LG-72-18-0102-18

Johns Hopkins University Data Rescue Infrastructure Toolkit

Overview

The Sheridan Libraries and Museums at Johns Hopkins requests $50,000 to bring together librarians and stakeholders from across the nation for an in depth planning session with the express purpose of creating a shared data rescue infrastructure toolkit and a network of support for coordinating government data preservation activities. The toolkit will contain the components needed for libraries and organizations to participate effectively in the widespread effort. Building on our pilot presented to ARL in summer 2017, this planning phase will provide the necessary steps to coordinate the items included in the toolkit and to structure the communication network to provide sustainability to the nation-wide effort.

Statement of National Need

Government Data Needs

In winter 2017 changes for scientific research agencies revealed the potential for depreciation or disappearance of agency-managed data if large changes to support and infrastructure were to occur. Data professionals, researchers, and those who rely on continued analysis of the data to inform decisions, took notice of this gap in our national capacity to provide sustained access, and are actively seeking ways to fulfill this need.

Summary of national efforts

Recent national contributions from diverse groups reflect varying levels of resources, infrastructure, and expertise. A major participant, Data Refuge[1], helped spread the word about hosting community events. Arising soon after, Libraries+ Network[2] engaged library professionals in the discussion. Many groups are producing specific tools for preservation activities and new projects continuously appear. Coordination of these efforts is essential in building a national capacity to ensure continued quality access for everyone involved.

Summary of JHU work so far

This spring, JHU Libraries connected with Data Rescue Boulder, a network of volunteers working with over 3 million datasets. They recognized a need for institutional infrastructure and JHU offered capabilities to transfer, describe, and preserve data. In May 2017 we presented a pilot workflow to the ARL community. The model leverages libraries as providers of infrastructure and guidance connecting potential participants to those who know the data best. In July 2017 JHU hosted a meeting of data professionals where we delineated the necessity of a unified system for preservation efforts across communities. The creation of a data rescue infrastructure toolkit is a way for libraries and organizations to participate in differing capacities and to guide and advise volunteers, both within and without their local communities.

Project Design and Deliverables

[1] http://www.ppehlab.org/datarefuge
[2] https://libraries.network/

Development of longer term solutions for government data requires steady, collaborative commitment. Our project will bring together existing communities dedicated to preserving government data to develop a toolkit to coordinate and aid in preservation efforts. We will hold a meeting of stakeholders in Fall 2018 to develop a suite of social and technical resources to facilitate libraries' participation in the network of government data preservation activities including: an inventory of existing tools; a directory of individuals and organizations; and information on metadata. It will provide information that libraries and other organizations need to connect users interested in volunteering with entry points into existing efforts.

During the meeting we will also investigate ways that libraries can partner with existing efforts to provide preservation infrastructure through a network of nodes utilizing existing technologies such as the Open Science Framework (OSF) and Fedora.[3] Collaborators at the planning meeting will consider use of the InterPlanetary File System (IPFS) for the transfer of data files.[4] After the meeting, a data management consultant at JHU will evaluate and test tools and workflows. Once we establish the viability of workflow and technical components, we will propose a National Leadership Project Grant to implement the plan.

July 2017 Meeting Collaborators

Brian Geiger - Center for Open Science; Joan Saez - Data Rescue Boulder; Brendan O'Brien - Environmental Data Governance Initiative; Jeremy Friesen & James Ng - University of Notre Dame; Justin Schell - University of Michigan; Lynn Yarmey - Research Data Alliance; Robert Olendorf - Penn State University; Matt Zumwalt - InterPlanetary File System/Data Together; David Wilcox - Duraspace; Ruth Duerr - Ronin Institute & Earth Science Information Partners

National Impact

We see a burgeoning need to develop and deploy a solution for the long-term storage and access of data produced and maintained by federal government agencies. The JHU library is well positioned to expand the National Digital Platform by facilitating coordinated involvement among libraries and other organizations to connect and support the robust efforts already underway. This planning proposal represents an integral step to sustaining access to essential data across disciplines. The outcome of this collaboration, where libraries provide technological and social infrastructure, has the potential to endure as a model for other data related needs. It arises from current uncertainty, but gets at deeper requirements of national data preservation that will continue over time.

Budget Summary: $50,000

Our estimated budget includes $20,000 to host the in-person meeting at JHU and cover travel and expenses for collaborators. This figure is generated from actual cost information of the July workshop hosted at JHU. It also includes $30,000 salary and indirect costs for the Data Services Manager and a Data Management Consultant to oversee and work on the project.

---

[3] https://osf.io; https://wiki.duraspace.org/display/FF/Design+-+API+Extension+Architecture

[4] https://ipfs.io