FA25 Foundations of Data Curation, UIUC MCS Online

CS598 Course Project Pilot: End-to-End Data Curation Workflow

Jeptha Davenport (wjd4@illinois.edu)

Overview Scientific examination of human consciousness spans diverse fields including philosophy, medicine, biology, and physics. Differing opinions reflect lack of consensus on many aspects of the problem, its formulation, framing, and relevance. It remains an unsettled problem in scientific pursuit. Publication in this area is similarly wide-ranging and voluminous. Recent attempts at describing existing work on consciousness demonstrate a broad and interesting corpus which remains to be organized.[1]

I have a longstanding personal interest in the question of human (and other) consciousness, even from childhood. My education (philosophy, medicine, health humanities, computer science) and career path (clinical neurology) continue to circle the question. Within the Foundations of Data Curation computer science course (CS 598, Fall 2025) of the UIUC Master of Computer Science program, I propose to apply relevant skills of data curation to the problem of consciousness:

Project Proposal

1. **Plan** The use case for this project is to curate a dataset of publications (periodical print, book, and electronic equivalents) related to the field of consciousness. Such a database will be useful for several reasons, including allowing an analysis of current and past research efforts, organizing the field and tracking its progress, and serving as a common point of reference for future studies.

2. A lifecycle model for this dataset will follow the first 4 steps of the United States Geological Survey (USGS) lifecycle model[2] of a course of actions (Plan, Acquire, Process, Analyze), deferring the steps of Preserve and Publish/Share for a post-project effort and preserving the USGS Data Lifecycle crosscutting elements of Describe (metadata, documentation), Manage Quality, and Back up and secure.[3] At the time of this proposal, this project is anticipated as part of coursework and represents the work of one student with advice from the Foundations of Data Science instructors, but the hope is that later work outside of the course can build on this foundation and continue to the Preserve and Publish/Share aspects of the lifecycle model. **Team** currently consists of a single member, Jeptha Davenport. I will be responsible for all workflow including data

---

[1] See Kuhn RL. A landscape of consciousness: Toward a taxonomy of explanations and implications. Prog Biophys Mol Biol. 2024 Aug;190:28-169. doi: 10.1016/j.pbiomolbio.2023.12.003. Epub 2024 Jan 26. PMID: 38281544

[2] J.L. Faundeen, T.E. Burley, J.A. Carlino, D.L. Govoni, H.S. Henkel, S.L. Holl, et al., "The United States Geological Survey Science Data Lifecycle Model," 2013, as cited in: Plale, Beth & Kouper, Inna. (2017). The Centrality of Data: Data Lifecycle and Data Pipelines. 10.1016/B978-0-12-809715-1.00004-3., p. 95

[3] Plale, Beth & Kouper, Inna. (2017). The Centrality of Data: Data Lifecycle and Data Pipelines. 10.1016/B978-0-12-809715-1.00004-3., p. 95

acquisition, curation, archive and process documentation (expected to be available as part of progress reports and in the final report and attached to the subsequent dataset in a combination of text documents and database).

3. (Dataset lifecycle model stage) PLAN Pre-search, I estimate there exist between 10^3 and 10^6 relevant records, a breadth which depends heavily on inclusion/exclusion criteria. The initial search will be of the PubMed Database from the US National Library of Medicine (PubMed). A broad search initially will be extended if search results are below the lower end of this estimated range, with GoogleScholar and Web of Science as optional citation databases for search extension. If search results are at the high end of the pre-search estimate, revisions to selection criteria will be considered. This will include an initial date range of 1996 to 2025, aiming to keep the record count below 10^6 records for this pilot project. Pre- and post-estimates of publication counts (records) will be gathered and refined, and the process of search and selection will be documented for review, discussion, and replication. (Quantitative and qualitative information will be gathered for this and subsequent steps.)

4. **Data sources** (Data lifecycle model stage) ACQUIRE PubMed will be queried programmatically to identify publications relevant to consciousness. Manual probing to refine or enlarge search criteria will be documented ad hoc. Metadata for publications will be collected, including, where available, DOIs, URLs, authors, author affiliations, author collaborations, abstracts, and licensing information (open access, public domain, licensed). The format for collected data will be Structured Query Language (SQL) accessible. The database format will be chosen using Findable, Accessible, Interoperable, and Reusable (FAIR) principles, with candidates to include Comma-Separated Values (CSV) or SQLite (both widely used and efficient for small- to medium-sized databases). Storage will be on a personal device with backup using a personal GitHub repository with restricted access initially. Beyond the scope of this pilot project would be the option of open-access database for subsequent stages outside of this coursework.

5. (Data lifecycle model stage) PROCESS Manual and automatic curation will be applied (and documented) with goals of classification, normalization, and metadata standardization. Initially metadata will be collected in the schema.org format mirroring DataCite metadata, with categories added to account for special aspects of consciousness, such as index species (human or other, including, for example, other primates, cetacean, cephalopod, avian, arthropod or cnidarian), overlapping fields of relevance (physics, cell biology, philosophy, and so on), and consciousness-specific labels (e.g., 'schools of thought' such as integrated information theory (IIT) or global workspace theory (GWT), panpsychism or intractable. Summary tables for overarching themes encountered during dataset construction will be created with classification schematics where useful.

6. (Data lifecycle model stage) ANALYZE The resulting dataset will be presented along with an overview of example analyses and data visualization which follow. Questions of reproducibility will be raised and discussed, as well as flagging ethical considerations on data sourcing (see also 8. Constraints). The process of this inquiry will be critiqued and summarized in a discussion of problems, solutions, and future directions. (Beyond the

scope of this project, peer review and publication may be sought as further contribution to the field.)

7. **Timeline** Subject to approval, the project will run over 160 hours (16 hours per week for 10 weeks) from September 22 through December 1, 2025, with an interim progress report and opportunity for corrections or refinement at the 5-week mark, October 20, 2025, edited as needed given feedback for the planned **progress report due date of October 27, 2025**. The final report draft on the project presentable for feedback and commentary from instructors (as available) will be made by December 1, 2025, with **final submission to occur by December 10, 2025**.

8. **Constraints (A)** This initial proposal supposes that adequate information for analysis will be contained within open access materials including especially abstracts of available studies/reports/publications. If important quantities of material are contained uniquely in sources which are hidden behind paywalls, or require institutional or personal subscriptions, this may present a barrier to best dataset construction. As copyright and/or closed sourcing is expected to occur to some degree, data regarding cost of access will be gathered when not clearly open access (OA). Considerations for acquiring needed data through institutional access through the University of Calgary (where I hold a clinical appointment) or through the University of Illinois (where I hold a position as a graduate student) for purposes of scholarly research or personal research will be made, cataloguing legal and ethical obligations as they appear and consulting with university copyright teams as needed (the Copyright Office at the University of Calgary; the University of Illinois Urbana-Champaign Fair Use Guidance copyright resource and Copyright Illinois[4]). For Canadian regulations which are applicable, the Canadian Copyright Act[5] and Universi**t**y of Calgary Acceptable Use of Material[6] will be consulted. **(B)** An additional constraint is that this project does not have external funding, which could limit longterm dataset curation and dissemination, but if successful as a pilot project, it could serve as a starting point for grant applications or be incorporated into an existing consciousness study program or other hosting sites for a dynamic repository.

9. **Gaps (A)** Another important potential shortcoming for this project beyond data access is that large texts may prove difficult to classify programmatically through metadata alone. The degree to which this occurs will be noted and discussed, with a view towards possible solutions using natural language processing or large language models, beyond the scope of this limited project but within the purview of larger aims and goals subsequently. **(B)** Finally, a high degree of data heterogeneity is anticipated for reasons discussed in the overview. This may interfere with a purely automated process and would entail manual classification, re-classification, and interim analysis interspersed with programmatic script modifications; nonetheless, clear documentation of such processes may allow for reproducibility in the future, for extension as the consciousness corpus

---

[4] https://copyright.illinois.edu/

[5] https://laws-lois.justice.gc.ca/eng/acts/C-42/Index.html

[6] https://www.ucalgary.ca/legal-services/acceptable-use-material-protected-copyright-policy

grows, for reproduction as part of scientific validation, and to serve as a template for parallel efforts in other multidisciplinary fields.

Reviewer Comments and Feedback