Building Infrastructure & Integrations for Open Data Preservation and Access

Abstract: The Internet Archive (IA) and the Center for Open Science (COS) propose a two-year National Leadership project grant in the National Digital Infrastructures and Initiatives category for $247,500 to prototype innovative social and technical work supporting Open Science data curation, preservation, and access by libraries and archives. The project will explore systems integrations for the preservation of open data, starting with the registrations data in Open Science Framework (OSF). OSF is COS's free, web-based platform that supports the data collection and collaboration that underpins research projects -- data of archival value to libraries. The project will also pursue technical work for distribution of this data across additional preservation networks to expand its availability to librarians for curation. Lastly, the project will provide related training to a cohort of data stewards, conduct exploratory work to include additional OSF datasets, and test library services supporting bulk access to this data for computational analysis by researchers. These deliverables will develop expertise, prototype interoperability, and expand access methods for an institutionally and technically distributed open data network. The work enables research reproducibility, distributed preservation, and perpetual access, with the goal of uniting researchers and data librarians in the broader mission of open data archiving.

Statement of National Need: University and research libraries are evaluated, in part, on their ability to attest to responsible stewardship of the research outputs of their institution. Significant headwinds and challenges complicate the ability of libraries to accomplish this mandate. Community debate continues over the efficacy of library institutional repositories (IRs) for stewarding research outputs. This focus on the role of library IRs comes at a time of growing commercial monopolization of the service ecosystems underpinning research and stewardship. For-profit commercial conglomerates have moved aggressively into purchasing repository providers like Bepress and consolidating their holdings of services around analytics, discovery, and access. The same cost-escalation of journal subscription pricing that has deleteriously impacted libraries now threatens to consume other critical services around research and data preservation and access. Likewise, requirements for researcher data deposit in locally-administered preservation and access systems are scarce or unenforced. Few integrations exist between popular open data publication and sharing platforms and library technologies while, at the same time, librarians' existing curatorial approaches struggle to meet the growing volume of research outputs. Exacerbating these challenges is the broad movement of core infrastructure to commercial cloud services, a trend that has the potential for infrastructure to follow the same financially unsustainable path as commercial journal subscriptions. Advancing open data preservation and access requires innovative, scalable, collaborative, social, and technical networks that leverage the joint power of integrating successful platforms to provide mission-aligned, non-profit services at the intersection of open research and open data stewardship. This project aims to address those needs through a mix of technical work, training, and prototyping new access mechanisms.

Project Design & Activities: OSF provides free tools for research registrations -- a technology solution that helps researchers to maintain research integrity. To date, over 19,000 projects have been registered on OSF. The Internet Archive is a non-profit, global-scale digital library that offers free and at-cost services and infrastructure for "universal access to all knowledge" via the preservation and online accessibility of billions of items of cultural and scholarly importance. This proposal aims to build the systems and expertise to pair these successful platforms in order to further the ability of librarians to steward open research data. As such, the project design features a mix of technical work, R&D to test additional joint services, and training to accelerate adoption by the community. All project code and training curriculum will be released under open-source license and shared publicly.

Phase 1: Archive registrations to Internet Archive: Phase 1 will focus on the engineering for automated transfer of registration data to IA's preservation repository, including metadata mapping and systems integration testing, and deployment. On the OSF, registration users will be able to archive their registration on IA. A stretch goal will be the matching of storage region locations on OSF side with IA's system to assist researchers with data storage locations to meet requirements.

Phase 2: Workflows & Training Development: COS and IA will move Phase 1 work into production and improve workflows to match any new data requirements. COS will develop training curriculum for research librarians on using OSF registration and connected preservation repositories so they can support researchers. At least two free trainings will be held around the U.S. alongside relevant conferences such as RDAP or ACRL. COS will also produce webinar content in support of librarians.

Phase 3: Data Distribution & Training Delivery: In this phase, IA will prototype methods for sharing the archived OSF data via additional preservation networks, such as LOCKSS and others. This work will test replication across a variety of additional preservation platforms with the intent of providing data librarians multiple curatorial options for harvesting-for-custody and sharing-for-reuse the open data under their stewardship. COS will lead the effort to organize and deliver the project's set of deliverables for in-person workshops and online training materials.

Phase 4: Computational Access, Continued Training, Exploratory Data: This phase will focus on publishing the project's technical documentation and training materials as well as additional in-person trainings. IA will facilitate bulk access to the OSF registrations data to promote innovative access methods in support of data mining and computational research services. Both COS and IA will roadmap and possibly pursue expanding the project's work to include other open data in OSF (such as preprints) to incorporate more of the research and preservation lifecycle.

Outcomes & National Impact: We expect this project to lay the groundwork, via both production- and prototype-level technical work, to build open infrastructure for open data archiving and to link this infrastructure to additional methods of distributed data preservation and large-scale computational analysis research services. The project will be structured for scalability to include additional data sources beyond the OSF registrations dataset and will be framed to lead to systemic change in how libraries approach preservation infrastructure, acquire preservation-mandated research data, and make it accessible for reuse. Conducting training, including the development of open education resources, webinars, and similar materials will ensure data librarians can incorporate the project deliverables in their workflows. The project's engineering work will establish an IA-COS integration that will continue beyond the funding period and the project's prototype development work will enable continued collaboration efforts and robust, open services that offer an alternative to for-profit services. Overall, the project will transform the practice of data stewardship in Open Science by building the networks and infrastructure that lower the barriers of preservation and access for researchers and librarians.

Budget Summary: The total amount requested is $247,500 and includes $185,500 in salary and benefit costs for engineering and project/training management ($90,200 for IA; $95,300 for COS). Travel costs are $12,000 to cover 12 total trips (6 each for IA and COS at $1,000 per trip) for IA and COS development meetings and outreach activities. Training development is budgeted at $25,000. Partnerships with distributed preservation service partners is budgeted at $20,000 and $5,000 is budgeted for the project's storage and infrastructure costs.