Beth Sandore Namachchivaya, PI, University of Illinois

National Forum: Data Mining Research Using In-copyright and Limited-access Text Datasets: Shaping a Research and Implementation Agenda for Researchers, Libraries, and Content Providers

Abstract

With the growth of digital scholarly publishing, online repositories of digitized texts, and increased interest in data sharing and re-use, text data mining and analysis has clearly emerged as a viable research method for scholars in an increasing number of subject domains. Open source text data mining tools such as Voyant and publicly-available services such as the HathiTrust Research Center (HTRC) have brought the potential of new research discoveries through computational analytics within reach of scholars. While the tools for mining and analysis of text datasets are increasingly accessible, the texts themselves are frequently protected by copyright or other intellectual property (IP) rights, or subject to license agreements that limit access and use. These IP-related considerations can complicate a researcher's efforts to access the dataset, incorporate it into analytical research, and communicate the output and related methods to a broader audience. Increasingly, academic libraries are engaging with content providers to facilitate access to text datasets for researchers.

"Data Mining Research Using In-copyright and Limited-access Text Datasets" is proposed as a one-year national forum grant. Our project will bring together experts and thought leaders to articulate an agenda that provides guidelines for libraries to facilitate research access, implement best practices, and mitigate issues associated with methods, practice, policy, security, and replicability in research that incorporates text datasets that are subject to IP-related restrictions. We propose to convene a 1.5 day national forum in March, 2018 in Chicago, Illinois that will assemble key stakeholders to explore issues and challenges for scholars performing data mining and analysis on in-copyright and limited-access text datasets. Such a national forum will shed light on the multidimensional aspects of these approaches and promote a sufficiently broad and deep perspective on a challenge where stakeholders include libraries, content providers (e.g., commercial publishers, government agencies, Google); scholars across multiple disciplines; policy makers and legal experts; software developers; and directors of data repositories, registries and data journals. The University of Illinois Library is poised to act as the organizing body for this event due to its ongoing commitment to developing innovative services around digital scholarship, curating research data, and pursuing text data mining research initiatives in partnership with the HTRC and the National Center for Supercomputing Applications.

Text data mining and analysis methods hold strong potential to enable transformative and significant scholarly inquiry. Libraries can clearly support and facilitate this research as part of digital scholarship services. No single agency or institution can develop the policy and best practices framework for libraries to facilitate access to text datasets for research data mining. And while a number of individuals and institutions have taken steps toward building components in this framework, this forum will provide an opportunity for a more collaborative effort. The project deliverables, which include an environmental scan, structured community-driven analysis, and ACRL-published white paper, will make recommendations for best practices and policy to guide libraries as they develop text data mining services. With the support of IMLS under the rubric of the National Digital Platform theme, the proposed forum can serve to catalyze, organize, coordinate, and synthesize the conversation into a cohesive agenda that will serve as a foundation for research and practice in libraries, and in the scholarly community.

National Forum: Data Mining Research Using In-copyright and Limited-access Text Datasets: Shaping a Research and Implementation Agenda for Researchers, Libraries, and Content Providers

Introduction

"Data Mining Research Using In-copyright and Limited-access Text Datasets" is proposed as a one-year national forum grant. Our project will bring together experts and thought leaders to articulate an agenda that provides guidelines for libraries to facilitate research access, implement best practices, and mitigate issues associated with methods, practice, policy, security, and replicability in research that incorporates text datasets that are subject to intellectual property (IP) rights.[1] This effort will be led by PI Beth Sandore Namachchivaya and Co-PIs Bertram Ludäscher and Megan Senseney, along with Investigator Eleanor Dickson at the University of Illinois in collaboration with campus-wide partners in the University Library, the HathiTrust Research Center (HTRC), the School of Information Sciences' Center for Informatics Research in Science and Scholarship (CIRSS), and the National Center for Supercomputing Applications (NCSA).

We propose to convene a 1.5-day national forum in March 2018 in Chicago, Illinois that will assemble key stakeholders among librarians, researchers, and content providers to explore issues and challenges for scholars performing data mining and analysis on in-copyright and limited-access text datasets where intellectual property rights associated with these texts require that:

1. researchers obtain permission for each use, or
2. researchers perform non-consumptive research where they do not have read access to the full text corpus[2], or
3. researchers, working through the library, identify whether the content provider's licensing terms and agreements allow for use.

This request to host a national forum addresses the IMLS "National Digital Platform" priority by identifying an agenda that recommends best practices and a policy framework for libraries and other stakeholder groups to support researchers' increasing demand to integrate in-copyright and limited-access text content into their text analysis and data mining. We propose to accomplish this by convening a group of key stakeholders and scholars to explore broadly applicable approaches that support computational research with IP-restricted data and to make recommendations for best practices and policy to guide libraries as they develop text data mining services, which may include:

[1] In most cases for text data mining, intellectual property will take the form of copyright, but access to text corpora for research purposes may also be partially or completely restricted due to the commercial or industry-based nature of the data, or contractual agreements related to terms of use and service. In many cases, the sheer scale of the data required to conduct analysis may present further challenges to access for research.

[2] For a full definition of non-consumptive research, see the Statement of Need (p. 5).

- models for working with content providers to provide researchers with access to text datasets, and
- models for hosting and preserving the outputs of scholars' text data mining research in institutional repositories and databanks.

The deliverable from this national forum will be a publicly-available white paper, published by the Association of College and Research Libraries (ACRL), that includes an environmental scan, an aggregated SWOT analysis[3] that takes into account the concerns and perspectives of multiple stakeholder groups, and a set of recommendations emerging from the national forum workshop. The proposed work addresses the Agency-level Goal of supporting exemplary stewardship of museum and library collections and promoting the use of technology to facilitate discovery of knowledge and cultural heritage. More specifically, the proposed national forum seeks to contribute to the efforts around the IMLS Strategic Plan Objective 4.3: support and extend a national digital information infrastructure that leverages libraries and museums as key partners and providers of reliable, persistent, and widely available access to digital information and services. The activities leading up to and including the forum will be focused on positioning libraries as essential partners with researchers and content providers in facilitating access to digital content to support text data mining and analysis, and making recommendations for how the research output can be curated.

Statement of National Need

With the growth of digital scholarly publishing, online repositories of digitized texts, and increased interest in data sharing and re-use, text data mining and analysis has clearly emerged as a viable research method for scholars in an increasing number of subject domains. In 2013 the International Federation of Library Associations (IFLA) issued a statement on Text and Data Mining followed by an ARL Issue Brief in 2015. That same year, the Association of College & Research Libraries' (ACRL) Environmental Scan 2015 identified data mining as an emerging area of focus for library support services, suggesting that research libraries provide "rich and deep content platforms with tools that facilitate discovery and analysis" and explore "options for providing data mining functionality in aggregated databases" (ACRL, 2015, p. 15). Open source text data mining tools such as Voyant and publicly-available services such as the HTRC have brought the potential of new research discoveries through computational analytics within reach of scholars. While the tools for mining and analysis of text datasets are increasingly accessible, the texts themselves are frequently protected by copyright or other IP rights, or subject to license agreements that limit access and use. These IP and licensing considerations can complicate a researcher's efforts to access the dataset, incorporate it into analytical research, and communicate the output and related methods to a broader audience. Increasingly, academic libraries are engaging with content providers to facilitate access to text datasets for researchers. The recent focus on text data mining combined with the relative lack of clear best practices for managing IP-protected datasets indicate that the challenges associated with these issues are both timely and of national significance.

[3] A SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis is a method for evaluation and planning.

Text data mining and analysis research involves multiple discrete activities, such as selecting and assembling subsets of a larger dataset, developing and executing data mining algorithms, and analyzing the summative output from the computation. In recent research, Williams et al. (2014) identify the tremendous potential benefits of text data mining with licensed journal literature in the biomedical sciences, and they describe the experience of one U.S. academic library to assist biomedical researchers in obtaining access to both in-copyright as well as publicly-available text datasets, recommending that journal and database licensing language address the use case of text data mining. In a recent study of the text analysis needs of humanities scholars, Dickson et al. (2016) reported that participants noted that access to in-copyright texts was a “frequent obstacle” in their ability to select appropriate texts for data mining. Grewal and Huhn (2016) identify several challenges to researchers who seek to use in-copyright or other limited-access text datasets in their research, citing the fact that rights language for protected content is often unclear, with licensing agreements further obscuring access, and the researcher is often responsible for negotiating directly with the content provider. Their analysis of license language in a selection of 32 licenses from 30 publishers across a variety of disciplines found that approximately 47% of the licenses both mention and allow text data mining, while another 47% of the licenses do not mention it, and a small percentage explicitly disallow text data mining with the licensed content.

In addition to noting the high transaction cost of negotiating multiple licenses and the restrictions on mining texts, a 2012 JISC (Joint Information Systems Committee, UK) report notes that the level of interest in text data mining may not match the current level of usage because many researchers do not possess the requisite technical skills and are ill-equipped to use the computational tools, methods, and resources necessary to achieve their research goals. Finally, few if any best practices exist for researchers to effectively curate, manage, and document the methods, results, and analysis involved in text data mining. Libraries are ideally positioned to develop services and programs to support scholars in this area, particularly through programs in digital scholarship centers and through subject liaison outreach. Moreover, libraries possess the necessary experience and the relationships with content providers to bring access negotiation to a successful outcome. Our project will build upon the findings of recent research by articulating the library’s role in text data mining service provision, capturing the needs of multiple constituents, building consensus around best practices for implementing library-based services, and identifying topics requiring additional research.

The Role of Libraries: The widespread availability of digitized books and electronic journals across numerous disciplines has stimulated the development and use of computational text analysis methods, software, and practice. Through programs in digital scholarship centers and the increased efforts of e-Resource librarians to facilitate and support text data mining, libraries are positioning themselves as partners in this emerging form of text analysis research. Many scholarly communications librarians hold dual degrees in law and information science, or are experts in legal issues surrounding copyright and intellectual property. Further, with the rise of library-based research data services (Tenopir, Birch, and Allard, 2012) and the federal emphasis on access to publicly funded research data (Office of Science and Technology Policy Memo, 2013), now is a critical moment to assess broad-based approaches to both policy and practices surrounding research using text data with IP restrictions. Libraries play an integral role in both access and use of text datasets by virtue of their mission to provide research services and support to a broad academic community. Increasingly, librarians function in bridging roles, connecting researchers with the content they need, and assisting them to integrate it into their research. Libraries are also poised to navigate the tensions between the need to work within the parameters of IP-related restrictions and the recognized value of promoting transparency and reproducibility through data sharing.

A national forum will shed light on the multidimensional aspects of these approaches and promote a sufficiently broad and deep perspective on a challenge where stakeholders include libraries, content providers (e.g., commercial publishers, government agencies, Google); scholars across multiple disciplines; policy makers and legal experts; software developers; and directors of data repositories, registries and data journals. The University of Illinois Library is well positioned to act as the organizing body for this event due to its ongoing commitment to developing innovative services around digital scholarship, curating research data, and pursuing text data mining research initiatives in partnership with the HTRC and the National Center for Supercomputing Applications.

Access to IP-Restricted Data: In the absence of a clear approach for efficiently and effectively managing this process, we are vulnerable to liabilities related to IP violations, which may adversely affect individual researchers, their affiliate institutions, and content owners. Even where no IP violation occurs, we run the risk of promoting incomplete (at best) and inaccurate (at worst) research through the proliferation of text data mining projects that exclude rights-protected data. In light of intellectual property considerations, many text data mining projects have focused solely on texts that have entered into the public domain or have been flexibly licensed through initiatives such as Creative Commons or Open Data Commons. Due to the absence of practical means to work with IP-sensitive texts, researchers are artificially limiting the scope of their work to texts that do not have access restrictions, but also do not contain the desired content. This practice can create a significant skew--not only in the text data that is the accessible target of research, but it could also render the findings from text data mining less broadly applicable. For many researchers, text data mining can only reach its full potential when conducted using in-copyright or other limited-access text datasets. Addressing the question of how libraries can support researchers to conduct effective and responsible research using such datasets requires the collective input of a widely-dispersed set of stakeholder communities.

Within the last few years organizations such as CrossRef[4] have launched pilot initiatives to address and mitigate access challenges. These and similar efforts would benefit from a concerted national dialogue. The HTRC focuses directly on providing secure, non-consumptive access to public domain and in-copyright digitized texts in the HathiTrust Digital Library (HTDL), and has established a policy framework, technical infrastructure, and a training and advanced collaborative support network for researchers who seek to integrate computational text analysis and mining into their research.[5] Elsevier has also recently made available an API as part of their policy to support text data mining across selected Elsevier journals.[6] From among these and related efforts, we have identified five distinct models for providing research access to data with IP restrictions, including:

[4] http://www.crossref.org/
[5] https://www.hathitrust.org/htrc
[6] https://www.elsevier.com/about/company-information/policies/text-and-data-mining

- libraries acting as intermediaries between the researcher and the content provider to secure access to datasets;
- research centers providing access through a secure virtual machine environment (e.g., HTRC Data Capsule);
- computing centers making datasets available for high performance computing on a competitive basis using a dedicated research network and compute resources (e.g., the Extreme Science and Engineering Discovery Environment, XSEDE);
- content holders providing controlled or mediated access to datasets through a dedicated API or portal (e.g. Elsevier and JSTOR Data for Research); and
- organizations that aggregate content or maintain registries that point to content (e.g., CrossRef) brokering access and permissions for research analysis.

We propose to examine access issues for both consumptive and non-consumptive research use cases. For the purposes of this project, we consider consumptive research to include situations in which researchers gain full access to text datasets but the content owner requires them to obtain permission for each use. Emerging policy from the HTRC defines non-consumptive research as "research in which computational analysis is performed on one or more volumes or textual objects, but not research in which a researcher reads or displays substantial portions of an in-copyright or rights-restricted work to understand the expressive content presented within that work" (Dickson et al., 2017).[7]

Use and Reproducibility: Beyond the technical skills and high transaction costs associated with accessing text for data mining, working with protected datasets introduces further challenges related to research documentation and dissemination. Scholars who secure access to text corpora from one or more sources are likely to be subject to terms of use that limit redistribution of both original datasets and their derivatives, and scholars conducting non-consumptive research never gain full access to the text. As such, we expect that scholars whose research involves both consumptive and non-consumptive text data mining and analysis will typically not have the right to re-distribute the data. In communicating their research results they may also be required to maintain the security of the content without exposing significant portions of it in such a way that would enable others to reconstruct the rights-protected data. These conditions rightfully protect intellectual property but stand at odds with the federal emphasis on sharing data from publicly funded research (Office of Science Technology and Policy Memo, 2013) and increased interest in publisher mandates requiring data deposit as a condition of publication (Piwowar and Chapman, 2008; Naughton and Kernohan, 2016). By articulating alternative approaches to documenting the research process, the national forum can identify existing best practices and seek community consensus for capturing research workflows and sharing research results in the absence of complete data, thus supporting compliance with data sharing mandates while respecting the terms of use that govern access.

Better understanding the most pressing use cases for text data mining will help establish areas requiring policy development and areas where further technical research is required. Input from stakeholders across multiple communities will help determine what additional metadata or

[7] For more information on the fair use aspects of non-consumptive research, see: Case No. 05 CV 8136-DC. Amended Settlement Agreement. 2009. § 1.93. 'Non-Consumptive Research'

documentation content providers might share with researchers to support transparency throughout the research process and reproducibility of research results, what constitutes sufficient documentation of research workflows, what an archival research object might contain, and whether researchers and publishers will require innovative repository services to support these research objects. As more scholarly and scientific journals create incentives for researchers to share data by providing outlets for peer-reviewed data papers and data publications, the project team also perceives a need to determine equivalent processes for peer review of data-intensive text mining projects where data cannot be deposited publicly.

Project Design

Goals and Project Design: The overarching goal of the proposed national forum event is to articulate a research agenda for the LIS community and to provide an action framework for libraries to facilitate research access, implement best practices, and mitigate issues associated with methods, practice, policy, security, curation, and replicability in research that incorporates in-copyright and limited access text datasets with IP restrictions. Deliverables from the national forum will include a publicly-accessible project web site and a white paper published by ACRL.This one-year project will commence on July 1, 2017 and be completed by June 30, 2018, with the forum targeted for March 2018.

We have developed three tracks of activities to guide the project team to achieve this goal within the one-year time frame of the project:

- **Environmental scan and stakeholder SWOT analysis** to ground the forum and its outcomes in appropriate theory, research methods, and practice;
- **Forum planning**, including identification of participants by expertise from the library, content provider, and research communities; iterative agenda development based on results of the environmental scan and the stakeholder SWOT analysis; logistics (preliminary and on-site) and coordination;
- **Development of final deliverables**, including the publication of a white paper and submission of a final project report to IMLS.

Work Plan: This project work plan incorporates the goals, projected outcomes, and deliverables of the proposed work. The project commences with a research component that aims to gather, synthesize, and share relevant research and practice with forum participants in the form of an environmental scan discussion paper. Forum participants will be selected from libraries, the text data mining and data curation research community, and the content provider community, and the project team with work with each forum participant to complete a SWOT analysis prior to the forum. The collected set of analyses will help to shape the agenda and form the basis of some of the forum activities. The environmental scan, SWOT analyses, and forum planning will take place concurrently, between July 1, 2017 and March 30, 2018. The final phase of the project, which includes the development of the white paper and the completion of the IMLS final report, is scheduled to occur between April 1 and June 30, 2018. Each activity is described in further detail below. Biweekly meetings will ensure that the project team sustains momentum and achieves desired outcomes. A detailed schedule of completion, with an associated timeline and responsibilities for key tasks assigned to individual members of the project team, is included as an appendix.

Selection of forum participants: In the first quarter of the project, we intend to identify (via snowball sampling) and to invite to up to 25 thought leaders in North America and Europe for a 1.5-day national forum on the topic of research and library community needs, best practices, policy recommendations, and near term challenges in data mining research using text datasets that have associated IP issues. The 1.5-day duration is targeted due to budget constraints, but also aims to encourage attendance by making the most effective use of participants' time, and to elicit input prior to and following the forum. Our selection process for forum participants will target those who can commit a reasonable amount of effort to shaping the agenda, and who can contribute key perspectives to shaping a profession-wide text data mining access agenda in the following areas: library practitioners, domain scholars whose research involves text and other forms of data analysis and data mining methods, library and information science researchers; content providers; and experts in copyright and intellectual property. Participants may be drawn from a number of organizations that have developed programs or services in this area, including but not limited to: CrossRef (Text and Data Mining for Researchers);[8] Elsevier (Text and Data Mining program); JSTOR (Data for Research);[9] the HathiTrust and the HathiTrust Research Center (HTRC); the Association of Research Libraries (ARL); the Research Data Alliance (RDA)[10] working groups (e.g., Practical Policy); German TextGrid project;[11] the Canadian NovelTM project;[12] the Association of College & Research Libraries (ACRL) Digital Scholarship Centers,[13] and Digital Humanities Interest Groups;[14] and the Modern Language Association (MLA).[15]

Environmental scan and stakeholder SWOT analysis: Prior to the forum, the project team will perform an environmental scan on the topic of text analysis and data mining using in-copyright and limited access text datasets within the academic research community. This work will be further complemented by an independent study conducted by Senseney and Ludäscher in Fall 2017 on this topic, and the results of this study will be integrated back into the environmental scan. The environmental scan, which we will share with forum participants prior to the event in the form of a discussion paper, seeks to document and support community understanding of current practice, researcher needs, and to identify facilitation and support roles for libraries. The project team will also interview each forum participant prior to the event to assist them in completing a SWOT analysis. Interviews for SWOT analyses with be based on a rubric using a "scaffolding" approach to draw out specific perspectives from the sub-communities represented by forum participants (libraries, researchers, content providers). The environmental scan discussion paper and the SWOT analyses will be shared with participants in advance of the

[8] http://tdmsupport.crossref.org/researchers/
[9] http://about.jstor.org/service/data-for-research
[10] https://www.rd-alliance.org/
[11] https://textgrid.de/en
[12] https://novel-tm.ca/
[13] http://www.ala.org/acrl/aboutacrl/directoryofleadership/interestgroups/acr-igdsc
[14] http://www.ala.org/acrl/aboutacrl/directoryofleadership/interestgroups/acr-igdh
[15] https://www.mla.org/

forum for feedback via a password-protected Commons in a Box site during an open comment period. These pre-forum research activities will enable the project team, working with the forum participants, to develop an in-depth agenda based on up-to-date information and perspectives across the stakeholder groups. A draft SWOT analysis rubric is included in *Supporting Document 1*.

Forum Agenda--Examining Key Issues and Challenges, and developing a framework for action: The national forum brings together experts in the key stakeholder groups (libraries, researchers, content providers) to examine and address the challenges in obtaining access to and using in-copyright and limited access text datasets, for both scholars and libraries. For libraries, facilitating access to and appropriate curation of text data mining with rights-protected content includes: communicating with content providers; crafting internal MOUs that require researchers to honor IP restrictions and refrain from re-distributing data; ensuring data security; developing guidelines and best practices for manipulating and using text data in research, documenting data provenance, and supporting reproducibility. For researchers, these challenges include: identifying and requesting the desired data; finding and using software packages and computing environments that support data mining; getting assistance to use data mining software and other tools; and using consistent and robust workflows to curate the research output and ensure transparency and reproducibility of results. The forum agenda will include participants providing informal use cases through a walk-through exercise that relates how they intended to incorporate in-copyright and rights-restricted text datasets into their research, along with the steps they took to do so, the challenges they faced, and the facilitative support they received. From prior experience with convenings and group projects, we have found that using this format is more engaging with a small group, and more productive than the traditional "frontal lecture" presentation style.

Forum planning, logistics, and coordination: The IMLS National Forum will be held in Chicago, Illinois in March 2018. The Project Coordinator will provide planning and logistics support: managing the event website, securing the forum venue and making travel and lodging arrangements, issuing invitations to participants, and ensuring that communication with participants occurs in a timely manner. A draft agenda for the 1.5 day forum is included in *Supporting Document 1*.

Deliverables and Dissemination: Post-forum, the team will synthesize the recommendations that emerge from the forum into a framework for action, integrating this with the environmental scan document and the SWOT analyses. The final product will be released as a public white paper, published by the Association of College & Research Libraries (ACRL) on the library's role in supporting the access, use, re-use, and curation of text data with IP restrictions, including recommendations for best practices and further research in the areas of data access, analysis and replicability, documentation and data sharing, and fair use. The white paper will be deposited in the University of Illinois' IDEALS Repository (Illinois Digital Environment for Access to Learning and Scholarship).[16] The team will also produce a public project report,

[16] http://www.ideals.illinois.edu

accessible on the project web site, containing the environmental scan discussion paper, and stakeholder SWOT analyses.

Key Personnel:

Beth Sandore Namachchivaya (PI) is Associate University Librarian for Research, Associate Dean of Libraries, and Professor at the University of Illinois Library at Urbana-Champaign. She is responsible for digital scholarship research programs and services, scholarly communications and publishing, and data management and curation services. Her research focuses on innovation in digital library discovery and access, new forms of scholarship, and the development of sustainable digital curation practices. Namachchivaya is a member of the Executive Management Team of the HathiTrust Research Center, and the HathiTrust Program Steering Committee. She is a co-PI and works with a team of librarians on an IMLS Laura Bush 21st Century Librarian grant ("Digging Deeper, Reaching Further: Libraries Empowering Users to Mine the HathiTrust Digital Library Resources") who are developing a collaborative cross-institutional training program and best practices for librarians and researchers to support and engage with scholars in text analysis and data mining research. Namachchivaya will contribute 2% effort serving as administrator of the project, and overseeing the project team's collaborative arrangements and interactions partners, budget, and dissemination of the white paper and related materials to the larger professional community.

Bertram Ludäscher, (co-PI) is Director of the Center for Informatics Research in Science and Scholarship (CIRSS), and Professor at the School of Information Sciences, National Center for Supercomputing Applications, and the Department of Computer Science. He conducts research in scientific data management, scientific workflows, and data provenance. His research interests also include foundations of databases, knowledge representation, and reasoning. Ludäscher applies this work in a number of domains, e.g., biodiversity informatics and taxonomy. He will contribute 2% effort during the academic year to the project environmental scan, forum planning, and white paper development.

Megan Senseney (co-PI) is Research Scientist at the Center for Informatics Research in Science and Scholarship (CIRSS) in the University of Illinois School of Information Sciences. She received a Masters of Science in Library and Information Science from the University of Illinois at Urbana-Champaign (2008). Starting in the fall of 2017 she intends to pursue doctoral studies part time at the iSchool (beginning Fall 2017) while continuing her employment as Research Scientist. She brings practical expertise and research experience in data curation education and the intersection of data curation and the digital humanities. Senseney will contribute 30% of her time for six weeks each summer and 3% of her time during the academic year to conduct the environmental scan, assist in developing SWOT analyses, and compose the initial discussion paper and final white paper. In addition to her official allocations as a Research Scientist at the iSchool, Senseney also plans to conduct a synergistic independent study in the iSchool during fall 2017 on the topic of text mining with limited access datasets.

Eleanor Dickson (Investigator) is the Visiting HathiTrust Research Center Digital Humanities Specialist at the University of Illinois Library at Urbana-Champaign. She is leading a task force on

HTRC Non-Consumptive Use Policy for the HathiTrust Research Center. She is also engaged in education and outreach initiatives within the Scholarly Commons at the University of Illinois Library, and serves as an Investigator on the IMLS Laura Bush 21st Century Librarian grant, “Digging Deeper, Reaching Further: Libraries Empowering Users to Mine the HathiTrust Digital Library Resources.” Dickson will contribute up to 20% effort toward conducting the environmental scan, working with participants to assist in developing SWOT analyses, and composing the initial discussion paper and final white paper.

Project coordinator. The School of Information Sciences is planning a search for a full-time project coordinator in spring 2017. This project will be included in the successful candidate’s formal allocations. The Coordinator will assist with website development, general project administration, and event coordination throughout the project at an allocation of 15%. A draft copy of the job description is included in *Supporting Document 1*.

National Impact

As noted in the Statement of Need section, the recent increase in the use of text data mining methods by scholars has been highlighted as a key indicator in the growth of new library services, by individual institutions as well as library professional organizations. The proposed deliverables have the potential to enable the library, researcher, and content provider communities to identify areas where collaboration and cooperation can strengthen this emerging area of scholarly inquiry. The environmental scan, SWOT analysis, and national forum discussions will be synthesized into a white paper that communicates a comprehensive set of recommendations and a cohesive action and research framework for libraries to develop robust services and best practices and policies to facilitate the use of in-copyright and limited access text datasets with intellectual property (IP) restrictions. In developing the recommendations from the forum, the project will seek to identify institutions or groups that can assist with their implementation. The white paper, to be published by ACRL, will be both publicly accessible and highly visible through ACRL to its membership of more than 11,000 academic and research librarians and beyond.

Text data mining and analysis methods hold strong potential to enable transformative and significant scholarly inquiry. Libraries are ideally-positioned to facilitate this research as part of digital scholarship services. No single agency or institution can develop the policy and best practices framework for libraries to facilitate access to text datasets for research data mining, but a number of individuals and institutions have taken steps toward building components in this framework, as noted above. With the support of IMLS under the rubric of the National Digital Platform theme, the proposed forum can serve to catalyze, organize, coordinate, and synthesize the conversation into a cohesive agenda that will serve as a foundation for research and practice in libraries, and in the scholarly community.

References

See *Supporting Document 1*.

Letters of Support

See *Supporting Document 2*.

National Forum: Data Mining Research Using In-copyright and Limited-access Text Datasets:
Shaping a Research and Implementation Agenda for Researchers, Libraries, and Content Providers

Schedule of Completion

			1 July 2017 - 30 June 2018											
			jl	**ag**	**se**	**oc**	**no**	**de**	**ja**	**fe**	**mr**	**ap**	**my**	**je**
Category	***Activity***	***Personnel***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Environmental Scan and SWOT Analyses	Environmental Scan													
	Literature review and background research	MFS, ED, BSN, BL	▨	▨	▨	▨	▨							
	Independent Study: In-Copyright and Limited-Access Datasets	MFS, BL		▨	▨	▨	▨	▨						
	Draft initial discussion paper	MFS, ED, BSN, BL					▨	▨	▨	▨				
	Disseminate discussion paper	PC								▨				
	Pre-Forum SWOT Analyses													
	Develop interview protocols for forum participants	MFS, ED, BSN	▨											
	Apply for IRB approval	MFS, ED, BSN		▨										
	Create Commons in a Box space for sharing SWOT analyses	PC		▨										
	Conduct SWOT interviews with forum participants	MFS, ED, PC			▨	▨	▨	▨						
	Deadline for intial SWOT analyses	PC, Participants							▨					
Event Coordination	Create event website	PC	▨											
	Finalize forum date and location	BSN, PC	▨											
	Finalize participant list and send invitations	BSN, PC		▨										
	Coordinate travel planning and event logistics	PC			▨	▨	▨	▨	▨	▨				
	Conduct national forum	BSN, MFS, BL									▨			
White Paper Development	Conduct post-forum participant analysis by theme	PC, Participants										▨		
	Deadline for post-forum analyses	PC, Participants											▨	
	Draft recommendations for policy, best practices, and new initiatives	BSN, MFS, BL, ED										▨		
	Assemble final draft of white paper	BSN, MFS, BL, ED											▨	▨
	Submit final white paper to ACRL for publication	BSN												▨
	Prepare final project report for IMLS	BSN, PC											▨	▨

DIGITAL PRODUCT FORM

Introduction
The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (i.e., digital content, resources, assets, software, and datasets). The products you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. However, applying these principles to the development and management of digital products can be challenging. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions
You must provide answers to the questions in Part I. In addition, you must also complete at least one of the subsequent sections. If you intend to create or collect digital content, resources, or assets, complete Part II. If you intend to develop software, complete Part III. If you intend to create a dataset, complete Part IV.

PART I: Intellectual Property Rights and Permissions

A.1 What will be the intellectual property status of the digital products (content, resources, assets, software, or datasets) you intend to create? Who will hold the copyright(s)? How will you explain property rights and permissions to potential users (for example, by assigning a non-restrictive license such as BSD, GNU, MIT, or Creative Commons to the product)? Explain and justify your licensing selections.

Event website. A website for the national forum will be created through publish.illinois.edu under a Creative Commons Attribution (CC BY) license.

Commons in a Box space. Interim-phase versions of the discussion paper and SWOT analyses will be shared in a password protected Commons in a Box instance hosted by the Illinois Open Publishing Network through the University Library at the University of Illinois. Content shared here will be considered work in progress toward a final white paper that is not intended for redistribution. Only project team members and forum attendees will be granted access to this space.

White paper. The Association of College and Research Libraries has agreed to publish our public white paper. This white paper will include finalized versions of materials shared through Commons in a Box along with recommendations and next steps articulated as outcomes of the national forum. ACRL white papers are made freely available at http://www.ala.org/acrl/issues/whitepapers and we intend to negotiate an open license. A final author’s version of the manuscript will also be openly deposited in IDEALS after the print version of the white paper has been released by ACRL.

Supplementary materials. Any supplementary materials that are created during the course of the grant (presentations, interview protocols, etc.) will be deposited in IDEALS, the University of Illinois’ institutional repository under a Creative Commons Attribution (CC BY) license.

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

In October 2015 the University of Illinois at Urbana-Champaign campus approved a University Policy on Open Access to Research Articles. Authors will retain copyright to all materials and for the purpose of making scholarly articles widely and freely available in an open access repository, will grant the University a nonexclusive, irrevocable, worldwide license.

.

A.3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

To encourage open dialogue during planning phases, access to Commons in a Box will be password protected and limited to the project team and forum attendees. While we do not anticipate an privacy concerns, protocols for interviewing participants as part of SWOT analysis planning will be reviewed and approved by the University of Illinois Institutional Review Board. We will also communicate intellectual property rights and permissions with all forum participants who contribute content to the final white paper.

Part II: Projects Creating or Collecting Digital Content, Resources, or Assets

A. Creating or Collecting New Digital Content, Resources, or Assets

A.1 Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and format you will use.

Event website. A website for the national forum will be created through publish.illinois.edu, powered by WordPress.

Commons in a Box space. Interim-phase versions of the environmental scan, discussion paper and (~20) SWOT analyses will be shared in a password-protected Commons in a Box instance hosted by the Illinois Open Publishing Network through the University Library at the University of Illinois. Content shared here will be considered work in progress toward a final white paper that is not intended for redistribution.

White paper. The Association of College and Research Libraries has agreed to publish our public white paper. This white paper will include finalized versions of materials shared through Commons in a Box along with recommendations and next steps articulated as outcomes of the national forum. ACRL white papers are made freely available at http://www.ala.org/acrl/issues/whitepapers.

Final Report. A final project report will be submitted to IMLS, and a copy of the report will be made publicly available through IDEALS, the University of Illinois' institutional repository

Supplementary materials. Any further supplementary materials that are created during the course of the grant (presentations, interview protocols, etc.) will be deposited in IDEALS, the University of Illinois' institutional repository.

Final versions of all materials will be deposited and disseminated in PDF format.

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

The event website will be created using a WordPress platform hosted at publish.illinois.edu. Further visuals for the website may be created using tools such as the Adobe Creative Suite. Word processing for the environmental scan, discussion paper, SWOT analyses, white paper, and report may be completed in one of several applications (e.g., Microsoft Word, Google Drive, etc.), and visuals will be created as needed in Microsoft Excel, the Adobe Creative Suite, or other applications suitable for communicating data. Final, public versions will be saved in PDF format.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to use, along with the relevant information about the appropriate quality standards (e.g., resolution, sampling rate, or 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
- Microsoft office application files (DOCX, XLSX, PPTX)
- PDF files for final versions of publications

B. 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 coordinator will be responsible for monitoring and evaluating workflows and projects. The project coordinator will report to the project PI, who assumes ultimate responsibility for resolving problems related to quality control. The project coordinator's responsibilities will include providing sufficient templates and scaffolding to ensure continuity across participant-supplied content; providing editorial support for web content, presentations, and papers; and ensuring compliance with depositing materials in appropriate formats with complete metadata to IDEALS.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

During the project period, public materials will be shared via a project website for the national forum will be created through publish.illinois.edu, powered by WordPress. The website will remain publicly available for at least three years after the date of the forum. Interim phase materials will be shared with forum participants via a password-protected Commons in a Box instance hosted by the Illinois Open Publishing Network through the University Library at the University of Illinois. Content shared here will be considered work in progress toward a final white paper that is not intended for redistribution. Interim-phase materials will be preserved for at least one year after the completion of the project, and all final versions of all project materials will be deposited in IDEALS, the University of Illinois' institutional repository for long-term storage and preservation.

C. Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata. 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).

For all final project materials, the project team will comply with the IDEALS metadata policy, which includes a selection of required and optional Dublic Core metadata elements. A copy of the policy is available at https://wiki.illinois.edu/wiki/display/IDEALS/Metadata+Policy.

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

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, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable 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 the Association of College and Research Libraries, the HathiTrust, and the CrossRef membership organization. We will also leverage press releases and social media strategies to publicize the national forum and all related publications.

D. Access and Use

D.1 Describe how you will make the digital content, resources, or assets 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. Final versions of project disseminations will be publicly available online for anyone interested in issues concerning text data mining and intellectual property rights, and a copy of the white paper will be published by ACRL and shared online at http://www.ala.org/acrl/issues/whitepapers. Interim versions of materials leading up to the final version will be shared with forum participants via a password-protected Commons in a Box instance hosted by the Illinois Open Publishing Network through the University Library at the University of Illinois.

D.2 Provide the name(s) and URL(s) (Uniform Resource Locator) for any examples of previous digital content, resources, or assets your organization has created.

Namachchivaya has been instrumental in developing University Library's digitization and digital preservation programs. The Library's digital collections are are linked and described at:

- http://guides.library.illinois.edu/digitalcollections

Namachchivaya and Senseney have been involved with several projects associated with the HathiTrust Research Center. Below are a few samples of web-based resources from our prior work:

- http://teach.htrc.illinois.edu/
- http://worksets.htrc.illinois.edu/worksets/
- More information about the HTRC is available at: https://analytics.hathitrust.org/

Senseney is also currently involved with a Mellon-funded digital publishing project called Publishing Without Walls:

- http://publishingwithoutwalls.illinois.edu/

Ludäscher's current projects include the NSF-funded Whole Tale, SKOPE, and Kurator projects:

- http://wholetale.org/
- https://www.openskope.org/

- http://wiki.datakurator.org/wiki/web/Kurator

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

Part III. Projects Developing Software

A. General Information

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

A.2 List other existing software that wholly or partially performs the same functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software and explain why you chose them.

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

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

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

B.5 Provide the name(s) and URL(s) for examples of any previous software your organization has created.

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under open-source licenses to maximize access and promote reuse. What ownership rights will your organization assert over the software you intend to create, and what conditions will you impose on its access and use? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain and justify any prohibitive terms or conditions of use or access and detail how you will notify potential users about relevant terms and conditions.

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

C.3 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:

URL:

Part IV: Projects Creating Datasets

A.1 Identify the type of data you plan to collect or generate, and the purpose or intended use to which you expect it to be put. Describe the method(s) you will use and the approximate dates or intervals at which you will collect or generate it.

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

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

A.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.

A.5 What methods 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).

A.6 What documentation (e.g., data documentation, codebooks) will you capture or create along with the dataset(s)? 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?

A.7 What is your plan for archiving, managing, and disseminating data after the completion of the award-funded project?

A.8 Identify where you will deposit the dataset(s):

Name of repository:

URL:

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