University of California
CDL
California Digital Library

UC Open Data Initiative

Information technology and resources thoroughly permeate all aspects of the University's manifold activities in pursuit of its longstanding research, teaching, learning, and outreach mission. Ensuring that these valuable assets remain available for use by scholars – both now and in the future – is therefore an important component of the responsible stewardship of the University's intellectual capital. These scholarly assets exist across a wide spectrum of material types, from traditional publications to research data. The University has already taken an important affirmative step towards ensuring permanent access to the scholarly literature through its open access policy, under which faculty can submit their publications to CDL's eScholarship or other open access repositories subject to pro-active preservation care and public retrieval.[1] Unfortunately, no analogous open data mechanisms are currently available, placing access to the non-publication outputs of UC research at significant risk.

The ongoing curation and sharing of research data offers significant benefits. Foremost, the independent assessment and verification of research results lies at the heart of the integrity of the scholarly enterprise. However, without access to the data underlying published research claims, verification is difficult, if not impossible. Additionally, the widespread availability of data minimizes the needless duplication of research efforts, leverages prior financial and intellectual investments, and often spurs novel new avenues of investigation and scholarly advancement. The importance of the public availability of research data is being widely recognized within the scholarly community, as evidenced by increasing publication requirements regarding data sharing, such as those of *F1000*, *Nature*, and *PLOS and a number of disciplinary journals*. Similarly, data sharing has received attention at the highest governmental levels, as evidenced by the 2007 promulgation of the OECD's open data principles, and in the US, by the 2013 OMB open data and OSTP data management policy memorandums, which place an obligation on Federal agencies and grantees to exercise responsible data curation.[2]

As with the open access policy, UC researchers can only comply with responsible practices if they are provided with appropriate options for the ongoing pro-active management of their data. The CDL has long operated the Merritt curation repository as a centrally-hosted service for the long-term preservation and citation of, and access to, the University's digital assets. Recent enhancements to

---

[1] Academic Senate of the University of California, *Open Access Policy for the Academic Senate of the University of California*, July 24, 2013 <http://osc.universityofcalifornia.edu/wp-content/uploads/2013/09/OpenAccess_ adopted_072413.pdf>.

[2] Organization for Economic Co-operation and Development, *OECD Principles and Guidelines for Access to Research Data from Public Funding*, 2007 <http://www.oecd.org/science/sci-tech/38500813.pdf>;

Office of Management and Budget, *Open Data Policy – Managing Information as an Asset*, May 9, 2013 <http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf>;

Office of Science and Technology Policy, *Increasing Access to the Results of Federally Funded Scientific Research*, February 22, 2013 <http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_ 2013.pdf>.

Merritt have significantly streamlined the process of individual researcher data contribution and search, which are now presented under the new service name "Dash".[3] As the University's data stewardship obligation is ongoing, it must be accompanied by sufficient technical and financial resources also on an ongoing basis. The geographic replication of managed data is one of the fundamental strategies Merritt uses to ensure the effective preservation of its managed assets and represents one of Merritt's largest operational cost components. To continue operating Dash as a sustainable core service offering, CDL must be able to recover the cost of providing the underlying preservation storage, which relies on fee-for-service private cloud infrastructure operated by UCLA and UCSD/SDSC IT groups.

UC prizes its role as one of the world's preeminent public academic institutions. This leadership role should be extended into the arena of responsible research data stewardship to its own benefit and the benefit of its stakeholder communities. Many of UC's peer institutions are pursuing similar capabilities. For example, both Indiana University, through its ScholarWorks, and the Pennsylvania State University, through its ScholarSphere, provide unlimited preservation services to their research communities; the Purdue University Research Repository (PURR) provides 10 GB of storage and 3 years of preservation service to its faculty, staff, and graduate students; and the University of North Carolina LifeTime Library offers students 1 GB of storage in perpetuity.[4] At a relatively modest expenditure, the University of California could join these institutions in supporting its own research community and assert an influential leadership role.

The best way for the University to enable and encourage responsible data stewardship by the UC community, as is increasingly necessary by funding agency policies, pre-publication requirements, and disciplinary best practices, is to allocate permanent funding for provision of Dash services for eligible researchers. At a cost of $79,300 per year, the University could provide all ladder rank faculty, staff researchers, and doctoral students with 10 GB of service capacity in Dash.[5] While enactment of a formal system-wide open data policy would require Faculty Senate approval, providing Dash capacity now addresses the immediate needs of early adopter researchers and positions the University for the future expansion of open data curation policies. The investment in Dash is modest, but the return is significant: the UC research community would be at the forefront of a paradigm shift towards more efficient, effective, transparent, and sustainable scholarship, and furtherance of the University's mission.

---

[3] California Digital Library, *Dash: Data Sharing Made Easy*, June 23, 2014 <https://github.com/CDLUC3/dash/ wiki>.

[4] Indiana University, *IU ScholarWorks: Preserve and Share IU Research*, 2014 <http://scholarworks.iu.edu/>;
Pennsylvania State University, *PennState ScholarSphere*, 2014 < https://scholarsphere.psu.edu/>;
Purdue University, *Purdue University Research Repository*, 2014 <https://purr.purdue.edu/>;
University of North Carolina, *A Perpetual Resource for Digital Life*, 2014 <https://lifetime-library.ils.unc.edu/>.

[5] The associated spreadsheet provides details of the cost derivation; the spreadsheet is parameterized to facilitate cost comparisons with different eligibility requirements and storage allocations.