACTICES: ON OVERVIEW

A glance at the literature shows that the issue of documenting and describing processes and practices has been discussed in different parts of the information science and technology field from diverse perspectives and covering such a variety of contexts from scientific (Leipzig et al., 2020; Thomer et al., 2018) and scholarly information (Huvila, 2020; Trace & Karadkar, 2017) to workplace studies (Trace, 2011), game heritage (Prax et al., 2019), digital preservation (e.g. Mayer et al., 2013a) and curation (Chao, 2014), and archives and records management (Trace, 2020). In spite of the varied contexts and perspectives, it is possible to identify certain trends. A part of the literature emphasises an explanatory and often prescriptive, structured modelling-oriented approach whereas others stress descriptions and documentation as means to provide a contextualised understanding of processes and practices. The explanatory work appears to be to a certain extent more inclined to focus on processes whereas the latter is perhaps more inclined to use practices and related concepts to refer to the undertakings they describe. The different strands of work do also seem to demonstrate somewhat different disciplinary influences. Explanatory research is more often stemming from knowledge organisation and information-modelling (e.g., Greenberg et al., 2021; Leipzig et al., 2020; Mayer et al., 2013b) background whereas in descriptive research it is possible to trace influences of heritage studies, anthropology and science and technology studies (e.g. Huvila, 2020; Sköld, 2015; Trace & Karadkar, 2017). In practice, however, the division between the two principled approaches and conceptual spheres is not watertight. Modelling can be done for descriptive purposes and schemes for retroactive documentation often have normative and prescriptive ambitions to steer information work. In parallel, it is possible to see that different parts of the information science and technology field conceptualise the documentation and description of processes and practices in somewhat diverging, although at the same time, often overlapping terms. The same applies to what is being documented i.e. whether processes and practices are considered as, for instance, professional or scholarly practices or processes.

A key concept pertaining to the documentation of processes and practices alike is the archival studies notion of provenance that lacks a consistent definition but in broad terms, is used to refer to the creator, records, and custodial history of archival records (Tognoli & Guimarães, 2019). Preserving the integrity of archival records and an understanding of their history requires that provenance, their original order and the *fonds*, or the original body, of records are respected and maintained intact throughout their chain of custody (Douglas, 2010). In practice, however, the complexity of the processes pertaining to archival information means that the order and contexts in which records are created and used change and it is often close to impossible to determine and maintain them in an 'original' order that would cover their custodial history as a whole. Therefore, it is often necessary to produce explicit descriptions to complement what information and its organisation tell about its custodial, or in broader terms, processual history.

Descriptions of the processual history of information and generally, information processes and practices, have many names and come in different forms. Data preservation literature refers often to the umbrella term provenance metadata to refer to structured and unstructured documentation of (data) provenance. Provenance metadata is an element in several metadata standards even if different standards tend to conceptualise the notion of provenance in somewhat diverging terms and focus on, for instance, custodial activities or provenance history of information (Bountouri, 2017). In addition, there are also specific schemes for provenance metadata including PROV Ontology (Moreau & Groth, 2013) and Open Provenance Model (Moreau et al., 2008). A related concept of paradata (roughly, data on data-related processes and practices, see e.g. Couper, 2000 versus metadata that describes data; Pomerantz, 2015) has gained prominence especially in survey research (Goodwin et al., 2017) and cultural heritage visualisation communities (Bentkowska-Kafel & Denard, 2012), and more recently, for instance, in archaeology (e.g. Huvila et al., 2021). A key aspect of both provenance metadata and paradata, underlined in the recent literature (e.g. Huvila et al., 2021; Michetti, 2017; Sköld, 2017), is that they can take many different forms and be embedded in the data itself—especially when the perspective to information-related processes and practices is extended beyond interactions pertaining to specific technical information objects to their broader stakeholder contexts.

Besides provenance metadata and paradata that are often oriented towards providing curatorial descriptions for conveying understanding of information processes, artefacts, and their contexts, there are parallel approaches to documentation and description of information work and practices that are explicitly geared towards reproducibility of processes. Systems engineering and management literature refers to lifecycle documentation to describe documentation of information, events, and activities in the course of the lifecycle of a system or service. In critical operational settings such as the energy and aviation industries, keeping a meticulously structured logbook is a crucial safety measure (Schmidt & van Hoof, 2012). Significant progress has also been made in computational semantic labeling of data (e.g., Greenberg et al., 2021; Zhao et al., 2020). In the context of data-intensive scientific work, calls for reproducibility of research and the reusability of research data have contributed to a comparable interest in formalising the documentation of work procedures. Digital preservation research has emphasised the need to preserve complete computer systems through meticulous documentation of the technical system and its use (Mayer et al., 2013a) together with its social, including for instance, legal context (Mayer, et al., 2013b). Several different conceptualisations of procedures and their representations exist. A part of the literature refers to process curation (Rauber, 2012) or preservation of processes (Mayer, et al., 2013a) understood as consisting of the collection, pre-processing and use of data. Pipelines refer to a series of computational transformations performed on data. The largely synonymous term (scientific) workflows are used sometimes in a somewhat broader sense to also cover non-automated aspects of processes (Leipzig et al., 2020). In contrast to sciences, non-computational scholarly research generally lacks an equivalent to scientific workflows and pipelines. Proposed models for scholarly workflows exist (e.g. Antonijevic & Cahoy, 2018; Chiquet, 2020) but due to the practical and epistemological conditions of research in scholarly disciplines, it is possible that for instance argumentation (Stead & Doerr, 2015; Vatanen, 2004), process-modeling (Thomer et al., 2018) or events-based (D'Andrea & Fernie, 2013) approaches may end up being more applicable to milieus with high epistemic variety such as scholarly work but also in other comparably heterogeneous contexts such as everyday-life information processes and practices. It is conspicuous that descriptions differ not only in how they describe doings but also in how specific descriptions are useful in different contexts and what aspects of doings they document.

Independent of the contexts and means of documenting and describing processes and practices, the different approaches share common challenges. Many of them are well-documented, for instance, in earlier knowledge organisation and information behaviour research (e.g. Greenberg et al., 2021; Huvila, 2020; Sköld, 2017; Thomer et al., 2018; Trace, 2020), others in the work stemming from science and technology studies (STS) and in the growing body of information science research that combines information and STS perspectives (e.g. Borgman et al., 2015; Borgman et al., 2019; Gregory et al., 2020; Huvila, 2016; Huvila et al., 2021). As a whole, it is, however, apparent that there is room for a higher degree of cross-fertilisation. Processes and practices are complex and often difficult to distinguish and demarcate from the settings in which they are carried out—and sometimes from each other. Consequently, a description is in practice always a simplification and it either may or may not represent the original process or set of practices in adequate detail. Similarly, it is difficult to determine what type and amount of information would be enough for different purposes, especially ones that are not known in advance. Understanding different types of descriptions requires different competences and literacies. Further, while simple flat descriptions can be relatively easy to assemble, a comprehensive understanding of processes and practices often requires rich documentation that can be arduous and expensive to produce.

LAYOUT OF THE PANEL

The panel starts with a presentation by the moderator (IH) that introduces documentation and description of information processes and practices as a topic of research and practice and briefly introduces central concepts of the field. After the 10 min introduction, all panellists give a 5-minute lightning talk of their work relating to documentation and description of information processes and practices with a specific focus on its theoretical and empirical insights and implications to their area of information science and technology research and practice. After the lightning talks, each of the panellists presents a short commentary on their colleagues' presentations with a focus on pointing out commonalities and differences in the approaches and the relationship of their different takes on documentation of processes and practices. After the commentaries, the panellists give short, one-minute reflections of how they would push the state-of-the-art of research on documentation of processes and practices on the basis of their experience. During the final 30 minutes of the panel, the audience is asked to join the discussion with panellists. The discussion is led by the moderator and facilitated by a set of questions based on the panellists' presentations. The panel closes with an invitation from the moderator to contribute to the discussion started at the panel and a round of proposals and ideas for future work in the field from the panellists and the audience.

The presentations combine two parallel approaches to engage with the research on documenting and describing processes and practices. All presentations explicate how processes and practices are conceptualised in five different contexts across the information field, how these different understandings coincide and diverge, and what implications it has on their describability and documentability. At the same time, all presentations engage with the question of different means of documenting and describing practices and what they imply for their usefulness in different contexts and for different purposes. By bringing these two parallel perspectives together, the panel explicates the disciplinary nexus of documenting information processes and practices and its implications to related different areas of the information field, including archival studies, knowledge organisation, game studies, research data and information management. Further, the panel delves into the practical and theoretical implications of different approaches to documenting information processes and practices.

PANELISTS AND THEIR CONTRIBUTIONS

Isto Huvila, Uppsala University

Isto Huvila presents empirical findings of his research on how archaeologists document their information making in archaeological field reports. The presentation draws on an ongoing empirical research project on archaeological paradata and documentation of information making and manipulation practices in archaeology. Huvila shows the diversity of ways and means of how information making becomes explicitly and implicitly documented in report-writing and how the functionality of specific types of paradata can be explained in terms of their epistemic distance to the situations where they are taken into use. The findings call into question simplistic ideas of paradata as straightforward acontextual descriptors and suggest ways forward for resource description research and practice.

Professor Isto Huvila holds the chair in information studies at the Department of ALM (Archival Studies, Library and Information Science and Museums and Cultural Heritage Studies) at Uppsala University in Sweden. His primary areas of research include information and knowledge management, information work, knowledge organization, documentation, and social and participatory information practices.

Xintong Zhao and Jane Greenberg, Metadata Research Center, Drexel University

Xintong Zhao and Jane Greenberg will present collaborative work on how materials science literature documents key processing methods. This embedded knowledge, along with the names and properties of materials, capture the process of materials discovery. Knowledge extraction work, analyzing these processes and associated properties of materials provide data for the development of ontological knowledge systems. The presentation will report on empirical results, and how this work can assist with further computational research. Current efforts have implications for the development of neural networks and automatic processes sharing knowledge and workflows.

Xintong Zhao (Doctoral candidate) and Jane Greenberg (Alice B. Kroeger Professor and Director of the Metadata Research Center) are both at the College of Computing & Informatics, Drexel University. Zhao’s research interests focus on knowledge discovery, information extraction and natural language processing, and Greenberg’s on metadata, big metadata, knowledge organization and extraction, ontologies, linked data, and data science.

Olle Sköld, Uppsala University

Olle Sköld's contribution consists of delving into the social micro-processes of information creation in the serious-leisure setting of videogame play. Videogames and videogame play comprise a presently very impactful domain of cultural production that is only fractionally understood, explored, and theorized from the viewpoints of information and archival science. Sköld reports on empirical research of information creation in two online videogame communities active in distinct videogame and social-media software environments. The findings offer insights into how online videogame communities create and document information and how these processes are interlinked with videogame lifecycles, community self-management practices, and software affordances. Sköld's presentation also discusses proven (practices) and promising (paradata) conceptualizations of information-creation processes in the videogame domain and outlines their implications for and varying usabilities in SSH scholarship and ALM institutions invested in the curation and dissemination of videogames and videogame documentation.

Olle Sköld is an assistant professor at the Department of ALM at Uppsala University in Sweden and a researcher in the CAPTURE (ERC 818210) and Labour’s Memory projects (RJ IN20-0040). Sköld's research is characterized by an extensive focus exploring matters of documentation and preservation in the videogame domain and by a broad interest in the GLAM field, digital humanities, and knowledge production.

Andrea Thomer, University of Michigan

Andrea Thomer reflects on work studying data practices in the natural sciences, and the many ways that provenance is captured (and sometimes, is not captured) at these sites. While traditional modes of documenting the broader context of data collection like field notes are still important, there’s a growing need for more computationally-friendly modes of capturing and sharing these chains of provenance. Thomer will particularly discuss a) the challenges of translating qualitative field notes into semantic data, and b) the need to infer provenance in long-lived natural science databases.

Andrea Thomer is an assistant professor at the University of Michigan School of Information. She conducts research in the areas of long-term data curation and knowledge infrastructure sustainability; database curation; integrative data reuse; and the collaborative use and curation of natural science data.

Ciaran B. Trace, The University of Texas at Austin

Ciaran B. Trace presents findings of her research on how cultural heritage professionals document their work to stabilize and inscribe context for the documentary evidence under their care. The presentation draws on current and extant research projects on provenance metadata and documentation of the information processing that lies at the heart of archival practice. Trace shows the way information contextualization is documented in physical and intellectual manifestations of processed collections and how diverse types of provenance metadata demonstrates shifting understandings of what it means to anchor collections to contexts both originary and emerging. The findings seek to create connections among disciplines for whom mapping and documenting information lineage is a central concern.

Ciaran B. Trace is an associate professor at the School of Information at The University of Texas at Austin where she serves as co-editor of the journal *Information & Culture*. Her primary areas of study include knowledge and research infrastructures, theories of information and information work, personal and disciplinary information practices, and materiality in the digital age.

ACKNOWLEDGEMENTS

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

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

Bentkowska-Kafel, A., & Denard, H. (2012). Introduction. In A. Bentkowska-Kafel, H. Denard, & D. Baker (Eds.), *Paradata and transparency in virtual heritage* (pp. 1–4). Farnham: Ashgate.

Borgman, C. L., Darch, P. T., Sands, A. E., Pasquetto, I. V., Golshan, M. S., Wallis, J. C., & Traweek, S. (2015). Knowledge infrastructures in science: Data, diversity, and digital libraries. *International Journal on Digital Libraries*, *16*(3), 207–227.

Borgman, C. L., Scharnhorst, A., & Golshan, M. S. (2019). Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. *JASIST*, *70*(8), 888–904. https://doi.org/10.1002/asi.24172

Bountouri, L. (2017). *Archives in the digital age: Standards, policies and tools*. Oxford: Chandos.

Chao, T. C. (2014). Enhancing metadata for research methods in data curation. *Proc. Am. Soc. Info. Sci. Tech.*, *51*(1), 1–4. https://doi.org/10.1002/meet.2014.14505101103

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.* https://doi.org/10.5281/ZENODO.3754263

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

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

Douglas, J. (2010). Origins: Evolving ideas about the principle of provenance. In T. Eastwood & H. MacNeil (Eds.), *Currents in archival thinking* (pp. 23–43). Santa Barbara, CA: Libraries Unlimited.

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.

Greenberg, J., Zhao, X., Adair, J., Boone, J., & Hu, X. T. (2021). HIVE-4-MAT: Advancing the ontology infrastructure for materials science. In E. Garoufallou & M.-A. Ovalle-Perandones (Eds.), *Metadata and semantic research* (pp. 297–307). Cham: Springer.

Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Lost or found? Discovering data needed for research. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.e38165eb

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. https://doi.org/10.1080/01972243.2016.1177763

Huvila, I. (2020). Information-making-related information needs and the credibility of information. *Information Research*, *25*(4), paper isic2002. https://doi.org/10.47989/irisic2002

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

Leipzig, J., Nüst, D., Hoyt, C. T., Soiland-Reyes, S., Ram, K., & Greenberg, J. (2020). *The role of metadata in reproducible computational research*. arXiv. Retrieved from https://arxiv.org/abs/2006.08589

Mayer, R., Guttenbrunner, M., & Rauber, A. (2013a). Evaluation of preserved scientific processes. In *Research and advanced technology for digital libraries* (pp. 434–437). Berlin: Springer.

Mayer, R., Pröll, S., Rauber, A., Palma, R., & Garijo, D. (2013b). From preserving data to preserving research: Curation of process and context. In T. Aalberg, C. Papatheodorou, M. Dobreva, G. Tsakonas, & C. J. Farrugia (Eds.), *Research and advanced technology for digital libraries* (pp. 490–491). Berlin: Springer.

Michetti, G. (2017). Provenance in the archives: The challenge of the digital. In F. Smit, A. Glaudemans, & R. Jonker (Eds.), *Archives in liquid times* (pp. 229–246). Gravenhage.

Moreau, L., Freire, J., Futrelle, J., McGrath, R. E., Myers, J., & Paulson, P. (2008). The open provenance model: An overview. In J. Freire, D. Koop, & L. Moreau (Eds.), *Provenance and annotation of data and processes. IPAW 2008*. Berlin: Springer.

Moreau, Luc, & Groth, P. (2013). *Provenance: An introduction to PROV*. San Rafael: Morgan & Claypool.

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

Prax, P., Sjöblom, B., Eklund, L., Nylund, N., & Sköld, O. (2019). Drawing things together: Understanding the challenges and opportunities of a cross-lam approach to digital game preservation and exhibition. *Nordisk Kulturpolitisk Tidsskrift*, *22*(2), 332–354. https://doi.org/10.18261/issn.2000-8325/-2019-02-08

Rauber, A. (2012). Data quality for new science: Process curation, curation evaluation and curation capabilities. *NSF Workshop on Data Quality, 10-11 September 2012, Arlington, USA*.

Schmidt, J., & van Hoof, A. (2012). Towards a cooperative life cycle documentation for distributed renewable energy power plants. *2012 7th International Conference on System of Systems Engineering (SoSE)*, 32–37.

Sköld, O. (2015). Documenting virtual world cultures. *Journal of Documentation*, *71*(2), 294–316. https://doi.org/10.1108/JD-11-2013-0146

Sköld, O. (2017). Getting-to-know. *Journal of Documentation*, *73*(6), 1299–1321. https://doi.org/10.1108/JD-11-2016-0145

Stead, S., & Doerr, M. (2015). *CRMinf: The argumentation model—An extension of CIDOC-CRM to support argumentation* (Version 0.7). Purley: Paveprime.

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

Tognoli, N., & Guimarães, J. A. C. (2019). Provenance as a knowledge organization principle. *Knowledge Organization*, *46*(7), 558–568. https://doi.org/10.5771/0943-7444-2019-7-558

Trace, C.B. (2011). Documenting Work and Working Documents: Perspectives from Workplace Studies, CSCW, and Genre Studies. *System Sciences (HICSS), 2011 44th Hawaii International Conference On*, 1–10. https://doi.org/10.1109/HICSS.2011.170

Trace, Ciaran B. (2020). Maintaining records in context? Disrupting the theory and practice of archival classification and arrangement. *The American Archivist*, *83*(2), 322–372.

Trace, Ciaran B., & Karadkar, U. P. (2017). Information management in the humanities: Scholarly processes, tools, and the construction of personal collections. *JASIST*, *68*(2), 491–507. https://doi.org/10.1002/asi.23678

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)*. Oxford: Archaeopress.

Zhao, X., Greenberg, J., Hu, X., Meschke, V., & Toberer, E. (2020). Scholarly big data: Computational approaches to semantic labeling in materials science. *Proceedings of JCDL 2020*.