FA25 Foundations of Data Curation, UIUC MCS Online

CS598 Course Project Progress Report: End-to-End Data Curation Workflow

Team: Jeptha Davenport (wjd4@illinois.edu)

Project Summary Scientific examination of human consciousness has increased publication volume across fields and remains to be organized.[1] The project undertaken for this course aims to create a database of relevant publications in the field of consciousness using the first four steps of the USGS lifecycle model[2] for a cumulative dataset.

Status Stages of Plan, Acquire, Process, and Analyze have begun. This interim progress report describes the status of deliverables. Artifacts generated so far are accessible publicly at the following repository: https://github.com/wjdavenport/FDC-Project-Backup

N.B. One-time dataset storage occurs in a pinned submodule, for archival purposes: https://github/wjdavenport/FDC-Project-Data

Please refer to the Project Proposal document previously submitted and archived in the repository for additional preceding details.[3]

Plan Progress: in a modification to the Project Proposal, the Acquire stage increased the date range for an initial search of PubMed (1996 to 2025) to 1843 to 2025 (retrieval date fixed at 2025-09-27). The reason for the increase in range was that the effort required to capture extra publications dated from 1843-1995 was not excessive given the long leftward tail of an exponentially increasing annual publication rate, and the earlier date marks the current limit of PubMed's earliest publications (Figure 1). With this modification, the total number of publications captured in the PubMed search was just over 65,000, which was well within the Project Proposal estimates (< 10^6).

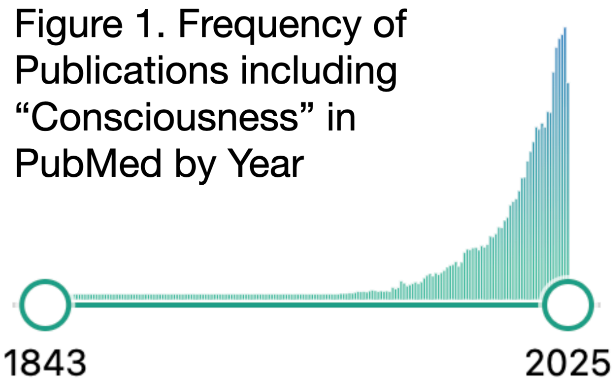


Figure 1. Frequency of Publications including "Consciousness" in PubMed by Year

[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] https://github.com/wjdavenport/FDC-Project-Backup/blob/main/Proposal%20and%20Guide/FA25%20FDC%20Proposal%20wjd4%20v03.pdf

Plan - Next Step: finalize decision re: pilot add-on or core focus in **Analyze** section (November 10, 2025)

Acquire Progress: a programmatic acquisition was accomplished following the National Library of Medicine Entrez Programming Utilities (E-utilities) guidelines.[4] Limits followed included not making more than three URL requests per second and performing the requests during off-hours (weekends and 21:00-05:00 Eastern Time). Although the search was performed with an active ORCID (available through no-cost registration[5]) through an NCBI api key (also available at no-cost with registration[6]), it would be possible for subsequent researchers to perform the same search without registration. Finally, to avoid overly large download blocks (i.e., respecting the 10k idlist limit), a helper routine was created with an output log of the date-count ranges used to keep within such a limit. The R code used is kept at: acquire/r-project/01_download_pubmed_consciousness.R within the public GitHub repository[7] as well as on a personal storage device.

Acquire - Next Step: export final R script manifest plus or minus search refinements and re-run of PubMed Search.

Process Progress: a Python script was created to convert the .medline file to .csv format (https://github.com/wjdavenport/FDC-Project-Backup/consciousness-ezr/scripts/01_export_pubmed_csv.py). As anticipated manual abstract analysis showed that a subset of flagged and acquired references was not relevant to the field of consciousness as opposed to the use of a MeSH as an indicator of level of arousal of human or animal study subjects. For example, many flagged studies referred to pharmacological, physiological or other interventions upon "conscious [rats, mice, sheep, monkeys, subjects]" while others referred to the presence or absence of consciousness as an indicator of coma or as a sequela to traumatic brain injury.

Based on the presence of such articles which ideally would be excluded from subsequent analysis, a system of parsing was developed beginning with a random sample of 300 flagged articles which were manually reviewed and annotated as 'irrelevant=0' or 'relevant=1' in an expanded .csv file (https://github.com/wjdavenport/FDC-Project-Backup/consciousness-ezr/scripts/02_make_seed_labels.py).

Process - Next Step: Refresh .medline to .csv conversion with any fixes and regenerate checksums.

Analyze Progress: seed labels were applied in a "bag-of-words"[8] linear text classifier implemented in Python (consciousness-ezr/scripts/03_train_baseline_and_ezr.py) using text

[4] https://www.ncbi.nlm.nih.gov/books/NBK25497/#chapter2.Usage_Guidelines, retrieved 2025-10-26

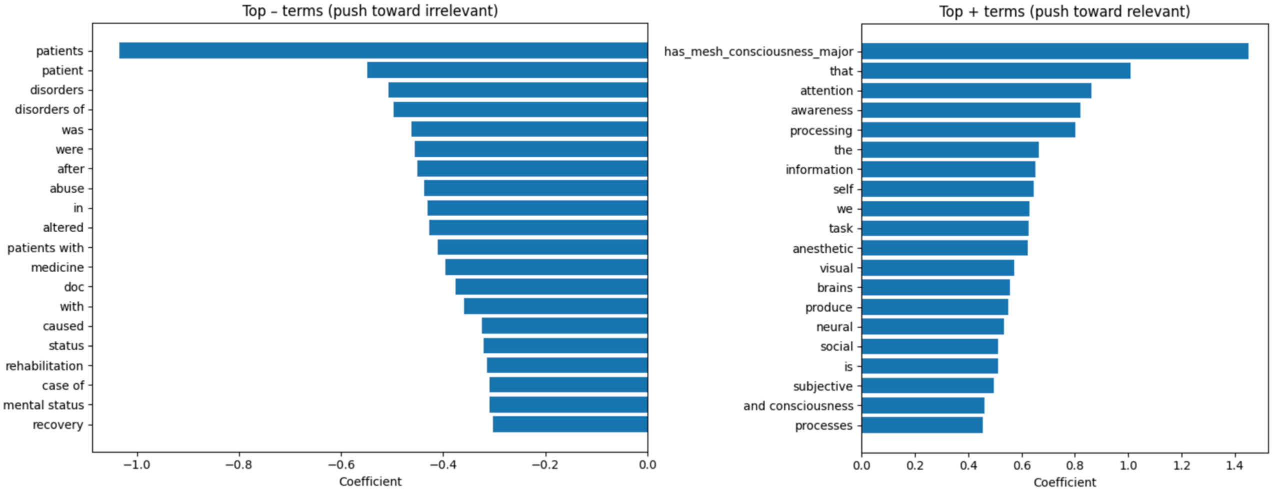
[5] https://orcid.org, retrieved 2025-10-26

[6] https://support.nlm.nih.gov/kbArticle/?pn=KA-05317, retrieved 2025-10-26

[7] https://github.com/wjdavenport/FDC-Project-Backup/

[8] https://en.wikipedia.org/wiki/Bag-of-words_model, retrieved 2025-10-26

Figure 2. Interim analysis of weighted n-grams for relevance of flagged articles including “consciousness”



feature of titles and abstracts as available (TfidfVectorizer, a scikit-learn module)[9] and modelled via binary logistic regression (LogisticRegression, also a scikit-learn module)[10]. The choice of this type of classification scheme was to preserve interpretability (since positive- and negative-weighted n-grams are produced and preserved, for example as in Figure 2).

Analyze **Future action**: A second round of manual review of random sampling of retained and rejected (‘relevant’- and ‘irrelevant’-labelled) references will be conducted after training to estimate accuracy of the training model. Further refinements can be considered if needed. Such refinements could include the explicit inclusion of known proposed named models of consciousness (e.g., Global Workspace Theory, Mind-Brain Identity Theory, Orchestrated Objective Reduction, Panpsychism, Monism, Dualism, Electromagnetic Theory, Neural Correlates, and so on) and further tests of elimination of “stop words” within the TfidfVectorizer. Additional inclusion features could include proper names of authors associated with theories of consciousness (e.g., Edelman, Crick and Koch, Dennett, Baars, McFadden, Searle, Chalmers) in abstracts, titles, or authors. It would also be possible to apply additional mini-models, including decision trees[11] or a previously considered rule-based learner (ezr[12]).

Analyze - Next Step: Second manual audit; report AUC results with precision confidence intervals; lock model artifacts.

[9] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html, retrieved 2025-10-26

[10] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html, retrieved 2025-10-26

[11] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier, retrieved 2025-10-26

[12] https://github.com/timm/ezr/blob/main/README.md, retrieved 2025-10-26

Progress update on **Constraints**: Thus far manual classification has been possible without recourse to bulk downloading of copyrighted references, and where questions of relevance arose (less often than 2% in sampling), single-reference review was available through institutional access (either University of Illinois Urbana-Champaign or University of Calgary) for the <10 references reviewed) for research and educational purposes of this project.

Progress update on **Gaps**: The metadata available through PubMed does not include publicly available resources including books, recorded lectures, interviews, conference talks, preprints, dissertations and theses, and liberal arts fields (humanities and philosophy) of relevance. At this checkpoint in the Project, focus will remain on refining the current PubMed search results and analysis so that a similar selection process can be applied to such search extensions beyond this initial project.

Future action and decision: I had originally hoped to include book publications by this stage of this project, but I am concerned that merging multiple source data before refining search criteria within PubMed and machine learning classifier tools may require more than 4-5 weeks.

Based on reviewer comments of this progress report, I will set a decision point on November 10, 2025:

Path A (default and included regardless of Path B implementation) Focus on PubMed data alone to deliver a cleaned dataset, model weights and interpretation figures, and a documented selection pipeline (estimated 32-40 hours remaining)

Path B (pilot add-on) Implement a books pilot using 1 to 3 curated bibliographies cross-matched to GoogleScholar/GoogleBooks to test a merger schema. **Scope of pilot add-on** (estimated 16-24 additional hours beyond Path A): I anticipate the number of published books relevant to the topic to be 10^2 (2 orders of magnitude less than journal references). The pilot would produce:

1. merger schema (book_id, title, authors, year, publisher, ISBN/DOI, subject keywords, citations (connected to PubMed PMIDs where available)

2. provenance log (source bibliography, retrieval date, query strings)

3. sample knowledge-graph sketch (author-work-cites relationships) to demonstrate a lineage and cross-pollination of ideas in the field of consciousness studies.

Internal criterion for success: If the books pilot is successful, 80% or more of sampled books will match the merger schema with (ISBN/DOI/Title+Author+Year) and show overlap with 20% or more of the PubMed dataset (i.e., journal articles cite books or books cite journal articles). If the books pilot is not successful, I will document this and defer the incorporation of books into this project timeline. The rationale I propose is that although books may require less time to sort manually (being significantly fewer in number), they will likely present fewer metadata clues to relevance as they will lack abstracts. A knowledge graph may address this gap. The pilot add-on would test the feasibility of a merger with the main deliverable dataset (PubMed journal articles).

Progress update on **Timeline** and **Future action**: 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. I apologize that an earlier draft of this document was not available as I had hoped by October 20, 2025; I will begin my final project report the 3rd week of November, 2025, (one week earlier than I in the original proposal) so that I will have an additional week to prepare a final project report for comments before the ultimate due date.

Reviewer Comments and Feedback