WHITE PAPER

Developing the Data Set of Nineteenth-Century Knowledge

Peter Melville Logan, Temple University, Project Director
Jane Greenberg, Drexel University, Co-Project Director

Digital Humanities Advancement Grant
HAA-261228-18
Submitted: 31 May 2021
# Table of Contents

INTRODUCTION ............................................................................................................................... 1

1. ACTIVITIES ...................................................................................................................................... 1

2. ACCOMPLISHMENTS ................................................................................................................... 4

3. AUDIENCES ...................................................................................................................................... 6

4. EVALUATION .................................................................................................................................. 6

5. CONTINUATION OF THE PROJECT ........................................................................................... 7

6. LONG TERM IMPACT .................................................................................................................. 7

7. AWARD PRODUCTS ...................................................................................................................... 7

Appendices ................................................................................................................................................. 11

Appendix A: OCR Output Format ..................................................................................................... 11

Appendix B: Fine-Tuning the OCR Process ....................................................................................... 12

Appendix C: Project Scope ................................................................................................................... 13

Appendix D: Support Challenges ........................................................................................................ 15

Appendix E: Long-s Correction .......................................................................................................... 17

Appendix F: Hyphenation Correction ................................................................................................. 19

Appendix G: Knowledge Project Web Site ......................................................................................... 20

Appendix H: OCR Production ............................................................................................................ 21

Appendix I: OCR Word Accuracy ...................................................................................................... 22

Appendix J: Processing .......................................................................................................................... 23

Appendix K: TEI with Subject Metadata ............................................................................................ 25

Appendix L: Automatic Subject Metadata Workflow ........................................................................ 27

Appendix M: HIVE2 Online Vocabulary Server .............................................................................. 31

Appendix N: HIVE Search Result in 1910 LCSH .............................................................................. 32

Appendix O: Authority File for 1910 LCSH ...................................................................................... 33

Appendix P: Image Sources .................................................................................................................. 34

Appendix Q: Metadata Research Center Web Page on the Knowledge Project ............................... 35

Appendix R: Knowledge Project Public GitHub Repository .............................................................. 36
INTRODUCTION

“Developing the Data Set of Nineteenth-Century Knowledge” proposed to study the system of knowledge in the nineteenth century and to identify how it changed over time. The chief goals were to: 1) create digital text versions of four historical editions of the *Encyclopedia Britannica* and make them suitable for computational analysis; 2) add subject terms, to make the data set easier to analyze; and 3) conduct a preliminary analysis of the data set. Our final objective was to make the full data set freely available to the public.

We are pleased to report that we have substantially completed these goals. We are on track to release the data set to the public in November 2021, following additional cleaning, and we will update the NEH with the results of our preliminary analysis at that time. This first section elaborates on our activities during the grant and explains how some of our original plans changed with experience. Project outcomes are detailed in the second section.

As part of our accomplishments reported here, we would like to acknowledge the contributions to the project that were funded through other sources. Dr. Jane Greenberg’s Metadata Research Center at Drexel University funded two doctoral students to work on problems of automatic subject metadata generation for historical materials. The IMLS also sponsored these students for two summers through the LEADS summer internship program, administered by Drexel. Don Kretz, MBA, from Distributed Proofreaders, volunteered his expertise to solve fundamental problems in data management and quality control. The Digital Scholarship Center in Temple University Libraries (renamed the Loretta C. Duckworth Scholar’s Studio in 2019) provided research assistance and a work facility for the first two years of the project.

1. ACTIVITIES

a. OCR Activity

During the grant term we developed an advanced knowledge of Optical Character Recognition (OCR) techniques to produce accurate text at scale while maintaining a high rate of production. We used the proprietary OCR program ABBYY FineReader (AFR) because open-source alternatives were unable to match its accuracy. It proved to be a highly customizable program which we successfully tailored to our historical documents.

The processing was almost exclusively done by undergraduate students, who we trained extensively in the workflow and the use of AFR. Every step in the process was described in our online manual, so that all operators processed the materials alike.

1. Volunteers with the crowd-sourcing organization Distributed Proofreaders clean most of the content in Project Gutenberg’s online archive. See [https://www.pgdp.net/c/](https://www.pgdp.net/c/).
During the course of the work, we made three substantive changes to the workflow. Each is explained in more detail in the appendices. Briefly, these changes were:

1. **Changing the OCR Output Format:** Switching the OCR output from HTML to DOCX. We discovered a bug in AFR’s routine for generating HTML files. Switching to DOCX improved our results in the output. (See appendix A.)

2. **Fine-Tuning the OCR Process:** Successfully automating the OCR process depended on fine-tuning the program’s recognition routine and limiting use of an automated verification feature to problem pages. (See appendix B.)

3. **Tailoring the Project Scope.** We revised our definition of what page elements we would capture in the initial OCR process. Since the project is designed to reproduce the text for computational analysis, we omitted images from the project scope. Later, we omitted tables and formulas since their data have limited value for text mining.

4. **Switching to better source images.** Late in 2019 the National Library of Scotland (NLS) released a new digital collection of the first eight editions of the *Encyclopedia Britannica*. Their images for the 7th edition proved to be an improvement over our selected image source, so we shifted to the NLS originals early in the OCR effort. Unfortunately, their OCR text was uncorrected and proved disappointing. (See appendix P. for a complete list of image sources.)

**b. Support Challenges**

The value of flexibility and the ability to adapt to changing circumstances were on full display during the award. We confronted three unexpected challenges to the on-going OCR work. These included:

1. The move of the project worksite to a new facility that was still under construction and unequipped to support our student workers.
2. The closure of the new worksite due to the Pandemic four months later.
3. Acquiring equipment and creating a new workflow for at-home work.

For more detailed reporting on these changes and impacts, see appendix D. Overall, while the disruptions were unwelcome surprises, our team met the challenges and continued the work successfully. Additional NEH support significantly eased the burden posed by these events.

**c. Processing Activities**

Once the OCR text of a page is created, that output needs to be processed before they become part of the project data set. Our automated steps include:

1. Transforming the OCR output for each page to valid TEI-XML (an archival format pioneered by the Text Encoding Initiative in eXtensible Markup Language).
2. Isolating each entry in the data set and converting them to individual files.

---

2 The project scope is online at [https://tu-plogan.github.io/source/c_editorial_standards.html](https://tu-plogan.github.io/source/c_editorial_standards.html).
3. Cleaning the text to remove OCR noise, correct common OCR errors, identify missing or misnamed entries, correct hyphenation, and modernize the “long-s” in the 3rd edition.

4. Creating subject metadata for each entry (discussed below).

These processing activities conclude with the creation of the “Master” files—valid TEI files with subject metadata for the entry. The Master files are used to generate usable data in any desired format: plain text (for data analysis) and HTML (for web publication).

During the grant period we also developed the Python scripts that automated the processing and modified the XSLT (eXtensible Stylesheet Language Transformation) files needed to transform OCR output to TEI, so that the final encoding met our standards. We improved the initial clean up routine with a collection of regular expressions (Regex) that could correct common encoding errors. We also created a new Python script to detect OCR operator errors in formatting footnotes.

Two particularly innovative processes addressed the “long-s” problem in the 3rd edition and correcting hyphenations. Both were created by Don Kretz and are further detailed in appendices E and F.

d. Automatic Metadata Generation

Our goal of generating subject terms for each entry became one of the most productive research foci of the project. We worked closely with the Metadata Research Center at Drexel University throughout the course of the grant period refining our approach. The Center maintains an automated vocabulary server, HIVE (Helping Interdisciplinary Vocabulary Engineering), a linked data, automatic indexing application that dynamically integrates controlled vocabularies (like the Library of Congress Subject Headings) and expedites semantic tagging of digital resources.3 We developed a dedicated workflow that could run our entries through HIVE and write the subject term metadata into the TEI header for entry file. (For a sample TEI header with subject metadata, see appendix K.) The workflow and the research steps we took to develop it are summarized in the “Accomplishments” section of this report.

Subject term indexing is key to our analysis of knowledge change over time. In March 2021, we successfully generated subject metadata for a pilot set equivalent to 20% of the data set. The pilot test results will support our next evaluation, leveraging subject metadata to segment the data set into broad knowledge domains that can be used in a comparative study of change over time. A supplemental report on our analytical efforts will be submitted to the NEH late in 2022.

e. XML Database

In addition to the original work plan, we have set up an XML database to store the complete data set. Unlike a traditional relational database, which structures data as tables, XML databases are centered on collections of documents rather than tables. Instead of SQL (Structured Query Language), it uses a query language designed for XML documents, called “XQuery.” This makes it much easier to check the

3 https://cci.drexel.edu/mrc/research/hive/
data for consistency and we plan to rely on database queries for some analytical procedures. We use eXist-db, an open-source XML database that has additional affordances for querying and publishing TEI documents.4

f. Website
We established a public website with information about the Knowledge Project early in the grant period.5 (See appendix G.) We acknowledge the NEH on the cover page with a link to the award. The site includes introductory information on the importance of the project, its scope, and its personnel. It explains our workflow, like the techniques used to generate index terms. A reference section explains our editorial standards, provides a TEI style manual, and acknowledges image sources. All material is indexed for ease of use. The site includes 117 discrete pages. The Metadata Research Center also maintains a webpage on the project, with publication information.6 (See appendix Q.)

2. ACCOMPLISHMENTS
Thanks to the award, we were able to accomplish the objectives set out in our original proposal.

- We completed 100% of the OCR for the 7th, 9th, and 11th editions of the *Encyclopedia Britannica*, and 65% of the 3rd edition.
- We integrated material from Project Gutenberg and converted everything to TEI, as originally proposed for the 11th edition.
- We successfully processed the OCR output to create a standards-compliant data set of individual entries from all four editions.
- We attained exceptionally high accuracy rates for all four editions, making this the first data set of historical *Encyclopedia Britannica* text suitable for computational analysis.
- We developed an original process to automatically add subject metadata to historical humanities material. The metadata includes linked open data, to improve accessibility. We successfully demonstrated the process by indexing a sample of 20% of the data set.

a. OCR
Our goal was twofold:

1. *Complete the OCR work for all four editions.* We reached a 94% completion for the data set by completing three of four editions and 2/3 of the fourth. Given the unpredictable disruptions of the worksite change and the pandemic, we believe this is an exceptional outcome. (Quantitative details are given in appendix H.)

---

5 [https://tu-plogan.github.io/](https://tu-plogan.github.io/).
6 [https://cci.drexel.edu/mrc/research/nineteenth-century-knowledge](https://cci.drexel.edu/mrc/research/nineteenth-century-knowledge).
2. **Achieve high accuracy rates.** As promised, we successfully achieved a word accuracy that is equal to or greater than (≥) 99.5% for three editions, and ≥ 98.5% for the 3rd edition, which includes the “long-s.” (See appendix I.) This accomplishment is particularly satisfying, because it solves the problem that motivated us in the first place. We originally proposed this project to the NEH because the poor quality of existing digital text for these editions made them unsuitable for computational analysis. The new digital text will allow us and other research teams to move ahead with their own research.

**b. Processing**

“Processing” refers to the steps needed to clean, encode, and validate the OCR output. The OCR output from scans of about 74,000 print pages has been successfully converted to TEI. This includes transforming all of the Project Gutenberg material from its original HTML format to TEI and integrating it into the data set. In order to optimize the material for users, we create individual files for each encyclopedia entry. We have individual files for 63,000 entries, or over 90% of the total. (See appendix J for further details.)

**c. Automatic Metadata Generation**

The key to any workflow for creating subject metadata is the quality of the results. Much of our success depended on identifying appropriate controlled vocabularies for use with our historical text. A 21st-century vocabulary can index older materials, but if it lacks the antiquated terms of the text, specific topics will be missed. And if it conflates modern terms with older ones, it runs the danger of adding an anachronistic subject term to an entry.

One of the most significant finding of our research is that using an older controlled vocabulary, one closer in time to the source material, can actually produce better results than current vocabularies will. (See appendix L.) This has not been documented previously, and doctoral student Sam Grabus’s recent publications and conference talks (supported by this award) are the first to make the case for this method.

In order to be added to HIVE, older vocabularies need to be formatted in the language used by automated vocabulary servers, SKOS (Simple Knowledge Organization System). To test their efficacy, we created two new SKOS vocabularies by using OCR to generate text from page images and converting the text to a structured SKOS format. These vocabularies are now online and freely available to the public through the HIVE2 online server. (See appendix M.)

---

8 SKOS (Simple Knowledge Organization System) was developed by the W3C for knowledge organization systems. See https://www.w3.org/2004/02/skos/.
9 https://hive2.cci.drexel.edu/.
Other findings concern the use of blending the Natural Language Processing method of Named Entity Recognition (NER) with subject term metadata. By processing NER results through HIVE, we were able to create subject metadata derived from NER that uses a controlled vocabulary and includes linked open data, connecting the subject terms to online authority files. This connects the subject metadata to the semantic web and increases the accessibility of the data to users doing internet searches.

Appendix L contains more detail on the process used to identify the optimal vocabularies blended with NER results to generate subject metadata for the entries with an unusually high level or relevance.

d. Analysis

We are planning a preliminary analysis of the data set in late 2021 to establish findings on the changes to the structure of knowledge in the nineteenth century. The results of the analysis will be appended to this Final Report at that time.

3. AUDIENCES

Prior to the COVID-19 Pandemic, the research team gave presentations on the project at American academic conferences in Washington D.C.; Philadelphia, PA; and Charlottesville, VA. We also presented at international conferences on three continents: Mexico City, Mexico; Tokyo, Japan; and Utrecht, The Netherlands. The audiences for these conferences were faculty members and graduate students in university humanities, computer science and information science departments, and members of the library, archive, and museum communities.

To date, published research based on the project has been limited to professional journals in Information Science. Future publications are planned for journals in History and Digital Humanities.

Once our data set is made publicly available, the audience of users will become more diverse. This is particularly true for the Oxford Text Archive (OTA), which remains committed to integrating the data set as individual entries linked to other resources in the Archive. As one of the largest repositories of free digital text editions in the English-speaking world, OTA has a broad user base, and we are delighted that they will make these materials so widely available.

We had planned a public symposium to introduce digital methods and historical materials to college and university students in the Philadelphia area. Unfortunately, the scheduled event had to be cancelled due to the Pandemic.

4. EVALUATION

A peer review of the project is currently in process by NINES: Nineteenth-Century Scholarship Online (https://nines.org/about/scholarship/peer-review/). When complete, we will add their evaluation to the Final Report.
5. CONTINUATION OF THE PROJECT
Dr. Logan and Mr. Kretz have agreed to continue work on the project. Dr. Greenberg and the Metadata Research Center group at Drexel University also wish to continue working with our team and with Temple University. We will pursue additional grant funding, as appropriate.

6. LONG TERM IMPACT
Some spin-off projects have already begun, as spelled out below, but the broadest impact will only be evident once the project is known and freely available to teachers for classroom use in secondary and undergraduate courses, students for class assignments, and researchers for further exploration. Existing work is as follows:

- Publication of the workflow needed to automatically generate subject metadata for humanistic archival materials has broad implications for accessibility practices in the library, archive, and museum communities.
- Creation of the SKOS version of the 1910 Library of Congress Subject Headings is the first historical controlled vocabulary ever made available for automated indexing. Its incorporation into the online HIVE2 site (https://hive2.cci.drexel.edu/) makes it freely accessible for use by other researchers interested in using historical ontologies to generate subject metadata for historical material.
- Researchers in artificial intelligence at the Alan Turing Institute, the British Library, have requested and been given early access to the data set, which they want to mine for the “Living with Machines” project.¹⁰
- The data set has already had significant interest among Information Science researchers and led to a new look at the value of historical controlled vocabularies for generating subject terms for archival materials.¹¹ One Drexel doctoral student, Sam Grabus, is writing a dissertation on the topic, using the Knowledge Project’s data set as the main corpus.¹² A second Drexel doctoral student is working on the ontology of the Ephraim Chambers *Cyclopaedia*, an interest that grew out of work for the Knowledge Project.

7. AWARD PRODUCTS

Data Sets (late 2021)
Note: the data set is released with a Creative Commons Attribution-ShareAlike 4.0 International, which lets others remix, adapt, and build upon our work even for commercial purposes, as long as they credit the Knowledge Project and license their new creations under the identical terms.

---

¹⁰ https://www.turing.ac.uk/.
¹² Dr Greenberg chairs the dissertation committee; Dr. Logan is a committee member.
1. **Oxford Text Archive** ([https://ota.bodleian.ox.ac.uk/repository/xmlui/](https://ota.bodleian.ox.ac.uk/repository/xmlui/)). A major archive of digitized texts freely available to the public. The text from the data set will be incorporated as individual entries.

2. **Humanities Commons CORE Repository** ([https://hcommons.org/core/](https://hcommons.org/core/)). The leading repository in the humanities for digital data sets. The text will be incorporated as batch (ZIP) files.

3. **TU ScholarShare** ([https://scholarshare.temple.edu/](https://scholarshare.temple.edu/)). The public-facing repository for Temple University researchers to make their data set publicly available. The text will be incorporated as batch (ZIP) files.

4. **GitHub** ([https://github.com/TU-plogan/KnowledgeProject](https://github.com/TU-plogan/KnowledgeProject)). The leading source for repositories in computer science and across the digital humanities. Users can download the entire data set, individual editions, or individual entry files. We will also post all Python and PHP scripts used in the project. (See app. R.)

### Databases

Both of the databases below require publishing named authority files for every term in the controlled vocabulary. The Metadata Research Center has established a webspace to host the authority files, which will use Archival Resource Keys (ARKs) as persistent identifiers. A temporary version of the authority files is online for our use. (See appendix O.) All authority files reference the source image files at Hathi Trust. (See appendix N for sample search result with the 1910 Library of Congress Subject Headings.)

1. 1910 Library of Congress Subject Headings, HIVE2 ([https://hive2.cci.drexel.edu/](https://hive2.cci.drexel.edu/)).

2. Ephraim Chambers Cyclopaedia, HIVE2 ([https://hive2.cci.drexel.edu/](https://hive2.cci.drexel.edu/)).

### Articles


3. Dr. Logan and Dr. Greenberg plan a journal article on the relationship between findings from the data set and its use of subject metadata. Doctoral students will participate in these and other outputs.

---

13 [https://arks.org/](https://arks.org/)
Conference talks


Posters


Future plans

1. Short-term goals (completion by 30 Nov. 2021)
   - Run OCR on the remaining portion of the 3rd edition (5,000 pages). This work will led by the Dr. Logan through an internal grant from Temple University, or by crowd-sourcing the transcription at Zooniverse.14
   - Additional automated cleaning of the data set. This work is co-led by Dr. Logan and Mr. Kretz and is already in process, as we continue to refine our cleaning scripts.
   - Add subject metadata to the remaining data set. Dr. Logan and Dr. Greenberg are evaluating the results of the pilot. Sample relevancy testing is promising, with results of >85% (the norm for automated indexing is about 60%). Once we have verified these findings, we will generate subject metadata for the remaining 80%.

14 Dr. Logan is participating in the NEH Institute, “Building Capable Communities for Crowdsourced Transcription” (HT-272556-20), based at the University of Minnesota, which teaches project leaders the theory of crowd-sourcing and the practice of using Zooniverse. The Institute runs from 2020 to 2023. https://www.zooniverse.org/.
• **Release v1.0 of the data set.** Dr. Logan and Mr. Kretz will do a formal release of the first version of the data set by publishing it on the project’s public GitHub site in TEI and TXT formats.¹⁵ (See appendix R.) We will also add it to the repositories listed above.

2. **Long-term goals**

The team wants to continue refining the data set and to release a second version (v.2.0) with the following changes:

- Dr. Logan and Mr. Kretz plan to create a crowd-sourced proofreading operation to improve text accuracy. Mr. Kretz’s experience with Distributed Proofreaders makes this a realistic option.

- Dr. Logan and Mr. Kretz plan to continue to refine the TEI encoding as follows:
  - add sections to entries with multiple parts.
  - add author names to entries signed with initials and incorporate linked open data for authors in the metadata for each entry.
  - seek additional funding to support the encoding of formulas and to incorporate original illustrations.
  - Set up a dedicated crowd-sourcing project at Zooniverse to involve volunteers interested in transcribing the many tables in the four editions.

- Dr. Logan will continue working with the data set to publish findings about how knowledge changed in the nineteenth century.

- Dr. Logan and Dr. Greenberg envision multiple articles based on the problems the data set poses for information science professionals.

- Dr. Greenberg’s Metadata Research Center is committed to making historical vocabularies accessible for automatic metadata generation and scholarly research. Although an NEH proposal submitted in spring 2020 was not funded, Dr. Greenberg will be seeking additional funding to continue this effort in collaboration with Dr. Logan and other researchers affiliated with the MRC.

¹⁵ [https://github.com/TU-plogan/KnowledgeProject](https://github.com/TU-plogan/KnowledgeProject)
Appendices

Appendix A: OCR Output Format

OCR can generate text in a variety of document formats. Some projects, such as Project Gutenberg, output everything as plain text and later add markup for text features like italics, bold, and small caps. Our smaller project needed to capture these features automatically. They have semantic significance as markers for entry subheadings, citations, and other structural components that we wanted to preserve.

1. **Bold**

   Early in the award period, we discovered that capturing bold in historical documents is unadvisable, because as paper ages it darkens, leading the OCR engine to add bold formatting to large blocks of normal text. In the end, we removed all bold formatting in a post-OCR cleanup process.

2. **Italics**

   While AFR generally captures italics well, OCR accuracy declined, so more post-OCR cleanup is needed.

3. **Small Caps**

   We discovered a bug in AFR’s routine for saving its results as Hyper Text Markup Language (HTML, the format used for web pages). While the program correctly identified small caps on the page, the HTML output routine was inconsistent in including it.\(^{16}\) No date for a fix was available from ABBYY. Small caps particularly mattered to our output because the titles of encyclopedia entries included them, and we needed the formatting to be able to automate the identification of the full title. After testing, we discovered that the routine for saving output as Microsoft Word documents (DOCX) worked properly, so we switched output formats from HTML to DOCX.

   DOCX proved to be the better format for our needs in other ways. Both HTML and DOCX were transitional formats for us, since the data set is in TEI. We initially chose HTML thinking it would be the easiest file format to convert to TEI. We also took advantage of the conversion program provided by the TEI Consortium to transform HTML to TEI. However, DOCX is itself a native XML format (unlike either its predecessor, DOC, or HTML). The TEI Consortium provides an excellent routine to convert DOCX to TEI, and it preserves every feature that AFR could capture, as far as we could discover. The TEI conversion routines are written as eXtensible Stylesheet Language Transformations (XSLT), one of the standard elements of XML technology. We spent several weeks customizing the XSLT routine for converting DOCX to match our specific needs, and it performed flawlessly throughout the project.

---

\(^{16}\) ABBY Customer Support was able to reproduce the problem and confirmed the bug.
Appendix B: Fine-Tuning the OCR Process

Initially we used a feature of AFR that identifies OCR characters it is unsure of and allows the operator to keep or change them. This verification routine allowed us to increase OCR accuracy to >99.5% without proofreading. However, in tests we learned that it tripled the time it took to OCR operators to process our pages. We used this for the first year of the grant before deciding that we would never complete the OCR work if we continued. Instead, we adopted a hybrid approach that involved several changes.

1. **Fine-tuning of the OCR process.** AFR allows users to customize its recognition algorithm by adding images of letters to its “User Pattern.” The information they provide on the process, however, is inadequate for projects with historical materials. After trial-and-error efforts, we discovered that increased recognition accuracy depended on adding the **fewest** letters possible to the user pattern. Previously we had added multiple examples of letters. AFR is excellent at recognizing most letters, and there is no need to add samples of letters that it gets right by default. Similarly, we learned to never add a letter that is deformed or broken. Instead, only add images that represent the ideal shape of letters that AFR has trouble recognizing, such as “I,” “i,” and “l.” These can be easily confused in historical documents where the ink can bleed over time and letter shapes lose their crisp edges. Finally, we discovered that three should be the maximum number of examples for any given letter.

2. **Limit Verification.** The increased accuracy from the improved user patterns made it possible to bypass verification for most pages, knowing that OCR errors were at a minimum. AFR shows the percentage of “low confidence” characters on each page, and this serves as a reliable guide to the relative accuracy of the OCR. After refining the user patterns, most of our pages had low-confidence ratings of <1%. We implemented a new procedure to only verify pages that were >2%. This meant verifying about one in every thirty pages (3%), and our production rate increased significantly as a result. In addition, we added automated cleaning routines to the post-OCR workflow to correct common OCR errors. This obviated the earlier trade-off between speed and accuracy, and I recommend every new project plan to include such a routine in their workflow.

---

17 An OCR program cannot know what it does not know, of course, so the low-confidence rating is not to be confused with actual OCR accuracy. But it can tell you which pages are more difficult to recognize than others, and that is what we mean by “relative accuracy.”
Appendix C: Project Scope

The following editorial principles are employed in creating this digital edition.

The sources texts we edit contain many different kinds of material, including multiple languages, scientific formulas, illustrations, tables, footnotes, marginal subheadings, and so forth. These standards evolved with the goal in mind of creating as accurate a reproduction of the encyclopedia text as possible and allowing for the incorporation of further details at a later date, if desired. This list describes the essential principles adhered to in creating version 1 of this edition, which focuses on the accurate reproduction of the text.

1. Captions
   - Table captions are included.
   - Image captions are not included.
   - An exception to this rule is the material in eb11 reproduced from Project Gutenberg, which includes image captions.

2. Diacritics and Alphabets
   - All original language characters with their diacritical markings in the source text are reproduced.
   - Some languages, such as Hebrew and Sanskrit, require special treatment and are noted for further review during a cleaning operation.

3. Notes
   - All note text is preserved and moved in line with the body text, at the site of the original note anchor.
   - Unanchored note text is moved in line with the body text, at the inferred point of reference.
   - Source note numbers are replaced with a continuous sequence of note numbers restarting with 1 for each entry.
   - The placement of footnotes, marginal and table notes are indicated in <note> encoding, as a value of @place.

4. Formulas
   - Simple one-line mathematical and chemical formulas are reproduced.
   - Multi-line formulas are not included in version 1 of this project.
   - An exception to this rule is the material in eb11 reproduced from Project Gutenberg, which includes multi-line formulas.

5. Images
   - Images are outside the scope of version 1.
   - An exception to this rule is the material in eb11 reproduced from Project Gutenberg, which includes images.

6. Headings
   - The text of headings and subheadings within the normal flow of the text are included.
7. Ligatures
   - Ligatures are reproduced as multiple letters, not a unique symbol.

8. Page Layout
   - The two-column format of individual pages is not preserved. All type is presented as a single column capturing the flow of the semantic content.
   - Page breaks are indicated in place in the metadata.
   - Column breaks are not indicated.

9. Tables
   - Tables are outside the scope of version 1.
   - Table data is preserved in the OCR process for inclusion in version 2.
   - Lines and graphical attributes of the source table are not reproduced.

10. Typographical errors
    - Obvious misspellings in the source text are noted in the TEI as a choice between the misspelling and the correct spelling.
    - Illegible text is rare. When it occurs, it is indicated with bracketed ellipses: [...].

11. Spelling
    - Original spelling is preserved throughout.

12. Type style
    - The following font formatting is preserved:
      - Italic
      - Small Cap
      - Strike through
      - Subscript
      - Superscript
      - Underline
    - The following formatting is ignored:
      - Bold
      - Font size
      - Typeface

URL: https://tu-plogan.github.io/source/c_editorial_standards.html
Appendix D: Support Challenges

1. Library relocation

Our worksite in the Library was critically impacted by the move to a new building in summer 2019, leading to a need to redesign our data storage system and to more than a month-long loss of our worksite.

Loss of server. While they prepared for the move, the Digital Scholarship Center shut down their internal server, where our data was safely secured. This necessitated a major change in our data storage policy as everything had to move to a cloud-based system that student workers could clone on the local machine. Because the project has a high volume of both image and data files, the solution involved two separate online services.

- Google Shared Drive: Temple University makes shared drives freely available to the University community. Google Drive boasts "unlimited storage," making it ideal for the large image files used in OCR work. In practice, repositories have storage limits, but researchers can have any number of repositories. We created six private repositories for the project. Google offers a desktop application that integrates the files into the computer file system. However, working directly on files stored in the cloud is too slow for OCR production, so we adopted the practice of cloning the OCR files to the local hard drive and running a daily backup to the online site.

- GitHub: Temple University makes private GitHub accounts available to faculty members. Limits on file size and storage capacity make this option unsuitable for image files, but its versioning capabilities makes it perfect for high-value text files. We moved our post-OCR material into a single repository, including OCR output files, processed entry files, metadata, and other project material. The presence on GitHub has also facilitated the sharing of preliminary data with researchers at other institutions. But it also meant training student workers in the GitHub workflow, because it is not intuitive.

Loss of worksite: We expected a one week period to transition to the new facility in August 2019. Instead, we lost over month and were never able to establish a satisfactory worksite in the new Loretta C. Duckworth Scholars Studio. Some of the reasons were out of the staff’s control. Due to an ordering mistake, new computers for the Studio did not arrive until November 2019. In the interim, we setup our team on four Studio laptops. Unfortunately, the Studio never provided us with a stable work location. The building was still under construction for all of Fall 2019, and the team moved from one room to another. Sometimes, our dedicated laptops were borrowed for other purposes and had to be hunted down. By November, the PI was able to negotiate an agreement with the Studio to reserve two of the 25 new desktop computers for our RAs. Setting up the machines with Google Drive, GitHub, AFR, and related software took time because of security settings controlled by a single person in the Studio, who was not always available. By the end of term, the desktops were ready but the students were...
on break. Student workers finally switched to them successfully in January 2020, after a challenging Fall 2019 semester.

Other problems caused by the move were wholly in control of the Library and the Studio staff. The new facility was never designed to support grant-funded research projects by faculty members. After the PI stepped down as Academic Director of the Digital Scholarship Center in summer 2019 (due to retirement), the Studio shifted its focus to serving undergraduates, with research support limited to serving the graduate students and faculty members with annual awards from the Studio. Support for the Knowledge Project from the Studio ceased to live up to its earlier promises, so that there was an end to all graduate research support and minimal provision for worksite support.

2. Pandemic Problems

One month after starting the team on the new desktop arrangement in the Studio, the University and the Library closed completely due to the Pandemic. While Dr. Logan requested the Studio to loan us the same laptops used two months earlier, the Studio could not accommodate us. To keep the students employed, Dr. Logan reassigned the RAs to non-OCR tasks, such as data-cleaning, that they could do from home on their own computers. This was useful, but it cost us several more months of planned OCR work.

3. Enabling at-home work

Thanks to an NEH CARES Supplement awarded at the end of May 2020, we were able to lease three laptops and purchase short-term licenses for AFR. After recreating the Google Drive and GitHub programs, students resumed OCR work in June 2020 in their homes. Another RA worked for the Fall 2020 semester, when the University was slated to reopen. The Studio could not provide a safe workspace for the RA, so Dr Logan set up a dedicated space in his campus office with his own laptop. When classes moved off campus again two weeks later, the RA continued to work safely and productively in the office for the semester.
### Appendix E: Long-s Correction

Books printed before the 1820s typically used the long-s character, “ſ”, which we change to the short-s for modern readers. There is no semantic significance to the “ſ”, and updating it makes the material more accessible. AFR struggled with recognizing the character in the originals of the 3rd edition; even though we added it to the alphabet and trained it carefully, the similarity between “ſ” and “f” caused frequent misrecognitions.

Corrections must account for the possibility of misrecognition and decide whether “ſ” should actually be “s” or “f.” We first created a series of regexes to correct “ſ” in unambiguous situations, like letter combinations where it can only be one or the other. For remaining instances, we created a PHP script that isolated all words with “ſ” in the 3rd ed. and compared them to word frequencies in the 7th ed. for the same word with both an “s” and with an “f”.

**Comparison of long-s words in 3rd edition with probable matches in 7th edition.**

<table>
<thead>
<tr>
<th>Long-s Word</th>
<th>Matches in 3rd Edition</th>
<th>Matches in 7th Edition</th>
</tr>
</thead>
<tbody>
<tr>
<td>cf</td>
<td>28865</td>
<td>98073,”os:1821”</td>
</tr>
<tr>
<td>for</td>
<td>35291</td>
<td>“for:98266,”sor:7”</td>
</tr>
<tr>
<td>from</td>
<td>34166</td>
<td>“From:107973”</td>
</tr>
<tr>
<td>fo</td>
<td>12420</td>
<td>“So:35043,”fo:1091”</td>
</tr>
<tr>
<td>theſe</td>
<td>11469</td>
<td>“These:38961”</td>
</tr>
<tr>
<td>if</td>
<td>11203</td>
<td>“If:23244,”“If:24961”</td>
</tr>
<tr>
<td>frome</td>
<td>9879</td>
<td>“Some:32745”</td>
</tr>
<tr>
<td>after</td>
<td>8735</td>
<td>“after:19302,”aster:9”</td>
</tr>
<tr>
<td>thofe</td>
<td>7222</td>
<td>“Those:24420”</td>
</tr>
<tr>
<td>should</td>
<td>7046</td>
<td>“Should:13896”</td>
</tr>
<tr>
<td>fuch</td>
<td>7035</td>
<td>“Such:21034”</td>
</tr>
<tr>
<td>fame</td>
<td>6524</td>
<td>“Same:30105,”fame:642”</td>
</tr>
<tr>
<td>most</td>
<td>6751</td>
<td>“Most:26238”</td>
</tr>
<tr>
<td>else</td>
<td>6263</td>
<td>“Also:21593”</td>
</tr>
<tr>
<td>before</td>
<td>5842</td>
<td>“Before:12776”</td>
</tr>
<tr>
<td>firſt</td>
<td>5624</td>
<td>“First:24927”</td>
</tr>
<tr>
<td>found</td>
<td>5609</td>
<td>“Sound:1875,”found:13555</td>
</tr>
<tr>
<td>must</td>
<td>5596</td>
<td>“Must:14971”</td>
</tr>
<tr>
<td>thercfore</td>
<td>5364</td>
<td>“Therefore:11823”</td>
</tr>
<tr>
<td>different</td>
<td>4100</td>
<td>“Different:12875”</td>
</tr>
<tr>
<td>smail</td>
<td>4043</td>
<td>“Small:13439”</td>
</tr>
<tr>
<td>himſelf</td>
<td>3789</td>
<td>“Himself:10927”</td>
</tr>
<tr>
<td>severſal</td>
<td>3670</td>
<td>“Several:12787”</td>
</tr>
<tr>
<td>of</td>
<td>3573</td>
<td>“Of:350873,”os:152”</td>
</tr>
<tr>
<td>againſt</td>
<td>3522</td>
<td>“Against:9319”</td>
</tr>
<tr>
<td>four</td>
<td>3273</td>
<td>“Four:11195,”sour:107”</td>
</tr>
<tr>
<td>form</td>
<td>3018</td>
<td>“Form:12476”</td>
</tr>
<tr>
<td>feef</td>
<td>3068</td>
<td>“Feet:10658,”seat:17”</td>
</tr>
<tr>
<td>ufed</td>
<td>2991</td>
<td>“Used:80021”</td>
</tr>
<tr>
<td>off</td>
<td>2860</td>
<td>“Off:5059,”oss:12”</td>
</tr>
<tr>
<td>faied</td>
<td>2830</td>
<td>“Faid:7380,”fai:1”</td>
</tr>
<tr>
<td>left</td>
<td>2545</td>
<td>“Last:8580”</td>
</tr>
<tr>
<td>fame</td>
<td>2790</td>
<td>“Same:30105,”fame:642”</td>
</tr>
<tr>
<td>themſelvſe</td>
<td>2685</td>
<td>“Themſelves:6942”</td>
</tr>
<tr>
<td>leſs</td>
<td>2599</td>
<td>“Less:9810”</td>
</tr>
<tr>
<td>half</td>
<td>2535</td>
<td>“Half:6658,”hals:3”</td>
</tr>
<tr>
<td>foon</td>
<td>2567</td>
<td>“Scon:6517”</td>
</tr>
<tr>
<td>fhall</td>
<td>2550</td>
<td>“Shall:6852”</td>
</tr>
<tr>
<td>she</td>
<td>2547</td>
<td>“She:4627,”she:14”</td>
</tr>
<tr>
<td>ufe</td>
<td>2469</td>
<td>“Use:66071”</td>
</tr>
<tr>
<td>ftațe</td>
<td>2411</td>
<td>“State:15880”</td>
</tr>
<tr>
<td>formerly</td>
<td>2397</td>
<td>“Sometimes:7247”</td>
</tr>
<tr>
<td>becauſe</td>
<td>2360</td>
<td>“Because:72101”</td>
</tr>
</tbody>
</table>
For example, the word “fame” (line 25) occurs 6,825 times in the 3rd edition. The words “same” occurs 30,105 times in the 7th edition, while “fame” occur only 642 times. The script uses these frequencies to provide an educated guess as to which substitution is most likely. In this case, it corrects all instance of “fame” to “same.” While this approach will create a small number of errors, overall it produces text with a higher rate of accuracy than before.
Appendix F: Hyphenation Correction

While designed to correct them, OCR often leaves in place hyphens for words broken at the end of a line; cleaning up these “soft” hyphens is a standard piece of post-OCR correction. The difficulty is differentiating soft hyphens from “hard” hyphens used in words that are always hyphenated. In historical documents, we can see in action the linguistic transition of words from two separate words (“house work”) to one hyphenated word (“house-work”) to a single compound word (“housework”), as usage increased. Hard hyphenations thus have semantic significance and need to be preserved. Soft hyphens are artifacts of the print format and should be removed.

When AFR fails to correct a word break at the end of a line, it leaves a telltale space after the hyphen, but it does not tell us whether the word is normally hyphenated or not. We used PHP to identify every end-of-line hyphenation and then search the rest of the same edition to identify the frequency of the word with a hyphen and without. For example, we had 14 instances of “main- tained” in the data set. The program gave us the following results:

<table>
<thead>
<tr>
<th>index</th>
<th>from text</th>
<th>occurs</th>
<th>hyphenated</th>
<th>occurs</th>
<th>joined</th>
<th>occurs</th>
<th>result</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>main- tained</td>
<td>14</td>
<td>main-tained</td>
<td>2</td>
<td>maintained</td>
<td>2660</td>
<td>maintained</td>
</tr>
</tbody>
</table>

This tells us that there are 14 uncorrected word breaks at the end of a line for the word; the word occurs with a deliberate hyphen twice, and without one 2,660 times. Given the difference, the program treats it as a soft hyphen and corrects it to the joined word “maintained.” Where there are similar numbers of hyphenated and joined matches—presumably, different authors preferred one option or the other—we take a conservative approach; we leave the hyphen and remove the extra space.
Appendix G: Knowledge Project Web Site

Home page URL: https://tu-plogan.github.io/
**Appendix H: OCR Production**

During the grant period, we generated text for 52,716 pages of the *Encyclopedia Britannica*. The following table shows numbers for each edition and the portion of the 11th edition integrated from Project Gutenberg (with permission). For each edition, we list the number of pages scanned with OCR, the percentage of each edition completed, and the number of remaining pages (if any).

<table>
<thead>
<tr>
<th>edition</th>
<th>pages scanned</th>
<th>% completed</th>
<th>remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rd</td>
<td>9,525</td>
<td>65%</td>
<td>5,056</td>
</tr>
<tr>
<td>7th</td>
<td>16,182</td>
<td>100%</td>
<td>0</td>
</tr>
<tr>
<td>9th</td>
<td>20,495</td>
<td>100%</td>
<td>0</td>
</tr>
<tr>
<td>11th</td>
<td>6,516</td>
<td>100%</td>
<td>0</td>
</tr>
<tr>
<td><strong>Subtotal</strong></td>
<td><strong>52,718</strong></td>
<td><strong>91%</strong></td>
<td><strong>5,056</strong></td>
</tr>
<tr>
<td>11th (Project Gutenberg)</td>
<td>21,238</td>
<td>100%</td>
<td>0</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>73,956</strong></td>
<td><strong>94%</strong></td>
<td><strong>5,056</strong></td>
</tr>
</tbody>
</table>
Appendix I: OCR Word Accuracy

Accuracy is determined by comparing the original to the processed output for four random pages from each edition.\(^{18}\) We divide the number of correct words by the total number of words to arrive at an accuracy rate for each page. This rate is then averaged as our measure for the edition. The following table shows the averages in the last column, on the right. Pages are indicated with volume and page numbers separated by a colon.

<table>
<thead>
<tr>
<th>edition</th>
<th>p1</th>
<th>rate</th>
<th>p2</th>
<th>rate</th>
<th>p3</th>
<th>rate</th>
<th>p4</th>
<th>rate</th>
<th>avg. rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rd</td>
<td>1:765</td>
<td>0.97909</td>
<td>3:590</td>
<td>0.97642</td>
<td>16:797</td>
<td>0.99742</td>
<td>18:646</td>
<td>0.99557</td>
<td>0.98712</td>
</tr>
<tr>
<td>7th</td>
<td>2:009</td>
<td>0.99920</td>
<td>9:541</td>
<td>0.99480</td>
<td>13:284</td>
<td>0.99632</td>
<td>20:133</td>
<td>0.99773</td>
<td>0.99701</td>
</tr>
<tr>
<td>9th</td>
<td>2:826</td>
<td>0.99772</td>
<td>5:436</td>
<td>0.98634</td>
<td>15:051</td>
<td>0.99710</td>
<td>24:324</td>
<td>1.00000</td>
<td>0.99529</td>
</tr>
<tr>
<td>11th</td>
<td>3:337</td>
<td>1.00000</td>
<td>10:108</td>
<td>1.00000</td>
<td>23:991</td>
<td>0.99544</td>
<td>28:623</td>
<td>0.98885</td>
<td>0.99607</td>
</tr>
</tbody>
</table>

The chart shows that the word accuracy rate for the 7th, 9th, and 11th editions met our goal of $\geq 99.5\%$. The third edition contains the long-s; as predicted, its rate is less than the other editions, but at 98.7% it meets our target rate of $\geq 98.5\%$.

---

\(^{18}\) Page selections are random, but spread out across different volumes and excluding pages with out-of-scope elements.
### Appendix J: Processing

*TEI Pages:* This chart shows the number of OCR pages that have been transformed to valid TEI.

<table>
<thead>
<tr>
<th>edition</th>
<th>pages scanned</th>
<th>converted to TEI</th>
<th>complete %</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rd</td>
<td>9,525</td>
<td>9,525</td>
<td>100</td>
</tr>
<tr>
<td>7th</td>
<td>16,182</td>
<td>16,182</td>
<td>100</td>
</tr>
<tr>
<td>9th</td>
<td>20,495</td>
<td>20,495</td>
<td>100</td>
</tr>
<tr>
<td>11th</td>
<td>6,514</td>
<td>6,514</td>
<td>100</td>
</tr>
<tr>
<td>Project Gutenberg 11th</td>
<td>n/a</td>
<td>21,238</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>52,716</td>
<td>73,954</td>
<td>100</td>
</tr>
</tbody>
</table>

We have generated 73,954 TEI files of individual encyclopedia pages. Project Gutenberg files were already scanned, so they are not included in column 2. They were converted to TEI, and so they are included in column 3. As mentioned earlier, we have about 5,000 pages of the 3rd edition remaining. Thus the total number of TEI pages, when complete, will be about 79,000.

*TEI entries:* Entries (or encyclopedia articles) are different from pages. An entry may run from a few lines to hundreds of pages. Extracting the text for the entries from the data set and saving it to separate files is a critical step in the processing workflow. The following chart summarizes the number of entries successfully extracted for the data set. The second column gives the number of pages processed, while the third column shows the number of entries generated from those pages. The last column shows the total percentage of OCR pages converted to entry files. Project Gutenberg files are not included in this first calculation because they were already formatted as individual entry files and only had to be converted to TEI.

<table>
<thead>
<tr>
<th>edition</th>
<th>pages</th>
<th>entries generated</th>
<th>complete %</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rd</td>
<td>5,012</td>
<td>9,262</td>
<td>53</td>
</tr>
<tr>
<td>7th</td>
<td>16,182</td>
<td>20,934</td>
<td>100</td>
</tr>
<tr>
<td>9th</td>
<td>20,495</td>
<td>17,765</td>
<td>100</td>
</tr>
<tr>
<td>11th</td>
<td>6,514</td>
<td>8,250</td>
<td>100</td>
</tr>
<tr>
<td>Knowledge Project Total</td>
<td>48,203</td>
<td>56,211</td>
<td>91</td>
</tr>
<tr>
<td>Project Gutenberg 11th</td>
<td>n/a</td>
<td>6,969</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>63,180</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


This chart shows that we created files for approximately 56,000 encyclopedia entries during the award period. After integrating the Project Gutenberg files into that data set, the total grows to about 63,000 entries from the four editions. As you can see, 91% of our OCR output is fully processed to date, and we have more OCR pages for the 3rd edition waiting. The total number of entries will grow as we finish the OCR and processing for the 3rd edition.
Appendix K: TEI with Subject Metadata

TEI Header for “Seamanship” entry in the 7th edition. The first set of keywords (highlighted) contains subject terms from the 1910 LCSH; the second set is from FAST Topical; the third set is from FAST Geographical. (FAST is a faceted subject heading schema derived from the current Library of Congress Subject Headings.) The <prefixDef> field (highlighted) contains a regex pattern that browsers use to automatically replace the abbreviation in the URI for each subject term.

```xml
<term ref="lcsh1910:b4222rd9g">Robin</term>
<term ref="lcsh1910:b48w38662">Fluids</term>
<term ref="lcsh1910:b4fj29d00">Chronometer</term>
<term ref="lcsh1910:b4bk16p7r">Architecture</term>
<term ref="lcsh1910:b44j0b09t">Tactics</term>
<term ref="lcsh1910:b4sn0149t">Sails</term>
<term ref="lcsh1910:b4513v11w">Water</term>
<term ref="lcsh1910:b4707wx1r">Ships</term>
</keywords>
```
```xml
<keywords scheme="http://id.worldcat.org/fast/">
<term ref="fast:1103513">Sailing</term>
<term ref="fast:1103579">Sails</term>
<term ref="fast:1171139">Water</term>
<term ref="fast:1110335">Seamanship</term>
<term ref="fast:986676">Keels</term>
</keywords>
```
<term ref="fast:934209">French</term>  
<term ref="fast:839044">British</term>  
<term ref="fast:1098971">Robins</term>  
<term ref="fast:928067">Fluids</term>  
</keywords>  
<keywords scheme="http://id.worldcat.org/fast/">  
<term ref="fast:1210276">India</term>  
<term ref="fast:1239509">Africa</term>  
<term ref="fast:1219920">England</term>  
<term ref="fast:1205401">Switzerland</term>  
<term ref="fast:1206830">Brazil</term>  
<term ref="fast:1209332">Madagascar</term>  
<term ref="fast:1206073">China</term>  
</keywords>  
</textClass>  
</profileDesc>  
</teiHeader>  
<text>  
</TEI>
Appendix L: Automatic Subject Metadata Workflow

Our final workflow resulted from the following sequence of tests.

1. Initial results with Chambers

Early exploratory research with controlled vocabulary automatic subject indexing found that there are time-specific vernacular disparities between the historical encyclopedia entries and the contemporary Library of Congress Subject Headings (LCSH) used for indexing.¹⁹ This disparity manifested through two phenomena: 1) automatic metadata generation results that were anachronistic for the 19th-century text; and 2) the absence of concepts essential to the encyclopedia entry. For example, no concepts related to Islam and Muslims were represented in the metadata for a 7th edition entry on Yemen, despite the presence of several essential terms in the entry, such as *Wahabees* and *Saracens*. This finding informed the decision to explore the use of historical vocabulary versions to address this vernacular misalignment and extract essential terms that would otherwise be missing.

The first phase of exploring the use of historical vocabularies employed the ontological structure of Ephraim Chambers’s *Cyclopaedia* (1728).²⁰ This ontology included 47 top-level classes, as well as approximately 2,900 concepts under them. A small-scale study was performed to compare the relevance of automatic indexing results when using the 2018 LCSH versus indexing results when using the historical Chambers ontology. A pilot sample of 30 entries of varying lengths were randomly selected from entries that were available at the time. The results showed that there was an 8.12% increase in relevant results when using the Chambers ontology.²¹ Subsequent exploration of Chambers with shorter entries found a significant decrease in recall, due to the small size of the Chambers vocabulary. This limitation informed the decision to explore the first edition of the Library of Congress Subject Headings (1910-1914). The 1910 LCSH is a large vocabulary published around the same time as the 11th edition of the *Encyclopedia Britannica* (1911), so it allows for comparison between historical and contemporary LCSH vocabulary versions.

2. Preliminary findings with 1910 LCSH

A pilot study was performed to identify the potential value of automatically indexing these historical encyclopedia entries with contemporaneous controlled vocabulary versions. This study focused on identifying the effect of vernacular-specific drift in controlled vocabularies over time. A stratified random sample of ninety encyclopedia entries was selected across the four historical *Encyclopedia Britannica* editions in the data set. To account for the wide range of entry lengths across the sample, three sets of samples were used to represent short, medium, and long entries. In addition to comparing indexing results using the 1910 LCSH and FAST Topical vocabularies, two indexing approaches were

---


²⁰ We are grateful to Dr. Joseph Tennis for his help with the Chambers ontology.

used with the selected automatic indexing tool, HIVE. The first approach entailed automatic subject indexing of the full text encyclopedia entries. The second approach integrated intermediary named entity recognition (NER) with Stanza, Stanford’s current natural language processing toolkit for Python, which includes Stanford NER. Previous exploration of the intermediary NER approach demonstrated its efficacy for generating precise subject heading results that significantly differed from those that resulted from the full text approach.

Basic descriptive statistics were used to determine the percentage of terms that were only generated using the 1910 LCSH vocabulary, and which of these terms have fully deprecated or been removed from the contemporary Library of Congress Subject Heading vocabulary. Results showed that 30.99 percent of the 1910 LCSH results were terms exclusive to that vocabulary and did not appear in the results using FAST Topical. Overall, 7.24 percent of the total 1910 LCSH results represent subject headings that are no longer used in the contemporary vocabulary version as a preferred label. This phenomenon demonstrates conceptual drift over time. Table 1 below provides a few examples of 1910 LCSH indexing output terms and the 2021 LCSH equivalent term. This research identified the potential for linking these anachronistic terms to their contemporary equivalents for 1) inclusion in the final metadata record; and 2) to trace how changes in knowledge over time manifests through controlled vocabulary terms.

<table>
<thead>
<tr>
<th>Sample Entry</th>
<th>1910 LCSH Term</th>
<th>2021 LCSH Equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entry Term: Yemen</td>
<td>Wahabees</td>
<td>Wahhābiyyah</td>
</tr>
<tr>
<td>Edition: 7th (1842)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entry Term: Sanhedrim</td>
<td>Deformities</td>
<td>Abnormalities, Human;</td>
</tr>
<tr>
<td>Edition: 7th (1842)</td>
<td></td>
<td>Also as topical subdivision</td>
</tr>
<tr>
<td>Entry Term: Teeth</td>
<td>Elasmobranchii</td>
<td>Chondrichthyes</td>
</tr>
<tr>
<td>Edition: 11th (1911)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Preliminary findings with FAST Geographic

The geographical facet of FAST was integrated into the project after preliminary research showed that the combination of intermediary NER with FASTGeo was effective for extracting locations from the entries. Analysis compared FASTGeo automatic indexing results for both full text and intermediary NER approaches. The intermediary NER approach demonstrated a slight increase in the total number of terms retrieved, and a 15 percent average increase in the number of locations retrieved per entry. The comparative analysis also found that the intermediary NER approach identified a greater number of distinct locations and was more effective at extracting FASTGeo locations that were not identified using

22 FAST is a faceted subject heading schema derived from the current Library of Congress Subject Headings.
the full text approach. Exploration of this data also suggests that the NER approach more effectively disambiguated and eliminated non-location terms, such as Franc (the currency) and Main (e.g., water main). A comparative analysis of topical relevance across a random sample of 75 entries also found that the intermediary NER approach had a 6.5% higher precision score, with a precision score of 0.97.

4. Summary of the final method (NER, FAST, and 1910 LCSH)
The final method for metadata generation used a synthesis of automatic indexing approaches and controlled vocabularies. Three controlled vocabularies were used to automatically generate subject headings for the entries: 2021 FAST Topical, 1910 LCSH, and 2021 FAST Geographic. Both full text and intermediary NER approaches were also used. Figure 1 illustrates the final topical metadata generation workflows, and figure 2 illustrates the workflow for extracting locations using intermediary Stanza NER and FAST Geographic.

![Figure 1. Visualization of automatic indexing workflow for 1910 LCSH and FAST Topical, using both full text and intermediary Stanza NER approaches.](image-url)
5. Analysis of final metadata

In order to test the final method, we generated final metadata for all entries in the letters R-Z for all four editions. This sample size constitutes about 20% of the full data set. Analysis compared the output per entry, using the two topical vocabularies and two indexing approaches. Across all 1910 LCSH and FAST Topical indexing output, between 16.51 and 19.43 percent were only generated for that entry using the 1910 LCSH, and between 21.10 and 24.74 percent were only generated for that entry using FAST Topical. Comparing the output from each of the two indexing methods (across all 1910 and FAST results), between 14.30 and 30.34 percent were only generated for that entry using the NER-to-ontology approach, and between 53.10 and 73.14 percent were only generated for that entry using the full text indexing approach.
Appendix M: HIVE2 Online Vocabulary Server

The HIVE vocabulary server includes two vocabularies developed for the Knowledge Project: the “1910 Library of Congress Subject Headings,” and the “Ephraim Chambers Cyclopaedia.”

URL: https://hive2.cci.drexel.edu/
Appendix N: HIVE Search Result in 1910 LCSH

URL: https://hive2.cci.drexel.edu/
Appendix O: Authority File for 1910 LCSH

Knowledge, Theory of

URI
b4m32n94s

ALTERNATE LABELS/VARIANTS

NOTES

BROADER TERMS

NARROWER TERMS

RELATED TERMS

Appearance
Belief and doubt
Comparison (Psychology)
Experience
Faith, philosophy
Intellect
Intuition (Philosophy)
Intuitionism
Judgment (Logic)
Perception
Pragmatism (Philosophy)
Reality
Sufficient reason
Truth
Worth

SOURCES

  Vol. 1 (A-F), https://hdl.handle.net/2027/coo.31924050327042
  Vol. 2 (G-P), https://hdl.handle.net/2027/coo.31924050327059

  Vol. 4 (Q-R), https://hdl.handle.net/2027/mdp.39015014520715
  Vol. 5 (S-Z), https://hdl.handle.net/2027/mdp.39015014520731

URL: https://onemoreiteration.com/lcsh-1910/lcsh-1910-term.html?b4m32n94s
Appendix P: Image Sources

Bibliographical information for the print source is given, followed by the location, if known. The digital archive where the files are stored, and the file format is given next. Finally we acknowledge the organization that digitized the print volumes.

3rd edition


7th edition

vol. 1, vol. 2 (pp. 415-828), vol. 3-15

*Encyclopaedia Britannica. 7th ed. 21 vols. Edinburgh: Adam and Charles Black, 1830-1842.* Print source at the National Library of Scotland. NLS, PDF format. Digitized by the NLS.

vol. 2 (pp. 1-414), vol. 17-21


vol. 16


9th edition

vol. 1-13, vol. 14 (pp. 1-33, 36-868), vol. 15-25


vol. 14 (pp. 34-35)


11th edition

vol. 1-29, except as noted


vol. 24 (pp. 190-191)

Appendix Q: Metadata Research Center Web Page on the Knowledge Project

Developing the Data Set of Nineteenth-Century Knowledge

Abstract
A project to study the structure and transformation of nineteenth-century knowledge via computational analysis of several editions of the Encyclopedia Britannica from 1788 to 1911.

This project draws on historic editions of the Encyclopedia Britannica, a vital resource of knowledge to build one of the most extensive, open, digital collections available today for studying the structure of nineteenth-century knowledge and its transformation. The most comprehensive representation extant of what constituted official knowledge at the time, they also demonstrate changes in the nature of knowledge in the English-speaking world. The project creates the first accurate textual data for this corpus and extends its usability by applying innovative methods to automatically generate metadata for each of the 100,000 entries. Each entry will be tagged with both current and historical subject categories. At the end of the grant period, all of the data will be made freely available, and a series of experiments will be conducted to identify the feasibility of tracking concept drift across time within the corpus.

Read more about the 19th Century Knowledge Project.

Project Team
- Sam Grabus, Lead Research Assistant
- Peter Logan, Project PI
- Jane Greenberg, Project PI

Publications

URL: https://cci.drexel.edu/mrc/research/nineteenth-century-knowledge
Appendix R: Knowledge Project Public GitHub Repository