Data Capsule Appliance For Research Analysis Of Restricted And Sensitive Data In Academic Libraries

The School of Informatics and Computing (SoIC) at Indiana University requests $279,313 from IMLS to fund research that evaluates a secure environment for data analytics and stewardship and enables libraries to expand their services to restricted and sensitive data (RSD). The proposed project will evaluate the feasibility of a plug-in appliance solution by which libraries enable computational analysis across a wide range of their special collections that can be restricted or sensitive due to copyright limitations or can be difficult to use due to their volume and heterogeneity. The project will engage library and developer communities in the discussions on policy approaches to governing RSD access, analysis, curation, and publication.

Statement of National Need and Impact

In the recent years, university libraries began to survey the landscape of RSD at their institutions. As a result the secure data enclave pilots emerged to allow researchers to work with RSD (Bose, 2014; Stiles et al, 2013). Such enclaves, however, often are restricted to analysis of microdata through statistical packages. They support a limited set of licensed software packages and embed human verification of researchers' data output in the workflow. Such approaches are difficult to scale to big data mining; they restrict the types of tools that can be used, and work for a limited number of types of digitized materials.

In this project we aim to build upon the advances in secure analytics made by the HathiTrust Research Center (HTRC) through its early work on Data Capsules (Zeng *et al*. 2014). Data Capsules work within the HathiTrust by providing a researcher with a dedicated, secure virtual machine (VM) that runs within a secure trust ring at HTRC. A data capsule allows a researcher to install her analysis tools through the network, but when analysis begins, the ports are closed and extraction of full content of the HathiTrust digital library collections is disabled. The solution could be generalized to any digital collection if 1) the Data Capsule were packaged as a software appliance that can be plugged into the library environment, 2) connections to all data sources were encrypted so that libraries could use their own computational resources and their own hidden, restricted, or sensitive special collections.

Enabling analytical access to libraries' special collections through Data Capsule solution will increase innovative research and educational uses of such collections and showcase them via summaries and visualizations. The focus on stakeholders and community engagement will allow to achieve greater impact that will go beyond efforts of individual libraries. The national impact of this work will be in extensibility of the framework to many types of collections. The enhancements to the open source Data Capsule tools as well as the methods of usability testing will be available to all interested in implementing it for their collections.

Project Design

Drawing on the most pressing themes of trust, access, infrastructure, and skills in providing data services (Woollard, 2016), we seek to understand the current library needs and practices in providing a Data Capsule appliance for special collection services, develop mechanisms that enable smooth and simple and yet secure computational access to restricted data in libraries, identify gaps in skills needed for librarians to enable secure data analytics and provide resources that can address those gaps. The project's goals include empowering librarians to become partners in projects that involve RSD at their institutions and develop technical capacity and human expertise to transform academic libraries' research data services and create collaborative data-intensive spaces that support data throughout the full lifecycle, from storage to analytics to publishing and re-use.

We will work closely with the community and combine our technical, social science and librarian expertise in developing the framework through the following:

1. **Assessment** of RSD, policies that could restrict use and other requirements across academic libraries interested in a Data Capsule appliance through stakeholder identification and coordination workshops that include librarians, IT security and data security officers, IRB, and so on.

2. **Prototype and packaging** of the Data Capsule system into an appliance that enables user-driven analysis and maximizes flexibility of the researcher in their own tools. Key activities: packaging, software extensions in response to assessment and guaranteeing a strong trust model in this new domain.

3. **Usability study** of 3-5 academic libraries who will be beta testers of the packaged Data Capsule solution, using various types of RSD and analytical needs (e.g., statistical analysis, text mining, visualization). Key activities: profiling and identifying libraries as use cases, conducting usability tests.

4. **Participatory design and training** to engage librarians and technologists in adapting the framework to their needs, create mechanisms for training, participation and feedback, and assist in capacity building. Key activities: develop instructional materials and use cases, develop and conduct surveys for evaluation, conduct user workshops to involve communities early in the design processes and develop user agreements and other policy documents.

Key Personnel

Project directors will include librarians, archivists, and faculty at the Indiana University School of Informatics and Computing (SoIC) and the IU Libraries (IUL) as well as key personnel of the HathiTrust Research Center (HTRC). PI Beth Plale, Director Data to Insight Center, HTRC Co-PI and Professor of Informatics will lead the project. Project Co-PI Inna Kouper (Research Faculty) will lead project assessment and user participatory design methods from partner libraries. Sr. Personnel Robert H. McDonald, Assoc. Dean for Research and Technology Strategy (IUL) will coordinate with HTRC technical staff and partner libraries to integrate testing with appropriate born digital collections. Sr. Personnel Rachel Hancock (IU Archives-Modern Political Papers) will work to test digitized political papers collection held in the IU Archives Modern Political Papers Collection. Sr. Personnel J. Stephen Downie will coordinate expertise on the workset, a key notion in large-scale, distant reading analysis.

Outcomes and Relevance to the National Digital Platform

As libraries grapple with increasing amounts of digital content, much of it restricted by copyright or other limitations, and faculty asking for tools to explore massive digital collections, libraries need novel advances to satisfy this growing need. But a library needs reliable solutions that fit into their IT architecture. This study will explore whether the Data Capsule solution used in the HathiTrust Research Center can be packaged to be a pluggable appliance that is easy to install and use, while retaining its strong trust model and guaranteeing confidentiality, security, and trustworthiness of their contents. A National Digital Platform could be the home of such solutions and others on an ongoing basis.

The project's findings will contribute to a deeper understanding of the needs across libraries for large-scale text mining of library collections. The study carried out on the porting of data capsule will inform others looking at similar solutions. Recommendations from this project can be adopted in diverse library settings; the surveys and community building efforts can bring together many stakeholders in data, including researchers, librarians, university administrators, and funding agencies. Results of the project will be disseminated through multiple professional, academic, and social media channels.

Schedule of Completion and Budget Estimate

Researchers estimate a 24-month project (June 1, 2017 – May 30, 2019). The first summer of the effort will engage the IU library in assessment and the prototyping effort. The prototyping effort will then proceed alongside the participant study efforts for the next 15 months in a staged, results oriented process of rapid feedback between the thrusts. The final 6 months of the project will focus on preparation and presentation of results.

Total budget request is $279,313 comprised of indirect costs of $89,925 (56% on year 1; 57.5% on year 2) on direct costs of $189,388. PI, co-PI, salaries and fringes are $66,367. Graduate student and hourly student salary, fringe, health insurance and fee remissions are $79,266. Travel for collaboration and dissemination of research is $2000 and research/computer supplies are $1000. University of Illinois-Urbana Champaign is subcontractor with budget of $40,755 to cover salary, fringe, travel for collaboration meetings, and overhead.

References

Bose, R. (2013). Implementing a Secure Data Enclave with Columbia University Central Resources. Presented at IASSIST-13, May 29, 2013. Available at http://www.iassistdata.org/conferences/2013/presentation/3647

Stiles, J., Church, J., Smith, E., & Elings, M. UC Berkeley Library: Report of the Restricted Use Data Task Force. Available at http://www.lib.berkeley.edu/AboutLibrary/reports/Restricted-Use-Data-Task-Force-Report0914.pdf

Woollard, M. (2016). Embracing the 'Data Revolution': Opportunities and Challenges for Research. Keynote at IASSIST-16, May 31-June 3, 2016. Available at http://blog.ukdataservice.ac.uk/embracing-the-data-revolution-opportunities-and-challenges-for-research/

Zeng, J., Ruan, G., Crowell, A., Prakash, A., & Plale, B. (2014). Cloud Computing Data Capsules for Non-consumptive Use of Texts. Paper presented at *5th Workshop on Scientific Cloud Computing, co-located with ACM High Performance Distributed Computing (HPDC),* June 2014, Vancouver, CA.