Call for Proposal to Develop a HathiTrust Research Center

Submitted to the HathiTrust Executive Committee by

Kat Hagedorn
Jeremy York
Melissa Levine

Members of the Research Center working group

Steve Abney, University of Michigan
Jack Bernard, University of Michigan
Geoffrey Fox, Indiana University
David Goldberg, University of California Irvine
Robert McDonald, Indiana University
Qiaozhu Mei, University of Michigan
Shawn Newsam, University of California Merced
John Ober, California Digital Library
Beth Plale, Indiana University
Scott Poole, University of Illinois Urbana-Champaign
Sarah Shreeves, University of Illinois Urbana-Champaign
John Unsworth, University of Illinois Urbana-Champaign
John Wilkin, University of Michigan

December 7, 2009

Introduction: Computational Research and HathiTrust

The founding institutions of HathiTrust undertook the effort of building a repository of published content with the expectation that this content, in addition to serving needs of traditional reading and research, would serve as an extraordinary foundation for many forms of computing-intensive research, particularly in the areas of language and literature. As the amount of content in HathiTrust continues to grow, the value of exposing this content to computer-mediated methods of research grows as well. The pending Settlement between U.S. plaintiffs and Google includes terms that would explicitly sanction the use of millions of in-copyright works owned by HathiTrust institutions in "non-consumptive" computational research. The terms also provide for the establishment of up to two Research Centers that would enable this research across the entire body of Google-scanned content (currently numbering more than 10 million volumes).[1]

The time is right for HathiTrust partners to develop a long-term strategy to meet the increasingly diverse, computer-oriented, and processing-intensive needs of today's researchers. HathiTrust has a particular incentive, and opportunity, in taking action in this area. The process for determining which institutions are awarded Research Centers has not yet been determined. However, because of the large amount of Google-scanned content contributed by partner institutions, the members of HathiTrust can ensure that one of the Centers will be awarded to a HathiTrust institution. By positioning ourselves to take on the responsibility of managing a Settlement-enabled Research Center, in addition to supporting research capabilities across the HathiTrust Corpus[2] itself, the partners can ensure the vitality of HathiTrust as a data provider in this new research environment and further our efforts to ensure, through cooperative means, the efficient management, preservation and accessibility of the scholarly record over time.

In August 2009, a working group was assembled by the HathiTrust Strategic Advisory Board to explore the needs and requirements of hosting such a Research Center. Through a series of conference calls in August and September the working group members examined in detail issues surrounding infrastructure, legality, funding, and sustainability of a HathiTrust Research Center.

The findings of the working group are laid out in this document, structured around the following major issues:

A. Research and Demand
B. Structure of the Research Center
C. Research Results

[1] Representatives from Google have articulated the term "non-consumptive research" to describe the type of activity that the Settlement will allow on these materials: supporting analysis of a form that does not require (and does not *permit*) reading access to in-copyright materials.

[2] Definitions of terms are available in *Appendix A*. Accompanying documents can be found in *Appendix B*.

D. Legal and Security Concerns
E. Sustainability

These findings provide a base understanding of the needs and requirements for a Research Center and lay the foundation for a more detailed plan of how the Center will be accommodated at a HathiTrust institution. However, such a plan is still forthcoming.

Call for Proposal

The HathiTrust Executive Committee is calling for full proposals from interested HathiTrust institutions for the design, development and ongoing support of a HathiTrust Research Center. The Research Center should be designed to accommodate computational research of multiple varieties on two different bodies of materials:

1. The Research Corpus, which will be composed of public domain and in-copyright works digitized by Google. The in-copyright works are defined as Protected under the terms of the Settlement and carry restrictions over and above those imposed by copyright law (*Appendix A* contains a full list of definitions used in this document). The Research Corpus will only be created if the Settlement agreement is approved.

2. The HathiTrust Corpus, which is composed of Public Domain, Google Public Domain, Open Access, and In-copyright Data.[3] The HathiTrust Corpus will exist regardless of what happens with the Settlement. It should be noted that because of the large amount of content Google has scanned from HathiTrust institutions, a significant percentage (perhaps more than 80%) of the material in the HathiTrust Corpus would also be represented in the Research Corpus if the Settlement were approved. Under these circumstances, much of the HathiTrust In-copyright Data would be subject to the terms of the Settlement as Protected Data.[4]

Using the ideas and information in this document as a starting point, the proposal should address the following key areas:

1. Research and Demand

    a. All of the areas below should be tied to actual research uses, examples of which are given in Section A of the current document, with a view to accommodating those uses at scale

[3] HathiTrust will undoubtedly make some limited use of the in-copyright materials in support of research (e.g., by exposing SOLR indexes). Proposals need not include provisions for use of In-copyright content, but are encouraged to address this possibility. A possible scenario is for Research Center proposals to define certain conditions under which in-copyright materials would be used to support research and would include Research Corpus and HathiTrust Corpus in-copyright volumes as they met those conditions.

[4] Some portion of in-copyright materials would not be subject to the terms of the Settlement because they were not scanned by Google or did not meet conditions of the Settlement. Provision should be considered for In-copyright materials that exist regardless of whether the Settlement is approved.

2. Structure of the Research Center

    a. Technical specifications for infrastructure (i.e., machines, networking, bandwidth, system tools) supporting:
        i. Access to, and use and management of, HathiTrust and Research Center and Data
        ii. Tools for use by researchers inside the Centers

3. Research Results

    a. Access to, and use and management of, derived results

4. Legal and Security Concerns

    a. Security Standard (with significantly advanced plans for most elements)
    b. Audit requirements

5. Sustainability

    a. Governance and funding models and approaches to sustainability (including budget and source(s) of funding)

As a formal initiative of HathiTrust, the Center will report to the HathiTrust Executive Committee. However, the Research Center will require its own structures for funding, and the management of day-to-day operations. Proposals are requested that address these administrative needs in the context of HathiTrust governance and in relation to the Executive Committee in particular, as well as the architectural and functional needs of the Research Center.

The Executive Committee recognizes the potential for increased capability and functionality of the Research Center as a result of collaboration between HathiTrust institutions and encourages this for the development of a viable infrastructure. Proposals are welcomed from individual institutions or from institutions working cooperatively. However, Research Center proposals that are fully compliant with Settlement requirements must name a single institution as the primary Research Center Host.

Given the challenges the working group has identified surrounding the use of in-copyright materials, and particularly materials covered by the terms of the proposed Settlement, a viable solution could be to build separate infrastructures for the use of 1) Sensitive Data (In-copyright Data in the context of the HathiTrust

Corpus or Protected Data in the context of the Research Corpus)[5], and 2) Public Domain, Google Public Domain Data, and Open Access Data in the HathiTrust Corpus.

We encourage proposals that support the use of all available research materials, but recognize the difficulties involved in supporting a Settlement-compliant research environment for distributed, multi-user, and system-intensive processing using Sensitive Data. Therefore, proposals that provide Research Center functionality on Public Domain, Google Public Domain, and Open Access Data only, with an eye to future development for Sensitive Data once funding and researcher interest increase, will also be accepted for consideration.

The Executive Committee will select the proposal with the greatest potential for success and allocate all Research Center support to that proposal, including funding that may proceed from the Settlement if it is approved and an award is given to a HathiTrust institution.

[5] In order to reduce confusion throughout the document, the phrase "Sensitive Data" is used to encompass both Protected Data and In-copyright Data. It refers to what would be Protected Data in the Research Corpus if the Settlement is approved, or In-copyright Data in the HathiTrust Corpus if the Settlement is not approved.

A. Research and Demand

Whether composed of the HathiTrust Corpus alone or both the HathiTrust Corpus and the Research Corpus, the HathiTrust Research Center will provide opportunities for investigation and scholarship in areas, and on a scale, that have not previously been available. The Research Center will include portions (in some cases large portions, numbering in the millions of volumes) of the collections of some of the United States' top research libraries, carefully selected and curated over centuries of time. The HathiTrust Corpus by itself currently spans publication dates from the 11th century to the present, and covers a comprehensive array of subjects, with particular strengths in the humanities and social sciences. More than 180 languages are represented. Such a research corpus has not been available to date, and is of increasing interest to both researchers and federal funding agencies as evidenced by the surge in digital humanities computing projects in recent years and the emergence of initiatives such as the Digging Into Data Challenge, which is cosponsored by NEH, NSF, SSHRC,[6] and JISC.

Through discussions with researchers from different institutions and in different fields, the working group was able to identify a cross-section of these research types and needs the Research Center should aim to support. These include, but are by no means limited to:

1. *Aggregation/Distillation* – raw texts or abstracts covering particular topics or types of materials are reduced to subsets or databases of interest that can be used by one or multiple researchers. Examples:

    a. Assembling references to people, places, and things across time, languages, and locations to study "big" questions of history. This would include questions such as the influence of Plato on culture and society in multiple geographic locations over time or the ways Abraham Lincoln was perceived in different areas of the North and South. A Research Center composed of HathiTrust or Research Corpus data would make it possible to investigate questions of this kind and scale.

    b. Assembling information from multiple sources to create dictionaries or indices of information on particular topics. For example, a dictionary of an endangered language assembled from multiple sources, or a database of proteins and how they interact with one another. These resources can aid in preserving and communicating cultural history that is contained only in bits and pieces in dispersed sources, or be used in medical research to develop treatments and vaccinations.

[6] The Social Sciences and Humanities Research Council in Canada.

c. Aggregating and analyzing a set of parliamentary proceedings to determine who had political power in a particular time and place.

d. Studying the use of verbs and verb forms in a language over time (see http://www.nature.com/nature/journal/v449/n7163/full/nature06137.html).

e. Normalizing (for example, markup into TEI) a large body of heterogeneous material (150 million words or more) to improve searching capabilities or test machine markup capabilities.

f. Merging text sources from the social web (Twitter, Facebook) with scholarly literature to compare the way issues are addressed in particular domains or show the comparative value of each in certain topics.

g. Automatically identifying concepts, concept definitions, moods, and emotions in text to study how particular topics are discussed and perceived. This is an emergent area of study on the web and relates to example (f) above.

h. Using word-counts to categorize individual volumes and subsections of Research Center data for browsing purposes. There is much work going on currently in the area of topic modeling, including a current IMLS grant-funded project between Yale University, the University of California, and the University of Michigan.

i. Summarizing and identifying unique topics in newly available texts. The output might be displayed in a graphical interface.

j. Recognizing and parsing reference sections and citations in text to analyze citation networks in different domains (e.g., science, humanities, etc.).

2. *Development of Tools for Research* – much of the research above requires textual analysis, entity extraction, aggregation of data, and the representation and analysis of results. The Research Center must be able to accommodate research that develops the basic tools and processes needed to effectively use and manipulate data. Some of these might include:

   a. Tools to build collections (aggregations) of content.

   b. Tools to cite and reference works line-by-line.

c. Part-of-speech tools that will count the frequency of certain words or phrases in a subset of the data. The use of the tool and the results can be shared and reported out in the literature.

d. Tools for image processing and to produce and improve Optical Character Recognition text. These could be used for purposes such as facilitating reading of texts (by humans or machines), recognizing page numbers (e.g., to test completeness of digital volumes), associating images with text to improve search of images, or testing the quality and integrity of images themselves for preservation and other purposes.

e. Tools to visualize and publish research results – for example, to create a map depicting areas with positive and negative opinions of Lincoln, or a network diagram of citation references.

f. Tools for automated translation.

g. Research to improve search indexing and unsupervised machine processing of texts.

h. Research to explore modes of Non-consumptive Research.

3. *Collaboration* – the increase in availability of electronic textual and image data across subject areas and disciplines has dramatically increased the ability (and need) for collaboration among researchers, teachers, and students. To effectively serve the needs of Research Center users, the Center must offer the ability to share processes, results, and communication with individuals and groups in a secure manner. Some possible scenarios of collaboration include:

    a. Testing or verifying the work of a collaborator (reviewing his or her results). This functionality is important both for researchers who are working together, and for the management and security of the Research Center data. Section *C. Research Results* contains more information on Research Center content and derivative results.

    b. Finding out how other scholars have used the Research Center data, including what tools they employed, what texts they searched, what secondary data they created, and whether this data is accessible.

    c. Manually annotating works and correcting OCR. One of the greatest changes occurring in many fields right now as a result of an increase in the amount and availability of raw research data is the ability of researchers to make substantial contributions to scholarship at earlier points in their careers. An example of this in classical studies is the ability of undergraduate students to undertake translation, annotation, and interpretation of texts that few scholars have worked on because of their relative obscurity (their unavailability and difficulty of access) in the literary canon. The ability to support, manage, and publish these types of contributions will keep the Research Center up-to-date with current teaching methods and research, and dramatically increase the scholarly output of the Research Center itself. Ivanhoe, a game supporting the collaborative interpretation of texts, is an example of this in the humanities (http://www.ivanhoegame.org/wordpress).

4. *Miscellaneous* – the working group identified additional needs and concerns of researchers as well:

    a. The ability to include additional data. The Research Center data may not be sufficient for some researchers, as in the example above using text from the social web. Other investigators may want to include data from sources that fill gaps in the holdings of the Research Center in their research and analysis. Mechanisms are needed by which researchers can include additional data at the time of processing, and include results of research, and possibly the additional raw data themselves, in the Research Center. A similar issue exists with regard to metadata. See *C. Research Results* for further discussion.

    b. The ability to have access to both raw and pre-processed texts. Some research needs will require access to raw data of the Research Center, while others will be met by a pre-processed set of data (e.g., a set of word-counts or an index of data, as opposed to the original data itself). See the note on pre-processed datasets in this section, below.

The following are research environments that can be used as models and examples supporting the needs described above:

a. MAEVIZ earthquake analysis: http://rcp.ncsa.uiuc.edu/maeviz/about.html
b. LEAD portal: https://portal.leadproject.org/gridsphere/gridsphere
c. DSS (Digitized Sky Survey): http://stdatu.stsci.edu/dss/
d. Sloan Digital Sky Survey: http://www.sdss.org/
e. JSTOR data for research: http://dfr.jstor.org
f. TREC (Text Retrieval Conferences) datasets: http://trec.nist.gov/

For reference, the Settlement provides examples of categories of Non-consumptive Research. These include:

- Image analysis and text extraction: Computational analysis of the digitized image artifact to improve the image (e.g., de-skewing) or extracting textual or structural information from the image (e.g., OCR).

- Textual analysis and information extraction: Automated techniques designated to extract information to understand or develop relationships among or within the Corpus. Includes tasks such as concordance development, collocation extraction, citation extraction, automated classification, entity extraction, and natural language processing.
- Linguistic analysis: Research that performs linguistic analysis over the Corpus to understand language, linguistic use, semantics and syntax as they evolve over time and across different genres or other classifications of books.
- Automated translation: Research for techniques for translating works from one language to another.
- Indexing and search: Research on different techniques for indexing and search of textual context.[7]

Note on pre-processed datasets:
All, or nearly all, of the research scenarios described above involve some level of pre-processing of data to get it into a form that is usable (e.g., data must first be aggregated, parsed, merged, etc.). Pre-processing may be done by a researcher, but it may also be performed by a Research Center docent. In spite of the examples of Non-consumptive Research given in the Settlement (above), the working group determined that in the context of the Settlement as a whole the Non-consumptive Research requirement on Protected Data would require that in many (possibly all) cases, these data be pre-processed into an inherently non-consumable form before being made available for research. This is discussed further in *C. Research Results* below.

Proposal Guidelines – Understanding what the needs of researchers are, the kinds of research they conduct, and the demand that exists for a Research Center of this kind are essential to creating, supporting, and sustaining the Research Center over time. The use cases and scenarios assembled above provide a point of departure to understanding the range of issues the Center will need to address. Proposal submissions should demonstrate concrete knowledge of the demand that exists for the data, infrastructure, and services they describe. This may involve surveys of acting research individuals and communities, reviews of currently funded research initiatives, and other modes that tie Research Center functionality to existing needs and demands. This grounding in real-world use and practicality is critical for the ongoing funding, relevance and success of the Research Center.

With regard specifically to point 4.a above concerning the addition of content to the Research Center, proposals should strive to establish ideal conditions or criteria under which new content will be incorporated into the Research Center. Written policies for uses, routines, and results that contributed content may be used to support will educate rights holders and others about needs of research in a variety of contexts and expand the amount of Open Use content in the Center.

[7] Amended Settlement Agreement – The Author's Guild, Inc. Association of American Publishers, Inc., et al., Plaintiffs, v. Google Inc., Defendant, Section 1.93.

B. Structure of the Research Center

Laying aside legal and technical concerns for a moment and focusing on the needs of researchers, the structure of the Center depends on two factors:

1. Whether the research can be performed remotely or on an individual's computer (*virtual v. desktop*); and

2. Whether the resources for research are shared among multiple institutions or available only from a single location (*distributed vs. central*). [8]

Regardless of how these factors play out, the environment must be able to provide:

- Non-consumptive access to in-copyright works (in the context of the Security Standard – see *D. Legal* below)
- Reading (i.e., consumptive) access to in-copyright works, via an Institutional Subscription to Google Books
- Consumptive and Non-consumptive access to Public Domain, Google Public Domain, and Open Access Data
- Infrastructure to receive, share, store, and in some cases provide access to, results of research (see *C. Research Results* below)
- Appropriate security for all Research Center data (see *D. Legal* below)

The environment should additionally be able to:

- Channel processing jobs appropriately—e.g., accommodate smaller jobs that run in parallel with other, computing-intensive jobs, including jobs that may run on a researcher's own machine
- Handle processing jobs in a reasonable amount of time
- Perform indexing tasks to populate pre-processed datasets for use by particular and/or all researchers

Implicit in these criteria are the needs to determine, detect, and accommodate the processing needs of researchers. Also implicit is the principle that the Center will be designed to grow or shrink over time, as the need for computing power will be variable and can be based on grid technology and cloud computing availability.

Additionally, system-level tools will be necessary for:

- Workflow and provenance tracking tools that are deployed in the center for assembling and processing texts

[8] Researchers may find it sufficient to use and share data inside of the Research Center environment. They may also find it necessary to ingest data into their own desktop environments to run customized tools. Research using Sensitive Data will be difficult if that data needs to be distributed (e.g., to support processing needs) or used on a researcher's desktop. (See *D. Legal* below).

- Metadata management, including management of multiple versions of Research Center data as OCR and other aspects of Research Center data are improved (see also the discussion in *C. Research Results* below). Management strategies should be designed to facilitate efficient access, storage, and preservation of Research Center data
- Auditing tools for every process (see *D. Legal* below)
- Distribution of processing resources (depending on Research Center architecture)
- Benchmarking data about processing, storage, etc.

Application-level tools capable of accommodating the tools and processes listed in Section A will also be needed.

The staffing needs of a Research Center of this scale and complexity must not be underestimated. Staffing of highly skilled, specialized technicians, programmers, docents, and legal advisers is imperative. They will perform the following tasks:

- Walk researchers through requirements and liaise with core Center staff
- Create pre-processed datasets, including datasets containing Protected or In-copyright data
- Oversee the inspection and certification of routines, processes, and tools that will be used on Sensitive Data to be sure they are Non-consumptive and do not violate copyright
- Oversee the inspection and certification of datasets resulting from routines on Sensitive Data to be sure they do not violate terms of the Settlement or copyright
- Appraise Derivative Results for storage and possible inclusion in core Research Center Data (see *C. Research Results*)
- Coordinate with the Book Rights Registry and Google
- Plan for growth and additional needs of the Research Center

As stated earlier, the difficulty in providing access to Sensitive Data for Non-consumptive Research is recognized. However, some level of computation (e.g., using vectorized word counts) will be possible on these materials without concern of violating terms of the Settlement or copyright law; discussions with the Book Rights Registry could lead to an opening of additional types of research. We therefore encourage the development of a multi-tiered environment that includes access to both Open Use and Sensitive Data.

Proposal Guidelines – Proposals should present a clear architecture that identifies where and how data is stored, how users gain access, where processing occurs and what strategies are used for job allocation and management. The architecture should include strategies for metadata management, including managing multiple versions of Research Center data and Derivative Results (see also *C. Research Results* below). Proposals should describe in detail the hardware and software technologies used and the reasons those technologies were chosen. Submissions should include any benchmark testing, survey data or other sources of information used to estimate hardware and software needs. Particular attention should be paid to the selection of hardware and software for the specific kinds of research to be supported. A preliminary example architecture developed by the working group is linked to in *Appendix B*.

It is expected that proposals will coordinate with, if not indeed leverage, the storage infrastructure and capabilities of HathiTrust for Research Center data. Information about HathiTrust infrastructure is available on the HathiTrust website (see *Appendix B* for relevant links). If alternative storage strategies are implemented, proposals should detail long-term technical preservation needs of those strategies such as replacement of storage, metadata requirements, backup, etc., and include schedules of replacement and required measures to ensure the integrity of Research Center Data, including Derivative Results. If the architecture accommodates distribution of resources among multiple institutions or desktop access to resources (i.e., data download), it must designate a single host institution to be compliant with Settlement Research Center requirements.

All aspects of Research Center infrastructure should be crafted in close concert with the Security Standard in Attachment D of the Settlement Agreement (a link to the Settlement is provided in *Appendix B*). Proposals should also be cognizant of the uncertainty surrounding the nature and definition of “non-consumptive” research, and the penalties involved in instances of unauthorized access to Protected Data (see in particular Article 8 of the Settlement).

C. Research Results

Results produced by any research process that can be used in a separate workflow (by the same or a different researcher) may be considered Derivative Results. This includes results of all of the research processes outlined in Section A, which may be in the form of visualizations, tables, tab-delimited files, indexes, etc. For the purposes of this discussion, datasets that are prepared for researchers by Research Center docents (such as word counts of Sensitive Data) or researchers themselves, though not research results per se, can be considered in the same category as Derivative Results. When referred to separately, these are called pre-processed datasets below.

Some Derivative Results may be shareable outside of the Research Center regardless of the nature of the data researched (e.g., visualizations or aggregate data that do not reproduce objectionable portions of any Sensitive works involved). Results involving actual data from Sensitive Data content (raw data or indexes) will likely need to remain inside the Research Center, although there may be exceptions. The Settlement does not provide a definition of a fully shareable result. The following are examples of result sets involving Protected Data that will probably not be shareable outside the Research Center under the terms of the Settlement:

- Indexes of information that can be queried to yield information contained in one or more in-copyright works (such as an index of dictionary terms and their definitions). Providing location information about where to find information in those works may be permitted.
- Results that contain facts from Protected Data (such as the height of a building or information that is only available from a copyrighted source).
- Results that compete commercially in any way with a published work or works (such as a dictionary, concordance, index, etc.).

There are no Settlement-imposed restrictions on Derivative Results involving Public Domain, Google Public Domain, or Open Access Data. There may, however, be restrictions on data from Google or rights holders who have opened access to them. For instance, Google Public Domain Data may not be re-hosted or used in search services.[9] HathiTrust has not determined what uses of In-copyright Data it may undertake to enable if the Settlement is not approved.

There are several important questions proposals should address surrounding Derivative Results. One is how it will be determined that a result is shareable or not. As mentioned in previous sections, processes must be in place to allow inspection, as well as appropriate storage and management, of Derivative Results. If Derivative Results are not shareable outside the Research Center, it does not necessarily mean that they are not shareable inside (e.g., results obtained from Sensitive Data may

[9] HathiTrust strives to avoid, when possible, the accession of Research Center data that is restricted. See the final portion of the Proposal Guidelines in *A. Research and Demand.*

contain factual or other information that in some way prevents it from being distributed more broadly, but be Non-consumptive and therefore freely available to researchers in the Center). A process is needed to address questions for which the Settlement, and in some cases copyright law, provides no clear answer.[10]

Proposals should also address the question of how to ensure appropriate levels of access to Derivative Results. It is conceivable that results from Sensitive Data are consumptive in a way that is permitted for researcher and collaborator review (e.g., reviewing sentences or paragraphs of data), but not for others inside or outside of the Research Center. Controls to prevent improper distribution of data must be in place.

A third question is the way Derivative Results will be managed so they remain useful to researchers as the original Corpus grows and changes over time. For instance, it will not be possible to merely keep the routines that produced a dataset if the OCR text and other characteristics of the data change over time. Appropriate management of Research Center Data will keep the original derivative or pre-processed dataset from being invalidated. Workflow management tools will appropriately maintain copies/versions of datasets and data. These datasets are invaluable as a means to re-validate results, run a process with different parameters, and potentially settle disputes such as academic misconduct.

A fourth question is how and whether Derivative Results should be incorporated into original Research Center data. Scenarios where this might occur include OCR correction, metadata correction as a result of research analysis (e.g., publication date), or metadata addition as a result of research analysis (e.g., gender or other information about authors, such as death date), among others. Strategies for managing and incorporating these kinds of data should also be addressed.

A fifth question, related to the fourth, is how to determine which Derivative Results will be saved (assuming the Research Center will not take responsibility for all Derivative Results). Issues of copyright and liability may come into play here, but appraisal processes and policies need to be defined to address this issue. Collection and use of provenance data will be invaluable in this area.

Finally, because of possible restrictions surrounding Derivative Results, the ability to house them with the same level of control over provenance, auditing, security, validation, quality control and storage, as for Sensitive Data, must be in place.

Proposal Guidelines – Proposals should contain strategies and policies for receiving, storing, accessing, sharing, appraising, and incorporating (if applicable) Derivative Results. This should include components addressing the inspection of

[10] Another example of this is the use of in-copyright data in tools used outside the Research Center (e.g., parsers). It is unclear if this is a violation of copyright law, though it may violate terms of the Settlement.

results (for auditing purposes and for review by other researchers, including authentication mechanisms where appropriate), capture and management of provenance information, management of multiple versions of Derivative Results (applicable also to Research Center content in general – see *B. Structure),* mechanisms enabling appraisal of results and facilitating inclusion of results in core Research Center data, and any additional metadata needed to facilitate access, storage, preservation, etc. of the results.

Proposals should also outline a process for addressing issues surrounding the creation and use of Derivative Results (and issues in other areas that come up) that are not provided for in the Settlement or copyright law. This might include the formation of a council to coordinate on issues with the Book Rights Registry or the creation of additional policies to enumerate steps to be taken when issues arise. A body devoted to advocacy for the Research Center in general should be considered, to raise awareness among rights holders and the Registry about the importance of making their materials available for research and adding to the body of freely available research material.

D. Legal and Security Concerns

As mentioned above, the Settlement includes the award of up to two Research Centers, each containing the Research Corpus, to Google partner library institutions. One of these is likely to be a HathiTrust institution. Any institution that takes on the management of one of these Centers will have both a tremendous opportunity, and an enormous responsibility due to the sensitivity of the data and the accompanying legal and security concerns. It will be a challenge to balance these duties without losing sight of the essential purpose of the Research Corpus: research and scholarship. This section attempts to outline what some of these challenges will be and the measures a Host institution will need to have in place to be eligible to receive the Settlement Research Corpus.

It is difficult to boil down into a few bullet points the legal issues that relate to the Research Center, but in broad terms, the Settlement forbids using Protected data in the following ways:

- Using Protected Data in a consumptive manner (see *B. Structure* above)
- Using Protected Data for services that violate the Settlement or copyright restrictions
- Allowing Protected Data to leave the repository and be used "inappropriately"

It requires us to:

- Have an audit trail for every action that touches Data in the Research Center
- Create and file an approved security plan
- Create a perimeter beyond which Sensitive Data cannot pass unencrypted
- Have appropriate authentication and access controls on the Data
- Have authority and permission trails for every researcher and docent who accesses and uses the Research Center
- Have a mechanism for documenting responsibility for granting access

In addition, the Settlement introduces some conditions that are more stringent than traditional copyright law. For instance, the Settlement does not allow "fair uses" of Research Center Data.

Apart from the Settlement, Google forbids:

- Re-hosting Google Public Domain Data or using them in search services
- Allowing systematic download of materials digitized by Google

The Settlement includes a framework for a Security Implementation Plan that each Research Center host must meet, and which must meet the requirements of the Settlement Security Standard. As mentioned in *B. Structure*, all aspects of the

Research Center infrastructure should be planned with this Standard and its requirements firmly in mind. There are stiff penalties for breaches of a Host Site's Security Implementation Plan.[11]

In addition to technical security requirements, there are additional policy requirements that need to be in place in order to make use of Protected Data. Researchers will need to:

- Sign an agreement understanding their liability for use of the Corpus
- Understand penalties surrounding use of the Research Corpus in a non-permissible way
- Have their research vetted before, and likely after performing the research, by authorities in the Center

An individual who is sponsored by an institution will need to have that institution sign its own agreement with the Center. If the individual is sponsored by a granting agency, the liability for the individual's research falls on the Center. It is also possible that the granting agency could be the Center itself.

Use of Google Public Domain Data will require a limited agreement with Google that forbids the researcher from re-hosting the Data or the results.

Proposal Guidelines – Security concerns of the Research Center have been described only briefly here. However, addressing these concerns and requirements will likely pose the greatest challenge to the creation of a Center that provides access to Protected Data—which will comprise a significant percentage of the Research Center Data if the Settlement is approved. Proposal submitters are encouraged to read carefully through the terms of the Settlement regarding the Research Corpus (largely enumerated in Article 7 section d, though elsewhere as well) and become familiar with the Security Standard. These requirements will have an impact on every stage of management and access for Protected materials.

[11] See the Settlement Agreement, Section 8.3.

E. Sustainability

Governance

At the center of the establishment of the HathiTrust digital repository was the idea that the long-term integrity and usefulness of the individual partner collections could be most effectively ensured through an organization that was *co-owned* and *co-managed* by all of the partners. This same principle is evidenced in the large NSF-funded DataNet program and many smaller preservation-oriented projects currently underway in the library community. While not touted as a preservation initiative, a Research Center established jointly by the HathiTrust partner institutions has the opportunity to leverage the long-term relationship the HathiTrust partners have entered into for the curation of their digital content, in the governance and management of the Research Center. Efficient operations, effective policies, and skillful allocation of resources will depend on bodies close to, and knowledgeable of, the particular interests and needs of researchers and the Research Center. A close connection between these bodies and the HathiTrust Executive Committee will ensure the sustained alignment of priorities between the Center and the digital repository over time. Proposals are encouraged to view governance in this way as both a means of ensuring the efficient operation and management of the Research Center and ensuring the long-term viability and sustainability of the Center over time.

Funding

Equally important to the long-term vitality of the Research Center are realistic strategies for securing funding to support the Center's ambitious aims. The Google partners expect to initially receive between 5 and 10 million dollars to fund up to two Research Centers. As mentioned above, it is likely that one of these will be awarded to a HathiTrust institution. Under the terms of the Settlement, Google may also host a Research Center at the request of one of the Google Participating Libraries, but Google has not, to-date, shown an interest in doing so. The one-time funding provided by Google will clearly not be sufficient to sustain the activity and management of the Research Center on an ongoing basis, and will need to be supplemented (or subsumed) by other funding options. If the Settlement is not approved, the Google funding will not be available for a Research Center involving the HathiTrust Corpus alone.

Research Center funding will need to support:

- An adequate processing infrastructure to offer multiple researchers the ability to perform research at the same time and at different levels of intensity
- An adequate storage infrastructure to manage this research over time
- The capability for growth in processing speed, power and storage

- Staffing of the Center as noted in *B. Structure*
- System development and optimization tools, including networking overhead
- Development or incorporation of workflow, provenance, auditing, and access tools
- Development or incorporation of an application to store, manage and make available metadata about the data
- Legal counsel to handle agreements and disputes
- Grants and fellowships for individuals unable to utilize an institution's funds

The Research Center can be funded through a variety of public and private one-time funds, as well as sustained contributions (possibly a partner model). Some strategies to maximize funding include:

- Tapping into existing large national initiatives such as NSF, Teragrid, IMLS, DARPA, and NEH
- Targeting initiatives such as the Digging into Data challenge jointly sponsored by NEH, NSF, SSHRC and JISC (http://www.diggingintodata.org/)
- Developing a "hosting" model whereby institutions contribute funds to support their own researcher's use of the Research Center
- Leveraging grants in specific areas (e.g., text mining, image analysis, OCR improvement) to build Research Center capabilities over time
- Taking an incremental approach to building the Center that focuses on serving particular needs, rather than building something large and complicated that is targeted towards *perceived* needs and not *actual* needs

The Research Center has significant potential to grow quickly in the number of researchers interested in using it. If it is successful in meeting the needs of initial researchers news will likely spread and lead to further funding.

Proposal Guidelines – Although sustainability is addressed in other portions of the Call, proposals must include detailed strategies for governance and funding of the Research Center in a sustainability context. Submissions are encouraged to consider multi-institutional participation in governance and operations management where appropriate, as well as periodic review of governance processes (e.g., every three years). In all cases, proposals must define clear channels of communication between the Research Center and the HathiTrust Executive Committee to coordinate on and enact final decisions.

Funding strategies for the Research Center should be detailed and consider short-term, medium-term, and long-term needs of the Center. Proposals must provide a budget, including sources of funding. If an institution is relying on funding from a central administration, it should provide information on the commitment to that funding. Proposals must contain a statement on how the Research Center will be sustained financially over time.

Appendix A: Definitions

General

Data – UTF-8 text files derived from images of digitized books and journals. Data may in some cases include the images themselves.

HathiTrust Corpus – The complete set of works in HathiTrust, including Public Domain, Google Public Domain, Open Access, and In-copyright Data.

Derivative Results – Results produced by any research process that can be used in a separate workflow (by the same or a different researcher).

Non-consumptive Research (simplified) – Research that does not allow the Corpus to be Read as the research is being performed; see also the full definition under Settlement Definitions.

Read – Human eyes on the Data, with the intent to understand the Data.

In-copyright Data – Data protected by copyright law that provides exclusive rights (Section 106) to the copyright owner except for the limitations on those rights (Sections 107-122).

Protected Data – Data protected by the Settlement that provides exclusive rights (Section 106) to the copyright owner without the limitations on those rights (Sections 107-122); see also the full definition under Settlement Definitions.

Public Domain Data – Data that according to copyright law, in its source country through expiration of term of protection, has made it available for access and use.

Google Public Domain Data – Google-digitized public domain Data made accessible to everyone—these cannot be re-hosted.

Open Access Data – Data that has been opened by the rights holder for access and use by everyone—these can be re-hosted.

Open Use Data – All of Public Domain, Google Public Domain, and Open Access Data.

Sensitive Data – Refers to Protected Data in the event that the Settlement is approved, and In-copyright Data in the HathiTrust Corpus in the event that it is not.

Settlement Definitions

Book – "Book" means a written or printed work that as of January 5, 2009 (a) had been published or distributed to the public or made available for public access as a set of written or printed sheets of paper bound together in hard copy form under the authorization of the work's U.S. copyright owner, (b) was subject to a Copyright Interest, and (c) (1) if a "United States work," as defined in 17 U.S.C. § 101, was registered with the United States Copyright Office, and (2) if not a United States work, either (x) was registered with the United States Copyright Office, or (y) had a place of publication in Canada, the United Kingdom or Australia, as evidenced by information printed in or on a hard copy of the work. Relevant information printed in or on a hard copy of the work may include, for example, a statement that the book was "Published in [Canada] or [the UK] or [Australia]," or the location or address of the publisher in one of those three countries. The term "Book" does not include: (i) Periodicals, (ii) personal papers (*e.g.*, unpublished diaries or bundles of notes or letters), (iii) written or printed works in which more than twenty percent (20%) of the pages of text (not including tables of contents, indices, blank pages, title pages, copyright pages and verso pages) contain more than twenty percent (20%) music notation, with or without lyrics interspersed, (for purpose of this calculation, "music notation" means notes on a staff or tablature), (iv) written or printed works in, or as they become in, the public domain under the Copyright Act in the United States, or (v) Government Works, or (vi) calendars. References in this Settlement Agreement to a Book include all Inserts contained in the Book, except where this Settlement Agreement provides otherwise.[12]

Digital Copy of a Book or Insert – a set (or portion thereof) of electronic files created by or for Google or provided to Google in connection with GBS, including the image files of the individual pages of the Book or Insert along with text (currently generated from OCR technology), coordinate information for the text, information about the ordering of pages along with page-level metadata such as page number and other similar information, regardless of the means or technology used to prepare such copy, whether now known or hereafter developed, and any digital copy of such set of electronic files.[13]

Expression – either (a) Protected expression, which, in the case of text, means no fewer than three (3) contiguous words, or (b) any contiguous set of ten (10) or more words from a Book or Insert, not counting expression that is not Protected.[14]

Host Site – an institution authorized under the Settlement Agreement to host the Research Corpus pursuant to the requirements of Section 7.2(d)(ii) (Host Sites).[15]

Non-consumptive Research – research in which computational analysis is performed on one or more books, but not research in which a researcher reads or displays

[12] Amended Settlement Agreement Section 1.19.

[13] Ibid., Section 1.48.

[14] Ibid, Section 1.55.

[15] Ibid, Section 1.70.

substantial portions of a book to understand the intellectual content presented within the book.[16]

Protected Data – works, material, Expression or content as to which a Person has a Copyright Interest under Section 106 of the Copyright Act, without giving effect to Sections 107 through 122 of the Copyright Act.[17]

Public Domain Book – a written or printed work that would be a "Book" but for the work being in the public domain under the Copyright Act in the United States, without regard to whether such work contains an Insert; provided, however, that, if the work is a "United States work" as defined in 17 U.S.C. § 101, it need not have been registered with the United States Copyright Office to be considered to be a Public Domain Book.[18]

Research Corpus – a set of all Digital Copies of Books made in connection with the Google Library Project, other than Digital Copies of Books that have been Removed by Rightsholders on or before April 5, 2011 pursuant to Section 3.5 (Right to Remove or Exclude) or withdrawn pursuant to Section 7.2(d)(iv) (Right to Withdraw Library Scans), which Google provides to a Host Site or that Google, if and as a Host Site, uses.[19]

[16] Ibid, Section 1.93.

[17] Ibid, Section 1.116. See Amended Settlement Agreement (Appendix B) for definitions not included here.

[18] Ibid, Section 1.118.

[19] Ibid, Section 1.132.

Appendix B: Accompanying Documentation

Working Group documents (attached PDF)

Core (central) environment example.
Distributed environment example.

Relevant Links

HathiTrust Infrastructure

- http://www.hathitrust.org/technology
- http://www.hathitrust.org/preservation
- http://www.hathitrust.org/papers - see, in particular, the presentation and notes for "From Access to Ingest: A Day in the Life of a HathiTrust Digital Object"

Google Settlement Agreement

- http://books.google.com/booksrightsholders/