Research Protocol*

Articulating epidemiological data harmonization initiatives as practical and collaborative experiences

Zack Batist

2025-01-18

> **Note**
>
> This document is still a work in progress.

Overview

This study investigates the social, technical, administrative and epistemic factors that mediate data-sharing initiatives in epidemiological research. It takes to heart the notions that data are media that facilitate communication across different research contexts, that data are created with specific intent, and that data are bounded by the social, practical and material circumstances of their creation. In light of these facts, I approach data-sharing as means of reconciling the varied circumstances of datasets' creation — both among themselves, and in relation to contexts of reuse. I therefore frame data-sharing as efforts to foster a series of collaborative ties beyond a project's original indended scope.

Data harmonization is one means of drawing the observations that constitute data into a formal schema, whose structure is driven by specific objectives and more general underlying suppositions and values. Although schemas are arrived at through

*This is an automatically generated PDF. Refer to the project website for continuous updates at https://zackbatist.info/CITF-Postdoc.

discussion, compromise and consensus-building, with an eye toward practical outcomes afforded by the data model, these socially-mediated interactions are generally under-recognized as important factors contributing to data-sharing initiatives' success relative to their potential impact.

My goal is to survey what factors are being priorized within various data harmonization initiatives, the rationales behind these decisions, and the relative efficacy of these variable approaches. More specifically, the project seeks to adress the following research questions:

- What are the objectives of data-sharing initiatives, how were they established, and what progress has been made to achieve them?
- What strategies do data-sharing initiatives employ to ensure they are able to meet their objectives, and how effective are they?
- What values underlie these strategies, and can they be linked with effective outcomes?

To be clear, my intent is not to pit various approaches against each other, but rather to ascertain what actions specific strategies entail, the circumstances in which each is adopted, the value that they bring, and the trade-offs involved. In other words, my goal is to reveal the diverse ways in which data-sharing occurs, and how different approaches impact the outcomes in different ways.

Background

Still in progress...

```
- About open data infrastructures in general
  - Discrepancies between aspirations and effective outcomes
  - Based on models of science that diverge from practical experience
- About open data in epidemiology
  - What is driving it?
  - Introduce Maelstrom as a key player
- Challenges and opportunities in the Canadian context
```

Methodology

Approach

This study is informed by a set of theoretical and methodological frameworks formed within a more interdisciplinary "science studies" tradition, which contribute to a more sociological outlook on science as cultural practice (cf. Pickering 1992). In practical terms, I will document the social and collaborative experiences involved in various research practices, which ultimately bind the many ways in which scientists do science.

I will specifically focus on how people contribute to and extract from information commons, which comprise both formal documents and mutually-held and information-laden situated experiences. This involves examining the ways in which participation in disciplinary or even more specialized communities of practice fosters mutual understanding about the potential and limitations pertaining to other people's data; and how this communally-held knowledge is accessed and re-produced. This approach aligns with the situated cognition methodological framework for examining the improvised, contingent and embodied experiences of human activity, including science (cf. Suchman 2007; Knorr Cetina 2001).

The situated cognition framework prioritizes subjects' outlooks, which are contextualized by their prior experiences, and enables scholars to trace how people make sense of their environments and work with the physical and conceptual tools available to them to resolve immediate challenges. Situated cognition therefore lends itself to investigating rather fluid, open-ended and affect-oriented actions, and is geared towards understanding how actors draw from their prior experiences to navigate unique situations.[1]

Situated cognition is especially salient in explorations of how people who are learning new skills learn how to work in new and possibly unfamiliar ways, and in this sense is closely related to Lave and Wenger's (1991) theory of situated learning (or 'communities of practice' approach), which focuses on how individuals acquire professional skills in relation to their social environments. In such situations, situated cognition enables

[1] I expand on this in my extended note on efforts to frame the plurality of research experiences as a continuum of practice.

observers to examine how people align their perspectives as work progresses, and to understand better how people's general outlooks may have changed under the guidance of more experienced mentors. In other words, situated cognition enables researchers of scientific practices to account for discursive aspects of work, including perceived relationships, distinctions or intersections between practices that professional or research communities deem acceptable and unacceptable, and the cultural or community-driven aspects of decisions that underlie particular actions.

In taking on this theoretical framework, I frame epidemiology as a collective endeavour to derive a coherent understanding of population-level health trends, which involves the use of already established knowledge in the validation of newly formed ideas, and which relies on systems designed to carry information obtained with different chains of inference. These systems have both technical and social elements. The technical elements are the means through which information becomes encoded onto information objects so that they may form the basis for further inference. The social elements constitute a series of norms or expectations that facilitate the delegation of roles and responsibilities among agents who contribute their time, effort and accumulated knowledge to communal goals.

As such, in constructing the arguments of this study and in carrying out the fieldwork that grounds it, I will rely upon both realist and constructivist viewpoints. In one sense, I rely on documenting how people actually act, including the longer-term and collaborative implications that their actions may have on other work occurring throughout the continuum of practice. To accomplish this, I identify research activities from the perspective of an outside observer. I also ascribe meanings to things (such as physical or conceptual tools, or objects that captivate subjects' interests) in ways that conform to my own perspective as an investigator of scientific research practices. On the other hand, a constructionist perspective enables me to consider how individual agents make components of information systems suit their needs to facilitate communication or interoperability among actors who hold different situated perspectives. By listening to participants' views about the systems with which they engage, including explanations as to why they

act in the ways that they do, I am able to trace the assumptions and taken-for-granted behaviours that frame their perspectives. Moreover, these insights are useful for developing a better understanding of how participants identify with particular disciplinary communities and their perception of their roles within broader collective efforts.

Ultimately, this study is about the social order of scientific research, i.e. the frameworks, mindsets or sets of values that humans adopt to carry out their work in specific ways. Human beings rely upon physical and conceptual apparatus to do this work but, in order to understand how they do science *in ways that conform to the epistemic mandates of the scientific enterprise,* it is necessary to prioritize attention to human intention, drivers and pressures. I am emphasizing the agency of human drivers — as opposed to tools and procedures[2] — since humans are the ones who (a) identify problems that need to be resolved; (b) imagine, project or predict potential outcomes of various kinds of actions that they may select to resolve the challenges; and (c) learn from prior experiences and change their behaviours accordingly. By highlighting how pragmatic actions are conducted in relation to broader social and discursive trends and tendencies, I consider scholarly practices in terms of potential, certainty and desire from the perspectives of practitioners themselves.

[2] Human and non-human agents are considered on equal footing under the Actor-Network Theory (ANT) framework, which has become very popular since its origins in the late 1980s, but which may not be suitable for this approach. See my extended note on this for further details.

Data

This project will draw from around 12-15 interviews with epidemiological researchers who lead and carry out initiatives to share and harmonize data. Interviewees will comprise members of several projects that partner with the Maelstrom research group, a service that supports data harmonization efforts in the field of epidemiology. See the case selection strategy document for further information on how cases are decided upon.

Interviews will be oriented by my goal to document processes of reconciling different stakeholders' interests as they converge in the formation of a common data stream. Specifically, interviews will focus on motivations for their initiatives, the challenges they experience, how they envision success and failure,

their perceptions of their own roles and the roles of other team members and stakeholders, the values that inform their decisions, how the technological apparatus they set up enables them to realize their goals and values, and ways in which they believe data-sharing could be improved. See the interview protocol for further details on the questions I will ask and how their responses will contribute to the project's findings, as well as logistical considerations.

I will transcribing the interviews and edit the transcripts so that they conform with transcript notation standards (e.g. GAT-2, see Selting, Auer, and Barth-Weingarten 2011) and so they are optimally formatted for application in qualitative data analysis software. I will use automated speech recognition and natural language processing to create preliminary transcripts, which I will then manual edit.

I will collect and handle all data in full compliance with the ethics protocol. I will present an overview of the research objectives to all participants and obtain verbal consent prior to each interview before proceeding. Interviewees will be given the option to obfuscate any identifying information in processed records, and I will reiterate this option immediately after the interview and in a follow-up email one week following the interview.

Data will be curated according the data management plan.

Methods

I will perform qualitative data analysis (QDA) methods to highlight collaborative aspects of data harmonization work, as elicited in the interviews. QDA involves encoding the primary sources of evidence in ways that enable a researcher to draw cohesive theoretical accounts or explanations. This is done by tagging segments of a document (such as an interview transcript) using codes, and by embedding open-ended interpretive memos directlly alongside the data. Through these methods, I am able to articulate theories based on empirical evidence that reflect my informants' diverse experiences.

Coding — which involves defining what specific elicitations are about in terms that are relevant to the theoretical frameworks that inform my research — entails rendering instances within a text as interpreted abstractions called codes (Charmaz 2014, 43). Codes can exist at various levels of abstraction. For instance, I may apply descriptive codes to characterize literal facets of an instance within a text, and theoretical codes to represent more interpretive concepts that correspond with aspects of particular theoretical frameworks. I tend to create codes on the fly as "open codes" when prompted by encounters with demonstrative instances in the text. However, as I create new codes, I situate them within a hierarchical code system that provides me with a rough taxonomic structure to help organize my work and to enable me to more effectively query the data. Coding in this manner involves synthesis of concepts that speak to my understanding of the phenomena of interest, while forcing me to remain receptive to limits imposed by what is actually contained in the text. In other words, coding involves applying a precise language to segments of transcribed interviews that serve to bridge the gap between what participants said and the theoretical frameworks I apply to explore them as epistemic activities, interfaces and values (cf. Charmaz 2014; Saldaña 2011, 95–98).

Memoing entails more open-ended exploration and reflection upon latent ideas in order to crystallize them into new avenues to pursue (Charmaz 2014, 72). Constructing memos is a relatively flexible way of engaging with data and serves as fertile ground for honing new ideas. Memoing is especially crucial while articulating sensitizing concepts, which Charmaz (2003, 259) refers to as the "points of departure from which to study the data". Memoing allows me to take initial notions that lack specification of well-defined attributes, and gradually refine them into more cohesive, definitive concepts (Blumer 1954, 7; Bowen 2006). Exploring the main features, relationships or arrangements that underlie a superficial view of a sensitizing concept through memoing helps me to identify what kinds of things I need to locate in the data in order to gain a full understanding of the phenomena of interest. Memoing is also very important in the process of drawing out more coherent meaning from coded data (cf. Charmaz 2014, 181, 290–93). By creat-

ing memos pertaining to the intersections of various codes and drawing comparisons across similarly coded instances, I am able to form more robust and generalizable arguments about the phenomena of interest and relate them to alternative perspectives expressed by others.

Throughout my analysis, I will follow the approach that Nicolini (2009) and Maryl et al. (2020, para. 30) advocate, who suggest "zooming in to a granular study of particular research activities and operations and zooming out to considering broader sociotechnical and cultural factors." This involves "magnifying or blowing up the details of practice, switching theoretical lenses, and selective re-positioning so that certain aspects are fore-grounded and others are temporarily sent to the background" (Nicolini 2009, 1412). This approach is useful for me because research projects all start from different positions but share common practices and tendencies that vary according to those contextual circumstances. I am therefore able to tactfully switch between those lenses to understand the interplay between circumstances and practical implementations, which vary across cases, which have their own histories, memberships, sets of tools, methods, and social or political circumstances.ships, sets of tools, methods, and social or political circumstances.

I will perform all of this work using MaxQDA, a QDA software suite that stores all of these connections within a centralized database (VERBI Software 2021).[3] This allows me to retrieve segments of text from across various documents that have been assigned the same sets of codes, and perform more complex queries that search along different parameters of overlap, intersection, and exclusion. I will then be able to query the integrated dataset to produce elaborated accounts of specific kinds of activities, decisions, values and sentiments.

See the QDA protocol for further details on the code system and memoing guidelines, as well as specific QDA procedures.

[3] I am most familiar with MaxQDA from my doctoral research. NVivo is another popular proprietary software suite with overlapping functionality, and QualCoder, OpenQDA and QCoder are open source implementations. In my experience, MaxQDA is the most feature-rich option and has a well-crafted interface that suits my needs well.

Expected Outcomes and Impact

Scholarly outputs

This project will produce three articles. One of these will be published in a journal concerned with scientific practice or research data management (e.g. Computer Supported Co-operative Work). Another will be published in a journal dedicated to advancing research practices in epidemiology (e.g. Epidemiologic Methods) Finally, I will publish a more practically-oriented set of guidelines deriving from the findings, as either an editorial or as a "10 simple rules" style article.

I will also present this work at conferences and workshops, but specifics have yet to be determined.

Practical outcomes

While this research is critical, it is intended to be constructive. It is therefore necessary to ensure that the findings may be put to practical use so as to enhance and improve data-sharing initiatives. While a "10 simple rules" article will be a handy brief of my findings, I would also like to contribute to the development of data-sharing infrastructures and policies in a more active manner. Specifically, I will reach out to leading stakeholders directly by highlighting our common interest in developing more effective data-sharing infrastructures and presenting a series of practical ideas for improving their services. I will ensure that these meetings conclude with a series of action items which will enable us to foster an outcome-oriented and productive relationship directed toward achieving tangible goals.

Still in progress...

```
- Support efforts to enhance data reuse potential of the CITF Databank specifically
- Submit comments to CIHR with regards to their ongoing revisions to federal open science polic
```

Broader outreach

Still in progress...

```
- Submit something to [The Conversation](https://theconversation.com/ca)
- Collaboration with the [School of Information Studies](https://www.mcgill.ca/sis/)
```

Timelime

Date	Milestone
2025/01/14	Finalize ethics protocol
2025/01/31	Finalize case selection and invite members to participate
2025/01/21	Finalize data managemenrt plan
2025/02/14	Finalize interview protocol
2025/02/14 — 2025/03/31	Conduct interviews
2025/04/01 — 2025/04/30	Prepare interview transcripts for analysis
2025/05/01 — 2025/05/31	Prepare code system and develop sensitizing concepts
2025/06/01 — 2025/09/30	Qualitative coding and memoing
2025/10/01 — 2025/11/30	Draft manuscripts and submit for publication
2025/11/01 — 2025/12/31	Write accessible and constructive reports

Blumer, Herbert. 1954. "What Is Wrong with Social Theory?" *American Sociological Review* 19 (1): 3–10. https://doi.org/10.2307/2088165.

Bowen, Glenn A. 2006. "Grounded Theory and Sensitizing Concepts:" *International Journal of Qualitative Methods* 5 (3): 12–23. https://doi.org/10.1177/160940690600500304.

Charmaz, Kathy. 2003. "Grounded Theory: Objectivist and Constructivist Methods." In *Handbook of Qualitative Research*, edited by Norman K. Denzin and Yvonna S. Lincoln, 2nd ed., 249–91. Thousand Oaks, California: SAGE.

———. 2014. *Constructing Grounded Theory*. 2nd ed. SAGE.

Knorr Cetina, Karin. 2001. "Objectual Practice." In *The Practice Turn in Contemporary Theory*, edited by Theodore R. Schatzki, Karin Knorr Cetina, and Eike von Savigny, 175–88. London; New York: Routledge.

Lave, Jean, and Etienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge, United Kingdom: Cambridge University Press.

Maryl, Maciej, Costis Dallas, Jennifer Edmond, Jessie Labov, Ingrida Kelpšienė, Michelle Doran, Marta Kołodziejska, and Klaudia Grabowska. 2020. "A Case Study Protocol for Meta-Research into Digital Practices in the Humanities." *Digital Humanities Quarterly* 14 (3). https://www.digitalhumanities.org/dhq/vol/14/3/000477/000477.html.

Nicolini, Davide. 2009. "Zooming In and Out: Studying Practices by Switching Theoretical Lenses and Trailing Connections." *Organization Studies* 30 (12): 1391–1418. https://doi.org/10.1177/0170840609349875.

Pickering, Andrew. 1992. "From Science as Knowledge to Science as Practice." In *Science as Practice and Culture*, edited by Andrew Pickering, 1–26. University of Chicago Press.

Saldaña, Johnny. 2011. *Fundamentals of Qualitative Research*. Understanding Qualitative Research. New York: Oxford University Press.

Selting, Margret, Peter Auer, and Dagmar Barth-Weingarten. 2011. "A System for Transcribing Talk-in-Interaction : GAT 2." *Gesprächsforschung : Online-Zeitschrift Zur Verbalen Interaktion* 12:1–51. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/166.

Suchman, Lucy. 2007. *Human-Machine Reconfigurations: Plans and Situated Actions*. 2nd ed. Cambridge University Press.

VERBI Software. 2021. "MaxQDA 2022." Berlin. https://www.maxqda.com.