Research Protocol*

Articulating epidemiological data harmonization initiatives as practical and collaborative experiences

Zack Batist

2025-01-22

Overview

This study investigates the social, technical, administrative and epistemic factors that scaffold 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, the study approaches data-sharing as a means of reconciling the varied circumstances of datasets' creation — both among themselves, and in relation to contexts of reuse. It therefore frames data-sharing as efforts to foster a series of collaborative ties beyond a project's original indended scope.

Data harmonization is one means of data-sharing that draws multiple studies' recorded observations into a unified formal schema, whose structure is driven by specific objectives and more general underlying suppositions and values. Although these schemas are arrived at through discussion, compromise and consensus-building, with an eye toward practical outcomes afforded by alignment of complementary records, these socially-mediated interactions are generally under-recognized as important factors contributing to data-sharing initiatives' success relative to their potential impact. The goal of this study is to survey how various factors are prioritized in harmonization activities, the rationales behind these decisions, and the relative efficacy of different approaches to data-sharing.

The project's 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 different 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?

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

- 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, the 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, the goal is to reveal the diverse ways in which data-sharing occurs, and how different approaches impact the outcomes in different ways.

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, the study documents the social and collaborative experiences involved in various research practices, which ultimately bind the many ways in which scientists do science.

The study 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 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,

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

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, the study frames 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 interviews that grounds it, the study will rely upon both realist and constructivist viewpoints. In one sense, the study relies 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, the study identifies research activities from the perspective of an outside observer. The study also ascribes meanings to things (such as physical or conceptual tools, or objects that captivate subjects' interests) in ways that conform to the analyst's own perspective as an investigator of scientific research practices. On the other hand, a constructionist perspective enables the author 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. The study emphasizes the agency of human drivers — as opposed to tools and procedures — 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.[2] By highlighting

[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.

how pragmatic actions are conducted in relation to broader social and discursive trends and tendencies, the study considers scholarly practices in terms of potential, certainty and desire from the perspectives of practitioners themselves.

To this end, the study follows an abductive qualitative data analysis (QDA) methodology to construct theories founded upon empirical evidence, which relates to, but is distinct from, grounded theory. Grounded theory consists of a series of systematic yet flexible guideline for deriving theory from data through continuous and reiterative engagement with evidence (Charmaz 2014: 1). The approach taken for this study draws from what Charmaz (2014: 14-15) calls the "constellation of methods" associated with grounded theory that are helpful for making sense of qualitative data. However, it differs from grounded theory as it is traditionally conceived in that I came to the project with well defined theoretical goals (as described above) and did not make a concerted effort to allow the theory to emerge through the analytical process. Proponents of a more open-ended or improvised approach, as grounded theory was originally applied, argue that researchers should be free to generate theories in accordance with their own creative insights and their intimate engagements with the evidence. We can evaluate the quality of such work in terms of the dialogical commitments between researchers and their subjects, and between researchers and those who read their work (Glaser and Strauss 1967: 230-233). Others view grounded theory more as a means of clarifying and articulating phenomena that lie below the surface of observable social experiences (Strauss and Corbin 1990; Kelle 2005). Proponents of this approach are very concerned with ensuring that concepts, themes and theories are truly represented in and limited by the data, and therefore prioritize adherence to systematic validation criteria to ensure the soundness of their claims.

Another view, known as constructivist grounded theory, most resembles the approach taken for this study. It recognizes that it is impossible to initiate a project without already holding ideas regarding the phenomena of interest, and that the ways that one ascribes meanings to the data represent already established mindsets or conceptual frameworks (Charmaz 2014). It encourages reflection on the researcher's standpoint as they pursue an abductive approach rooted in their own preconceptions (Mills, Bonner, and Francis 2006).

All of these approaches rely on a core set of methods of coding and memoing. Coding, which involves defining what data are about in terms that are relevant to the theoretical frameworks that inform the research, entails rendering instances within a text as interpreted abstractions called codes (Charmaz 2014: 43). These methods, which are described in more detail below, are particularly useful for examining the broad assemblage of evidence comprising various kinds of media and spanning multiple case studies. The abstraction of specific instances as conceptual codes enables comparisons across documents that would otherwise prove difficult to compare, due either to the analyst's own preconceptions (drawn from internalized narratives or biases) that might have framed their attitudes, to disproportionate volumes of evidence that might obscure parallels between case studies, or to difficulties experienced when examining different kinds of documents that call for different lenses or perspectives.

Data

The study draws from semi-structured interviews with 12-15 individuals from 4-5 cases who lead, support or participate in epidemiological data-sharing initiatives. These individuals include profesional researchers, research trainees and administrative and technical support staff affilited with epidemiological projects that coordinate, support or participate as a member of data-sharing initiatives.

Only data-sharing initiatives that partner with the Maelstrom Research project, which facilitates collaborative epidemiological research through rigorous data documentation, harmonization, integration, and co-analysis, will be considered to serve as cases for this study. Maelstrom is a well-established entity in this field and has established a broad network of partner projects to select cases from. Maelstrom serves as a "fixed point" which ensures that the researchers and study participants share a common frame of reference. Maelstrom established generalizable nomenclature and toolsets across all its partner projects, which reduces the overhead of matching different terms and and practices across cases. This is especially valuable in the context of interview-based research, wherein it is crucial to remain focused on obtaining information that is relevant to the themes the project seeks to address during the limited time alloted. Additionally, Maelstrom's principal investigators and partners have written extensively about the values and practical challenges concerning data-sharing in epidemiology, which provides a rich foundation upon which the analysis may be based (cf. Demir and Murtagh 2013; Fortier et al. 2017; Bergeron et al. 2018; Fortier et al. 2023).[3] See the case selection strategy document for further information on how cases are decided upon.

The project seeks to interview 3-4 people from each case, including individuals who work in specific roles, such the leaders of data-sharing consortia, support staff, and leaders of contributing projects. Interviews are oriented by the study's goal to document processes of reconciling different stakeholders' interests as they converge in the formation of a common data resource. 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 participants' responses will contribute to the project's findings, as well as logistical considerations.

The number of cases reflects the capacity to draw adequate comparison across unique circumstances while also complementing the meaningful number of individuals who may serve as interview participants. Breadth of perspective is of greater concern than sample size in qualitative research following the constructivist grounded theory methodological framework. Since this study is rooted in case study methods, and the goal is to articulate the series of inter-woven factors that impact how epidemiological researchers coordinate and participate in

[3] See my extended notes on these works and on additional related studies: zackbatist.info/CITF-Postdoc/notes/maelstrom-readings.

data-sharing initiatives while explicitly accounting for and drawing from the unique and situational contexts that frame each case, the goal is not to define causal relationships or to derive findings that may be generalized across the whole field of epidemiology. As such, statistical representativeness is not an objective of this research.

At the same time, the number of individuals also reflects the practical constraints that this work affords, namely the time-consuming nature of transcribing interviews and conducting qualitative data analysis. My experience leading projects of similar scale will enable me to collect, process and analyze such a comprehensive corpus in the relatively short period in which funding has been allotted. The number of participants therefore represents a careful balance between a meaningful sample size and the amount of work required to collect, process and analyze the data within the project's one-year timeframe.

Interviews will be transcribed, and transriptions will be edited to optimize them for use in qualitative data analysis software. Secure and locally-hosted automated speech recognition software may be used to create preliminary transcripts, which will then be manually edited. All data will be collected and curated in full compliance with the ethics protocol and in accordance with the data management plan.

Methods

The study will implement qualitative data analysis (QDA) methods to highlight collaborative aspects data-sharing in epidemiology, as elicited in the corpus of transcribed 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, a researcher is able to articulate theories based on empirical evidence that reflect the informants' diverse experiences.

Coding — which involves defining what specific elicitations are about in terms that are relevant to the theoretical frameworks that inform the 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, an analyst 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. This project will primarly implement an "open" coding protocol, which entails creating codes on the fly when prompted by encounters with demonstrative instances in the text. As new codes are generated in this manner, they are situated within a code system that affords greater taxonomic structure to encoded observations, thereby facilitating more effective queries. Coding in this manner involves synthesis of concepts that speak to the analyst's understanding of the phenomena of interest, while forcing the analyst to remain receptive to limits imposed by what is actually contained in the corpus. 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 that the analyst applies 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 the researcher 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 the analyst to identify what kinds of things they 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 creating memos pertaining to the intersections of various codes and drawing comparisons across similarly coded instances, an analyst is able to form more robust and generalizable arguments about the phenomena of interest and relate them to alternative perspectives expressed by others.

Throughout the 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 in the context of this study because the research projects that represent the cases start from different positions but share common practices and tendencies that vary according to those contextual circumstances. It is therefore possible 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.

This work will be performed using computer assisted qualitative data analysis software that enables analysts to retrieve segments of interview transcripts and identify patterned distributions of codes from across the entire corpus. Querying the dataset in this way enables the analyst to articulate elaborated accounts of specific kinds of activities, decisions, values and sentiments that cut across various informants' perspectives.

Statistical methods will play a limited role in this study. Basic summary statistics (e.g. cross tabulation) will be used to represent the distribution of codings across individual interviews or ranges of interviews, which will help to identify trends and associations as they pertain to their limited scopes. This will be used to support theory-building but will not be used to infer generalizable causal relationships.

See the QDA protocol for further details on the code system and memoing guidelines, as well as specific QDA procedures, and my methodology notes that situate this's project's methods in relation to alternative approaches.

Outcomes

This project will produce insights regarding the practical benefits and challenges involved in epidemiological data-sharing. It will identify how relevant stakeholders *actually* engage with the systems that scaffold data-sharing initiatives, which may differ from modelled behaviours specified in aspirational plans and procedural documents. In effect, by articulating how these systems succeed or fail to account for their users practical needs and disciplinary values, this study will provide constructive feedback that will inform their further development.

The study will produce three peer-reviewed articles. One of these will be published in a journal concerned with scientific practice or research data management (e.g. Computer Supported Cooperative Work; Scientific Data); Another will be published in a journal dedicated to advancing research practices in epidemiology (e.g. Epidemiologic Methods) A third paper will comprise a more practical 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, as opportunities arise.

Moreover, this work is intended to be constructive, and it is therefore necessary to ensure that the findings may be put to practical use so as to enhance and improve data-sharing initiatives. I will therefore draft policy briefs and reach out to leading stakeholders at the helm of major data-sharing infrastuctures and policy frameworks (such as the ongoing revisions to federal open science policies) so that the findings may directly inform efforts to improve these systems.

Additionally, I will promote this work publicly. This will entail publishing an article in The Conversation, a website with a broad public following and which specializes in showcasing specialized research for the general public. I may also share the findings on podcasts about open science and science policy with broad interdisciplinary appeal. Moreover, owing to my broad pan-disciplinary background, I am plugged into a diverse network of scholars, and my work will reach a very broad audience through active my engagement on social media and posting regular updates on my professional blog.

Ethics

The study will be conducted according to ethical principles stated in the Declaration of Helsinki (2013). Ethics approval will be obtained before initiating the study. Consent forms will take into consideration the well-being, free-will and respect of the participants, including respect of privacy. The practices undertaken to ensure adherence to these principles are described in the ethics protocol.

Timelime

Date	Milestone
2025/01/14	Finalize ethics protocol
2025/01/21	Finalize data management plan
2025/01/31	Finalize case selection and invite members to participate
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

References

Bergeron, Julie, Dany Doiron, Yannick Marcon, Vincent Ferretti, and Isabel Fortier. 2018. "Fostering Population-Based Cohort Data Discovery: The Maelstrom Research Cataloguing Toolkit." *PLOS ONE* 13 (7): e0200926. https://doi.org/10.1371/journal.pone.0200926.

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, 3rd ed., 249–91. Thousand Oaks, California: SAGE.

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

Demir, Ipek, and Madeleine J. Murtagh. 2013. "Data Sharing Across Biobanks: Epistemic Values, Data Mutability and Data Incommensurability." *New Genetics and Society* 32 (4): 350–65. https://doi.org/10.1080/14636778.2013.846582.

Fortier, Isabel, Parminder Raina, Edwin R Van den Heuvel, Lauren E Griffith, Camille Craig, Matilda Saliba, Dany Doiron, et al. 2017. "Maelstrom Research Guidelines for Rigorous Retrospective Data Harmonization." *International Journal of Epidemiology* 46 (1): 103–5. https://doi.org/10.1093/ije/dyw075.

Fortier, Isabel, Tina W. Wey, Julie Bergeron, Angela Pinot de Moira, Anne-Marie Nybo-Andersen, Tom Bishop, Madeleine J. Murtagh, et al. 2023. "Life Course of Retrospective Harmonization Initiatives: Key Elements to Consider." *Journal of Developmental Origins of Health and Disease* 14 (2): 190–98. https://doi.org/10.1017/S2040174422000460.

Glaser, Barney G., and Anselm L. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. https://doi.org/10.4324/9780203793206-1.

Kelle, Udo. 2005. ""Emergence" Vs. "Forcing" of Empirical Data? A Crucial Problem of "Grounded Theory" Reconsidered." *Forum: Qualitative Social Research* 6 (2): 133–56. https://doi.org/10.17169/FQS-6.2.467.

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.

Mills, Jane, Ann Bonner, and Karen Francis. 2006. "The Development of Constructivist Grounded Theory." *International Journal of Qualitative Methods* 5 (1): 25–35. https://doi.org/10.1177/160940690600500103.

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.

Strauss, Anselm L., and Juliet M. Corbin. 1990. *Basics of Qualitative Research*. Sage Publications.

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

World Medical Association. 2013. "World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects." *JAMA* 310 (20): 2191. https://doi.org/10.1001/jama.2013.281053.