

ABSTRACT

“Digging Deeper, Reaching Further: Libraries Empowering Users to Mine the HathiTrust Digital Library Resources” is a three-year project (October 1, 2015-September 30, 2018) that will be led by Harriett Green at the University of Illinois. The project will leverage formal collaborations with Angela Courtney (Indiana University), Geoffrey Morse (Northwestern University), Neil McElroy (Lafayette College), and Stewart Varner (University of North Carolina). Beth Sandore Namachchivaya and J. Stephen Downie, members of the HathiTrust Research Center Executive Management Committee, will also serve as co-PIs.

The project proposal is being submitted under the Learning Spaces in Libraries project category as the primary deliverables are a shared curriculum for use in academic libraries and a Train the Trainer series designed to assist librarians in getting started with the tools, services, and related research methodologies of the HathiTrust Research Center (HTRC). This project also contributes substantially to the National Digital Platform funding priority by increasing access to an important digital service through instructional interventions that will ultimately improve users’ experience conducting cutting-edge explorations in a unique research environment.

We propose to situate the academic library as a learning space for encountering the “big data” tools and methodologies that are being made more broadly accessible through HTRC, and adapting them for use in research and instruction. By developing a suite of curricular materials disseminated through a Train the Trainer program, the project will:

1. Provide librarians with new content for instructional services while also empowering them to become active research partners on digital projects at their institutions; and
2. Provide the foundation to transform academic libraries’ scholarly commons and digital humanities centers into more data-intensive collaborative learning spaces, both physically and virtually, through use of this curriculum and engagement in community dialogues on text and data analytics.

The project team will develop a curriculum to be piloted at five academic institutions. Once fully developed, the curriculum will be shared widely through a Train the Trainer program comprising a minimum of 10 events in Year Three. Program attendees will then disseminate the curriculum to their home institution, embedding new services and initiatives at some 200 institutions and providing training opportunities for roughly four million students. By the conclusion of the grant period, the team will also release an instructional webinar and will package the curriculum as an open education resource for use by some 120,000 libraries in the United States.

The webinar and curriculum will be made publicly accessible through the HathiTrust website, thus extending our potential reach beyond the 90 current HT members to libraries everywhere and the general public seeking access to HathiTrust resources. The HTRC tools and services are publicly available online, and are not limited to use by HathiTrust member institutions. As such, the HTRC strives to serve as a training platform for anyone interested in learning about “big data” tools and methodologies. The open education resource could be adapted to further uses in secondary school classrooms or public library workshops, thus addressing a critical shortage of data analysis skills and experience in the United States.

The 2015-2016 academic year is a critical juncture for HTRC as we shift from experimental prototyping toward hardening production services and increasing user engagement and outreach. The HTRC has just hired a Visiting Digital Humanities Specialist whose roles and responsibilities include programmatic outreach and instruction, beginning this year. Her efforts constitute a large portion of project cost share, and the direct funding this award will extend the scope of projected center activities to ensure deep engagement with selected user communities and expand outreach activities to a national scale.

Introduction

“Digging Deeper, Reaching Further: Libraries Empowering Users to Mine the HathiTrust Digital Library Resources” is a three-year project (May 1, 2015-April 30, 2018) that will be led by Harriett Green, English and Digital Humanities Librarian at the University of Illinois, in collaboration with partners from Indiana University, Northwestern University, Lafayette College, and the University of North Carolina.

Our project proposes to build a suite of curricular materials and training programming around the HathiTrust Research Center that seeks to address the professional development needs emerging as academic libraries expand their services and resources for digital scholarship. This initiative will arm academic librarians in diverse institutions across the U.S. with skillsets in text mining methods and knowledge of critical analytic tools relevant to addressing researchers’ needs in digital scholarship and data mining research.

The proposal is being submitted under the Learning Spaces in Libraries project category, as the primary deliverables are a shared curriculum for use in academic libraries and a Train the Trainer series designed to assist librarians in getting started with the digital tools, services, and related research methodologies of the HathiTrust Research Center (HTRC). This project also contributes substantially to the National Digital Platform funding priority, by increasing access to an important digital service through instructional interventions that will ultimately improve users’ experience in conducting cutting-edge explorations in a unique research environment.

The 2015-2016 academic year is a critical juncture for HTRC as we shift from experimental prototyping toward hardening production services and increasing user engagement and outreach. The HTRC has just hired a Visiting Digital Humanities Specialist whose roles and responsibilities include programmatic outreach and instruction, beginning this year. Her efforts constitute a large portion of project cost share, and the direct funding this award will extend the scope of projected center activities to ensure deep engagement with selected user communities and expand outreach activities to a national scale.

Statement of Need

Modern academic librarianship demands reconfigured skillsets and expertise to meet the rapidly evolving, data-driven needs of today’s students, faculty, and researchers across multiple disciplines as they pursue digital scholarship and produce immense amounts of research data in a variety of formats. A recent McKinsey Global Institute report (Manyika et al., 2011) notes the dearth of data skills among today’s workforce, estimating that the shortage could reach nearly 200,000 workers by 2018. In the academy, data-intensive research is expanding exponentially, in the sciences as well as the humanities: digital humanities is transforming the ways that an increasing number of humanities scholars are pursuing scholarship today. Varner and Vandegrift (2013), Coble et al. (2014), and Sula (2013) are among those who have considered the reconceptualization of academic librarians’ roles and support services in connection to digital humanities research, and the newly released *Digital Humanities in the Library: Challenges and Opportunities for Subject Specialists* (Hartsell-Gundy et al., 2015) features an array of case studies written by librarians and teaching faculty who are building new services and programs in academic libraries to address digital humanities research needs.

These rapid developments in the research services of academic libraries speak to an urgent need to provide training for academic librarians to support digital scholarship. The Research Libraries UK’s *Re-Skilling for Research* report (Auckland, 2012) investigates strategies for instilling new skills in academic librarians to meet researchers’ needs today, noting that “the research environment is changing, driven not least by the power of technology to transform the way researchers work. Libraries are largely in uncharted territory, and have the chance to draw a new map of support and services for researchers” (p. 5). Ithaka S+R’s *Supporting the Changing Research Practices of Historians* (Rutner and Schonfeld, 2012) and *Supporting the Changing Research Practices of Art Historians* (Long and Schonfeld, 2014) both note in their findings that the rise in

digital resources and tools in humanities suggest that libraries must reexamine collections access, discovery tools and research support services offered.

In recent years, initiatives have been launched at academic libraries to respond to these scholarly needs for digital humanities, including programs at the University of Maryland (Munoz and Guiliano, 2014), Indiana University (Dalmau, 2014), Columbia University (<http://developinglibrarian.org>), and New York University (Vinopal and McCormick, 2013). The libraries at these institutions, among others, seek to re-train librarians in digital scholarship methods and tools, especially with the advent of digital scholarship centers and research commons being added to many libraries' infrastructures and services.

The recent emergence of digital scholarship centers in academic libraries catalyzes the need for librarians to acquire skills in areas such as data management and digital tools, and offers opportunities for librarians to apply these new digital skills in spaces where they can collaborate with researchers. This type of collaborative services model has been investigated and assessed in recent years, including reports from the Coalition of Networked Information (Lippencott and Goldenberg-Hart, 2014) and the Council for Library and Information Resources (Zorich, 2008).

But the infrastructure of a center is not always available: in a recent Association of Research Libraries survey and report (Bryson et al., 2011), only 24 percent of libraries had a general digital scholarship center and a scant 5 percent had a digital scholarship center focused on the humanities; but a dominant 48 percent of the surveyed libraries indicated that they provided ad hoc services for digital humanities projects. This suggests that individual librarians often provide digital humanities research support in libraries, and there is a need to train librarians in digital scholarship skills in a way that recognizes the decentralized and fluid nature of such research support.

The "train the trainer" model has been instrumental in developing new areas of expertise in librarians in order to enable libraries to provide lean yet effective services in new areas of research. Notable effective examples of this training method include the Association for College and Research Libraries' (ACRL) Scholarly Communications Initiative, which includes the Scholarly Communications Toolkit and the national Scholarly Communications Roadshow seminar series; and the IMLS-funded Data Information Literacy project (Carlson et al., 2013) that, in part, creates a training infrastructure for librarians to acquire data literacy skills. Using such a "train the trainer" model, this project seeks to address the need for inculcating knowledge of digital humanities methods and tool skills in librarians, focusing on the area of textual analytics and data mining.

Impact

The HathiTrust Research Center (HTRC) is well positioned to serve as a training ground for building digital research skills in librarians. The HTRC has created online services and tools that allow research access to the HathiTrust Digital Library's textual corpus in a large aggregated manner, rather than volume-by-volume basis. To maximize accessibility to the entire corpus, including in-copyright texts, the HTRC has been crafting models and tools to facilitate large-scale data mining analyses under a "non-consumptive research" paradigm. Under this paradigm, analytic algorithms are applied to the restricted data held by HTRC. Once the analyses are run, only results are returned to researchers. Thus, restricted material is never directly "consumed" by scholars.

This access to in-copyright material is a revolutionary development in text mining and resource access that humanities researchers across disciplines need and want (Underwood, 2014). The primary alternative to the HTRC suite of tools is a manual approach to large-scale text analysis, where a researcher has to find access to the full works and develop a technical workflow for analyzing them. In contrast, HTRC provides a space for immediate access to millions of works and lowers the technical barrier to entry for digital humanities research by offering tools for searching, compiling, and analyzing those works. Although the tools and services provided by the HTRC are designed to lower technical barriers to access, we find that students,

staff, and faculty in academic settings still face challenges in navigating an unfamiliar research infrastructure and adapting to emerging methods of discovery, selection, and data analysis.

We propose to situate the academic library as a learning space for encountering the “big data” tools and methodologies that are being made more broadly accessible through HTRC, and adapting them for use in research and instruction. By developing a suite of curricular materials disseminated through a Train the Trainer program, the HTRC seeks to transform library services and professional development. Our goals are to:

1. Provide librarians with new content for instructional services that address the curricular and research needs of students and faculty around digital scholarship and digital humanities;
2. Empower librarians to become active research partners on digital projects at their institutions;
3. Provide the foundation to transform academic libraries’ scholarly commons and digital humanities centers into more data-intensive collaborative learning spaces, both physically and virtually, through use of this curriculum and engagement in community dialogues on digital humanities resources.

To meet these goals, the project team will produce these key target deliverables:

1. Build a multi-part curriculum that provides instructional materials on the HTRC suite of tools and principles of large-scale text mining research;
2. Develop a suite of workshop program materials that will be piloted at the five partner institutions and iteratively revised to produce “train the trainer” workshop program materials, which can in turn be adapted by attendees at diverse types of institutions for teaching HTRC local workshops;
3. Conduct a nationwide “roadshow” of workshops that will train library and information professionals in teaching the HTRC tools and principles of large-scale text mining;
4. Create a webinar and finalized package of open educational resources that will be hosted in a trusted institutional repository, and will be promoted as fully accessible to all interested constituencies.

The project team has developed an explicit impact strategy that begins by piloting the new curriculum at five academic institutions. Once fully developed, the curriculum will be shared widely through a Train the Trainer program comprising a minimum of 10 events in the final year of the project. Program attendees will then disseminate the curriculum to their home institutions, embedding new services and initiatives at some 200 institutions and providing training opportunities for roughly four million students. By the conclusion of the grant period, the team will also release an instructional webinar and will package the curriculum as an open education resource for use by some 120,000 libraries in the United States.

The webinar and curricula will be made publicly accessible through the HathiTrust website and in a trusted institutional repository for long-term storage, thus extending our potential reach beyond the 90 current HT members to libraries everywhere and the general public seeking access to HathiTrust resources. The HTRC tools and services are publicly available online, and are not limited to use by HathiTrust member institutions. As such, the HTRC strives to serve as a training platform for anyone interested in learning about “big data” tools and methodologies. The open education resource could be adapted to further uses in secondary school classrooms or public library workshops, thus addressing a critical shortage of data analysis skills and experience in the United States.

The workshops and open educational resources produced by our project ultimately will reach a wide array of information professionals, technologists, teachers, and students across North America and the world and teach them, via the tools and resources offered by the HathiTrust Research Center, about text mining methods applied in digital humanities research.

Project Design

HTRC Tools and Services

The HathiTrust Research Center (HTRC) is a collaborative partnership between Illinois and Indiana to build resources and tools for enabling researchers to access and use the full HathiTrust textual corpora in digital scholarship. This project will leverage pre-existing HTRC tools to design curricular materials that will be assessed and revised collaboratively with project partners. There are four primary tools and resources created by the HTRC for which we will develop training materials:

- The **HTRC Portal**, a contained research environment for building customized worksets of volumes from the HathiTrust Digital Library collection in the **Workset Builder** tool, running HTRC-provided off-the-shelf algorithms against a workset, and interpreting the generated results
- The **HTRC Data Capsule**, which enables non-consumptive reading of in-copyright texts by bringing outside algorithms to the text corpus such that the algorithms are run against the text in a secure, encapsulated environment in which the text is strictly isolated from the user
- The **HathiTrust + Bookworm** tool, which permits the discovery of lexical trends within the HathiTrust Digital Library through interactive visualization and affords an analytical understanding of the constitutive elements contributing to such trends
- The **HathiTrust Extracted Features (EF)** dataset, which allows researchers to download page-level features — metadata and extracted informational content — derived from the text data associated with a researcher-created workset

Instructional Design and Training

This project includes two distinct training initiatives, each with slightly different training methods, audiences, and benefits. An initial set of pilot workshops are designed to assess and improve a shared suite of curricular materials. Following the pilot workshops, the curriculum will be integrated into a broader Train the Trainers program designed to assist academic librarians with integrating the new curriculum into their local instruction and outreach services. Descriptions of each training initiative follow below, and a sample curriculum is included as a supporting document.

Pilot Workshops

Pilot workshops are designed to assist the project team in assessing and revising the initial HTRC curriculum within a number of institutional contexts including a private research university, a small liberal arts college, and several large public universities. Each partner institution will host a series of workshops geared toward their core constituents, which may range from exclusively undergraduate populations to a diverse range of students, staff, and faculty. In all cases, the participants' learning objectives are to gain exposure to a suite of tools designed for computational analysis of large bodies of text; consider the implications of underlying data, algorithms, and content types in reviewing results of analysis; and explore opportunities to incorporate textual analysis in class projects, research, or teaching. The workshop will discuss scenarios and example use cases in which each of the HTRC tools shines, and will demonstrate the ways in which they can work in complementary fashion with each other. Attendees will learn new tools and approaches for the complex research questions that digitized text corpora are uniquely poised to help answer across disciplinary fields, including literary, historical, and sociological studies.

Train the Trainer Events

Train the trainer events will be geared toward integrating digital tools and techniques into educational experiences using services provided by the HathiTrust Research Center and grounded in best practices in digital pedagogy. Each workshop will comprise a set of lectures, demonstrations, hands-on practice with digital tools, and guided group discussions. Academic librarians with outreach and instruction responsibilities are the target audience for these events. While no workshop will be dedicated to a particular

pedagogical type, we expect participants' instructional responsibilities will include audiences of undergraduates, graduate students, and even faculty and staff. Our primary objective is to demonstrate how HTRC tools and services can serve as a useful platform for introducing key concepts related to digital scholarship and data analysis to library patrons. To achieve this goal, we will share the curriculum, which will have been developed and finalized over the course of the pilot workshops, in conjunction with strategies for approaching instruction. Participants' learning objectives will be to understand how to effectively incorporate digital humanities tools for text analysis into instruction services. Within the context of the Train the Trainer events, we will also provide tools and scaffolding for continued curricular innovation. Participants across all ten Train the Trainer events will be treated as a single cohort, which will allow for the ongoing exchange of curricular content as well as long-term community development. This approach will encourage scholars who work across institutional contexts to engage with a wide range of academic librarians.

Project Work Plan

The following work plan divides the project into a series of annual objectives, milestones, and activities. A schedule of completion, included in with this proposal, also charts project activities by month.

Year One

Objective: Develop an open suite of **instructional materials** that will range from introductory activities, such as how to build textual corpora and the steps for conducting basic text analysis with HTRC tools, to advanced tutorials such as text mining approaches for social science research or using the datasets derived from the HTRC feature extraction tools.

Milestone: The major milestone of year one will be sharing a finalized suite of curricular materials with project partners for use in local workshops.

Activities: In Year One, HTRC core staff will work in consultation with project partners to develop a set of training worksets, guided tutorials, and instructional activities for use in academic libraries.

Key activities include:

- Creating a shared, internal space for posting content under construction with all partners;
- Conducting monthly calls with partners and lightweight listserv communication;
- Developing use cases and learning scenarios for HTRC;
- Developing a set of hands-on activities for use during workshops;
- Creating a suite of worksets for instructional use during workshops that illustrate the problems, aspects, and challenges of underlying data and algorithms as well as particular types of texts (e.g., multi-lingual works);
- Developing feedback surveys for evaluation;
- Conducting first-run workshops at University of Illinois and Indiana University.

Year Two

Objectives: Conduct a series of **pilot workshops** at each of the non-HTRC partnering institutions (i.e., Lafayette, Northwestern, and North Carolina) and iterate curriculum based on partner feedback. Build scaffolding for Train the Train workshops that integrates digital pedagogy concepts, discussions of collaborative service models, and opportunities for revising and reusing HTRC workshops in a variety of local contexts.

Milestone: The major milestone of year two will be the completion of a partner-vetted and improved HTRC curriculum and a set of related sessions around digital pedagogy and collaborative service models. These will be delivered to the project advisory board for final review in advance of the annual conference call.

Activities: In addition to conducting workshops, team members will regroup monthly to evaluate workshop outcomes and iteratively revise the initial suite of curricular materials in response to information behaviors and user requirements encountered during instructional sessions.

Key activities include:

- Conducting feedback surveys from pilot workshops;
- Convening semi-regular conference calls with each partner;
- Revising the initial workshop materials;
- Developing text-based guided tutorials for workshop instruction and self study;
- Writing scripts for online tutorials and recording screencasts;
- Developing scaffolding and frameworks for teaching how to teach with HTRC;
- Planning and promoting the Train the Trainer roadshow events.

Year Three

Objective: Expand the reach of the HTRC instructional program through a **Train the Trainer** roadshow and share an **open educational resource (OER)** for general use.

Milestones: The major milestones of year three will be delivery of ten Train the Trainer events, final deposit of an OER, and creation of a guide for recommended practices based on participant feedback and general project evaluation.

Activities: Train the Trainer events will be aligned with conferences including the annual meetings of ALA and ACRL; DLF Forum; select regional conferences; selected THATcamps; and similar events around digital libraries and digital humanities, chosen strategically to maximize geographic coverage across the United States.

Key activities include:

- Coordinating and conducting ten Train the Trainer events;
- Evaluating project success and impact by circulating feedback surveys after each Train the Trainer session, and additionally following up with event participants in the final quarter of Year Three to learn what activities have been initiated locally;
- Developing an OER project site with instructional materials;
- Depositing training and instructional resources in a trusted institutional repository;
- Building participant cohorts through the creation of a community-based resource (e.g., wiki, forum, etc.) for sharing curricular revisions and continuing discussions around digital pedagogy and instruction;
- Promoting OER resources within the broader professional community.

Evaluation

A series of formative evaluations will be led by the project team at the University of Illinois to ensure that project objectives are being met, to identify areas for improvement, and to inform outreach strategies.

Evaluations will include working group meetings, participant surveys after each workshop and program to incorporate continuous and iterative assessment, and analysis of use statistics for online materials.

Project Resources

Project work will be overseen through weekly meetings at Illinois and monthly teleconferences with project partners for designing the curriculum, planning events, joint decision-making, and appropriate tracking of project milestones, deliverables, and evaluations. Teleconferences with the advisory board will be convened annually each June. Agendas, project documents, and all ongoing work of the team will be managed on a project wiki. A public discussion forum will be used for coordination with and among workshop participants.

The following section outlines institutional partners, PI expertise and key personnel, and the composition of the advisory board.

Partners

HathiTrust Research Center

The HathiTrust Research Center (HTRC) is the official research branch of the HathiTrust, which is a repository that centrally collects image and text representations of library holdings digitized by the Google Books project and other mass-digitization efforts. The HTRC is a unique collaboration between the University of Illinois and Indiana University, co-directed by Professor Beth Plale (Indiana) and Professor J. Stephen Downie (Illinois). By co-locating the HTRC at two separate institutions, the center benefits from drawing upon the cyber-infrastructure expertise of the Data to Insight Center and the collection, metadata, and content expertise of the Graduate School of Library and Information Science.

University of Illinois

Scholarly Commons

The Scholarly Commons is a technology enriched space for faculty, researchers, and graduate students to pursue research and receive expert copyright, data, digital humanities, digitization, scholarly communications, and usability consultation services. Scholarly Commons services are supported by experts in the Scholarly Commons, subject specialists at the University Library, and partners throughout campus. The Scholarly Commons represents a proven model of instruction and a collaborative culture of scholarly inquiry by providing an environment in which students and scholars can access, and contribute to available resources. The Scholarly Commons also provides library, information, and technology services that support traditional and emergent forms of scholarly inquiry.

Graduate School of Library and Information Science

Founded in 1893, the Graduate School of Library and Information Science (GSLIS) is a world leader in library and information science education, research and practice. Consistently ranked as one of the very best in the field, GSLIS has earned its reputation by creating pioneering and innovative educational opportunities, by leading groundbreaking research to advance preservation of and access to information in both traditional and digital libraries, and through its services and strong commitment to outreach and community development.

Indiana University Bloomington

IU Scholars' Commons

Designed on the cutting edge of research and scholarly publishing, the IU Scholars' Commons serves as a powerful academic service hub that offers easy access to experts and technology for every stage of their scholarship. The IU Scholars' Commons currently supports workshops, seminars and weekly consultation services for the HathiTrust Research Center and has been involved in outreach events for the HTRC at Indiana University and the Ohio State University.

Data To Insight Center

The Data To Insight Center is a research center at Indiana University. Its current projects are in data management, big data, data and text analytics, advanced cyberinfrastructure, tools for automated metadata and provenance capture, and the management and preservation of scientific data. Currently the D2I Center hosts operational services for the HathiTrust Research Center.

Northwestern University Library

As one of the leading private research libraries in the United States, Northwestern University Library serves the educational and information needs of its students and faculty as well as scholars around the world. Its

collection comprises 5 million volumes and 52 terabytes of unique digital content. It connects users to hundreds of scholarly databases and pursues an active role in educating faculty and students in how to make the best use of these resources.

Lafayette College Library

Lafayette College is a small liberal arts institution focused exclusively on undergraduate education. Lafayette's Libraries provide students with a wide range of information sources and services developed to support their educational pursuits. Lafayette Library was awarded the 2014 Excellence in Academic Libraries award from the Association of College and Research Libraries (ACRL). The award highlights the Library's commitment to digital scholarship, its pioneering of new models for collaborative content acquisition, and the impact of its Information Literacy and Special Collections programs on student learning.

University of North Carolina at Chapel Hill

The University of North Carolina at Chapel Hill Library provides collections, expertise, services, and facilities that are essential to achieving the University's mission of research, teaching, learning, and public service for the campus community, state, nation, and world. Building on a foundation of library excellence at the nation's first public university, the UNC Libraries are committed to defining and integrating new library roles, practices, partnerships, and technologies to be a leader among academic research libraries in meeting users' evolving knowledge-creation and knowledge management needs.

Investigators

PI, Harriett Green, is English and Digital Humanities Librarian and assistant professor, University Library at the University of Illinois at Urbana-Champaign. Her research focuses on the use and users of digital humanities tools and resources, development of digital scholarship services, digital pedagogy, and humanities data curation. Green earned her M.S. in Library and Information Science from the University of Illinois, and also holds a M.A. in Humanities from the University of Chicago and a B.A. in History and Literature from Harvard University.

Co-PI, Beth Sandore Namachchivaya, is the Associate University Librarian for Research, Associate Dean of Libraries and Professor at the University of Illinois Library at Urbana-Champaign, where she is responsible for research, scholarly communications, and publishing programs and services. Namachchivaya is a member of the HTRC Executive Management Committee.

Co-PI, J. Stephen Downie, is Associate Dean for Research and a Professor at the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. Downie is the Illinois Co-Director of the HathiTrust Research Center. He holds a BA (Music Theory and Composition) along with a Master's and a PhD in Library and Information Science, all earned at the University of Western Ontario, London, Canada.

Co-PI, Angela Courtney, is Head of the Arts and Humanities and Reference Services Departments, Librarian for English and American literature, and Director of the Scholars' Commons at the Herman B Wells Library, Indiana University, Bloomington. She is the co-editor of the *Victorian Women Writers Project* and *Indiana Authors*.

Co-PI, Geoffrey Welles Moore, is the Coordinator of Humanities and Social Sciences at the Northwestern University Library where he has worked since September 2005. Geoffrey is also the library liaison to the

Religious Studies department and the Linguistics department and chairs the library's collection development committee and Resource Discovery Operations Group.

Co-PI, Neil McElroy, is Dean of Libraries and co-director of the Digital Humanities Initiative at Lafayette College in Easton, PA. He is active in national organizations such as the Digital Library Federation (DLF) and the Coalition for Networked Information (CNI) and speaks often on issues pertaining to open access publishing and digital scholarship services.

Co-PI, Stewart Varner, is the Digital Scholarship Librarian at the University of North Carolina, Chapel Hill. He earned his Ph.D. in American Studies from Emory University and his Master of Library and Information Science from the University of North Texas.

Key Personnel

Eleanor Dickson is the Visiting HathiTrust Research Center Digital Humanities Specialist. She works on education and outreach initiatives for the HTRC and within the Scholarly Commons at the University of Illinois Library.

Sayan Bhattacharyya is a postdoctoral research associate (CLIR research fellow) at the HathiTrust Research Center, University of Illinois, Urbana-Champaign. His interests are in the application of digital humanities in the fields of teaching and pedagogy.

Crystal Sheu is the E-Learning Specialist at the Office of Information Literacy in the University of Illinois Library. Her professional foci are in the areas of educational technology, instructional design, training and development, and distance education.

In addition to the above-mentioned personnel, the project team will also include a project coordinator and master's level graduate research assistant. The project coordinator will assist with coordinating project activities across multiple collaborators and will serve as the key point of contact for event coordination in relation to the Train the Trainer Roadshow. The graduate research assistant will assist with finalizing curricular materials, project evaluation activities (e.g., analysis of feedback surveys), and event coordination as needed.

Advisory Board

The project team has assembled a four-member advisory board with deep expertise regarding library outreach and instruction, scholarly communication, digital humanities, and digital scholarship. The Advisory Board will be comprised of the following members:

- **Char Booth**, Director of Research, Teaching, & Learning Services at the Claremont Colleges Library
- **Francesca Giannetti**, Digital Humanities Librarian, Rutgers University Libraries
- **Miriam Posner**, Coordinator and Core Faculty, Digital Humanities Program, University of California, Los Angeles
- **Claire Stewart**, Associate University Librarian for Research and Learning, University of Minnesota Libraries

Diversity Plan

The project is critically grounded in collaborations across multiple types of institutions that range from small liberal arts colleges to large public research universities, and includes both well-resourced institutions and those with more financial constraints. The collaborating institutions have very diverse constituencies that will be targeted in the attendee recruitment for the workshop events. The events in Year Three will be promoted to listservs and organizations for underrepresented groups and other key institutions for underrepresented constituencies in digital scholarship. When distributing the final Open Educational Resource curriculum to an expected 200 or more institutions, the project leaders will make sure to target outreach to HBCUs, women's colleges, and other key institutions for underrepresented constituencies in digital scholarship.

Communications Plan

We will leverage existing HathiTrust and HTRC communications tools, including the Twitter and Facebook accounts, the HathiTrust monthly newsletter, and the HTRC listservs, which include an active HTRC User Group listserv for all researchers and students interested in working with HTRC tools. We also will create a project website for posting regular updates, workshop materials as they're produced, and workshop schedules. As the workshops are carried out, a community-based resource (e.g., wiki forum, listserv) will be hosted at Illinois to enable trained attendees to continue discussions about training at academic libraries, share experiences, and revise/reuse existing materials.

The internal communications of the project team will include use of the HathiTrust Research Center wiki for exchanging documents and communications, scheduled monthly Skype conference calls among team members, and a project webpage for posting updates for both internal and for external audiences.

We will recruit attendees by promoting events on a broad range of listservs and community discussion groups in the LIS and digital humanities communities. Targeted organizations include HASTAC, code4Lib, DLF, TEI, SHARP, and Humanist. We will also contact ACRL sections and interest groups containing members who may be interested, including the Digital Humanities Interest Group, Digital Curation Interest Group, and ACRL sections such as Literatures in English, Western European Studies, and Instruction. We will also work with conferences' organizers to promote events that are held in conjunction with major conventions and conferences, such as the DLF Forum and ACRL National Conference.

Sustainability Plan

All teaching and training materials will be deposited in a trusted repository as open educational resources that will be accessible to all. The links to materials will be maintained on a static website up to five years following the conclusion of the project. Training is part of HTRC mission and we are committed to continuing workshops and outreach beyond the grant period.

A community-based resource (e.g., wiki forum, listserv) will be hosted at Illinois to enable trained attendees to continue discussions about training at academic libraries, share experiences, and revise/reuse existing materials. We will plan to host regularly scheduled convening meet-ups, modeled on currently existing HTRC user group meetings, that will be focused on training and outreach.

References are provided as a supporting document.

DIGITAL STEWARDSHIP SUPPLEMENTARY INFORMATION FORM

Introduction:

IMLS is committed to expanding public access to IMLS-funded research, data and other digital products: the assets you create with IMLS funding require careful stewardship to protect and enhance their value. They should be freely and readily available for use and re-use by libraries, archives, museums and the public. Applying these principles to the development of digital products is not straightforward; because technology is dynamic and because we do not want to inhibit innovation, IMLS does not want to prescribe set standards and best practices that would certainly become quickly outdated. Instead, IMLS defines the outcomes your projects should achieve in a series of questions; your answers are used by IMLS staff and by expert peer reviewers to evaluate your proposal; and they will play a critical role in determining whether your grant will be funded. Together, your answers will comprise the basis for a work plan for your project, as they will address all the major components of the development process.

Instructions:

If you propose to create any type of digital product as part of your proposal, you must complete this form. IMLS defines digital products very broadly. If you are developing anything through the use of information technology – e.g., digital collections, web resources, metadata, software, data– you should assume that you need to complete this form.

Please indicate which of the following digital products you will create or collect during your project.

Check all that apply:

Every proposal creating a digital product should complete ...	Part I
If your project will create or collect ...	Then you should complete ...
<input checked="" type="checkbox"/> Digital content	Part II
<input type="checkbox"/> New software tools or applications	Part III
<input type="checkbox"/> A digital research dataset	Part IV

PART I.

A. Copyright and Intellectual Property Rights

We expect applicants to make federally funded work products widely available and usable through strategies such as publishing in open-access journals, depositing works in institutional or discipline-based repositories, and using non-restrictive licenses such as a Creative Commons license.

A.1 What will be the copyright or intellectual property status of the content you intend to create? Will you assign a Creative Commons license to the content? If so, which license will it be? <http://us.creativecommons.org/>

The primary digital content created during the project are curricular materials, which will be licensed under a [Creative Commons Attribution 4.0 International License](#) with the goal of encouraging the broad reuse that is in the educational and community-building spirit of open educational resources and the Train the Trainer program. Participants in workshops and events will also be encouraged to share derivative curricular materials with community members, and the project team will recommend adoption of the 4.0 license while also outlining alternative options for Creative Commons licensing.

We anticipate that the HTRC curriculum will refer to digital content with various access restrictions. For example, the HathiTrust data is under the purview of the HathiTrust and not available for release. Where appropriate, source code and documentation for HathiTrust Research Center tools and services are made available publicly and freely online through GitHub for download, distribution, and continued contributions and modifications. Copies of reports and other grant-funded publications will also be deposited in IDEALS, the institutional repository of the University of Illinois. The rights to any resulting publications (e.g., peer-reviewed journal articles) will be subject to negotiation with publishers, and we will advocate for authors to retain rights to maintain and share open access versions of grant-funded content.

A.2 What ownership rights will your organization assert over the new digital content, and what conditions will you impose on access and use? Explain any terms of access and conditions of use, why they are justifiable, and how you will notify potential users of the digital resources.

Copyright owners of curricular content will release materials under a [Creative Commons Attribution 4.0 International License](#), and no further ownership rights or use and access conditions will be asserted. A project website will include a public discussion forum for issues pertaining to incorporating the HTRC curriculum into library education, outreach, and training. No restrictions will be placed on who may participate in the forum. This project site will also be the home of all new curricular content generated during the course of the project.

Optimal use of curricular materials does require interaction with the HathiTrust Research Center's online environment. To use the HTRC Portal and Workset Builder, individuals must sign up for an account that will be linked to an email account affiliated with an educational institution. As part of the sign up process, users must also agree to the following statement: "I acknowledge that I am being granted access to HathiTrust (HT) texts and/or data and will not knowingly copy HT page text, HT page images, or primary bibliographic metadata (MARC catalog records) from my workspace to another computing system, nor will I remove these from the SHARC (Secure HathiTrust Analytic Research Commons) environment. I also understand that some of the data may be determined at a later date to be in copyright and that the HathiTrust Research Center (HTRC) will delete these items and any copies that have been made upon notification from the HathiTrust." The goal of the HTRC is to provide non-consumptive computational access to data that might otherwise remain unavailable. At present, the HTRC currently supports computation against digital content in the public domain, but the center expects to integrate non-consumptive access to works in copyright over the course of the project. This statement alerts users to access and use restrictions related to the underlying data.

Increasing the reach of HTRC resources is a key component of this grant proposal. Currently, HTRC contributes to a monthly HathiTrust newsletter for member institutions, and engages the public through Twitter and Facebook using the HathiTrust accounts, which have 7427 followers and 919 likes respectively. While we will continue to engage this core network through regular project posts and updates, we will also plan to engage new communities through participation at professional conferences and deeper engagement with practitioners who provide library services through scholars' commons and library-based digital humanities centers.

A.3 Will you create any content or products which may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities? If so, please describe the issues and how you plan to address them.

While we do not expect new curricular content will involve privacy concerns or raise cultural sensitivities, we will build these questions into project evaluation processes. We will also review content for proper attribution, acknowledgement, and re-use permissions prior to posting.

Part II: Projects Creating Digital Content

A. Creating New Digital Content

A.1 Describe the digital content you will create and the quantities of each type and format you will use.

Curricular materials will include a set of 5-6 training worksets (natively stored in the HTRC environment and exported as a CSV list of volume IDs), no fewer than four text-based guided tutorials and demonstrative screencasts that introduce each tool within HTRC to prospective users (shared in an easily edited textual file format and MP4 video format), and documentation for a set of 5-6 instructional activities for use in academic libraries (also shared in an easily edited textual file format).

A.2 List the equipment and software that you will use to create the content or the name of the service provider who will perform the work.

All curricular content will be constructed and distributed in Open Educational Resource formats. The University Library will setup a virtual learning environment via Blackboard Collaborate to deliver and record the project's workshops and events, using the common cartridge standard for open educational resources. Demonstrative screencasts that introduce tools within HTRC will be made using Camtasia. Any group can utilize workshop content to recreate and/or remix this content for inclusion in other curricular offerings. Additionally, all of our OER content will be highlighted within the HTRC instructional wiki for those interested in basic pedagogy around HTRC tools and services.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to create, along with the relevant information on the appropriate quality standards (e.g., resolution, sampling rate, pixel dimensions).

We anticipate that curricular materials may include the following file formats:

- CSS files for webpage customization
- HTML files for basic information and web links to resources
- JPEG or PNG files for images
- MP4 files for video
- Microsoft office application files (DOCX, XLSX, PPTX)
- MP3 files for audio recordings

Additional formats will be considered based on evaluation feedback and selected based on considerations for maintaining quality and access over time.

B. Digital Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

The project PI and the project coordinator will work together to monitor and evaluate day-to-day aspects of project activities and ensure the project team is meeting its designated milestones for deliverables. The project itself is designed to incorporate several iterations of curriculum development, delivery, assessment, and revision. One of our primary goals is to test the curriculum within a range of instructional environments, and we selected our project partners with an eye toward geographic and institutional diversity. Project partners will convene monthly project meetings, and then advisory board will also meet annually to provide outside expertise on instructional design and approaches to instruction in the digital humanities. We anticipate that the multiple layers of evaluation and revision built into the project design will provide ample opportunity for ongoing quality control. As workshops are delivered, we will also seek feedback from workshop participants to benefit from the perspective of both the instructors and their workshop attendees. Only after curricular materials have undergone rigorous assessment and revision will content be posted and distributed online.

B.2 Describe your plan for preserving and maintaining digital assets during and after the grant period (e.g., storage systems, shared repositories, technical documentation, migration planning, commitment of organizational funding for these purposes). Please note: Storage and publication after the end of the grant period may be an allowable cost.

HathiTrust Digital Library is a digital preservation repository and highly functional access platform. It provides long-term preservation and access services for public domain and in copyright content from a variety of sources, including Google, the Internet Archive, Microsoft, and in-house partner institution initiatives. The HathiTrust Research Center enables computational access for nonprofit and educational users to published works in the public domain and, in the future, on limited terms to works in copyright from the HathiTrust using a non-consumptive paradigm that protects the security of sensitive in-copyright data.

This project provides an opportunity to develop and share educational resources that leverage HathiTrust data and HTRC tools for library instruction. We will provide ongoing access to all curricular content, which will be constructed and distributed in Open Educational Resource formats, for a minimum of 5 years, but the HTRC envisions making this this a permanent component of its service model. All project documentation, working papers, technical reports, and other research material will be deposited in the IDEALS institutional repository (<http://ideals.illinois.edu>). We will also provide access to these materials from the HTRC and HT websites.

C. Metadata

C.1 Describe how you will produce metadata (e.g., technical, descriptive, administrative, preservation). Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

The Common Cartridge builds upon the widely implemented IMS Content Packaging specification, adding support for the following metadata standards:

- ISO 15836:2003: Dublin Core Metadata Element Set (mapped to the corresponding elements in LOM)
- IEEE 1484.12.1-2002: Learning Object Metadata
- IEEE 1484.12.3-2005: LOM Schema binding (loose binding)

C.2 Explain your strategy for preserving and maintaining metadata created and/or collected during your project and after the grant period.

Metadata will be preserved and maintained as part of our broader strategy for digital assets outlined in B.2

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content created during your project (e.g., an Advanced Programming Interface, contributions to the DPLA or other support to allow batch queries and retrieval of metadata).

While the project team will select a set of overarching keywords for the project website in support of search engine optimization (SEO) to facilitate online discovery, our primary strategies for facilitating discovery and use rely on deeper community engagement through our Twitter network, resources like DH+Lib, and participation in professional conferences and events.

D. Access and Use

D.1 Describe how you will make the digital content available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

The University of Illinois Library will implement a university-managed instance of the WordPress.com content management platform, made available at publish.illinois.edu, for a project website that meets all web accessibility requirements. The project curriculum will be publicly available online for anyone interested in issues concerning library outreach and instruction for digital scholarship, digital humanities, or textual analysis.

D.2 Provide URL(s) for any examples of previous digital collections or content your organization has created.

Below are a few samples of digital tools and services implemented by the HTRC:

- HTRC Portal: <https://sharc.hathitrust.org/>
- HTRC Workshop Builder: <https://sharc.hathitrust.org/blacklight>
- Extracted Features Dataset: <https://sharc.hathitrust.org/features>
- HTRC Bookworm: <http://bookworm.htrc.illinois.edu/>
- HTRC Introductory Guide: <http://uiuc.libguides.com/htrcguide>

Parts III and IV are not applicable to this grant proposal.

Part III. Projects Creating New Software Tools or Applications

A. General Information

A.1 Describe the software tool or electronic system you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) the system or tool will serve.

A.2 List other existing digital tools that wholly or partially perform the same functions, and explain how the tool or system you will create is different.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your new digital content.

B.2 Describe how the intended software or system will extend or interoperate with other existing software applications or systems.

B.3 Describe any underlying additional software or system dependencies necessary to run the new software or system you will create.

B.4 Describe the processes you will use for development documentation and for maintaining and updating technical documentation for users of the software or system.

B.5 Provide URL(s) for examples of any previous software tools or systems your organization has created.

C. Access and Use

C.1 We expect applicants seeking federal funds for software or system development to develop and release these products as open source software. What ownership rights will your organization assert over the new software or system, and what conditions will you impose on the access and use of this product? Explain any terms of access and conditions of use, why these terms or conditions are justifiable, and how you will notify potential users of the software or system.

C.2 Describe how you will make the software or system available to the public and/or its intended users.

Part IV. Projects Creating Research Data

1. Summarize the intended purpose of the research, the type of data to be collected or generated, the method for collection or generation, the approximate dates or frequency when the data will be generated or collected, and the intended use of the data collected.

2. Does the proposed research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity already been approved? If not, what is your plan for securing approval?

3. Will you collect any personally identifiable information (PII) about individuals or proprietary information about organizations? If so, detail the specific steps you will take to protect such information while you prepare the research data files for public release (e.g. data anonymization, suppression of personally identifiable information, synthetic data).

4. If you will collect additional documentation such as consent agreements along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

5. What will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

6. What documentation will you capture or create along with the dataset(s)? What standards or schema will you use? Where will the documentation be stored, and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

7. What is the plan for archiving, managing, and disseminating data after the completion of research activity?

8. Identify where you will be publicly depositing dataset(s):

Name of repository: _____

URL: _____

9. When and how frequently will you review this data management plan? How will the implementation be monitored?