National Forum: Responsible Research Practice and Policy for Using Restricted Access Datasets: Shaping a Research and Implementation Agenda for Researchers, Libraries, and Content Providers

Introduction

"Responsible Research Practice and Policy for Using Restricted Access Datasets" is proposed as a one year national forum grant ($98,667) to bring together experts and thought leaders to articulate a collective research agenda and current set of best practices around methods, practice, policy, security, and replicability for scholars and organizations invested in research using restricted access datasets. This effort will be led by PI Beth Sandore Namachchivaya and Co-PIs Bertram Ludäscher and Megan Senseney at the University of Illinois in collaboration with campus-wide partners in the University Library Research Data Service, 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).

The investigators are requesting a national forum grant to bring together key stakeholders to explore issues and challenges for scholars performing data mining and analysis on restricted access datasets, with a particular emphasis on "non-consumptive" research environments in which researchers never have full access to the data in question (for more information, see: Case No. 05 CV 8136-DC. Amended Settlement Agreement. 2009. § 1.93. 'Non-Consumptive Research'). The proposed forum will bring together domain, information science, and legal scholars; library and archives practitioners; and content providers to develop a shared research agenda and best practices for working with restricted datasets. It will also explore content and service provider models for provisioning access to scholars in light of restrictions as well as library and archives models for hosting and preserving the outputs of scholars' non-consumptive research in institutional repositories and databanks. This request to host a national forum addresses the IMLS "National Digital Platform" priority by convening a group of key stakeholders and scholars to explore broadly applicable approaches to supporting computational research with restricted data and making recommendations for best practices and policy based on the complex landscape surrounding these data.

Statement of Need

Access to a wide range of data fit for research purposes may be partially or completely restricted due to copyright or intellectual property concerns, protective policies around health and human subjects research, the commercial or industry-based nature of the data, or the sheer scale of the data required to conduct analysis. Addressing the question of how to conduct responsible research using such datasets requires the collective input of a widely-dispersed set of stakeholder communities. Several models exist for providing research access to restricted data, including: the use of an isolated physical facility (e.g., census data centers); provision of access through a secure virtual machine environment (e.g., HathiTrust Research Center Data Capsule); libraries as intermediaries, securing access to datasets from commercial content providers on behalf of researchers; computing centers making datasets available for high performance computing on a competitive basis using a dedicated research network and compute resources (e.g., XSEDE); and content holders providing controlled or mediated access to datasets through a dedicated API or portal.

We are particularly interested in situations where the content owner requires researchers to "mine" the data without access to the complete dataset over the course of the research project. While data mining and analysis research is necessarily complex, scholars whose research involves the analysis and mining of restricted datasets face stringent requirements to maintain the security of the content without re-distributing or exposing significant portions of said content in such a way that would enable others to reconstruct the rights-protected data. 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 the case of text data mining, for example, a 2012 JISC report makes a strong case for the benefits of data mining to support research, but notes that the level of interest may not match the current level of usage because many researchers do not possess the requisite technical skills, nor do they have the skill to deal with the high transaction cost of negotiating multiple licenses and the restrictions on mining texts.

With the rise of library-based research data services (Tenopir, Birch, & Allard 2012) and the federal emphasis on access to publicly funded research data (OSTP Memo 2013), now is a critical moment to assess broad-based approaches to both policy and practices surrounding research using restricted data. A national forum will shed light on the multi-dimensional aspects of these approaches and enable us to convene key stakeholders across disciplines and data types to develop a sufficiently broad and deep perspective on this challenge. Stakeholders would include content providers (e.g., commercial publishers, government agencies, Google); scholars across multiple disciplines; policy makers and legal experts; software developers; and data repositories (e.g., ICPSR). Libraries are inherently interdisciplinary, and librarians function increasingly 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 protecting sensitive content and promoting transparency through data sharing. The University of Illinois Library is ripe to act as the organizing body for this event due to its ongoing commitment to developing innovative services around research data and pursuing data-intensive research initiatives in partnership with the HTRC and the NCSA.

Project Design

This one-year project will commence on 1 July 2017. The project team will begin by identifying and inviting to up to 25 thought leaders in North America and Europe for a **1.5-day national forum** on the topic of responsible best practices in the use of restricted access datasets in research. Participants will be identified in the following areas: domain scholars, whose research involves text and other forms of data analysis and data mining methods working with large corpora; library and information science researchers; library and archives practitioners; content providers; and experts in copyright and intellectual property. Prior to the forum, the project team will perform an **environmental scan** of the landscape of restricted access data and its use within the academic research community and share a discussion paper with attendees. Attendees will, in turn, be asked to prepare a brief **SWOT analysis** from the perspective of the stakeholder community they represent. During the forum, participants will examine and address the challenges in using restricted access datasets in research including: challenges to ensuring data security; guidelines for data discovery and manipulation; best practices for documenting data provenance; identification of best practices for replicability of research. Following the forum, the team will release a **public white paper** on the use, re-use, and curation of restricted access data in non-consumptive research environments, including recommendations for best practices and further development in the areas of data access, analysis and replicability, documentation and data sharing, and fair use. The team will also produce a **public project report** containing the environmental scan, discussion paper, and stakeholder SWOT analyses.

Goals, Outcomes, and Impact

The goal of this national forum is to develop an agenda that will guide future research and development related to the creation of secure infrastructures and services for mining restricted datasets using standards-based technologies, data, and metadata. We intend for this meeting to result in a set of recommendations and an agenda for: developing best practices for access to content in restricted datasets; documenting reproducibility of non-consumptive research through consistent, recognizable workflows and results; establishing secure computing and analysis environments; and outlining the legal challenges of handling, access to, and use of restricted access datasets in a research setting. We anticipate a range of research, technical development, and service implementation projects to emerge from the recommendations and research agenda established during the national forum.

We anticipate this work will be of interest not only to those involved in curating and managing scholarly output, but also to scholars who use non-consumptive data mining as a core component of their research as well as organizations that maintain and manage restricted data. The Association of Research Libraries, in its Code of Best Practices in Fair Use, describes non-consumptive research and other uses as an "emerging phenomenon at many libraries" and considers the development of databases that support non-consumptive research to be "highly transformative" in the research process. To ensure full participation and engagement within the community of research librarians, we also propose to coordinate a remote participation option for interested academic librarians and archives professionals.

Budget

We request $98,667 to fund this initiative. Salaries and wages ($21,920) include .1 and .25 FTE allocations for key personnel Megan Senseney and Eleanor Dickson to conduct an environmental scan in advance of the workshop, prepare and distribute a discussion paper to stakeholders, and compile final report based on forum outcomes; a .15 FTE project coordinator to support planning logistics; and two graduate hourlies for on-the-ground forum assistance. Fringe benefits ($9,371) will be applied at a rate of 44.45%. $36,720 will support travel for project team members and 25 stakeholders to attend the forum. $8,900 will cover event space rental, A/V, and event materials and supplies. Indirect costs are budgeted at 28.6% which is $21,756 of modified total direct costs.

A complete list of references and project team biosketeches are available at http://bit.ly/2bTdBVD