ADVANCES IN KNOWLEDGE ORGANIZATION
Proceedings of International ISKO Conference - ISSN: 0938-5495 - 2026 - São Paulo

Exploring Indexicality and Origination Constructs in Data about Data

Joseph T. Tennis

University of Washington Information School, United States of America

Abstract

Motivated by definitional ambiguity and inspired by Peircean and archival concepts, this paper offers a proposed definitional framework for data about data. First, I lay out methodology, definition and context. Then, I propose a relational framework for concepts of data, metadata, and paradata. Through this exploration new terms arise and are defined. We close with a proposal for a mixed model of vocabulary related to data about data.

1.0 Introduction

Data, metadata, and paradata are contested concepts in knowledge organization and cognate fields. Metadata has a large definitional literature (e.g., Gorman 2003; Greenberg 2005; Furner 2020). Data has figured in our literature as well (e.g., Renear et al. 2010; Furner 2016). And more recently, paradata definitional work has entered the conversation (e.g., Huvila 2022; Huvila et al. 2024).

As Huvila notes, "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" (2022, 32).

As part of a larger project, this paper explores ways we can build on ideas in the literature and move toward two things: (1) a formal language that may be useful in the research literature that discusses these topics, and (2) an acknowledgment of contextual and paradigmatic nature of these constructs in information systems.

2.0 My Methodology

As I have argued earlier, "[m]ethodology is the combination of epistemic stance and methods of investigation. Methods of investigation, what I will call *techniques*, form a practice that carries with it the knowledge needed to have a result aithful to the chosen epistemology" (Tennis 2008, 106).

For much of my work, I operate from a postpragmatist epistemology (e.g., Piacente 2024; Rorty 1982; Rorty 1999). "Rather than advancing pragmatic theories of inquiry, truth, justification, meaning, normativity, etc.—as per pragmatism traditionally and typically—postpragmatism sees little use in this and stops advancing such theories" (Piacente 2024,

191). Instead, the discussion below is an attempt to construct a *vocabulary* to *cope* with data, metadata, and paradata (Rorty 1982, 202-203).

My method here is arguing for a particular perspective through *writing*. Writing as a method or technique is well documented by Richardson (2004).

3.0 The Problem Space

As introduced above, there are a range of definitions for data, metadata, and paradata. These terms have proliferated with the move of information systems from novelty to infrastructure in business, research, and learning. These concepts, and the inscriptions that flow from them permeate every aspect of life. This means that the *pragmatics* of the terms affect their *semantics*. That is, how they are used is what they mean.

Acknowledging that these concepts undergo redefinition in common use means we, as a scholarly community that studies them as first-order objects would benefit from strict definitions. And here lies the problem space: what parameters do we need to declare, and assumptions do we need, to lay bare, in order to craft definitions of data, metadata, and paradata that are useful in basic research on these concepts?

4.0 Parameters and Assumptions

I will start with the assumptions first. Assumption (1): we (humans) constructed these concepts to do certain work for us. Assumption (2): it is possible that someone's metadata can and will be someone's data. Assumption (3): Assumption 2 is transitive between all three concepts under consideration.

Taking the assumptions above, and in an effort to move toward precise definitions I offer two concepts at play in defining data, metadata, and paradata. The first is *indexicality* and the second is *origination*.

Indexicality is, in its simplest definition is, pointing to an entity in a particular context. This concept is born out of Charles Sanders Peirce's work, but debated heavily in semiotics (Sonesson 1998). One could argue the value we derive from the simple definition of metadata as data about data is because that definition is indexical. We can see the *meta-* in metadata is pointing to an entity, namely, data. What we are lacking in this case is the context. Because of their value, concepts like data and metadata are in a wide range of contexts and lose a unique or even strong indexical relationship because of that. I propose that origination can provide much needed context for the purposes of scholarly investigations on data, metadata, and paradata.

Origination, in the sense I am deploying it here, is inspired by archival theory. An originator is a "person assigned the electronic address where the record has been generated" (InterPARES 2025, 37). Inspired by this concept and for our purposes here, an originator is

the agent who is inscribing. This happens at a particular point in time and different originators can inscribe what we are defining here across the lifecycle of the resource.

5.0 Indexicality and Origination and our Terms

With these parameters and assumptions in place, we can use them to provide more precise definitions of data, metadata, and paradata. I will do this by starting with an example context.

In the construction of an online survey, the author (one originator in the lifecycle of the resource) writes questions (data). They also construct the manner in which the respondent will answer those questions. For example, they will use dropdown menu choices, radio buttons, or short of long form narrative answers in the survey (potential metadata). I propose here that the author, in this case, has created both a kind of data and potential metadata for downstream use. The author originated both of these kinds of inscription. There is, what we might call a *primary indexical relationship,* between this originator and the data and potential metadata. And the author here is, by extension, the *primary originator.*

To continue my definitional proposal, the aforementioned *potential metadata* is made *metadata proper* 1) by a different *secondary originator* and derived from conventional information retrieval systems design and implementation. For example, while the author has provided a title of the survey, the information professional or their proxy (a machine) has decided it is metadata. Because of this, that data can and will be used downstream by a human or a machine to populate the "title" field of a database that stores this survey, becoming metadata proper.

This human or machine, at this stage, also has an indexical relationship now to the document. This is a *secondary indexical relationship* to the data, but a *primary indexical relationship* to the metadata.

We now come to the third point in this baton race. When someone fills out this online survey, they are providing *data* to the author of the survey. Origination here is not from the author, but rather from the participant in the survey. They are creating something new – a *completed* survey. A different set of data for the author. Here we see how instructive the concepts of indexicality and origination are in defining our terms. Whereas the survey author created data forming a *primary indexical relationship* with the survey, when it is disseminated to respondents, each of them then creates a new *primary indexical relationship* to the resource by adding their responses. Each creates a new instantiation (Smiraglia 2005) of the resource (new data) and each of them is a new and unique originator.

After this happens, the author of the survey or any other individual can examine *how* the respondent(s) filled out the survey. From the perspective of the *primary originator* of the survey, this becomes secondary or even tertiary data. This has been termed *paradata* by Couper (2000).

However, paradata, like metadata, has traveled from computer-assisted data collection research (Couper's field) to many other fields, and has taken on many other meanings.

6.0 Paradata Definitions

Couper defined paradata as "auxiliary data describing the process" of administering a survey captured by computers" (2000, 393). In a partial and mixed extensional definition and assessment Couper says sources of paradata, in the context of computer-assisted data collection, are "case management information such as response rates, number of calls per case, average interview length, and so on. A most useful type of paradata for usability evaluation is keystroke or trace files (also called audit trails)" (Couper 2000, 393-394).

Fritz Scheuren asserts there are two types of paradata, macro and micro. "Macro paradata is very common and widely used. Most of the familiar examples of macro paradata measures are in the form of global process summaries, like overall response or coverage rates" (1999, 2C-1). And micro paradata are "process details known on each case (such as how many attempts it took to get an interview, whether the interview was in English or Spanish, etc.)" (Scheuren 1999, 2C-1).

Data curation picks up the term, paradata, to be "roughly, data on data-related processes and practices," (Huvila et al. 2021, 605). Huvila categorizes the definitional landscape here to have three types of definitions of paradata: broad, narrow, middle-ranged perspectives.

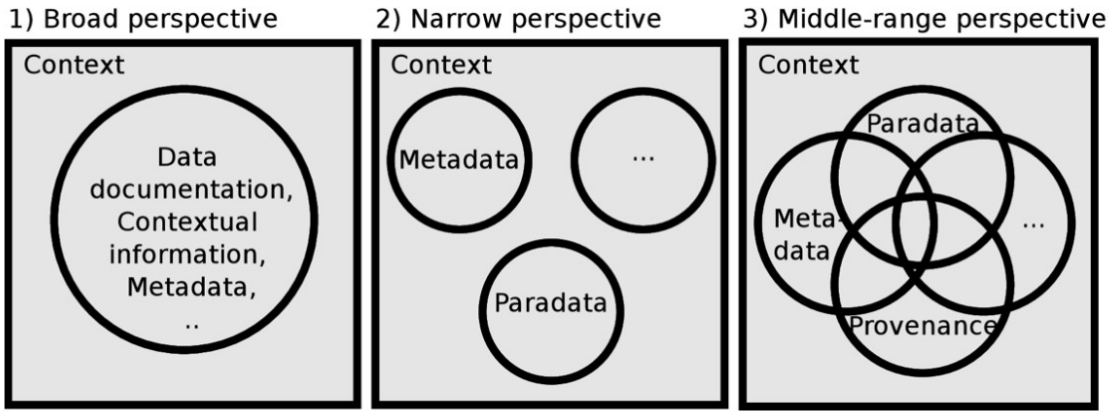

Figure 1. Huvila's (2022) visualization of definitions of paradata

In this paper, Huvila also says, of examples of process data, "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" (2022, 32). By my lights this points to the ontological instability of these terms. This observation also motivates this work.

The term has gained purchase in archival studies as well. The InterPARES Trust AI Terminology Database defines paradata as, "Information about the procedure(s) and tools used to create and process information resources, along with information about the persons carrying out those procedures" (2024).

Here we see the need to contextualize creation and processing of information resources. Were we to create a conference paper we would be an author, hence *author* would be paradata by this definition. This fact, the context of naming, also reinforces the need for indexicality and origination parameters on particular instances of paradata.

7.0 Examples of Deploying Parameter Definitions of Data about Data

If we return to the example of the computer-assisted data collection research, we can perhaps provide an example of how indexicality and origination play out. In this case, an author of the survey is the primary originator. There is also a primary indexical relationship between the author and the survey. Once this resource is created, we then expect a human or machine to derive and assign metadata to this survey. The title which was data in the first relationship becomes metadata by the intervention of an information professional or machine identifying it as such. Now the originator is secondary, they interact with the resource *after* it is created. This creates a temporal indexical relationship a second originator is creating a new set of data about data, in this case metadata. Someone, other than the author or the secondary originator, who interacts with this resource is a tertiary originator AND potentially creates paradata. This unintentional tertiary originator in this instance is the individual engaged in the process of filling out the survey.

8.0 Potential Definitional Ramifications of this Approach

If this is an acceptable way to describe data about data (i.e., data, metadata, and paradata) then one ramification is that the concept of *process* is removed from the ontology or nature of paradata. That is the concept of *process* can be ascribed to, perhaps, both metadata and paradata.

There are also drawbacks to this approach that need to be explored. Are we underspecifying when we take this approach. This is clear by removing *process* as a characteristic of the definition of paradata. Does this mean we would benefit another column of definitional exercises that we must engage in? This invokes Gilliland's (2016) categorization of metadata. Would we combine these in order to clearly specify both the context and semantics of these concepts?

Both of these point to future explorations in this space.

9.0 Summary and Conclusion

In this paper I have argued that data about data can constructively be specified as metadata or paradata based on indexicality and origination – that is the context of creation. The design of this approach is to achieve a particular level of precision. That is, by prescriptively contextualizing where data becomes metadata or paradata and by whom. However, it is possible, that by working toward that we have jettisoned meaningful semantics in the service of context. Conversely, we have proposed removed confusing concept of process from the term paradata because, as noted in the literature, it throws ambiguity into the definitional landscape. Acknowledging the chameleon nature of these terms throws into relief the *regional* or *hyper-contextual* work that needs to happen to communicate, at a foundational level, the object of research. To label clearly, we need data about the data about data.

AI Usage Statement: No AI was used in researching or writing this paper.

References

Couper, Mick P. 2000. "Usability Evaluation of Computer-Assisted Survey Instruments." Social Science Computer Review 18: 384–396.

Furner, Jonathan. 2016. "'Data': The Data." In Information Cultures in the Digital Age: A Festschrift in Honor of Rafael Capurro, edited by Matthew Kelly and Jared Bielby, 287–306. Wiesbaden: Springer.

Furner, Jonathan. 2020. "Definitions of 'Metadata': A Brief Survey of International Standards." Journal of the Association for Information Science and Technology 71: E33–E42. https://doi.org/10.1002/asi.24295

Gilliland, Anne J. 2016. "Setting the Stage." In Introduction to Metadata, 3rd ed., edited by Murtha Baca. Los Angeles: Getty Publications. https://www.getty.edu/publications/intrometadata/setting-the-stage/

Gorman, Michael. 2003. "L'Authority Control nel Contesto del Controllo Bibliografico in Ambiente Elettronico." In Authority Control: Definizione ed Esperienze Internazionali: Atti del Convegno Internazionale, Firenze, 10–12 Febbraio 2003, edited by Mauro Guerrini and Barbara B. Tillett, 17–26. Firenze: Firenze University Press; Roma: Associazione italiana biblioteche.

Greenberg, Jane. 2005. "Understanding Metadata and Metadata Schemes." Cataloging & Classification Quarterly 40, nos. 3–4: 17–36. https://doi.org/10.1300/J104v40n03_02

Huvila, Isto, Jane Greenberg, Olle Sköld, Andrea Thomer, Ciaran B. Trace, and Xintong Zhao. 2021. "Documenting Information Processes and Practices: Paradata, Provenance Metadata, Life-Cycles and Pipelines." Proceedings of the Association for Information Science and Technology 58: 604–609. https://doi.org/10.1002/pra2.509

Huvila, Isto. 2022. “Improving the Usefulness of Research Data with Better Paradata.” Open Information Science 6: 28–48.

Huvila, Isto, Lisa Andersson, and Olle Sköld, eds. 2024. Perspectives on Paradata: Research and Practice of Documenting Process Knowledge. Knowledge Management and Organizational Learning 13. Cham: Springer.

InterPARES. 2025. InterPARES 2 Project Glossary. https://www.interpares.org/display_file/ip2_glossary.pdf

InterPARES Trust AI. 2024. Terminology Database. https://interparestrustai.org/terminology/search

Piacente, Albert. 2024. “Postpragmatism: Quine, Rorty, and a Thoroughgoing Atheoreticism.” Contemporary Pragmatism 21: 190–216. https://doi.org/10.1163/18758185-bja10087

Renear, Allen H., Simone Sacchi, and Karen M. Wickett. 2010. “Definitions of Dataset in the Scientific and Technical Literature.” Proceedings of the American Society for Information Science and Technology 47: 1–4. https://doi.org/10.1002/meet.14504701240

Richardson, Laurel. 2004. “Writing: A Method of Inquiry.” In Approaches to Qualitative Research: A Reader on Theory and Practice, edited by Sharlene Nagy Hesse-Biber and Patricia Leavy, 473–495. New York: Oxford University Press.

Rorty, Richard. 1982. Consequences of Pragmatism. Minneapolis: University of Minnesota Press.

Rorty, Richard. 1999. Philosophy and Social Hope. New York: Penguin Books.

Scheuren, Fritz. 2001. “Macro and Micro Paradata for Survey Assessment.” In 1999 NSAF Collection of Papers, 2C-1–2C-15. NSAF Methodology Report 7. Washington, DC: Urban Institute.

Smiraglia, Richard P. 2005. “Instantiation: Toward a Theory.” In Data, Information, and Knowledge in a Networked World: Proceedings of the Canadian Association for Information Science Annual Conference, June 2–4, 2005, London, Ontario, Canada, edited by Liwen Vaughan. https://doi.org/10.29173/cais310

Sonesson, Göran. 1998. “Index, Indexicality.” In Encyclopedia of Semiotics, edited by Paul Bouissac, 306–311. New York: Oxford University Press.

Tennis, Joseph T. 2008. “Epistemology, Theory, and Methodology in Knowledge Organization: Toward a Classification, Metatheory, and Research Framework.” Knowledge Organization 35: 102–112.