Narrative Section of a Successful Application

The attached document contains the grant narrative and selected portions of a previously funded grant application. It is not intended to serve as a model, but to give you a sense of how a successful application may be crafted. Every successful application is different, and each applicant is urged to prepare a proposal that reflects its unique project and aspirations. Program guidelines also change and the samples may not match exactly what is now required. Please use the current set of application instructions to prepare your application.

Prospective applicants should consult the current Office of Digital Humanities program application guidelines at https://www.neh.gov/grants/odh/digital-humanities-advancement-grants for instructions.

Applicants are also strongly encouraged to consult with the NEH Office of Digital Humanities staff well before a grant deadline.

Note: The attachment only contains the grant narrative and selected portions, not the entire funded application. In addition, certain portions may have been redacted to protect the privacy interests of an individual and/or to protect confidential commercial and financial information and/or to protect copyrighted materials.

**Project Title:** Developing the Data Set of Nineteenth-Century Knowledge

**Institution:** Temple University

**Project Directors:** Peter Logan and Jane Greenburg

**Grant Program:** Digital Humanities Advancement Grants, Level II
Participants

Senior Personnel
Logan, Peter M., Temple University, Professor of English and Academic Director of the Digital Scholarship Center
Greenberg, Jane, Drexel University, Professor of Informatics and Director of the Metadata Research Center

Research Collaborators
Bingenheimer, Marcus, Temple University, Associate Professor of Religion
Shoemaker, Matt, Temple University, Librarian and Coordinator of Digital Scholarship Service Development

Advisory Board
Flanders, Julia, Northeastern University, Professor of Practice in English and Director of the Digital Scholarship Group
Jockers, Matthew L., University of Nebraska-Lincoln, Associate Professor of English
Tennis, Joseph T., University of Washington, Associate Professor of Information

Letter Writers
Flanders, Julia, Northeastern University
Lacey, David, Temple University, Director of Library Technology and Knowledge Management Services
Lucia, Joseph, Temple University, Dean, Temple University Libraries
McNamee, Robert V., Oxford University, Director Electronic Enlightenment Project and Oxford Text Archive
Mandell, Laura, Texas A&M University, Professor of English and Director of the Initiative for Digital Humanities, Media, and Culture
Piper, Andrew, McGill University, Professor of Languages, Literatures, and Cultures, and Director of txtLAB @ McGill
Developing the Data Set of Nineteenth-Century Knowledge

Enhancing the Humanities

This project draws on 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 nineteenth century marked a dramatic transformation in the nature of what constituted legitimate knowledge. This was the era that brought us the Industrial Revolution and that saw cultural authority for knowledge shift from religion to science, as the former declined in influence and the latter increased. The different editions of the *Encyclopedia* chronicle that transformation in their changing treatment of many topics. The entry for “History” in the seventh edition (1842) asserts the biblical narrative from Genesis as fact. The next major edition, the ninth (1889), eliminates the biblical narrative, replacing it with a treatise on historiography that reflects Auguste Comte’s Positivism. It also limits itself to recorded human history, and we see human prehistory migrate to a new entry, “Anthropology,” written by the central figure in British anthropology, Edward B. Tyler. The early history of life on earth also migrates to a new entry, “Evolution,” written by Thomas Huxley. Such individual changes are part of a larger pattern. The Encyclopedia was designed to include all knowledge, and in that comprehensiveness, it represents a system of knowledge, rather than discrete topics in isolation. By creating a data set of multiple editions, we create a resource with which to study changes in that system over time.

Tracking these changes is complex and interpreting them even more so, in part because, at approximately twenty-five million words each, these editions exceed our ability to read them from cover to cover. That same scale makes them good candidates for the use of quantitative text-mining procedures. This project enhances the usability of the Encyclopedia corpus for digital humanists by creating accurate textual data from key historic editions and making it available for analysis. A new collaboration with the Metadata Research Center at Drexel University will add innovative subject metadata to each Encyclopedia entry, further enriching it for researchers.

*Britannica* produced fifteen print editions, beginning in 1768. Not all of them were equally important or original. The relatively small (3 vol.) first edition consisted mainly of material cut and pasted from other sources. We focus on four major editions consisting primarily of new material, dating approximately from the French Revolution to WWI: the third (20 vols., 1788-97), seventh (22 vols., 1830-42), ninth (25 vols., 1875-1889), and eleventh (28 vols., 1910-11). When complete, all data will be made freely available through two public repositories: the Oxford Text Archive (OTA) and the Humanities Commons Open Repository Exchange (CORE).

Systems of Knowledge and the *Encyclopedia Britannica*

In the nineteenth century and well into the twentieth, *Encyclopedia Britannica* was the general reference source of record for the English-speaking world. Beginning with the third edition of 1788-97, its editors began a programmatic effort to recruit notable contributors as authors and paid enough to attract them. The list of authors made it a virtual congress of intellectuals in the arts and sciences. Political economist Thomas Malthus wrote the entry on “Population.” Astronomer John Herschel wrote on “Meteorology” and the “Telescope.” Utilitarian philosopher James Mill penned entries on “Government” and “The Law of Nations.” Novelist Walter Scott authored “Drama” and “Romance. Thomas Young’s 1818 entry on “Egypt” was the first published description and partial translation of the Rosetta Stone. Because it used prominent scholars, the Encyclopedia reflected changes both large and subtle in current ideas, creating its reputation as the most authoritative summary of knowledge at the time. It continues as an online publication today and is the only encyclopedia to survive that 250-year period.

The value of historical editions of the Encyclopedia to humanists is unique and its history well documented.¹ Sociologists have examined the evolution of particularly sensitive cultural topics, such as

suicide, by studying their changing treatment in different editions. But so far we have been unable to move beyond qualitative research on isolated subjects like these to consider what this curated data set as a whole might reveal about the social construction of knowledge in the nineteenth-century English-speaking world.

Decisions to include entries and the sizing of entries were based on assumptions at the time about what counted as legitimate knowledge. Many of these are assumptions we no longer share. The editors excluded forms of knowledge rooted in folk and tribal cultures. Their approach to religion changed over time, but initially privileged Christianity. Articles on India and Africa reflect the perspective of the British Empire rather than that of indigenous populations. Racism is evident in many entries. These prejudices reflect the social beliefs of the authors and editors, of course, and as such, they illustrate the profound degree to which knowledge in the nineteenth century was socially constructed. Rather than dismissing these entries because of their biases, we value them because of what they tell us about the social conditions within which knowledge was produced. The curated contents of the Encyclopedia are the best representation we have in the English-speaking world of that system in its totality.

Individual editions represent the condition of knowledge at a given point in time, but multiple editions document its systemic transformation across time. Because this knowledge tradition is the also the predecessor of current knowledge in the English-speaking world, identifying the patterns of change within this corpus bears directly on humanistic scholarship then and today.

Existing Digital Resources

Because of the scale of this corpus, quantitative analysis is essential to examining these questions. While high-quality image scans of the older editions are freely available, the extant textual data derived from these images is so error-prone that it is unusable (see Appendix A). The absence of accurate data poses the main obstacle to extending the value of this material to digital humanists.

The current textual data was extracted from page images by Google Books, using a fully automated procedure that was not adapted to the complex layout of the two-column pages (see Appendix B and Appendix C). In 2012 a group of NEH-funded researchers attempted a large-scale analysis of historic editions of the Encyclopedia but truncated their goals to consider only proper names because of the poor quality of Google’s textual data. They calculated the error rate in the transcriptions of the third and ninth editions as between one-in-twenty and one-in-ten words (5.3-9.6%). With one or more errors in every sentence, the current textual data precludes drawing valid conclusions from the existing data set.

Enhancing the Data

During the grant period, we will create accurate textual data for four historical editions of the Encyclopedia, using the equipment and resources of the Digital Scholarship Center (DSC). The data set will be enriched with both standard and innovative subject metadata for each entry, by Drexel’s Metadata Research Center (MRC). Dr. Logan and Dr. Greenberg will run preliminary tests comparing different methods of tracking concept drift in the data set, network analysis and online topic-modeling, and publish their results. At the end of the grant period, the data set will be deposited with the Oxford Text Archive, who will make it freely available to the public in a user-friendly form. The complete data set will also be freely available in bulk form through Humanities CORE.

---

The project aims to satisfy three qualitative criteria for the final corpus. Each will enable us and other researchers to identify specific patterns of change: (1) high-quality data; (2) a flexible storage format; (3) and rich metadata.

**High-quality data.** We use a semi-automated method to improve accuracy in OCR to $\geq 99\%$. The process includes manually identifying the text areas for each page, as well as isolating note anchors and note texts. The textual data preserves information on the entry title, entry position on the page, and page breaks (see Appendix G). Other tagged elements of the text include:

1. Names, places, and dates, using the Stanford NER (a named entity recognition system).
2. Author attribution for each entry, beginning with the seventh edition, when the Encyclopedia first included that information.
3. Tables, mathematical and engineering formulae.
4. Italics. While unimportant in themselves, we will use them to identify citations later.

**Flexible storage.** OCR generates files in HTML format, with one print page per file. Using an XSLT script in Oxygen XML Editor, we transform the HTML files into valid TEI-XML. We then run a Python script to make the handling of footnotes conform to TEI guidelines; it removes the note text from the foot of the page and relocates it within the body at the anchor point for the footnote reference. The same script performs an additional step of concatenating the separate files; it then identifies the beginning and end of each entry and generates a new file for each. The entry is the basic storage unit for all of the encyclopedia data. We use TEI-XML as the basic data storage format because XML accommodates detailed metadata and content tagging, but it also converts quickly into a variety of other formats used in specific analytical procedures and for online display.

**Rich metadata.** In collaboration with the Metadata Research Center at Drexel University, we intend to curate the full data set and work with the Helping Interdisciplinary Vocabulary Engineering (HIVE) (http://cci.drexel.edu/mrc/research/hive/) technology to assign an appropriate topical vocabulary. Funded by the Institute of Museum and Library Services, HIVE enables automatic metadata generation, using multiple vocabularies and linked data technology. We will use the Library of Congress Subject Heading (http://id.loc.gov/authorities/subjects.html) and other selected vocabularies, as part of the metadata creation process. Given that knowledge ontologies change in sync with the times, we will use the HIVE technology to experiment with applying historical ontologies. Our objective is to develop sets of subject terms that reflect categorizations of knowledge at the time each edition was published. This avoids the problem of historical anachronisms in the metadata, like categorizing the 1889 entry on “Anthropology” as a social science, when at the time it was still classified as a life science by the British Association for the Advancement of Science. Adding both historical and contemporary subject tags makes it possible to parse the data in the way that is critical to investigating the research questions we seek to support. Our approach with standardizing metadata across the collection will also enable greater tracking of the historical provenance of concepts, as well as interoperability and better precision in searching. It should allow us and others to reconstruct larger patterns in the way some topics originated within one subject field but were later picked up by others, creating concept drift.

**Community engagement.** In addition to the quality criteria for the corpus, this project also serves the local scholarly community as a training ground for graduate student researchers and post-doctoral fellows from both Temple and Drexel. To further encourage an awareness of advanced coding techniques, we are planning a symposium in the final months of the grant, in fall 2019, for regional scholars to discuss the intellectual challenges of text encoding and analysis, with presentations on preliminary analyses of subsets of the Encyclopedia data and discussions of advanced metadata techniques.

---

6 Oxygen is proprietary software but there is no comparable open-source product available for editing XML-related formats and running XSLT.

7 On linked data, see Tim Berners-Lee, “Linked Data,” W3C (2006), www.w3.org/DesignIssues/LinkedData.html; see also www.w3.org/standards/semanticweb/data.
Environmental Scan


Work in improving OCR technology continues to improve accuracy. The Early Modern OCR Project (eMOP), managed by Laura Mandell at Texas A&M, focused on texts from two large collections Early Modern and eighteenth-century collections, EEBO and ECCO, and achieved an 86% accuracy rate on ECCO books and 68% on EEBO. Unlike those collections of multiple sources, the only pre-1820 edition of the Encyclopedia, the third edition, presents fewer problems because it is a single print source with high-quality images, rather than a collection of mixed print sources. This has made it feasible for us to train the recognition engine to achieve results that exceed eMOP’s 97% target accuracy rate. Two European consortia are also working on Early Modern digitization problems, and their findings to date have informed our own procedures: Linguistic DNA and IMPACT digitization.eu. ActiveOCR is proof-of-concept project at the University of Maryland designed to crowd source OCR text correction for eighteenth-century texts, which may also prove useful to us once it is fully operational.

History of the Project

Work on the project began in the new DSC in August 2015. We acquired archival-quality, high-resolution image files for three of the four editions in the project: the third, ninth, and eleventh. The highest quality images of the seventh edition are large Adobe Acrobat files (.pdf) in the Hathi Trust collection, and we acquired those as well. (See Appendix D for image sources.)

*Testing program.* In AY 2016-17, we ran pilot trials to optimize OCR techniques on the four editions. ABBYY FineReader (AFR) was purchased for project use. With training of the recognition program, we achieved accuracy rates > 99% for the seventh, ninth, and eleventh editions. Printed matter before 1820 poses known difficulties for modern OCR engines, which the third edition exhibits; it uses an older font (Caslon), the “ſ” character, and ligatures. After adding the correct font and the “ſ” character to AFR and improving its built-in dictionary, we managed a 98% accuracy rate. Using a modified version of Ted Underwood’s Python script for correcting the most common of these errors, we were able to further increase accuracy to 99.0%.

The OCR process is semi-automated. Operators identify the text boundaries on each page for the OCR engine to analyze; they also flag footnote text and note references, to avoid OCR inconsistencies. The print pages use a two-column layout with a running header, but each edition differs. The third edition uses callouts in the left- and right-hand page margins for subheads and references to images, and these page elements confuse the OCR engine (see Appendix B). The seventh edition limits the use of callouts to subheadings but adds a new region for footnotes running the width of the page. In both of these early editions, images are restricted to separate pages. The ninth and eleventh editions did away with the

---

8 Laura Mandell, Matthew Christy, and Elizabeth Grumbach, "eMOP Mellon Final Report," (Texas A&M University, 2015).
11 AFR is proprietary software, but it currently has a higher accuracy rate than either Tesseract or OCRopus, the two open-source OCR engines.
callouts in the margins, but they inserted cutouts for small images directly into the running text, causing irregularly shaped text regions that wrap around the images (see Appendix C). Operator involvement in the OCR process averages 30 seconds/page (120 pages/hour) when following the guidelines detailed in the online Project Manual (https://tu-plogan.github.io/).

In spring 2017 we developed an XSLT script to transform AFR’s HTML output into TEI-XML. We also created a Python script to identify footnote text in each TEI page and position it within the body at the note anchor point as TEI <note> data. The Python script also concatenates individual page files into a single text variable. It then identifies the start of each entry and outputs a single XML file for each entry, preserving page breaks and adding the TEI header data.

In December 2017, Temple’s DSC and Drexel’s Metadata Research Center agreed to collaborate on the project, creating a new and promising relationship between the two centers. The Drexel Center agreed to develop new knowledge ontologies from the nineteenth century and to refine their existing automated metadata technology for the project. Both institutions are conveniently located in Philadelphia.

Work Plan

Work on creating the data set will be completed in AY 2018-19 at Temple’s Digital Scholarship Center. This is a laboratory space equipped with eight high-end PC’s and four Macs. All of the PC’s are equipped with AFR, Oxygen XML Editor, and Python.

The full corpus of four editions totals 74,000 print pages. Excellent data already exists for part of the eleventh edition from Project Gutenberg (equivalent to 11,622 pages), and we have permission from them to use the material (see Appendix E). This leaves 62,447 pages to be scanned. Prior to the grant period, the DSC will have completed processing 26,600 pages, or 42.6% of the total. During the first year of the award, GRAs will generate text for 35,847 pages, completing the data set.

Timeline

In the following timeline: (1) OCR processing is calculated at 60 pages/hour; this is one-half of the actual maximum of 120 pages/hour, so it allows time for initial training, breaks, and unexpected delays; (2) contributions of GRAs funded by Temple University are indicated by an asterisk: *GRA; all others are funded by the Grant; (3) “DSC” refers to Temple University’s Digital Scholarship Center, and “MRC” to Drexel University’s Metadata Research Center;

**Fall 2018**
- DSC: Dr. Logan trains new graduate student assistants and supervises their work.
- DSC: One *GRA and one GRA (combined 20 hrs./wk.) process 18,000 print pages. Cumulative total is 44,600 pages (71.4% of total to be scanned).
- DSC: Dr. Logan and Matt Shoemaker collect Project Gutenberg data.
- DSC and MRC: Dr. Logan and Dr. Greenberg make initial plans for fall 2019 Symposium.
- Consultations with Advisory Board members

**Spring 2019**
- DSC: One *GRA and one GRA (combined 20 hrs./wk.) process 17,847 print pages. Cumulative total is 62,447 pages (100% of total to be scanned).
- DSC: Dr. Logan and Dr. Bingenheimer write XSLT script to generate metadata for all entry files.
- DSC: Dr. Logan and Matt Shoemaker integrate Project Gutenberg material into the dataset.
- MRC: Dr. Greenberg and one GRA refine metadata tagging project.
- Symposium: Dr. Logan and Dr. Greenberg finalize schedule and begin circulating publicity.
- Online meeting of Advisory Board with Dr. Logan and Dr. Greenberg.

**Summer 2019**
- DSC: One GRA to complete any unfinished OCR work and transform all files into TEI-XML.
- DSC: Dr. Logan uses Python to create the individual entry files.
- MRC: Dr. Greenberg and one GRA complete the metadata tagging project and trial methods for identifying concept drift.
- Conference: Dr. Greenberg and Dr. Logan present talk at DH2019 on the Encyclopedia data set.
- Data Evaluation: Dr. Logan submits the data set to NINES for peer review.14

Fall 2019
- Symposium: One-day symposium in October.
- DRC: Dr. Logan and one *GRA finalize the dataset of 100,000 files by cleaning the data and prepare it for deposit with OTA and Humanities CORE.
- Dr. Logan and Dr. Greenberg prepare journal article for submission in January 2019.

Final product and dissemination

The final product will consist of approximately 100,000 XML files encoded in TEI. These contain all of the textual data and complete metadata for the third, seventh, ninth, and eleventh editions of the Encyclopedia, with one entry per file. This complete data set will be distributed by two repositories; Oxford Text Archive (http://ota.ox.ac.uk/) will make individual entry files available in a user-friendly format, while the Humanities CORE (https://hcommons.org/core/) will preserve the bulk data for researchers to download. The two primaries on the grant will also present papers on the project at the 2019 conference of the Alliance of Digital Humanities Organizations (http://www.adho.org/), July 9-12, in Utrecht, Netherlands and write a peer-reviewed article for Cultural Analytics or a comparable venue.

The Oxford Text Archive (OTA) was established in 1976 to collect, catalogue, preserve, and distribute electronic literary and linguistic resources for use in higher education research and teaching. Since 2016, it has been located in Oxford’s Bodleian Library. The OTA is also involved in the development of standards and infrastructure for electronic language resources. It hosts about 60,000 texts in more than twenty-five languages, including copies of all of the Eighteenth Century Collections Online (ECCO) and Early English Books Online (EEBO). The bulk of their collection, more than 97%, is TEI-encoded, and TEI texts are made available in a variety of formats. Among other projects, it is a member of CenterNet, the international network of digital humanities centers and is an active member of CLARIN, the pan-European consortium dedicated to building a joint infrastructure for language resources. It also works closely with the TEI, the Text Encoding Initiative consortium. The OTA is actively working on new ways to link together the content of its resources with other collections held in the Bodleian Libraries (e.g. the Electronic Enlightenment, a collection of high-quality digital editions of correspondence) and elsewhere.

Dr. Logan will work with the OTA team, Dr. Robert V. McNamee, Martin Wynne, and Mark Rogerson on how best to prepare the corpus for deposit in OTA and to link it with items in their other collections. All material will be deposited with a Creative Commons CC BY-SA 4.0 license (https://creativecommons.org/licenses/by-sa/4.0/), which allows the material to be freely shared and reused as long as its creator is credited and any new material built upon it is shared under the same licensing terms. OTA will make the dataset available in a user-friendly form, with options to view or download individual articles in HTML, PDF, XML, and TXT formats.

The complete data set will also be uploaded to the Humanities CORE, a non-profit, open-access repository created by the Modern Language Association in 2013. TEI-XML entry files will be bundled as ZIP or RAR archive files, of about 10 MB each, representing entire print volumes. This is the optimal way to keep the metadata fully intact while sharing the complete data set with other researchers. It will also be deposited under the same CC BY-SA license. Using Humanities CORE ensures that the bulk data set is safely archived and made widely available to others. It makes it highly discoverable by giving each file a permanent identifier, or DOI, creating a persistent link and citable metadata. Materials in the repository are indexed by Google, Google Scholars, SHARE, Altmetric, and BASE-OA.

14 http://www.nines.org/about/scholarship/peer-review/.
Biographies
Senior Personnel


Dr. Jane Greenberg is the Alice B. Kroeger Professor and Director of the Metadata Research Center (http://cci.drexel.edu/mrc/) at the College of Computing & Informatics, Drexel University. Her research activities focus on metadata, knowledge organization/semantics, linked data, data science, and information economics. She serves on the advisory board of the Dublin Core Metadata Initiative (DCMI) and the steering committee for the NSF Northeast Big Data Innovation Hub (NEBDIH). She is a principal investigator (PI) on the NSF Spoke initiative, 'A Licensing Model and Ecosystem for Data Sharing,' and the lead PI the Metadata Capital Initiative (MetaDataCAP'T'L) and the Helping Interdisciplinary Vocabulary Engineering (HIVE) linked data project. She is also a co-PI for Drexel's NSF Industry/University Collaborative Research Center (NSF-I/UCRC), Center for Visualization and Decision Informatics (CVDI). Her research has been funded by the NSF, NIH, IMLS, Microsoft Research, National Library of Medicine, Library of Congress, OCLC Online Computer Library Center, among other organizational and private sponsors. She has received numerous awards and honors for her research and leadership; most recently she was recognized as a 2016 ELATE at Drexel® Fellow and in 2014, and a Data Science Fellow at the National Consortium for Data Science, Chapel Hill, North Carolina.

Research Collaborators

Dr. Marcus Bingenheimer is Associate Professor of Religion at Temple University and specializes in Buddhism. He was responsible for the Chinese Localization of TEI. He has published extensively in the Digital Humanities with a focus on markup technologies and currently serves as consultant to multiple digital projects in the field of Buddhist and Asian Studies. He received an M.A. in Sinology and a Ph.D. in the History of Religion from Würzburg University, as well as an M.A. in Communication Studies from Nagoya University. From 2005 to 2011 he taught Buddhism and Digital Humanities in Taiwan, where he supervised various projects concerning the digitization of Buddhist culture, and he continues to publish work on Buddhist Studies, with a focus on Chinese Buddhist history.

Mr. Matt Shoemaker, M.L.I.S. is Coordinator of Digital Scholarship Service Development at Temple University Libraries. He oversees the daily operation of the Digital Scholarship Center. He holds an M.L.I.S. with a concentration in archives and a M.A. in history, focused on French empire in North Africa, both received from the University of Wisconsin Milwaukee. He has worked to build the digital scholarship program on Temple's campus since 2013. He has created and given workshops in several areas of digital scholarship including making technologies (3D printing, 3D scanning, physical computing, and photogrammetry), basics of GIS for digital scholarship, creating digital exhibitions, textual analysis, data cleanup, project design, games for education and as historical models and other digital scholarship areas. Prior to coming to Temple University, Shoemaker worked for the Historical Society of Pennsylvania where he led the development of HSP’s digital library, digitization program, and co-authored several successful grant proposals for digital projects.
Advisory Board
The Advisory Board was established to lend expertise to the project in three key areas: TEI-encoding best practices, analytical procedures for testing concept drift, and familiarity with knowledge organization and subject vocabularies in the nineteenth century. The project’s senior personnel consult every semester with the Advisory Board members to ensure that the data set meets the highest standards possible.

Dr. Julia Flanders, Professor of Practice in English and Director of the Digital Scholarship Group, Northeastern University. She is Director of the Women Writers Project at Brown University and the founder and editor-in-chief of Digital Humanities Quarterly. A specialist in TEI, she will advise the project on all aspects of text encoding for the project.

Dr. Matthew L. Jockers, Associate Dean for Research and Partnerships, College of Arts & Sciences, and Susan J. Rosowski Associate Professor of English, University of Nebraska-Lincoln. A former research scientist and software development engineer for Apple Computer, Dr. Jockers has published widely on textual analysis methods and will advise the project on preparing the data for research and analytical techniques such topic modeling and network analysis.

Dr. Joseph T. Tennis, Associate Professor and Associate Dean for Faculty Affairs, School of Information, University of Washington. Currently President of the International Society for Knowledge Organization, he specializes in classification theory, information provenance, and comparative metadata analysis. He is advising the project on the history of knowledge ontologies and will assist in constructing historical ontologies for use adding subject-area metadata.
Appendices

Appendix A: Existing Textual Data

Example of Google Books OCR for the third edition, showing how marginal callouts are erroneously interpreted as part of the body text. The sample also shows that Google Books always substitutes the letter “f” for “f” (long “s”), leading to a high transcription error rate.

RAN

Rain, \( T_3 \) ANA, the frog, in zoology; a genus belonging to the order of amphibia reptilia. The body is naked, furnished with four feet, and without any tail. There are 17 species. The most remarkable are,

1. The temporaria, or common frog. This is an animal so well known, that it needs no description; but some of its properties are very singular.

Its spring, or power of taking large leaps, is remarkably great, and it is the best swimmer of all four-footed animals. Nature hath finely adapted its parts for those ends, the fore members of the body being very lightly made, the hind legs and thighs very long, and furnished with very strong muscles.

While in a tadpole state, it is entirely a water animal; the work of generation is performed in that element, as may be seen in every pond during spring, when the female remains oppressed by the male for a number of days.
Appendix B: Page Layout for Third Edition

Typical page layout for the third edition, with large number of callouts in the margin.

Logan 17
Appendix C: Page Layout for Eleventh Edition

Complex page layout with cutouts from the eleventh editions. ABBYY FineReader has difficulty recognizing all text blocks and frequently sequences them wrong in the text output.
Appendix D: Image Sources

Image Sources

Third Edition

Seventh Edition

Ninth Edition

Eleventh Edition
Appendix F: Data Organization

Data Organization Scheme

The chart shows the hierarchical structure of XML files, truncated to display two entries, two volumes, and two editions only. Each block in the diagram represents one XML file. The actual scheme has four editions of 20-28 volumes each, and 20,000 entries per volume. The hierarchical structure allows all metadata for each level of the hierarchy to be inherited by the levels below them, eliminating duplication and the possibility of error, while preserving the display of full metadata when any individual entry is output. The structure is also extensible. Should additional volumes be added in the future, they are easily integrated into the existing scheme.
Appendix G: Coding Sample

The example shows an entry from the ninth edition that begins near the end of the page and has two footnotes before the page break. The code is from the XML master file. It shows the treatment of the entry term (the `<label>` tag), the placement of note text (the `<note>` tag), and the page break encoding (the `<pb>` tag). In the next step, we add metadata and tag named entities. This note placement reflects the TEI guidelines; notes are automatically numbered and links created between the anchor point and note text when the file is transformed for output.

```
ROOK (Anglo-Saxon Hrōc, Icelandic Hrókr,1 Swedish Råka, Dutch Rock, Gaelic Rocas), the Corvus frugilegus of ornithology, and throughout a great part of Europe the commonest and best-known of the Crow-tribe. Besides its pre-eminently gregarious habits, which did not escape the notice of Virgil (Georg. i. 382)2 and are so unlike those of nearly every other member of the Corvidae, the Rook is at once distinguishable from the rest by commonly losing at an early age the feathers from its face, leaving a bare, scabrous, and greyish-white skin that is sufficiently visible at some distance. In the comparatively rare cases
```

1 The bird, however, does not inhabit Iceland, and the language to which the name belongs would perhaps be more correctly termed Old Teutonic. From this word is said to come the French `<hi rend="i">Rêch</hi>`; and others, but the bird is generally known in Germany as the `<hi rend="i">Saat-Krehe</hi>`; and in English as the `<hi rend="i">Crow</hi>`.

2 This is the more noteworthy as the district in which he was born and educated is almost the only part of Italy in which the Rook breeds. Shelley also very truly speaks of the "legioned Rooks" to which he stood listening "mid the mountains Euganean."

...
Data Management Plan

Roles and Responsibilities

This data management plan will be implemented and managed by Matt Shoemaker, under the project supervision of Peter Logan. Mr. Shoemaker will oversee maintenance, backup, and archiving of data generated by the project. And he will manage the transfer of data for project research to Drexel’s Metadata Research Center. He will also be responsible for final project data transfers to the OTA and Humanities CORE repositories. All repository data will be publicly accessible. If Mr. Shoemaker leaves Temple University during the course of the grant, his role will be taken over by his successor as Coordinator of the Digital Scholarship Center.

Data Storage Hardware

During the production phase, all data is stored locally in the DSC on its internal server, consisting of a networked pair of 6TB hard drives in a RAID1 configuration. Access to the DSC server is limited to project participants authorized by Dr. Logan. Project files are automatically archived daily to a separate 4TB external hard drive. Files are also synchronized daily with an online Box service provided by Temple University, with remote access limited to team members.

Data Formats

1. Image files. These are duplicates of external image files downloaded from the Hathi Trust or the Internet Archive (see Appendix D).
2. AFR Project Files. These are file folders generated by ABBYY FineReader to organize large amounts of page images for scanning.
3. AFR User Files and Dictionary Files. These are proprietary files that store custom information used to fine-tune the OCR process.
4. HTML files. There are two types:
   a. Those output by AFR, with one file for each print page of the Encyclopedia.
   b. Final files generated by XSLT from the project’s TEI-XML data files with full metadata, for uploading to the OTA at the completion of the project.
5. XSLT scripts.
6. TEI-XML files. These contain the core textual data of the project and the generated metadata.
7. Oxygen Project Files. These are small data files for use by Oxygen XML Editor to organize large numbers of XML files.
8. TXT files, generated by XSLT from the project’s TEI-XML data files for internal use in testing the viability of using online topic-modeling to identity concept drift.
9. Project spreadsheets documenting the creation of all files and any modifications made to them.

Data Organization Plan

Core Data and Metadata Organization

In order to output consistent metadata with the textual data, our 110,000 individual entry files are organized within a hierarchy of file dependencies, with entry files at the bottom level (see Appendix F). Above them are container, or “wrapper” files for each volume, followed by wrappers for each edition, and finally a corpus wrapper containing basic encoding to all files. Files at lower levels of the hierarchy inherit the attributes of those above them. They also have unique metadata describing the content of each entry. This structure allows us to keep the encoding of entry files as simple as possible, while the series of dependencies means that each entry will have a comprehensive set of metadata associated when output into its final format for repository storage.
Production File Organization

A comprehensive data organization plan for use by project participants is spelled out in the “Data Organization” section of the online Project Manual, specifying the storage location for all forms of project data (https://tu-plogan.github.io/#source/data_organization.html).

Expected Data for Preservation

The project will generate approximately 110,000 different files of textual data. Each file represents one entry in the four Encyclopedia editions. These files serve as “master files” that are used to generate output in other file formats for end-users, including researchers. From the files, we will generate final “digital edition” files TEI-XML format that include all of their critical metadata within each file, and so serve the needs of researchers working from file metadata, rather than the textual data alone.

Permanent Preservation and Access

One copy of these “digital editions” files will be uploaded to Humanities CORE in bulk form, with all files for each print volume contained within a single ZIP or RAR archive. The four editions contain a total of 89 volumes, so the archive will consist of 89 ZIP or RAR files. The textual data will also be output in two other formats: HTML and TXT, the most useful formats for researchers who want to work with the complete data set. The data set in these alternate formats will be uploaded to Humanities CORE in the same ZIP or RAR archive form. Humanities CORE promises permanent storage and open access for data deposited with them.

A second copy of the data will be made publically accessible in perpetuity by the Oxford Text Archive, when they are uploaded at the end of the project. OTA will make them publically available for free as HTML files readable online. Additional details will be negotiated with them, such as whether or not they wish to supply alternate formats, like EPUB or TXT, and the DSC can provide them with those formats.

Five-Year Preservation

All AFR Project, User, and Dictionary files, plus the XSLT and Python scripts used in the production process, will be preserved internally in the DSC for a minimum of five years, to allow us to regenerate the raw textual data at any time. The XSLT and Python scripts used to modify that data and convert it to TEI-XML will also be uploaded to the project GitHub site, for free download by anyone interested (https://github.com/TU-plogan/encyclopedia-project) during the same time period.

Test Data

In the final stage of the process, Dr. Logan and Dr. Greenberg will trial two different methods for identifying concept drift in the data set and write a journal article explaining their results. The data for those tests will be preserved and posted on Humanities CORE, for use by readers of the journal article or others interested in the outcomes. Some of it will also be shared at the DH2019 conference during our presentation.

Continuing Research

The full textual data set and master files will also be retained by Dr. Logan for continuing research on nineteenth-century knowledge. Dr. Greenberg also will retain a copy of the TEI-XML digital edition files for use in her future teaching.