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 aims to leverage the intersection between open data underpinning the research process, the long-term stewardship activities of libraries, and collaborative, distributed data sharing and preservation networks. By focusing on these three areas of work, the project will test and implement infrastructure for improved data sharing and access in further support of open science and data curation.

Few integrations exist between popular open data platforms like the Open Science Framework (OSF) of COS, large scale digital archives like IA, and collaborative preservation networks. At the same time, librarians' existing curatorial approaches struggle to meet the growing volume of research outputs -- struggles that better systems integrations and automations, along with targeted training, can help alleviate. This project aims to pair mission-aligned, non-profit services working in the sectors of open research and open data stewardship to develop tools and workflow to give libraries the power to scale their ability to select, preserve, and provide access to research data. The project will ensure the outputs of open science are discoverable in the collections maintained by libraries. OSF supports the research lifecycle by enabling researchers to produce and manage registrations and data artifacts for further curation to foster adoption and discovery. Preservation of these artifacts on Internet Archive and other networks allows for long term stewardship. Building on IA's services for data preservation and global access expands this project's works to test collaborative infrastructure across a broader community of libraries. Explicit goals of the project are:

- Further the ability of librarians to steward open research data and provide enhanced discoverability of data for researchers.
    - Develop systems integrations for the preservation of open data, starting with having the registrations data in OSF mirrored at Internet Archive.
- Accelerate adoption by the community and increase the volume of data published in the sciences.
    - Provide 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.
    - Integrate the long-term storage of research projects into their publication processes.
- Reduce the cost and risk of data disappearance and test networked infrastructure.
    - Prototype a distributed design wherein open science data is replicated from IA to a number of partner libraries to test distributed, networked preservation approaches.

We expect this project to lay the groundwork, via both production-level and prototype-level technical work along with professional training, to build open infrastructure for open data archiving and to link this infrastructure to additional methods of distributed data preservation and and access. 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 in different ways. Conducting training, including the development of open education resources, webinars, and similar materials will ensure data librarians can incorporate the project deliverables into their local workflows.

Supporting Open Science Data Curation, Preservation, and Access by Libraries

Statement of Need

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 $248,247 to prototype innovative social and technical work supporting open science data curation, preservation, and access by libraries. The project will bring together the data underpinning open research, the long-term stewardship activities of libraries, and the formation of collaborative, distributed data preservation networks. By focusing on these three areas of work, the project will test and implement infrastructure, tools, and training in support of open science and data preservation. This proposal pairs two non-profit organizations with expertise in their respective fields and a broad network of partners to explore how technical integrations and collaboration can scale accessibility of digital content vital to accelerating research and can expand the role of libraries in supporting open science and the stewardship of its data.

There has been increased attention in the scholarly and academic communities on the importance of open infrastructure supporting the production, distribution, and stewardship of research.[1] Much of this discussion has been driven by the increased centralization of many parts of the research ecosystems into the hands of a few commercial entities.[2] As Heather Joseph, of SPARC, notes in her article, *Securing community-controlled infrastructure*, "the threat posed by commercial lock-up of crucial infrastructure has implications that transcend libraries and extend across the operations of higher education institutions. This is not a problem that libraries can solve alone."[3] Furthermore, centralization and commercial monopolization are not the only challenges faced by researchers, scholars, and librarians in building community infrastructure. Researchers often rate "discoverability of data" as one of their prime concerns[4] and many journals, including the Public Library of Science,[5] require that authors publish their data. Yet publishers don't specify how or where. Many data sharing platforms exist, yet their proliferation has led to a fragmented environment in which data, other research outputs, as well as curation and long-term stewardship, are all hampered by an ecosystem that lacks technical integration, reliability of access, and coordination with local library expertise and services.

This fragmented environment impedes research reproducibility and scientific resilience and also creates a barrier to the essential preservation functions of libraries in ensuring ongoing access to the knowledge produced by their institution's scholars. As Joseph notes, the community needs to find "ways to position our repositories, libraries, and research institutions as the foundation for a distributed, globally networked infrastructure for scholarly communication."[6] This proposal seeks funding to pursue the technical, social, and professional development work needed to prototype such distributed networked infrastructure in support of platforms and libraries critical to data preservation.

---

[1] Bilder G, Lin J, Neylon C (2015), "Principles for Open Scholarly Infrastructure," http://dx.doi.org/10.6084/m9.figshare.1314859"

[2] "Elsevier acquisition highlights the need for community-based scholarly communication infrastructure," https://sparcopen.org/news/2017/elsevier-acquisition-highlights-the-need-for-community-based-scholarly-communication-infrastructure/

[3] Joseph, Heather. "Securing community-controlled infrastructure" College & Research Libraries News [Online], 79.8 (2018): 426. Web. 4 Mar. 2019. https://crln.acrl.org/index.php/crlnews/article/view/17246/18986.

[4] "Practical challenges for researchers in data sharing," https://figshare.com/articles/Whitepaper_Practical_challenges_for_researchers_in_data_sharing/5975011

[5] PLOS Data Availability, https://journals.plos.org/plosone/s/data-availability.

[6] Joseph, Heather. "Securing community-controlled infrastructure."

The Challenges of Open Research

Reproducibility, the ability to independently obtain evidence supporting scientific findings, is a central tenet for research because it places the burden for "truth" on the quality and repeatability of the evidence itself rather than the authority or prestige of its originator. Yet, the openness upon which reproducibility is dependent is not a standard practice among research communities and many of the incentives that drive researchers' behavior do not promote these values. Recent investigations across the sciences, suggest that reproducibility of published findings is lower than expected or desired.[7] For researchers, the currency of reward is publication. Whether the research is open or reproducible is rarely relevant to publication success. Instead, publication depends on achieving novel, positive, clean outcomes. In a competitive marketplace, researchers may make choices, even unwittingly, that increase the likelihood of publishable outcomes even at the cost of accuracy. Without access to open data there is scarce opportunity for self-corrective processes to identify findings that are robust and accumulate evidence of their limiting conditions.[8] Many factors contribute to irreproducibility and, as a consequence, create friction that slows the pace of discovery. For research to succeed, it requires openness and reproducibility, which in turns requires open infrastructure and distributed access -- services few institutions can provide alone. Providing a trustworthy network for perpetual availability of research data is critical to ensuring reproducibility. This project will employ shared, open infrastructure while working in collaboration with multiple library partners to improve access to research outputs, support reproducibility, and advance open research.

The Challenges of Open Research Stewardship

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 hinder the ability of libraries to accomplish this mandate. Community debate continues over the efficacy of library institutional repositories (IRs)[9] for stewarding research.[10] This focus on the role of library IRs comes at a time of growing commercial monopolization of the services around research and access. As with the complex incentive structure for researchers outlined above, publishers have a similar lack of incentive to provide infrastructure for open research data. For publishers, the dominant monetization strategy is to provide access to the research process and outcomes via purchase or subscription business models. Publishing conglomerates have moved aggressively into purchasing repository providers like Bepress and consolidating their holdings of services around analytics, discovery, and access.[11] While these business models provide profitable income streams, they also limit access to this information to those who can pay. The same cost-escalation of commercial journal subscription pricing that has deleteriously impacted libraries now threatens to occur with other critical services around research data accessibility. The increasing monopolization of the scholarly infrastructure poses significant risk to the essential mission of libraries to provide open access to information and to responsibly steward this information. This project proposes open, community-oriented infrastructure for local curation and stewardship of

---

[7] Open Science Collaboration. (2015). Estimating the Reproducibility of Psychological Science. Science, 349(6251), aac4716. DOI: 10.1126/science.aac4716. .

[8] Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. Science, 348, 1422-1425. Doi: 10.1126/science.aab2374.

[9] "Time to Re-Think the Institutional Repository: Q&A with CNI's Clifford Lynch," https://richardpoynder.co.uk/Clifford_Lynch.pdf

[10] "Where Are We Now? Survey on Rates of Faculty Self-Deposit in Institutional Repositories," *Journal of Librarianship and Scholarly Communication*. 5(1), p.None. DOI: http://doi.org/10.7710/2162-3309.2203

[11] Operation Beprexit, https://beprexit.wordpress.com/about/

research data. A mission-aligned, locally-integrated approach to data stewardship can help libraries fulfill their mission as critical service providers in advancing open science and knowledge production.

The Challenges of Distributed Preservation and Access

Along with the challenges of open research reproducibility and research data stewardship, a third challenge complicates perpetual access to content: changes in library preservation networks. Preservation is a critical area of research library activity also impacted by centralization and termination of services as well as by a general move of campus IT services to the commercial cloud. The recent final report from the shutdown of the Digital Preservation Network illuminated many of these issues.[12] Key factors in its shutdown included that "confidence in the viability [of] cloud-based solutions increased" among members as well as that broad-based engagement was difficult to achieve for a service starting from scratch with no existing integrations with members' local systems. The difficulty of creating new, bespoke technical workflows was a barrier to entry for many libraries planning on depositing content. DPN was, in a sense, an idea and an argument before it was ever a functioning service -- its protracted "go to market" timeline was highlighted as another lesson learned -- pointing out the need for agile prototyping, both technical and social, prior to building sustainable infrastructure and the need to scale existing collaborations and integrations instead of starting entirely new services. Similar challenges, as well as the need to achieve better economies of scale for library repository services, has also been highlighted by the recent merger/acquisition of Duraspace by Lyrasis.[13] And the larger move of university campus IT services to the commercial cloud has elicited a range of concerns about the impact on long-term preservation and access, as thoroughly explicated by David Rosenthal, co-founder of LOCKSS, who highlights vendor lock-in and economics as a few of the critical issues and also notes, "it is very unlikely that more conservative institutions could approach the Internet Archive's cost per TB/month" and the "advantages of on-demand scaling are so large that institutions lacking the Internet Archive's audience cannot compete with the major cloud platforms for access."[14] This proposed project intends to leverage IA's scalable infrastructure to support open science data preservation and also to build on existing collaborations and systems integrations to prototype a more bottom-up, community-aligned network of institutions dedicated to supporting perpetual access to open research data.

The challenges outlined evoke a range of needs: the need to support open, reproducible research; the need to maximize data discoverability through distributed access; the need for innovative partnerships to seed collaborative work; and the need for training to support these activities within libraries. The Center for Open Science (COS) can address these needs through its mission-aligned open data and research infrastructure, its knowledge of the research process and research data management requirements, and its experience with training. Internet Archive (IA) can address these needs with its scalable digital archiving and global access technologies, as well as its experience in providing archiving services, support, and training as part of its work with hundreds of library, archive, and heritage organizations. Combining the capabilities of these two organizations, and leveraging each institution's network of partners, will help advance the initial work to build sustainable community infrastructure and further the adoption of "open access first, perpetual access forever" behaviors by researchers and libraries alike.

---

[12] Digital Preservation Network (DPN) Final Report, https://osf.io/md9yk/.
[13] https://duraspace.org/amplifying-impact-lyrasis-and-duraspace-announce-intent-to-merge-2/
[14] Rosenthal, David. "Cloud For Preservation," February 7, 2019. https://blog.dshr.org/2019/02/cloud-for-preservation.html

Addressing These Challenges

IA and COS propose to address these challenges by providing cooperative infrastructure to ensure long-term access and connection to research data; defining curation standards among repositories to improve data sharing and reuse, and by supporting and promoting adoption of open science practices to enhance research reproducibility. Long-term stewardship depends on multiple copies of critical data being archived at many institutions. A platform like OSF enables researchers to directly create and manage data about their research projects; however local archiving of this data by the library and institution supporting that researcher is not built into the system. Additionally, digital preservation best practices require that multiple copies of this master dataset of research outputs be mirrored at many institutions. Such distributed replication relies not just on technical integrations but on collaborations across institutions and upon giving local institutions the tools to be able to seamlessly curate and take ownership of data that falls within their mandate to archive and make accessible. This project will accomplish work, both production-grade and prototype-level, to make this possible.

First, the project will build interoperability between COS and IA to allow critical research data to be archived outside of COS's OSF system. Preliminary work will focus on OSF registrations data (described below), as this data is both critical to research, fully open access, and generally immutable, making it ideal for long-term preservation and access given its open license, criticality to the research process, and overall lack of restrictions. The mirroring of this data, as far as technical work and user workflows, will be informed by the needs and use cases of researchers and librarians. Alongside the technical integration, significant training will be developed and delivered, in-person and virtually, on multiple aspects of this process, for data and preservation librarians, including education on topics such as research archiving best practices, supporting open science research, and local data curation and archiving activities. Furthermore, the project will then explore work with multiple preservation partners for additional replicas of this data to be shared across a prototype network and a full copy of the complete data archive will be made available for data mining by researchers to encourage novel analysis for further scholarship and reuse. The lifecycle of the full project thus is sequenced to:

- Form innovative collaboration between two non-profit partners working at scale to provide infrastructure for open scholarship;
- Technical integration between these partners to archive and provide replicated access to a specific corpus of valuable open science research data;
- Extensive training materials and workshops for users of both platforms to support the above at a professional level and for such work to be performed beyond the project
- Further archiving of this data amongst a set of pilot partner libraries to prototype a data sharing/preservation network;
- Exploration of further data archiving between all partners and support of new access models.

The data in scope for archiving in this proposal is that of research registrations. Registration, the process whereby a researcher captures specific information about a research study, are time stamped and immutable once submitted and approved by the research and their collaborators and are made discoverable, either immediately or after embargo period. Specific, relevant, and comprehensive registrations are essential for constraining researcher degrees of freedom that will ultimately produce the most robust outcome reports for maximizing the rigor, reproducibility, and transparency of the conducted research. COS built and maintains OSF Registries, a platform that provides research communities an easy, efficient workflow for registering studies. Each OSF registration is backed by a

project, where users can upload materials either to support the registration or to share the outputs of the study. This facilitates research transparency and reproducibility; researchers can find the registered study plan but also the data output and code. For the IA/COS integration, both existing registrations and new projects that use OSF to create their registration will be in scope for this project's work. Once completed and made public, each registration can be archived.

Few integrations exist between popular open data platforms like OSF, large scale digital archives like Internet Archive, and preservation networks. At the same time, librarians' existing curatorial approaches struggle to meet the growing volume of research outputs -- struggles that better integrations and automations, along with targeted training, can help alleviate. Advancing open data preservation and access requires bold, collaborative social and technical networks that leverage the joint power of integrating successful platforms. This project aims to pair mission-aligned, non-profit services at the intersection of open research and open data stewardship to give libraries the power to scale their ability to select, preserve, and provide access to research data.

Demonstrated Expertise of Partners

Project Leads: The Center for Open Science (COS) is a non-profit technology and advocacy organization with a mission to increase openness, integrity, and reproducibility of research. COS acquires evidence to encourage change, provide incentives and training to embrace change, and creates infrastructure to enable change. COS developed and maintains the Open Science Framework (OSF), a suite of cloud-based applications, which enables rigorous, reproducible science by providing collaboration, registration, and data management support across the entire research lifecycle. The Internet Archive (IA) is one of the world's largest digital libraries, with the the demonstrated commitment and capability for making public data globally available, free of charge, forever. As a nonprofit library and technology organization, IA provides web, research, and data preservation services to hundreds of libraries and archives, governments, civil society organizations, research organizations, and others, with over 1 million users visits per day to its collection of over 45 petabytes of archived information and its well-known Wayback Machine. IA brings unparalleled expertise to the field of data preservation and access in its mission of "Universal Access to All Knowledge." More information on each organization can be found in the included *Organizational Profile*.

This project pairs the expertise of COS in advancing research resiliency and open science infrastructure with IA's technical scale, global and access experience, and close work with libraries in preservation and collaborative technology development. COS and IA can draw on their work with researchers and libraries to establish the skills and interoperable technologies to improve access to key data underpinning open knowledge research, which in turn can help position librarians as critical to the open research process and essential to data stewardship. This project also builds on IA's ongoing work supporting preservation and distributed access to open scholarship and digital content. With California Digital Library, IA is pursuing a data sharing and preservation pilot that will inform this project's work[15] and IA is also working with print publishers, libraries, and data services to ensure the accessibility of

15 https://blog.archive.org/2018/06/05/internet-archive-code-for-science-and-society-and-california-digital-library-to-partner-on-a-data-sharing-and-preservation-pilot-project/

at-risk scholarly publications and datasets.[16] Past initiatives on systems interoperability[17] and training to librarians building born-digital archives will also inform this project.[18]

Project Partners: Four institutions have committed to participating in the project's distributed data sharing and access deliverables explicated in the below project activities. These institutions include: 1) LOCKSS (Stanford University), an established, distributed digital preservation service supporting a global network of universities and research institutions. LOCKSS and IA have established systems integrations, have worked on collaborative technology development, and are already prototyping data preservation networks. 2) Academic Preservation Trust (University of Virginia), is a "consortium of higher education institutions committed to providing both a preservation repository for digital content and collaboratively developed services related to that content" that has worked with both COS and IA on joint projects. 3) Los Alamos National Laboratory Research Library is an institution that has developed many technologies for open access scholarship and runs a version of COS's OSF platform internally. 4) University of Notre Dame, has worked on the IMLS-funded PresQT project, partnered with COS/OSF, and will provide additional expertise and capacity for data accessibility.

Project Design & Activities

Project Design & Goals

The project will ensure the outputs of open science are discoverable in the preserved public scholarly record maintained by libraries. Explicit goals of the project are below. The project is designed for an iterative approach to technical work, training, and network building. Goals and overall activities include:

- Further the ability of librarians to steward open research data and provide enhanced discoverability of data for researchers.
    - Develop systems integrations for the preservation of open data, starting with having the registrations data in Open Science Framework (OSF) mirrored at Internet Archive
- Reduce the cost and risk of data disappearance and test networked infrastructure
    - Prototype a unique distributed design wherein data is replicated from IA to a number of partner libraries to test distributed, networked preservation approaches
- Accelerate adoption by the community and increase the volume of data published in the sciences
    - Provide 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
    - Integrate the long-term storage of research projects into their publication processes

Project Activities:

Phase 1: Prototype Development and Implementation (July 2019- December 2019)

This initial phase will focus on data transfer decisions between technical teams, conducting community and user research on the preserved parts of an OSF registration with COS product team and building a prototype of an automated transfer of OSF registration data and metadata to Internet Archive.

1. COS and IA product teams work with the registration researcher community to determine which registration contents best meet the preservation use case and metadata mapping.

---

[16] https://blog.archive.org/2018/03/05/andrew-w-mellon-foundation-awards-grant-to-the-internet-archive-for-long-tail-journal-preservation/

[17] https://www.imls.gov/grants/awarded/lg-71-15-0174-15

[18] https://www.imls.gov/grants/awarded/re-85-17-0060-17

2. COS and IA technical teams collaborate to select an export system to transfer data and metadata efficiently. IA uses an S3-like API protocol for data ingest, which the OSF will use to transfer the registration data and metadata in the format required by the protocol.
3. Community engagement on data and metadata mapping and format to ensure downstream curation and discoverability of preservation collection content. Feedback will be targeted from librarians as primary stakeholders in curation and access to content. Other communities involved include researchers, funders, publishers, and open science service providers.
4. Generate user stories and requirements based on the efficient transfer format, data and metadata mapping of registrations and stakeholder feedback to ensure efficient curation and discovery.
5. COS and IA prototype the automated transfer of registration data from OSF to IA's preservation repository, including metadata mapping and integration testing, and deployment.

Deliverables:

1. OSF registration users will be able to preserve their registration on IA.
2. User stories, use cases, metadata mappings, from researchers and data librarians and documentation mapping these outputs to the project's technical and workflow products.
3. Matching of storage region locations on OSF side with IA's system to assist researchers with data storage locations to meet requirements.

Phase 2: Community Training and Workflow Integration (January 2020- June 2020)

This second phase will take the automatic transfer of OSF registration data and conduct a review with technical, QA and product teams. The teams will conduct user research with the community of research librarians and the OSF registration users to encompass broader researcher workflows as use cases with the transfer. From there, the prototype will be finalized and made ready for release on production environments and available to users. Further engagement with users and librarians will be done through training and education. The existing COS training curriculum and related workshop materials will be expanded to include a discussion of how registration data is mirrored by IA and other participating institutions and how this workflow embodies best practices in data preservation. This information will become a routine part of all COS trainings on open and reproducible research practices.[19]

1. Building on the COS Reproducible Research training, which covers reproducibility, research management, structuring projects, version control, sharing data, etc, and develop an additional training module on curation of preserved registration artifacts to support data librarians.
2. The COS Trainer will provide this training at no charge, once per year (with additional trainings offered in Phase 3) in coordination with professional research library conferences such as RDAP, ACRL, and use feedback from participants to improve the curricula for data librarians.
3. COS will host free webinars on Reproducible Research and Preservation Training each year of the grant and will also be offered as a stand alone free webinar for librarians.
4. COS will make the Preservation Training curriculum, workflows, and instructional material publicly available for reuse under open licence (CC0), hosted on OSF with webinars on its YouTube channel alongside addition resources for research librarians and researchers.
5. IA will review the data transfer prototype, test for end to end workflow, iterate on design and technical changes, and conduct user testing.

[19] OSF Curriculum for the Introduction to Open and Reproducible Research Workshop: https://osf.io/4b5du/

6. IA & COS will release into production the automated transfer of OSF registration data for mirroring in IA including release of user documentation and promotion of the service.
7. IA will conduct preliminary meetings with preservation partners on technical and workflow needs for the distribution and archival mirroring of the registration data at their institutions.

Deliverables:

1. COS: develop training curriculum for research librarians on using OSF registration and connected preservation repositories so they can support researchers.
2. At least two free trainings will be held around the U.S. alongside relevant conferences such as RDAP or ACRL and COS will produce webinar content in support of librarians.
3. COS and IA will release the service integration that allows for registration data to be archived.
4. IA will conduct preliminary technical and procedural work for the further distribution of this archived data to four preservation partners for prototyping a broader data sharing network.

Phase 3: Expanded Data Distribution, Training, Access (July 2020 - June 2021)

IA will advance methods for sharing archived OSF data with preservation partners in order to test approaches to networked data sharing. The project team will continue to foster stewardship through delivery of in-person and online workshops and training materials. Further data replication will be tested with additional OSF datasets and the archived data will be made available for computational analysis.

1. Develop methods for sharing the archived OSF data via additional preservation networks. IA will work with confirmed partners, LOCKSS/Stanford, AP Trust/UVA, Los Alamos National Lab (LANL), and Notre Dame. This work will test replication across a variety of preservation platforms with the intent of providing data librarians multiple options for custodial acquisition and stewardship of the open research data of their institution.
2. Publish the project's technical documentation, including lessons learned on partner integrations and guidance for participation by additional libraries for local archiving of OSF data.
3. Improve training materials, workflows and instructional materials based on prior training feedback and conduct additional in-person and webinar trainings.
4. IA will build on its work supporting computational access to big-data by making the OSF registration dataset available in "bulk access form" for data mining[20] to expand access methods and help promote innovate types of computation research enabled by the project's work.
5. Both COS and IA will roadmap and pursue expanding the project's work to include other open data in OSF (such as preprints) to facilitate additional open research outputs being included in the research, preservation, and data curation lifecycle and computational research activities.

Deliverables:

1. Project team will lead, organize, and deliver a free workshop at the ACLR (or similar) conference and host additional trainings via webinars and online training materials.
2. Replication of the archived OSF data to four other institutions by at least two different methods of data transfer for preservation.
3. Technical documentation and reporting on the distributed data sharing.
4. Research and development for supporting data mining of the archived OSF data.
5. Additional OSF data archived and potentially distributed across project-established networks.

[20] Internet Archive Research Services: http://webservices.archive.org/

Project Principles

Audience: Both COS and IA's development teams operate with an agile scrum development process. The product teams develops product vision, requirements, use cases, and user testing, and defines development priorities. The Engineering teams translate requirements into defined work plan. Scope management and planning is addressed via a priority list of features and development work. Delivery on tasks is managed through the JIRA ticketing system with bi-weekly sprints and daily stand-up meetings.
Evaluation: The project approach is based on an iterative, feedback-driven agile development cycle, meaning staff can respond very quickly to real world user testing and go through dozens of rapid release cycles as we move towards completion.
Communication: Openness and transparency of project activities, outcomes, and technical work is not just a key ethos of this specific project but a core value to building partnerships across institutions and user groups. IA and COS will make use of institutional and third-party blogs, microsites, newsletters, social media accounts, and other dissemination tools to ensure promotion of project outcomes. Both IA and COS will be jointly responsible for coordinating communication, outreach, and presentations.
Training: Both COS and IA regularly provide training support services through in-person training, webinars and consulting. For this project COS will host two free trainings in conjunction with relevant conferences such as RDAP or ACRL. Training will speak to how library staff can support the practice of registration to advance the reusability of data, what data is captured in registrations, and how to extract data from the aligned repository. Additionally, COS and IA will develop and maintain web-based instructional materials. All content will be openly licensed (CC0) to maximize opportunities for library staff to use and incorporate these resources into their institution's training. This model will help foster a sustainable community of trainers and lead to adoption of rigorous research and stewardship practices.
Code: All software and technical products created by the project will be released under open-source licenses and published on Github or other code-sharing platforms. Both IA and COS are committed to "data persistence" and access. The code itself is free and open source, meaning that any organization or individual could take a copy. Also, files are stored with long-term preservation in mind (e.g., multiple locations, integrity checks, etc) and are accessible via a public, documented API.
Advisement: To support meeting the goals of the grant, an Advisory Board of aligned research institutions include the following: project partners LOCKSS, UVA, a member of the Academic Preservation Trust, Los Alamos National Library, University of Notre Dame, the Johns Hopkins University, Massachusetts Institute of Technology Libraries, Arizona State University, and University of Houston. Board members will join in calls and meetings to provide guidance on the project's progress.

Diversity Plan

For the open science movement, open refers to both transparency and inclusivity. All infrastructure built and maintained by IA and COS is created with a commitment to open access and inclusivity. COS is currently enabling Web Content Accessibility Guideline 2.0 AA standards on all OSF pages and IA is involved in the "Federating Repositories of Accessible Materials for Higher Education" project.[21] OSF is being migrated into Ember which will support screen-readers for visually -impaired users. All OSF's tools and services are offered for free so that researchers and consumers of research content will able to easily search, access, and reuse, data. As well, IA has no charges for access to any of its preserved data

[21] https://news.library.virginia.edu/2019/01/11/federating-repositories-of-accessible-materials-for-higher-education-awarded-a-1000000-grant-from-the-andrew-w-mellon-foundation/

via the web or multiple APIs and allows researchers free storage for their uploaded content. Producers and users of any of the data in scope for this project do not need to be affiliated with a institution; they only need an internet connection. Accessibility of public infrastructure and content helps encourage inclusivity. COS will offer free in-person trainings at research librarian conferences as well as create and host webinars for library staff to maximise project deliverables. The project team will market all training opportunities and will specifically engage the 50+ universities and institutional partners linked to OSF tools (OSF Institutions), and members of the Historically Black Colleges and Universities Library Alliance to encourage broad participation in training. IA will ensure the project's work remains relevant to its network of non-academic research partners. The project will also ensure a variety of institution types, sizes, geographic location are considered in user stories/testing and that librarians of different professional skill levels, gender, and racial backgrounds are part of the cohort involved in training.

National Impact

This project will lay the foundation, 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 data and will be framed to lead to systemic change in how libraries approach preservation infrastructure, acquire research outputs that are in scope for local preservation, and enable new forms of bulk access and computational reuse. Conducting training, including the development of open education resources, webinars, and workshops, 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 development work will enable continued collaboration efforts and robust, open services that offer an alternative to the for-profit services currently threatening to impede open access and distributed preservation. Both the IA-COS collaboration, and their collaboration with preservation partners and data librarians, is expected to continue beyond the grant.

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. By involving an open research platform, a digital archive, and a set of custodial research libraries, this project seeds follow-on work by establishing the technical mechanisms and the social architecture necessary for formation of a larger data preservation network. The project's open tools and systems can scale the ability of librarians to provide ongoing access to open knowledge production, both within their institution and within the larger heritage and scholarly communities. The distributed approach at the heart of this work will keep the ownership of the data in the hands of the researchers and institutions -- not in the hands of platforms or commercial providers. The project's mission is to ensure that regardless where a researcher's data lives, a copy of the data will always be available, both from the Internet Archive and from a number of other institution. Librarians, along with researchers, will be the ultimate stewards of open data. The more copies available, the more their preservation is assured. Progress in science facilitates innovation and improvement of policy and activity across all sectors of society and the economy. The broader impact of this outcome is more efficient use of research dollars, more reliable and reusable research output, faster translation from insight to progress, and ensuring libraries are central in the stewardship, support, and perpetual accessibility of open research.

Internet Archive & Center for Open Science NLG Project Grant for NDII	Year 1 (July 2019 - June 2020)												Year 2 (July 2020 - June 2021)											
ACTIVITIES	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**
PHASE ONE: Prototype Development and Implementation																								
COS & IA: Preplanning and recruiting community and stakeholders to provide feedback on registration contents and format for preservation																								
COS: Engage registration researcher community to determine which registration data is best for preservation use case																								
IA: Prep repository for registration data ingest and syncing																								
COS & IA: Community engagement on the data and metadata mapping and format and preservation mandates																								
COS & IA: Generate user stories and requirements from community and advance stakeholder engagement																								
COS & IA: Prototype the automated transfer of registration data from OSF to IA, conduct inital testing to check data matches																								
PHASE TWO: Community Training and Workflow Integration																								
COS & IA: Review prototype, test for end to end workflow success, iterative design and technical changes; re-engage community and stakeholder members for feedback on prototype																								
COS & IA: Release the automated transfer of registration data from OSF to IA in production, review and testing to check data transfer																								
COS: Develop new training module for research librarians with IA input by expanding and supplementing existing training materials																								
COS: Deliver training at RDAP conference and provide additional free training opportunities via webinars																								
IA: Preliminary meetings with preservation partners to map technical workflows for data distribution across multiple institutions																								
PHASE THREE: Expanded Data Distribution, Training, Access																								
IA: Preliminary testing of prototype methods for sharing the archived OSF data to preservation partners																								
COS: Deliver training at additional conference(s), such as ACRL, and provide additional free training opportunities via webinars																								
COS: All training materials made available as OER																								
COS & IA: Publish technical documentation on COS-IA integration																								
IA: Implementation of technical approaches to sharing the archived OSF data to preservation partners																								
IA: Make the OSF registrations dataset available in “bulk access” and solicit research analaysis users																								
IA: Publish technical documentation and lessons learned white paper on data sharing activities and prototype preservation network																								
COS & IA: Explore expanding the project’s work to include other open data in OSF (preprints, etc)																								

DIGITAL PRODUCT FORM

Introduction

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (e.g., 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. 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

All applications must include a Digital Product Form.

☐ Please check here if you have reviewed Parts I, II, III, and IV below and you have determined that your proposal does NOT involve the creation of digital products (i.e., digital content, resources, assets, software, or datasets). You must still submit this Digital Product Form with your proposal even if you check this box, because this Digital Product Form is a Required Document.

If you ARE creating digital products, 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.

The research reports and technical documentation produced as part of this project will released under the Creative Commons Attribution 4.0 (CC-BY-NC-SA) license.

All software and code developed as part of the project will be released under open source licenses, either GNU-GPL, Apache 2.0, and other open-source licenses.

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.

The project will not be creating new digital content, but rather provides integrations between data sharing and data preservation platforms and services. We will notify users of these digital systems integrations via a number of methods, including publications (institutional blogs, journal articles), presentations at multiple conference (RDA, Open Repositories, DLF, CNI, others), listservs, and newsletters.

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.

This project will not create any content or products which involve privacy concerns, require obtaining permissions or rights or raise any cultural sensitivities. Any data included in this project is fully open access.

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 the format(s) you will use.

The project is not creating or collecting any new digital content, resources, or assets. All digital content within the project is pre-existing.

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.

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).

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan. How will you monitor and evaluate your workflow and products?

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).

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).

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

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).

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).

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.

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.

We intent to create scripts and small applications for two API-based digital storage, preservation, and access platforms and systems to interoperate. These applications will connect the Center for Open Science's OSF platform and the Internet Archives archival repository for the seamless transmission of open science data hosted on OSF to be automatically backed-up and mirrored on IA. The same data, or portions of it, will then be replicated at other institutions to prototype an open data preservation network. The primary audiences are 1) librarians tasked with curating open science data 2) researchers creating open science data 3) stewards and custodians supporting the open science data lifecycle.

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.

There is no software that performs the same functions as those proposed in this project. Open APIs can allow for systems integrations, usually at the metadata level. Services such as Duracloud offer middleware that connects upload functions with storage in cloud services like AWS. Projects like IA's WASAPI service have connected systems for generic transfer of web archive data. However none connect distinct, unique data hosting and digital preservation services and then go one step further to have the data replicated across a prototype, multi-institutional networks. This latter aspect is significant in that collaborative non-profit data sharing networks are increasingly necessary to provide a mission-aligned option to the for-profit commercial cloud and give libraries a low-cost option for distributed preservation.

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.

The programming languages used will primarily be Python, and Unix (bash) scripting, with some work done in Ruby and Java for front-end development and for integration of existing tools.

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

OSF supports the research lifecycle by enabling researchers to produce and manage data for further curation to foster adoption and discovery. Preservation of these artifacts on Internet Archive and other networks allows the long term stewardship of these collections. The software will allow OSF data to be preserved and more easily available for local curation and for IA, adding the OSF data to the archive will enrich the availability of material available for access and reuse.

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

There will be no software or system dependencies necessary.

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

Development documentation will make use of the Jira issue and project tracking software. Documentation will include functional/technical requirements research, systems architecture/mappings, basic user guides (for developers and non-developers), and we be further articulated and documented in training materials.

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

IA has created dozens of software tools and systems including widely-adopted open-source tools for web archiving and digital libraries. COS developed and maintains the Open Science Framework (OSF), a cloud-based applications, for research data management, as well as other tools that support open science.

COS: https://github.com/CenterForOpenScience/
Internet Archive: https://github.com/internetarchive

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.

IA/COS will release all project software and code with an open-source license, such as Apache 2.0, GNU GPL, or BSD licenses and all code will be published, updated, and available on Github.

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

All software and tools developed will be publicly available to users via Github, SourceForge, Google Code, or a similar public code-sharing platform. Both IA and COS primarily use Github.

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

Name of publicly accessible source code repository:

as of yet unnamed

URL:

https://github.com/CenterForOpenScience/ & https://github.com/internetarchive/

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.

This project is not creating additional datasets beyond the data in scope for preservation and replication; though the data involved in the project, can, like any data, be considered a dataset in aggregate for non-consumptive computational use. As this Part IV seems to have a more specific inference of "dataset" it seems not relevant to this proposal.

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?