GOOGLE DIGITAL HUMANITIES AWARDS RECIPIENT INTERVIEWS REPORT

PREPARED FOR THE HATHITRUST RESEARCH CENTER

VIRGIL E. VARVEL JR.
ANDREA THOMER

CENTER FOR INFORMATICS RESEARCH IN SCIENCE AND SCHOLARSHIP

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

CONTENTS

This report does not promote or imply endorsement of the HathiTrust Research Center or any other parties by the University of Illinois. It is intended for research purposes only.

GENERAL PROJECT OVERVIEW AND GOALS

As input into the development, design, and improvement of the HathiTrust Research Center (HTRC), recipients of Google's Digital Humanities Grants were interviewed to identify issues encountered during their projects. This project was guided by the following goals:

- Increase empirical understanding of how to identify materials for use by scholars.
- Increase empirical understanding of how to provide better access to materials for use by scholars.
- Identify meaningful characteristics of content that affect identification, retrieval, and other parameters.
- Identify data preprocessing and transformation issues encountered by scholars.
- Provide input to inform the architecture of the HTRC related to representation of collections, faceted browsing, identifiers, etc.

HATHITRUST RESEARCH CENTER OVERVIEW

The HTRC is a complex partnership intended to aid research and discovery. In their words, "[the] HTRC is a collaborative research center launched jointly by Indiana University and the University of Illinois, along with the HathiTrust Digital Repository, to help meet the technical challenges of dealing with massive amounts of digital text that researchers face by developing cutting-edge software tools and cyberinfrastructure to enable advanced computational access to the growing digital record of human knowledge."[1] The HTRC will enable access to public domain works and limited access to copyrighted works within the almost 9 million volumes (27% public domain) and 3 billion pages of archived materials maintained worldwide by HathiTrust institutions.[2]

Contributing partners in the HTRC at Indiana include: the Data to Insight Center[3]; Office of the Vice-President for Information Technology; Office of the Vice Provost for Research; and the IU Libraries[4]. Contributing partners at Illinois include: the Illinois Center for Computing in the Humanities, Arts, and Social Science[5]; Illinois Informatics Institute[6]; and the National Center for Supercomputing Applications[7].

[1] http://www.hathitrust-research.org/article/about-hathitrust-research-center
[2] For more on HathiTrust see http://www.hathitrust.org; For more on HTRC see http://www.hathitrust-research.org/
[3] http://pit.iu.edu/d2i
[4] http://libraries.iub.edu/
[5] http://www.ichass.illinois.edu
[6] https://www.informatics.illinois.edu/icubed/
[7] http://www.ncsa.illinois.edu

GOOGLE DIGITAL HUMANITIES GRANTS OVERVIEW

Google partnered (in their words) with institutions through their Humanities Research Awards to "expose the interconnections of the world's knowledge"[8] and to analyze the vast amount of digitized texts. As of 2011, Google had scanned over 12 million books in more than 400 languages comprising over 5 billion pages and 2 trillion words. For the Digital Humanities Research Awards, Google sought both quantitative research techniques on mass texts as well as qualitative in-depth analysis of texts in their studies. They selected a pool of 12 projects led by 23 researchers at 15 universities. The awards were for one year (with one-year renewal) and provided access to Google tools, technologies, expertise and access to selected Google books scans, text, and derived data. To date no renewal awards have been given to anyone interviewed in our study.

It is our understanding that the selected projects received $50,000 from Google to conduct their research. They then selected digitized books, text queries, or metadata content from Google upon which to conduct research. There were no costs in general to acquire the content since the transfers were all digital. The institutions and project name are listed below with highlighted entries participating in this research. Individual researcher names have been removed from the public version of this report.

1. University of Michigan. *Automatic Identification and Extraction of Structured Linguistic Passages in Texts.*
2. The Open University, University of California-Berkeley, & University of Southampton, United Kingdom. *Google Ancient Places (GAP): Discovering historic geographical entities in the Google Books corpus.*
3. George Mason University. *Reframing the Victorians.*
4. Tufts University. *Classics in Google Books.*
5. Graduate School of Library and Information Science, University of Illinois. *Meeting the Challenge of Language Change in Text Retrieval with Machine Translation Techniques.*
6. University of California-Riverside & Eastern Connecticut State University. *Early Modern Books Metadata in Google Books.*
7. Princeton University. *The Open Encyclopedia of Classical Sites.*
8. University of Oxford. *Bibliotheca Academica Translationum: link to Google Books.*
9. University of California-Los Angeles. *Hypercities Geo-Scribe.*
10. Universidad Complutense de Madrid. *Collaborative Annotation of Digitalized Literary Texts.*
11. University of Virginia. *JUXTA Collation Tool for the Web.*
12. University of California-Los Angeles & University of Washington. *Northern Insights: Tools & Techniques for Automated Literary Analysis, Based on the Scandinavian Corpus in Google Books.*

[8] http://googleblog.blogspot.com/, July 14, 2010

METHODS

This project was conducted using a case study method following an experiential knowing methodology as outlined by Stake[9]. For this method, naturalistic interpretations are made following direct observation and categorical aggregation of patterns within interviews and observations. These interpretations are then collated into a cohesive holistic report.

Dr. John Unsworth, a representative of HTRC, distributed invitations to participate in this study via email to the 22 researchers given Google Digital Humanities Research Awards. At least one follow-up email was sent to individuals who did not respond. Due to the variety of institutions, research foci, locations, etc., we hoped to obtain at least one interview at each institution. One individual did not agree to participate. Additional scheduling and logistical conflicts resulted in a final sample of fifteen different individuals at eight of the twelve award institutions.

Interviews were conducted via telephone, Skype®, or face-to-face, and all were audio recorded. All participants agreed to an IRB permission statement via email (see Appendix 1). A semi-structured interview protocol was developed with input from HTRC to elicit responses from the participants on the primary goals of the project (see Appendix 2). Slight adjustments were made to the protocol subsequent to the first pilot interview. Two researchers conducted the first 3 interviews to establish reliability in the process. Independent interviews were conducted with the remaining 12 participants. Following each interview, a debriefing document was composed to extract features that addressed this study's goals or uncovered new findings. In the future, these debriefing documents could be readdressed following transcription of the interviews and coded for a finer grained analysis. Independent of who was present at an interview, both researchers completed debriefing documents, with a non-present researcher using an audio recording as a reference. These documents highlighted key findings from each interview, which were then aggregated into this report.

Profiles of each project summarizing key aspects relevant to the goals of this report were generated from the debriefing documents. These are presented first to provide context for the findings. The findings are organized in categories that represent primary issues that emerged over the course of the interviews. This report concludes with a summary of findings with particular attention to recommendations for the HTRC and highlighted points of information from the findings.

[9] Stake, R. E. (1995). *The art of case study research*. Thousand Oaks, CA: Sage Publications, Inc.

PROJECT PROFILES

Interview results have been compiled by project and presented below. These profiles include project overviews; data retrieval; characteristics of the data; analyses or techniques employed on the data; data management processes; products if any from the project; and the primary data difficulty cited in interviews. These profiles offer background and context for the findings section that follows and provide insight into project goals regarding identification, access, and processing of materials used by researchers.

AUTOMATIC IDENTIFICATION AND EXTRACTION OF STRUCTURED LINGUISTIC PASSAGES IN TEXTS

Synopsis: This project created a linguistic database for language comparison in context. They used texts with multiple language usage in direct comparison to generate a database through manual and automated joining of the tokenized language instances. The heart of this project was code underwriting for language identification and matching / annotation.

Data Selection: Library of Congress Subject Headings (LCSH) and metadata elements such as GrammarOf were used to identify relevant books, and resulting titles were given to Google for retrieval. Their primary searches were carried out using the University of Michigan library and returned a list of approximately 100 relevant books.

Data Retrieval: Whole books were retrieved including title and URL. More than half of the books were retrieved from the University of Michigan library rather than Google Books, however, because the researchers were not allowed to crawl the Google texts. Only about 12 books were used in the final data set.

Data Processing: Processing was completed one page at a time from the OCR text. They tried running OCR on the images, but were unsuccessful. They ended up using the OCR text with errors rather than parsing the errors out, stating that they were unable to get them out. They selected the foreign text and then the translation (if available) using a web interface to train/test the system they were building. After identifying foreign versus English word tokens, they used a machine translation system to line up foreign and English word tokens to identify language blocks. Reappearing tokens would be compared to each other to help in identification. A translation dictionary was iteratively built to help aid the translation software refine its “guesses” that at first were built on nearby words. They did not substantially work with metadata other than the texts’ title and subject headings, but they did create their own tags to label whether or not an item was in a Roman script.

Data Management: This project was the only one to make use of versioning software, SVN. Files were specifically kept on secure servers with restricted access maintained, though security was not a high concern.

Primary Issues: Higher quality OCR was a priority for this project. The quality of the OCR limited what they could do. Text that was not in Latin or Roman style scripts did not process well, and accents sometimes caused errors.

Secondary Issues: They would like better metadata about text languages, particularly in multi-text documents and on language by sections within text. Automatic language identification functions would be helpful, but human-created metadata is preferred, particularly for documents with low OCR quality.

Additional Requests: They would appreciate a web API to access data.

GOOGLE ANCIENT PLACES (GAP): DISCOVERING HISTORIC GEOGRAPHICAL ENTITIES IN THE GOOGLE BOOKS CORPUS

Synopsis: This project integrated classic works into the Linked Open Data collective Pelagios with the aim of using data and visualization to show how different geographic places appearing within items in web-based collections are referenced within collections. The work is a combination of geoparsing and information visualization. The final data was released to the public under CCO public domain dedication and as N3 triples / RDF triples. This project resulted in a front-end interface for visualization of place identification within texts involving narrative timelines and a map with navigation elements.

Data Selection: Initially, they requested data using Library of Congress Subject Headings (LCSH) related to classics. Google provided them a redacted version of metadata focused on public domain classics in a spreadsheet from which a classicist identified 22-24 out-of-copyright items on which to focus.

Data Retrieval: Data files were retrieved from Google as TAR files containing OCR text and page images from a Google server space that was identified in an email. They had the fastest turn around time for data retrieval of those interviewed (1-2 weeks), perhaps due to the limited number of items and simplicity of the request.

Data Processing: HTML OCR was the primary data used following cleaning to remove tagging with the exception of Google tokenization linking text items to page scans. A geoparser retrieved place names from the texts with a 50% efficiency with the aid of a place names gazetteer, Pleiades+ (Pleiades with the addition of ancient places), when compared against a standard. Specifically, the geoparser analyzed sentence constructs to identify tokens that were proper nouns, including geonames. A human names gazetteer would have been helpful to remove proper people names mistaken for places. Outputs were harvested XML documents, which were placed in a relational database that could be queried against.

Data Management: Though the geoparser generated an output of thousands of XML documents, with one document per page of text and a database record for every token, they did not formally host or manage the data due to the small size of the project. They still recognized that data management would have been needed if the scale had been larger. The files for the project seemed scattered among participants with no thought toward long-term curation, although the final dataset will be deposited in their digital library and the updated Pleiades+ gazetteer is publicly available.

Primary Issues: OCR quality was a primary issue including skew, spacing, poor punctuation, changed characters, and noise. Also, the overall legibility of the scanned image affected tagging. HTML and metadata behind text was generally enough to meet their needs. Additional OCR issues affected interface development. For example, the place names have a “span” tag surrounding them that had to be properly identified to be clickable in the interface. Scan images also differed on how much text was on the page and on the readability of the scanned images, which could affect OCR quality and link identification in their interface.

Secondary Issues: Additional gazetteers or indices, such as one containing proper people names, would have been useful for improving the quality of their data processing. Also, the length of the books and narrative structure influenced development of the front-end web site for visualizing their results.

Additional Requests and Notes: They would like to see best practice information on data analysis, data manipulation, and coding. Also, they suggest HTRC support projects through a catalog of results, methods, or APIs developed. Further, a community of researchers is needed to aid information sharing. Finally, they believe that any data created from these works must also be in the public domain. Finally, they were the only group to specifically note that they received continued funding for the project from Google.

REFRAMING THE VICTORIANS

Synopsis: Project intent is to reinvestigate claims about Victorian Literature and frame-of-mind claims based in that literature. Using comparative text analysis, they wished to see if trends could be found in the Google books corpus during the Victorian era that would coincide with those in the baseline literary criticism standard, *The Victorian Frame of Mind*.

Data Selection: They desired the entire Google corpus in English from 1789-1914 and constructed queries accordingly. Query construction required learning how Google Books data were structured. Discussions were also held with Google on what was possible due to copyright restrictions, and issues such as how much text from each book they could obtain, and the commonality of words they could use in searches.

Data Retrieval: For every search term, such as "Christianity" or "industrious", they retrieved 50 characters on either side of the term, along with title, author, publication year, and Google book ID. Data was provided as a text file, in which every line was a different item. This text file was retrieved from a server space at Google, the location of which they received in an email. The data were retrieved in waves over several months, and were not retrieved in any order, which required them to group and query the snippets for processing. Although only snippets were retrieved, the metadata allowed them to look up the entire text if necessary.

Data Processing: They studied the comparative rise and fall of terms over time, with the overall intent of examining increasing secularization in Victorian culture by quantifying decreasing instances of biblical references over time. They also performed co-localization of terms along with the commonality studies. They referred to their process as having a “conversation with the data”. In terms of visualizations, so far they have only constructed graphs of word usage over time for terms in Excel and then in Google as the dataset got larger.

Data Management: The largest data file was 4-5 GB, but most were 100-200 MB, and there were 1-1.5 million books in the set. They did not consider data security, validity, versioning, or management. The only data management was organization of the files, particularly in Excel; however they did feel that special infrastructure was needed for the files that were being shared locally.

Primary Issues: They needed a better organizational structure for the files that the project generated, meaning that they needed versioning and data management as the project progressed. They had not considered incidental files, data products, or intermediate steps towards visualizations, nor a labeling scheme for their files. They also had to learn how the Google Books data were structured; however they did get a metadata map from Google to help in query construction.

Secondary Issues: Data retrieval was intermittent and took place over a long period of time, which hindered their research at times. Managing data within Amazon Web services involved a learning cycle, but proved manageable once they were proficient.

Additional Requests: They would like to find example projects from which to learn, and a methodology toolbase. They would also like to know how the import and export of data will work within HTRC, and did not have a good sense of the metadata for HTRC.

Other Notes: They made use of Amazon cloud services for queries and storage.

MEETING THE CHALLENGE OF LANGUAGE CHANGE IN TEXT RETRIEVAL WITH MACHINE TRANSLATION TECHNIQUES

Synopsis: This project seeks to automatically create information about topical relationships in texts via statistical text analysis in order to improve the searchability and browsability of a collection.

Data Selection: It took several months for the researcher just to figure out how to get the data. The initial request from Google was that the researcher download specific works, but the researcher wished to use particular call number ranges instead. In the end, an SQL statement was crafted to match desired criteria, such as authors and titles of interest. The researcher was interested in collections over a sufficiently broad time span where language use changes would be present in the corpus.

Data Retrieval: Data were retrieved as digitally compressed TAR files over the Internet. There were no formal limits; however the first request of 500,000 books was not returned. Three kinds of data were retrieved: HTML/OCR versions of scanned text; image files; and XML describing structure.

Data Processing: Only the HTML files stripped of the code were used. It was impossible to strip all of the code, so some noise had to be tolerated. It was desired that the XML file contained changes in language style such as from Middle English to Standard English, as these text relationships are not encoded in a machine readable way. During processing, self-created software repurposed for large bodies of text was used. The software makes use of LEMUR's code for indexing and core search. This code sits on top of LEMUR for statistical modeling. Statistical libraries used included SSJ and COLT.

Data Management: Large amounts of data (3 TB of images and 600 GB of text) were used. These were stored on local servers with a single USB backup drive with no real data management plan or archiving. All access was managed locally. No data security issues were noted and Google was apparently casual about these issues, perhaps due to the out-of-copyright nature of materials. No real version control was implemented; rather, the researcher simply kept abundant and redundant versions of the data.

Primary Issues: The hardest problem was deciding on the scope of the collection to use for the research and for retrieval by Google. Enumeration of the URLs of the books to be worked with required more work than expected. Ideally, the project would have asked for Library of Congress Subject Headings (LCSH) or call number ranges rather than to enumerate collections. By making container decisions, the researcher was putting fingerprints on the output in a way that was not wanted. The researcher, however, benefited substantially from people at Google during the process of acquiring data to make the problem more tractable.

Secondary Issues: The collection of materials lacked additional metadata, available test queries, and relevance judgments, which added work. This work added more person hours than was required for enumerating the data.

Additional Requests: The researcher would like to see the self-created code shared with other researchers via a reusable code base. Although the code is modular, there is no front-end, and the researcher does not wish to call it production quality.

EARLY MODERN BOOKS METADATA IN GOOGLE BOOKS

Synopsis: This project sought to integrate bibliographic records for Google digitized books from selected libraries' collections prior to 1901 into the institution's own library catalog. This process involved identifying digital surrogates of physical copies within the Google book archive, then matching them with existing bibliographic records in their local catalog, updating existing records to reflect Google books resources, or creating new records where an existing matching record did not exist. They believed this integration would improve access to the Google books resource.

Data Selection: They desired bibliographic records for Google Books published prior to 1901. The files were mainly MARC-like fields and sub-fields (although non-MARC format) that were imported when in-scope. They would like the availability of more explicit metadata and its processing history within the HTRC records if possible.

Data Retrieval: The bibliographic data were compiled by a Google employee as a CSV file and placed in an online drop box for retrieval. Although processing work began in June 2010, they did not receive any data until March 2011, and at the time of our interview, had received no recent return emails. Part of the delay was believed to be due to the type of data being retrieved, querying of the data, and Google workforce availability issues according to those interviewed. They currently have 5-6 files including records from Madrid, Oxford, and Munich.

Data Processing: They imported the records in order to match an existing record, generate a new record or holding, or to update an existing holding. Importing records required a machine-matching process that was 50% efficient using existing workflows, which is apparently normal for MARC records, followed by manual manipulation using an online tool. A site was built in Drupal to allow users to manually match records. Only 3% of total records obtained were in-scope, or about 2,000 items. For the University of Madrid, 73,000 records were out-of-scope for comparison. During analysis, they performed quick in-scope passes to parse the data on MARC fields. Types of analyses included comparison of author, title beginnings, place of publication, and date of publication. They reported the MARC-based records were very inconsistent, and that working with the records in ways that they were not necessarily intended to be used was difficult, such as managing data across collections. While they do not believe MARC was always the best format, they do not want to lose those fields and do not see Dublin Core as any better. Most of the bibliographic data were discarded following import except URL/URI.

Data Management: No discernible data management was performed beyond standard backups until records were imported. CSV files were loaded into a database on a single server, with each originating library's data managed as a single database file with no archiving. Files typically contained 48MB of text for a given library. Once data had been matched, it was entered into their main system and managed by others. The end result was a bibliographic record (either a new record or a modification to an existing record), which was added to their bibliographic OPAC. Data security was not a concern.

Primary Issues: The primary issue was retrieving the bibliographic records in usable form, unparsed by Google. This process took 10 months to design the queries and get the data.

Secondary Issues: They saw a need for easy data retrieval through a well-designed API. They also saw the need for a project manager to keep the project moving forward.

Additional Requests: There was a need for a method to help manage or process records across collections or a complimentary system of repositories. They would also like to see increased communication within the research community.

THE OPEN ENCYCLOPEDIA OF CLASSICAL SITES

Synopsis: This project dealt with large-scale methods for machine learning from text, in order to provide a view of Google books from the perspective of ancient archeology. They are interested in what patterns are found and interpretable within a text that will be of interest to a user. Specifically, they mapped places found in the texts to create a public website to browse the materials for classical site representation.

Data Selection: Books to include in the collection were found via a hand-generated index of book titles reviewed by the American Journal of Archaeology in the last 100 years. A specific list of query words and relevant books to search through was generated.

Data Retrieval: They used metadata from the above list of books to query Google to retrieve volume IDs, which were then entered into a search query containing a classical site search term to retrieve data items. Since they were working with copyrighted materials, "snippets" of text were retrieved rather than entire books. From the 50,000+ titles requested, they were able to retrieve approximately 27,000 volumes with some metadata, including title, author, publisher, issue, date of record, and location. Each item in the retrieval file included 120 characters before and after the search term. Approximately 2.7 GB of compressed text were returned. Google provided an internal tool for retrieving tab-delimited spreadsheets of the data. While Google did not impose a limit on the number of responses, it took from weeks to months to get data to the researchers.

Data Processing: Preprocessing was required to clarify the text due to issues with OCR quality, word splitting, etc. They did not have images or full text to help in OCR error-checking. Language identification was also required, especially when the text had mixed language usage. Place names and nouns were then identified in the texts and tokenized. The characteristics of the data, and limitations of the small amount of surrounding text made geographic disambiguation difficult in the statistical topic modeling, but it was still possible. The final dataset was then visualized through a web interface.

Data Management: As with other projects, very little data management was performed. The data were retained on a central server with little additional security, archiving, or version control other than post-processing iterations. Their public website likely has backups, but that was not discussed.

Primary Issues: The hardest problem was memory allocation during processing of large datasets. This was a problem with the scalability and efficiency of the method. Making sure that everything was running efficiently -- ARCOS, tokenization, running algorithms -- posed a problem. Also, the technique used was developed for 5-10,000 documents and collections of articles, rather than the longer books and larger numbers of items in this project. Their algorithms did not take into account the internal structure of the texts, resulting in the potential loss of the relationships between aspects of different parts of the texts, which has implications for how books are delivered to end users.

Secondary Issues: Another problem was OCR quality affecting algorithm function. Disambiguation of place names was a difficult problem, particularly with short snippets. They also mentioned disambiguation issues from a conceptual standpoint, not just technical.

Additional Requests: They suggested an API to allow people to directly query data on common algorithms.

Other Notes: More sophisticated work in digital humanities often requires close collaborations between computer scientists and humanists. This group was the only one working with copyrighted works. They were able to manage this nuance by using excerpts surrounding their search terms. They desire continued access to copyrighted materials. This group also has a re-usable research output in a public website listing keywords in topic and distribution over time, and critical books that relate to particular geographic regions.

JUXTA COLLATION TOOL FOR THE WEB

Synopsis: The project's general intent is to collate and compare different books in the Google books corpus, but they were still in the development phase.

Data Selection: No data had been selected yet from the Google books corpus.

Data Retrieval: They were still in the tool development phase and had not gone beyond 5 text comparisons. They had not retrieved any data from Google.

Data Processing: Length of texts, number of texts, and amount of markup was a limiting factor for their algorithm, which currently can work with text files tagged with TEI and various XML sources. Visualizations include a single text with changes, side-by-side comparisons of two texts, and histograms of a document visualizing areas of change.

Data Management: No data had been selected yet. There was no indication of data management being conducted on their prototyping.

Primary Issues: The current primary issue is the size and number of the documents, and the processing ability of the software not allowing multiple collations simultaneously. Files over 2 MB crash the program.

Secondary Issues: They would like to be able to identify similarities as well as differences, but this requires different algorithms/methods that have not been worked out.

Additional Requests: They would be interested in developing tools in cooperation with HTRC. Also, community feedback is important, as is user-centered design. These comments were provided as overarching principles that could be integrated into the HTRC report.

NORTHERN INSIGHTS: TOOLS & TECHNIQUES FOR AUTOMATED LITERARY ANALYSIS, BASED ON THE SCANDINAVIAN CORPUS IN GOOGLE BOOKS

Synopsis: Northern Insights sought to uncover how a large corpus could be used to determine the influence of one set of texts on another literary domain. They converted books into statistical patterns and then used algorithms to find these patterns in other books, focusing on Nordic books prior to 1923. They produced a visualizer of 1,600 Danish books using a publication slider reduced by a search box.

Data Selection: Since they were working with specific (Norwegian, Danish, and Swedish) language texts, they first had to reverse-engineer the Google interface to return books in a specific language for a given timespan. They built their own Google Books interface, which restricted by time period, by a given publisher, and/or by last name of known origin. They found that existing subject metadata alone was an insufficient finding aid for their purpose, and that language differences were not properly delineated, thus requiring the more complete query. In cases of overlap, they would perform manual fitting. It was an iterative process to obtain their final dataset. Of 36,000 books, only 1,600 were retrieved.

Data Retrieval: Books were retrieved from a Google server and downloaded sequentially. Page images were retrieved as either JPEG2000 or PNG files, which they discarded, and hOCR. Approximately 60 GB of text were retrieved. They received Dublin Core-like metadata from a Google books interface to add to these files. Unlike many of the groups, one participant reported that they were able to get their documents quickly; however, another reported this process was delayed (but they may have included their selection time in the total). The bottleneck in delivery was attributed to staff availability and workload at Google. As a side-note, this group mentioned during interviews that a group at Google is looking into automatically encoding books into TEI.

Data Processing: They stripped the positional and additional markup from the OCR files and looked only at the text. The text was chunked by page or paragraph into UTF-8 encoded text files with sequence information retained. Then they derived a topic model for Danish folklore. The next step was to determine where the topic model had the greatest saturation across the text corpus. To process the text, an implementation of LDA called MALLET (Machine Learning for Language Toolkit) was used to run probabilistic topic modeling algorithms. The MALLET files were visualized using another tool that allowed the data to represent itself based on statistical patterns of co-occurrence to elicit topic models from a given corpus. Each term in a topic had a weight and a count.

Data Management: The parsed data was managed in an SQL database. They did not think HTRC could run databases for everyone, since each database would be on the scale of those used in astrophysics. Their data management appeared to be minimal. (For example, they did not seem to know where data were stored, other than that the text data elements were in the database.)

Primary Issues: A primary issue was coming to terms with the advanced mathematics needed for this research. While these projects are still within the domain of literary criticism, the methods used in them extend to data mining and mathematics, which can be foreign to humanists. The other primary issue was overloading the system when working with such large amounts of data. Although opinions were split, some saw the time it took to get the data as a primary issue, as mentioned above.

Secondary Issues: They had to perform a lot of data normalization, because the printing processes were not standardized in the texts they were working with, which led to errors in the OCR. They also did not see the quality of the OCR being exposed within the Google metadata. Due to the OCR quality issues, they had to eliminate some books.

Additional Issues: Books' representations change over different print editions, and the chunking algorithm can therefore create different chunks for the same text over different editions. Having multiple copies of the same text in the corpus can therefore affect the topic model, and a way to distinguish these texts is needed.

Additional Requests: Latent semantic analysis as a tool for selecting text, such as a topic model saturation query, would be helpful for their purposes; in a way, topic modeling could lead to material selection. Also, although they did not use the metadata in their processing, having more metadata viewable in the search would aid in their source material discovery and retrieval.

FINDINGS

The findings are organized in categories that represent primary issues that emerged over the course of the interviews. These categories are: Identification and Retrieval of Materials; Characteristics of Content; Collaboration and Community; Policy; and Data Management. The first two categories directly address goals of this research, while the subsequent categories are grounded in the results and address concerns reported by the participants in relation to the particular projects they were undertaking.

It is important to note that the resources provided by Google were not developed for research purposes, but rather for more general public access. In their words, "Our ultimate goal is to work with publishers and libraries to create a comprehensive, searchable, virtual card catalog of all books in all languages that helps users discover new books and publishers discover new readers."[10] This analysis, however, reports on the perspectives and experiences of researchers using Google resources for computational analysis of text, bibliographic record integration, machine learning, and other highly technical methods.

IDENTIFICATION AND RETRIEVAL OF MATERIALS

BIBLIOGRAPHIC RECORDS AND METADATA

Clear understanding of data structures allows researchers to design their queries and research methods accordingly. Researchers were hindered by the opacity of metadata use by Google Books. Google did not explicitly state the format, standards, or contents of its descriptive or structural metadata, requiring researchers to spend time figuring it out for themselves. While it seems that Google has access to the original MARC records for their books, their licensing agreement with OCLC prohibits them from distributing the complete, unaltered records. However, the metadata they could share was not well communicated to the users. One group received Dublin Core-like metadata from Google, but little semantic information to clarify it. Another group wanted not just metadata but also the ability to process the metadata. It will be vitally important that the HTRC clearly and coherently communicate

[10] http://books.google.com/googlebooks/library.html

what researchers can and cannot get from the collection, particularly for researchers wishing to obtain bibliographic metadata. Delivery of existing metadata needs to be consistent, especially with complex standards like MARC. If metadata are being substantially transformed or transferred from one format to another then that transformation or crosswalking needs to be explicit. It is imperative that the HTRC make its metadata standards well known and distinguish between implicit and explicit metadata. Clear understanding of data structures will allow researchers to design their queries and research methods accordingly.

Several kinds of metadata were identified that would have been useful, based on the needs of the researcher. Metadata describing a text's language (or languages, in some cases) and metadata describing the script or font of the text would have greatly helped several researchers. Scans of texts that may have differed substantially at the "item" level – for instance, books from the hand-press era – would have benefitted from additional item-level description. In older texts it is important to remember that they need to be treated as individual items rather than homogenous print runs. Detailed language metadata needs to include time frame, language, location, and additional identifiers to completely establish identity. Researchers in one group had to reverse-engineer the Google interface to get books from a given language within a given time span. Another group additionally needed clear delineation of how multiple copies of the same book in different editions/manifestations would be described to control for effects on algorithms and resultant topic models.

Finally, multiple researchers relied on the Google-assigned identifiers for their own reference, and for projects that involved linking bibliographic records back to the Google Book. Persistent, resolvable IDs will clearly be necessary for each title in the HathiTrust's collection.

API/DATA DELIVERY CONCERNS

Researchers described the search, appraisal and retrieval process as "having a conversation with the data" and as involving iteration "between a close and distanced reading." Research is not done with a single download of data, but rather with iterative downloads and progressively tuned queries. Researchers do not necessarily need huge sets of data to do interesting work, but the implication is that they do need flexible data delivery services that can deliver different kinds of data, in different formats, based on different searches and for different kinds of research over the course of a project.

For those participating in the Google Digital Humanities Grants, data were not retrieved through a Google Application Programming Interface (API), but rather following compilation by a Google contact and then made available on a Google server for download. While HTRC already provides access to its collection with an API, very few of the researchers were aware of the nature of this interface, having not had experience with it.

One researcher specializing in front-end and visualization development recommended that "APIs should be built with specific data applications and visualizations in mind." A few participants wanted to be able to construct their own queries within the API. This would provide them with the necessary flexibility that a proprietary API might not allow. One participant also wondered if researchers would be able to directly query the data on common algorithms without first downloading a dataset.

HTRC should be prepared to alter its existing API for new use, or to build a new one if warranted, responding to the diverse groups of researchers (digital humanists, computer scientists, information scientists, etc.) who seek to access the materials for potentially very different research projects. Researchers did not necessarily need huge sets of data to do their research, but rather services for flexible data delivery that provide different kinds of data in different formats based on different searches for various purposes over the course of a project. Making existing APIs more visible is important as well. Beyond API development, HTRC could facilitate the processes necessary for researchers to construct workflows to get the data they need to run queries.

Text and data delivery options need to be as flexible as possible. Researchers worked with a wide range of text formats – everything from large numbers of n-grams, to "textlets" or "snippets" of text surrounding search terms, to full, marked-up text. Researchers also worked with different sizes of datasets; some worked with just a few texts, and small amounts of data, whereas others downloaded several thousand texts or textlets, amounting to 4-5 gigabytes of data. Some found it helpful to not have large datasets delivered to them all at one time, as some projects did their querying in a somewhat iterative fashion. This kind of back-and-forth querying of the dataset is discussed further in the "Research Methods" section. In addition, some software cannot work with particularly large files (Juxta, for instance), and many algorithms take prohibitively large amounts of memory to run over large datasets or files. It may therefore be necessary to support modularized file downloads or hosting.

Several projects experienced delays in progress due to delays in acquiring data. In some cases it was cited as due to personnel or priority issues at Google, but query construction, lack of a functional API, and the iterative process of discovery were also noted. Delays were not a problem when specific items were requested or when running simple queries.

CHARACTERISTICS OF CONTENT

OPTICAL CHARACTER RECOGNITION (OCR)

Numerous researchers reported issues with the quality of Google Books OCR. It is useful to bear in mind here, as noted earlier, that the resources provided by Google were not developed for research purposes, but rather for more general public use. This primary audience is likely not significantly affected if one out of every 20 words (as one participant noted could be the error rate in some foreign text) has an error, for the existing OCR quality is fine for reading. However, Hathi's audience is affected if one out of every 20 words has an error, and steps should be taken to improve OCR quality if and when possible. This correction could include allowing researchers to upload corrected texts, making use of computational methods that will correct the text based on n-gram frequency analysis, crowd-sourcing, or using new OCR engines to re-scan existing images. Scalability of scanned image viewing is also a helpful online tool for OCR correction and reference. It was unclear based on interview responses if Google metadata exposes the quality of the OCR within the text or if such metadata tagging could become standardized.

Issues with OCR were most pronounced with non-Roman texts, and in texts printed in unusual, archaic, or non-uniform typefaces, such as with Fraktur or texts from the hand-press era. Errors were common in mixed language texts and variable font passages as well as with hyphenation, accents, and word joining. OCR quality problems were noted that affected readability near margins of texts in particular.

In some cases, there is specialty OCR software that could be used with these texts. For instance, ABBYY produces software that can read texts printed in Fraktur, but Google does not appear to use ABBYY. The researcher working with these texts noted that it is quite expensive. If OCR quality is not addressed in some way, the HTRC may find that many of its texts will simply not be fit for use by their intended audience without preprocessing or acceptable error ratio allowances.

ADDITIONAL LANGUAGE ISSUES

The language metadata of the Google Books materials was insufficient for some researchers' purposes. There were two primary reasons: 1) the original materials were miscataloged by the contributing library (Danish misidentified as another Scandinavian language, for instance) and 2) the texts in question were primarily in a foreign language, but included a modern English preface, front matter, or glossary. In the latter case, the modern English text became problematic when the texts were being used as a whole in computational analyses; the undesired modern English text clouded some of the initial algorithms and needed to be compensated for. Several researchers developed their own ways of computationally differentiating one language from another. At least one relied on an algorithm that identified text based on the indefinite participles and pronouns. The HTRC will want to be aware of these algorithms for identifying languages when necessary and will want to build in services to support them or perhaps make use of them in their own databases.

COLLABORATION AND COMMUNITY

KNOWING THE COMMUNITY

Digital humanities research teams are highly diverse, and the problems they address require interdisciplinary collaboration and cooperation among humanists, computer scientists, information scientists and librarians. The group of interview participants was a good representation of the community, since it included literary scholars, archaeologists, linguists, information scientists, and historians. Because of the different disciplines within the user base, it will be important for the HTRC to be explicit about its contents, metadata standards, and any underlying algorithms it uses to make data available.

COMMUNICATION AND CONNECTIONS WITHIN THE COMMUNITY

Multiple researchers stated that they wished they had a way to communicate with others working in their research area. While there was awareness of other Google Digital Humanities Projects that were doing similar work, there seems to have been limited communication among these groups. Many of the researchers working on these projects were not trained as computer scientists, but rather as scholars in the humanities. There was interest in ways to identify methods, tools, and algorithms that researchers could make use of in their projects. Several researchers needed experts to consult on advanced mathematics needed in their work. The teams could benefit from a community message board and a code base, and other structures for communication and sharing of resources and expertise.

Connections with other systems were also discussed. One researcher spoke about the need to establish a "complementary system of repositories" and the need for Google Books and the HTRC to be able to interact with each other, and with other services including cloud-based services, and applications, such as Juxta and MALLET.

RESEARCH METHODS AND TOOLS

There was strong support for the sharing of algorithms and other complex computational analyses and tools, implying the need for a centralized place for researchers to share products, processes, and best practices, and to exchange information on how they are used and why. Many tools have their own file formats, workflows, and steering committees that researchers need to be aware of. There was seen to be a need for places where researchers could share these tools, questions, and successes; and for ways for the community to make explicit the function and implementation of these tools and methods, even so far as best practices toolkits. Furthermore, there are needs within the community for new developments in models and methods for determining patterns and structures in large volumes of text, and for analyzing relationships between different text parts and internal structures.

Many of the researchers described their work as still being "exploratory"; for instance a number of the projects focused on testing tools and algorithms to see how effective different methods were in answering different kinds of questions. The exploratory nature of much of the work underscores the need to provide researchers a way to explain and tweak underlying mechanics in algorithms, data visualization tools, and other computational resources. Researchers need to be able to change their methodology – and therefore, their tools and queries – on the fly, and with a minimal amount of additional help.

The HTRC should consult with a diverse group of researchers in the course of its development activities for everything from data standards to APIs to visualization tools. One researcher in particular, a front-end developer, stressed that data standards and their development also affect user interface development just as they affect research methodology and search and appraisal. It is worth noting that there were people in this group of interviewees that stated they would be interested in collaborating directly with the HTRC's development team, especially in the area of tool development, tool integration, and visualization development. It is important that resources developed by the HTRC work with established tools and user groups (MALLET and Juxta both came up); thus a more exhaustive look at these tools and groups would be beneficial.

POLICY

COPYRIGHT AND DISSEMINATION

Most of the researchers worked only with out-of-copyright materials, and therefore did not anticipate problems in disseminating the results of their research. They felt that data created from public domain documents should remain in the public domain. Those that did work with in-copyright materials either worked with smaller “snippets” of text that Google felt “comfortable” disseminating (e.g. a search term surrounded by 140 characters or so on either side), or worked with n-gram data. Those that worked with in-copyright materials felt that because they were generating “facts” about the materials, they should be able to disseminate the results of their research without a problem; however they were aware that this position may be a debatable one. Most of the participants would welcome the availability of copyrighted works with acceptable restrictions.

At least one researcher stated that it was imperative that he have access to in-copyright materials in some way, and stated that computational work that relies on large datasets will be hindered if researchers are not given at least some access to in-copyright materials. Providing researchers with n-gram data – even of in-copyright materials – may be one way to provide them with the broad dataset they require while staying compliant with copyright laws.

METADATA

As explained in the metadata section, some users were hindered by Google’s legal obligation to keep MARC records confidential. If the HTRC will encounter similar barriers to sharing, then rights of access and dissemination should be well and clearly articulated, and workarounds may be needed to provide researchers with bibliographic metadata.

DATA MANAGEMENT

DATA CURATION PRACTICE

Data curation was clearly lacking in all of the projects interviewed. Few, if any, researchers had any formal versioning, archiving, or data curation practices in place during the duration of their projects, and several stated that their datasets became close to unmanageable by the end of their work. Those working with numerous “textlets” and experimental topic models or algorithms seemed to have a particularly difficult time tracking different iterations of their research. Another did not consider versioning issues until creating visualizations resulted in a need for better data organization. None of the groups performed copy protection beyond local security of their workstations. Several researchers did have access to repositories or nightly back-up systems, but those were largely a result of their working within an academic computing environment; that is, all of their school's computers were backed up regularly, and therefore, their data were as well.

There is a clear need for data curation with this group of researchers. Quick options to begin to address this need might include integrating the University of Illinois at Urbana-Champaign's Guide to Data Curation in the Digital Humanities (recently released in beta version at http://guide.dhcuration.org/) or other best practice guides as a first step toward including best practices in data curation with the HTRC. Support for digital workbenches or workspaces would also be helpful, as would a way to track downloads, queries, and analyses. Researchers need a way to manage and process records across collections and to have all data easily available and searchable.

SERVICE PROVISION

One of the primary questions researchers had for the HTRC was "how do we get stuff out? And to where?" Several researchers already make use of cloud services for storage of not only working data, but also for their tools and programs. The HTRC may want to offer similar services as well.

DATA VISUALIZATION

Although not a major concern, several groups were reaching a significant level of data delivery and were considering how to make use of visualizations. HTRC should consider ways to implement visualization tools within its search capabilities (e.g. via word clouds as recommended by one participant). Finally, tools developed by users (as mentioned above under the Research Methods and Tools section) should be shared and could possibly be implemented into new data visualizations within the site.

SUMMARY OF FINDINGS & RECOMMENDATIONS

Based upon interviews of Google Digital Humanities Grants, the following recommendations and points of information are forwarded to the HathiTrust Research Center:

- Identification and Retrieval of Materials
    - Bibliographic Records and Metadata
        - Because of the interdisciplinarity of the research, and the diversity of the user base, it will be important for the HTRC to be explicit and clear about its contents, its metadata standards, its method of marking up text (if any), and any underlying algorithms it uses to make data available.
        - Researchers desired a consistent delivery of metadata with explicit statement of the format, standards, or contents of the descriptive or structural metadata.
        - The metadata needs varied based on the needs of the researcher (see specifics outlined in the Findings section); however persistent, resolvable IDs will clearly be necessary for each title in Hathi's collection.
        - Researchers desired an understanding of how multiple copies of the same book in different editions/manifestations would be managed to control for effects on algorithms and resultant topic models.
    - API/Data Delivery Concerns
        - Researchers do not necessarily need huge sets of data to do interesting work, but the implication is that they do need flexible data delivery services that can deliver different kinds of data in different formats based on different searches for different kinds of research at different times.
        - Participants want to construct their own queries within the API that directly query the data if possible.
        - Text and data delivery options need to be as flexible as possible. Researchers worked with a wide range of data formats and sizes of datasets.
        - Due to constraints of researcher software, it may be necessary to support modularized file downloads or hosting.
        - All forms of delivery options need to be at a speed conducive to researchers' needs so as to not delay their progress.
- Characteristics of Content
    - Optical Character Recognition
        - Steps should be taken to improve OCR quality if and when possible.
        - Scalability of scanned image viewing is necessary for OCR reference and correction.
        - Metadata should expose the quality of OCR.
    - Additional Language Issues - The HTRC will want to be aware that the language metadata is sometimes insufficient for researchers and to build in services to support researcher use of algorithms that identify text language when necessary.

- Collaboration and Community
    - Knowing the Community - Digital humanities research teams are incredibly diverse; they require interdisciplinary collaboration and cooperation between humanists, computer scientists, information scientists, librarians, etc.
    - Communication within the Community - Communication was clearly lacking within this research community, and the HTRC could serve as a conduit to support future communication pathways.
    - Complementary System of Repositories - Google Books and the HTRC need to be able to speak to each other, and to other cloud-based services.
    - Research Methods and Tools
        - The HTRC should consult with a diverse group of researchers when developing everything from data standards to APIs to visualization tools.
        - Places are needed where researchers can share their tools, questions, and successes; and ways are needed for the community to make explicit the function and implementation of their tools and methods, even so far as best practices toolkits.
    - Collaborations with HathiTrust Research Center - There are researchers interested in collaborating directly with HTRC, especially in the area of tool development, tool integration, and visualizations.
- Policy
    - Copyright and Dissemination
        - Most of the participants would welcome the availability of copyrighted works with acceptable restrictions.
        - Researchers feel that data created from public domain documents should remain in the public domain.
    - Metadata - Some researchers desire direct access to metadata, and the HTRC may encounter barriers to sharing and rights of access to these records.
- Data Management
    - Data Curation Practice
        - Data curation was clearly lacking in all of the projects interviewed.
        - The HTRC need a way to manage and process records across collections, links to curation best practice, and support for curation services.
    - Service Provision - HTRC may want to offer cloud services for storage of data and tools.
    - Data Visualization - HTRC should consider ways to implement visualization tools within its search, such as word clouds. User-developed tools for data visualization could be implemented into the HTRC site.

APPENDIX 1: INTERVIEW PERMISSION FORM

[The interview permission form is a little different because of the audio recording. There is no over 18 clause since that is assumed. There was initially no intent to report the findings beyond the departments involved and the funding agency, but an additional dissemination clause was included as an insured safeguard.]

Dear [Participant's Name]

Thank you for your interest in participating in this research being conducted by Dr. John Unsworth of the Graduate School of Library and Information Sciences at the University of Illinois. By participating in this research, you are helping us to provide better input into the architecture of the HathiTrust Research Center and improve the collections use to both scholars and the public.

Participation will involve a telephone interview of approximately forty-five minutes in length. We can arrange these interviews at your convenience, and we will incur the cost of the call. During these interviews, you will be asked approximately 16 [number may be reduced] questions concerning your experience as a recipient of the Google Digital Humanities Research Award. All answers will remain confidential. There are no risks involved in participating in this research other than those involved in ordinary everyday life.

Print and retain a copy of this consent form for your records. If you have any questions / comments about this study or are interested in the results, please direct your inquiry to Dr. John Unsworth (217.333.3281 or unsworth@illinois.edu). The results of this study will be delivered in departmental reports and to funding agencies and may be reported in papers, scholarly journals, or research conferences. If you have any questions about your rights as a participant in this study, please contact the University of Illinois' Institutional Review Board at 217-333-2670 (collect calls accepted if you identify yourself as a research participant) or via email at irb@illinois.edu.

By responding to this email, you verify that you have read and understood the above consent form and voluntarily agree to participate in this study. You also are agreeing to be audio-taped during your phone interviews. However, at any time during the interviews, you retain the option to end your participation, at which time any recordings of you will be erased.

Sincerely,

Dr. John Unsworth

Dean and Professor
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel St., Champaign, IL 61820

APPENDIX 2: INTERVIEW GUIDE

Hello, [insert name here]. Thank you again for taking the time to participate in this research. I wanted to begin by once again verifying your permission to record this interview. You are free to (a) discontinue participation in the study at any time, (b) request that the audio recorder be turned off at any time, and (c) pass on any question you do not want to answer. Do you consent to this interview?
[Brief greeting including additional verbal consent for voice recording. This consent is required by IRB when recording.]
[This protocol assumes that the research group has a basic understanding of the humanities research center for which the interviewee is a member. If not, an additional question should be added for formality demographics: **Q 0. Demographics.** Could you please describe your research center / institution / agency / project?]

1. **Data Lifecycle:** [Bold items are the areas and indicate why the questions are being asked. An introductory sentence can be added in interviewers own words. Really, any can be reworded as necessary based on the flow of the interview.]
    a. How were data or materials <u>identified</u> for inclusion in your research?

    b. What type of data was acquired?

    c. How were data or materials <u>retrieved</u> by your institution from Google?

    d. Were there limits on the amount of material(s) retrievable?

    e. How were the data or materials <u>used</u> by your project?

    f. In what ways did the <u>characteristics of the data</u> or materials affect their acquisition or use?

    g. What types of <u>analyses</u> were performed on the data or materials that were obtained from Google?

        i. What <u>techniques or tools</u> were employed in your work with these data or materials?

        ii. What <u>transformations</u> were required prior to or during analysis?

    h. Are there potentially re-usable research products that resulted from your project?

2. **Data Management:**

    a. How were data managed once they were obtained?

    b. How much data was managed?

    c. Where did you typically keep your data in terms of working copies and archival purposes?

    d. How was data access managed? Who had access to the data?
    [Who manages and how is that access managed? We are just referring to data access, not machine access. Issues include hacking security, free versus controlled, and vandalism protection. These lead into next question and may cover it.]

    e. How important was data security? [data acquired and data generated]
        i. What are the primary concerns and/or measures taken?
        ii. What more would you like to do?

    f. How did you manage different versions of your data?

3. **Data Difficulties:**

    a. What was the hardest problem that you face in terms of acquiring, developing, keeping, or using the data over time?

    b. Why was this difficulty the hardest problem that you faced?

    c. Were there other problems of similar difficulty?

4. **Generalizability:**

    a. Do you view your project as typical among the Google Digital Humanities Research Awards?

5. **Added Comments:** [Independent of how far you get above, ask the following question.] Thank you very much for your time. Before we end, are there any questions that you would like to ask or final comments that you would like to add.