The Data Collective

A proposal to establish a repository for faculty research data

Executive Summary

This is a proposal from the Library, working collaboratively with the Earth Research Institute (ERI), to undertake a one-year pilot project to establish the "Data Collective," a campus repository for faculty research data. The Collective would be operated and managed by the Library's Data Curation Program; ERI would initially provide the storage and related infrastructure.

Responding to a widely recognized need, the Library's Data Curation Program has researched repository and storage technologies, data curation services, policies, and repository development approaches taken by other research institutions to arrive at this proposal. The software platform hosting the Data Collective is proposed to be a lightly customized installation of an open source repository/data publishing solution (to be determined by the project, but likely Dataverse). The storage for the Collective will be procured and managed by ERI, and will consist of extensible, triply-redundant storage arrays offering 250TB net storage initially, physically spread across multiple datacenters on campus and augmented by cloud backup to UCSB's institutional Box.com account.

In the pilot phase the Collective will be open to ERI and Marine Science Institute (MSI) researchers and, in order to broaden the initial experience beyond Earth science data, researchers in the humanities and social sciences on a select basis. Library curators will work with researchers to ingest data in order to evaluate potential platforms for the Collective and to pre-populate the repository. At the conclusion of the pilot project the Collective will become a production service and be opened up to the campus at large.

The cost of the pilot project is $78,500. There are no personnel costs: ERI has agreed to waive its normal storage recharge rate, and the development and operation of the Collective will be covered by the Data Curation Program as part of its normal operations. A capital cost of $78,000 is required to purchase three 250TB storage arrays (750TB total); given the ERI donation, this cost is less than one-tenth what CDL would charge for comparable Merritt storage. An additional $500 is required for server hosting.

The pilot project will answer key curation-related questions, including the role that the Library and Library curators play in the curation of faculty data, and the relationship between the Collective and the Library's ADRL repository. Additionally, the experience gained from the pilot project will inform a subsequent discussion on future funding, management, and overall sustainability of the Collective and other campus data curation services.

Table of contents

Background

In 2012 the Office of Research, the Earth Research Institute (ERI), and the Library sponsored the Data Curation @ UCSB project[1] with the goals of investigating campus research data creation processes and curation needs and formulating a recommendation for future action; the latter would eventually be included in the campus cyberinfrastructure plan[2]. In a campus-wide survey conducted as part of that project[3], over 80% of the almost 300 respondents identified research data storage as an activity they needed help with.

In 2015 the Library conducted a series of pilot projects with select faculty to give Library staff experience working with data-related issues and to ascertain the roles the Library might play in the research data lifecycle. A key takeaway from those projects is that faculty found it difficult to consider *consultation* regarding data curation separate from *curation* itself; that is, faculty were surprised and confounded that curators could advise them but not actually curate their data. While this gap had already been recognized, the pilot projects emphasized it.

In 2016 the Library launched the Data Curation Program[4]. Operating with the resources available—staffing in the form of three part-time data curators and web-based resources such as best practice guides—the Program offers consultation services in such areas as data management planning and repository selection. In engagements with faculty to date, Program curators have found that the first and most insistently-asked question is, “Can you take my data?” This question has been encountered in every phase of the research data lifecycle, from researchers actively writing or thinking of writing grant proposals, to those just launching projects or working with existing datasets, to those concerned with archiving data following the end of a grant or project.

---

[1] https://people.eri.ucsb.edu/~gjanee/dc@ucsb/
[2] http://cio.ucsb.edu/resources/UCSBCyberinfrastructurePlan.pdf
[3] Greg Janée and James Frew (2013). Faculty/Researcher Survey on Data Curation. UCSB Library. https://doi.org/10.5062/F4PN93K4
[4] http://www.library.ucsb.edu/data-curation

In 2017 the Data Curation Program surveyed ERI and Marine Science Institute (MSI) principal investigators, asking questions related to an advance version of this proposal[5]. 93% of respondents indicated they would take advantage of a campus repository for research data.

From all these experiences it is clear that it is incumbent on UCSB to provide a local solution to the problem of curation of faculty research data. The need for data curation arises in different situations and throughout the research lifecycle. For some researchers, there are simply no appropriate discipline-specific repositories available for their work, or external repositories exist but are too costly or not a good fit for their data or goals. For others, discipline-specific repositories exist and deposit may even be required by granting agencies, but the form the repositories require the data to be deposited in, and the limited services the repositories provide, make external deposit a burden, not an asset. And for many researchers, external repositories do not fulfill the desire to maintain data locally in an accessible, working form to sustain threads of research across projects and grants, and to support collaboration among project teams both on campus and between UCSB and partner institutions.

Meanwhile, the lack of a local solution is harming research. For example, the Data Curation Program is aware of two researchers whose recent grant proposals were rejected in part due to deficiencies in their data management plans.

Proposal overview

This is a proposal to establish a repository for research data and other publishable digital artifacts produced by campus faculty and researchers. We dub it the “Data Collective” for two reasons. One, as will be seen, it represents a collective effort by multiple entities on campus: the Library, by integrating curators into research processes; and campus computing groups, by utilizing their expertise in supporting large-scale, locally-managed, cost-effective, robust storage. Two, the Collective will serve as an initial central gathering place for research data for which curation is required or desired.

The Data Collective being proposed will be a repository system broadly defined by four characteristics:

- It will **limit the data that may be deposited** by, for example, imposing requirements on metadata and restrictions on accepted file formats. Furthermore, content in the Collective will be actively managed by the Library, and ingest will be mediated by Library curators (although the mediation factors, the extent of curator participation, and the division of labor between researchers and curators is unknown at this point and will be determined by this pilot project). The Collective will not be unconstrained disk storage or

[5] Greg Janée (2017). Straw-man proposal for a “data collective”: Results of a preliminary requirements-gathering survey. https://people.eri.ucsb.edu/~gjanee/archive/2016/eri-msi-survey.pdf

a workspace for day-to-day computation; those needs are better satisfied (and in many cases already are) by existing computational centers and departments on campus. Because it is intended to directly support medium-term curation of stored data (up to 10 years; see below) and to facilitate subsequent longer-term curation, and in order to limit abuse in the form of storage overuse, the Collective will impose controls on deposits.

- It will **provide access to stored data**. The Collective will not be a dark archive or a backup system. Instead, following for FORCE11 FAIR principles[6], it will make data publicly available to support reuse and attribution. Additionally, data will be made available in forms and by mechanisms that the depositors themselves find useful. In this way, researchers who deposit data will come to see the Collective as an extension of their computational work environment, and not as a system they might deposit into but not use themselves.

- It will **provide value-added services** such as persistent identification and citability of stored data, discoverability of data and searchability within datasets, and citation metrics. Despite the well-documented desire for data curation services at UCSB and at other institutions, many past repository efforts have struggled. Indeed, conversations with a half-dozen contemporary data repository efforts have revealed that all are encountering lower-than-anticipated adoption rates. The challenge here is in providing a compelling value proposition to researchers while articulating a compelling business case to the Library, and in avoiding documented pitfalls such as, “[the] benefits [of repository storage] weren’t being disseminated in a manner that presented them as incentives”[7] and “the institutional repository became, in essence, a ‘roach motel’.”[8] In addition to offering value-added services, the pilot project will develop outreach materials and a departmental outreach program.

- It will **store data for a limited period of time only**; the proposed maximum retention period is 10 years. This period of time is by definition finite, but nevertheless is long enough to satisfy many granting agency’s data management plan requirements (for example, NSF’s Engineering Directorate requires a minimum of 3 years[9]). And in any case, 10 years is long enough to provide researchers continuity of storage across grants and projects. Data reaching the retention limit will be evaluated by Library curators, consulting with the original depositor, to identify a curation strategy for the data going forward. Options may include accessioning the data into the Library as a collection, if desired by the researcher and if the data meets the Library’s collection development and

[6] https://www.force11.org/group/fairgroup/fairprinciples
[7] David Scherer (2016). Incentivizing Them to Come: Strategies, Tools, and Opportunities for Marketing an Institutional Repository. In: Burton B. Callicott, David Scherer, and Andrew Wesolek (eds.), *Making Institutional Repositories Work* (West Lafayette, Indiana: Purdue University Press, 2016).
[8] Dorothea Salo (2008). Innkeeper at the Roach Motel. Library Trends 57(2):98-123. https://doi.org/10.1353/lib.0.0031
[9] https://nsf.gov/eng/general/ENG_DMP_Policy.pdf

faculty acquisition policies; transferring the data to another repository, on campus or off; returning the data to the researcher; removing the data from the Collective because the data's preservation is already being satisfied by a discipline-specific repository; or removing the data outright, because the data is no longer seen to be of sufficient value[10]. (Presumably the data could also be allowed to remain in the Collective longer than 10 years, depending on the Collective's ultimate policy.)

The principal reason for imposing a time limit is that such a policy provides an explicit place and time at which, and an explicit mechanism by which, data will be evaluated for long-term preservation. *Long-term* preservation in a repository (which we define as preservation for an unbounded, indefinite period of time) represents a significant undertaking by the institution operating the repository (typically the Library), for the connection to the original depositor is unavoidably lost over time and the Library must assume full responsibility for the data and absorb the entire burden of maintenance. It is also costly. After a long enough period of time, in order to keep data usable in contemporary contexts, data will need to be converted to new forms, described in new ways and using new tools, recontextualized, and made available by new technologies.

But after 10 years, some datasets will be found to no longer be appropriate for long-term storage for any of a variety of reasons: the data are found to no longer be useful, to be incorrect, to have been supplanted by better data, etc. In a sense, a time-bounded repository such as the proposed Data Collective will act as a vetting ground, culling data before more significant resources are expended preserving the data in a long-term Library archive or institutional repository. Additionally, because the Collective will impose requirements on ingested data, it will act as a staging area for the data, facilitating the data's subsequent transfer into another repository.

It goes without saying that UCSB does not have an institutional repository at present, nor is the Library's own repository, the Alexandria Digital Research Library (ADRL)[11], presently open to faculty or researcher contributions. But one or both of these limitations may change in the future. If and when that happens, UCSB will come to resemble other research institutions that have found it useful to offer multiple data management and storage options to their researchers [12], and have found it effective to implement mandatory review of stored data after a designated

[10] Technologies exist to support persistent identification of datasets even as they move from repository to repository. In addition, technologies exist to support citation and description of datasets that have been removed entirely.

[11] https://www.alexandria.ucsb.edu/

[12] Katherine McNeill (2016). Repository Options for Research Data. In: Burton B. Callicott, David Scherer, and Andrew Wesolek (eds.), *Making Institutional Repositories Work* (West Lafayette, Indiana: Purdue University Press, 2016).

period of time. Purdue[13], the University of Michigan[14], the University of Minnesota[15], and ETH Zürich[16] are all examples of institutions that have taken this approach.

And if nothing else, a maximum retention period gives UCSB time to develop a solution to the long-term preservation of research data—10 years in fact.

The next sections of this document lay out proposed approaches to implementing the Collective with respect to 1) the management platform that will provide the Collective's services and 2) the underlying storage architecture. In addition to establishing an operational repository, the pilot project will also answer these key questions:

- What role does the Library play in the research data lifecycle? What are its responsibilities? And, how do Library curators engage and interact with faculty and researchers? What services do they provide and what is the division of labor? There are at least two challenges here. One is balancing researcher appreciation for and need of curatorial assistance on the one hand, with the workload that can be sustained by the Library on the other. A second challenge is balancing the requirements and uniformity that are unavoidably imposed by centralized repository solutions on the one hand, with flexibility and accommodation of discipline-specific practices on the other.

- What will the cost of local data curation be? As noted previously, adoption tends to be less than anticipated. What will the growth rate of the Collective actually be, and how much data will be targeted for preservation beyond the Collective's limited timeframe? While the scope of the pilot project is limited, we believe that early numbers will be useful in gauging campus-wide behavior.

- What is the path to an institutional repository? If ADRL's scope were to be expanded to include faculty and researcher data, how could the Collective facilitate ingest into ADRL? Are there any ways that Collective requirements and policies should be aligned with those of an institutional repository?

Finally, by providing answers to these questions, the project will inform a campus-level discussion to be held at the conclusion of the pilot to determine how and by whom the Collective (and campus-based data curation services generally) will be funded, managed, and sustained.

[13] https://purr.purdue.edu/legal/digitalpreservation
[14] https://deepblue.lib.umich.edu/data/agreement
[15] https://conservancy.umn.edu/pages/drum/policies/
[16] http://www.library.ethz.ch/en/ms/Digital-Curation-at-ETH-Zurich/Research-data/Publishing-research-data

Management system

A management system (or “platform”) will be required to host the Collective, and in particular to control ingest of data, to control access to stored data, to provide value-added features, and to support the data’s curation. The specific requirements listed below are based on basic data curation and management needs as well as researcher desires and use cases as revealed in the aforementioned Data Curation Program's 2013 campus-wide survey and its subsequent 2017 survey of ERI and MSI principal investigators.

R1. **Data management.** Provide a means by which data can be ingested into the Collective, via direct upload by the researcher and/or via Library-mediated transfer; and, provide a means by which uploaded data can subsequently be managed, by the researcher, the Library, or both.

R2. **Visibility control.** Provide a means of distinguishing researcher-private data from public data, and a means of transitioning private data to public. The Collective is not intended to be used as a day-to-day project workspace, but support for private data nevertheless arises in at least two use cases: embargoing data before the associated publication appears; and assembling datasets that have complex structure or that are created or gathered over a period of time.

R3. **Data publication.** Support “publication” of public datasets, that is, persistent identification of datasets (specifically, by DOIs) and citation of datasets. Additionally, support rich description of datasets, including, in addition to standard abstract/scope/purpose metadata, the ability to bundle arbitrary documentation, contextual information, and links to external resources.

R4. **Branding and licensing.** Support UCSB branding of datasets, to enhance institutional reputation and clarify ownership and responsibility. Also, support, and make explicitly visible, dataset rights descriptions and licenses. Mandate a license for datasets (e.g., CC0) and/or offer licensing options (e.g., CC-BY for creative works)[17].

R5. **Search and browse.** Offer search and browse over stored datasets. Preferably, beyond offering these capabilities over the content in the Collective, offer the ability to serve as a general catalog of datasets, stored in the Collective or not, to allow researchers to more effectively gather project data and their collected works.

R6. **Large and complex data.** Support large datasets, in terms of individual file sizes, numbers of files, and total dataset size; and, support complex datasets, in terms of schema complexity and hierarchical structure.

[17] https://creativecommons.org/

R7. **Workflow access.** Provide programmatic means by which large datasets can be uploaded into the Collective; and, provide data access mechanisms that are sufficiently convenient and performant to support (re)use of data in existing workflow environments.

R8. **Administration.** Provide administrative and curatorial controls that allow Library curators to view, manage, and cull deposited content and to implement retention policies on content.

R9. **Customizability/extensibility.** Support multiple metadata formats and standards, including discipline-specific and geospatial formats; and, support discipline-specific access methods and search features. More generally, provide an extension mechanism that allows new metadata formats and standards to be added over time.

Past experience (with, for example, the original Alexandria Digital Library system and related services) has shown that it is difficult to develop and sustainably maintain a running service that is based on locally-developed (i.e., unique) software, for the maintenance burden is borne entirely by UCSB. Furthermore, attempting to reduce the maintenance burden by simplifying the software and services it provides is unlikely to suffice in this case. A “bare bones” or minimal platform (e.g., a basic filesystem with FTP server for data access) will provide too little support to ever implement the value-adding features listed above. Thus arriving at a platform for the Collective will not involve development, but will instead consist of selecting and installing an externally-developed platform solution, with local development limited to configuration and basic customization. This leads to two additional requirements:

R10. **Sustainability.** The platform should be open source and supported by a distributed and active community. There should be no significant platform lock-in in terms of file and metadata formats and database structures.

R11. **Scope**. The community should be responsive to the needs of data curation writ large, and not overly focused on any one discipline.

There are many potential data management platforms to select from. As will be described below, Dataverse is currently the leading candidate for the platform, but to give some context to the decision that needs to be made we list a few of the major platforms that exist today and note their strengths and weaknesses.

- The Dataverse Project[18] is “an open source web application to share, preserve, cite, explore, and analyze research data.” It was initially developed by Harvard, and Harvard continues to be the primary installation and backer, but today there are two dozen installations and contributions are starting to come from the Dataverse community.

[18] http://dataverse.org/

While Dataverse is novel in that it grew out of the social science community (and which also makes it unique in its extensive support for statistical data), support has been added by both Harvard and the community for other disciplines and types, notably geospatial, astronomy, and biomedical.

- HUBzero[19] is an "open source software platform for creating dynamic web sites that support scientific research and educational activities." It was developed at Purdue, and serves as the platform for the Purdue University Research Repository (PURR)[20], one of the earliest and leading efforts in this area. Locally, HUBzero was used as the platform for the NEES project[21]. While HUBzero supports data management and curation functions, its primary strengths are in providing a working environment in which datasets can be read and written and computational workflows can be assembled and run.

- Dash[22] is the California Digital Library's (CDL's) public interface to its Merritt repository[23]. Dash offers self-deposit of multi-file objects, but its support for large and complex datasets is limited and, as a pure software-as-service offering, provides no opportunities for curator engagement or administrative oversight. We note, however, that while Dash may not be a good fit for the Collective as presently constituted, in a recent UC-wide teleconference some of the requirements for the Collective listed above were presented, and they resonated with the other UC campus libraries. We anticipate that over the coming years CDL, with continued campus input, will enhance Dash along these lines.

- Hydra[24] "gives institutions a mechanism to combine their individual repository development efforts into a collective solution with breadth and depth that exceeds the capacity of any individual institution to create, maintain or enhance on its own." Technically, it consists of an ecosystem of components fundamentally based on the Fedora repository system[25]. Locally, Hydra is being used as the platform for ADRL. However, the experience of the Library and of other Hydra adopters is that significant development time is required to create a usable system. Nevertheless, given the potential for interoperability, Hydra merits additional review, particularly the Sufia and Hyrax front-ends to Hydra[26] which offer "common repository features and self-deposit and mediated deposit workflows."

[19] https://hubzero.org/
[20] https://purr.purdue.edu/
[21] https://nees.org/
[22] https://dash.ucop.edu/
[23] https://merritt.cdlib.org/
[24] https://projecthydra.org/
[25] http://fedorarepository.org/
[26] http://sufia.io, http://hyr.ax

- Metacat[27] is a "flexible, open source metadata catalog and data repository that targets scientific data, particularly from ecology and environmental science." Developed by the National Center for Ecological Analysis and Synthesis (NCEAS) for the DataONE[28] federation, of which UCSB is a coordinating member, it supports over twenty member nodes that host ecological data. Taking advantage of NCEAS's prior work would be advantageous, given its connection to UCSB, but further analysis is required to determine Metacat's applicability beyond its stated scope.

- The Open Science Framework (OSF)[29] is being developed by the Center for Open Science, which is currently grant-funded but seeking more sustainable funding. OSF offers team, project, data, and workflow managements functions, in addition to data publication features. OSF offers institutional branding, but otherwise operates as software-as-service and offers no administrative capabilities.

- Additional platforms include Figshare[30] (a free repository but that also offers institutional licensing at substantial cost); Islandora[31] (another Fedora-based content management system); and DSpace[32] (commonly used for institutional repositories and that supports ingest workflows).

In its background investigation the Data Curation Program has tentatively identified Dataverse as the platform that best fits the Collective's purposes. Dataverse supports all of the requirements to at least some degree, works "out of the box," offers many value-added features that we believe researchers will appreciate, would require little initial customization, and has an active and supportive community. Furthermore, a platform such as Dataverse would be complementary to ADRL (assuming ADRL policy is expanded in the future to encompass an institutional repository role). It should be noted that a number of institutions (Harvard, University of Virginia, MIT, and others) support Dataverse as one of the multiple data-related services they offer to their faculty, in the spirit that there is no one answer to the question of how to curate data.

The selection of the Collective's platform is not finalized, however, and the pilot project will convene a task force consisting of interested campus members both within and without the Library to evaluate Dataverse and other platforms to reach a final decision. In doing so, the pilot project will install, exercise, and investigate the features, performance, and reliability of the platforms, and will outreach to researchers to gain experience in populating the Collective with a wide variety of actual datasets. The end result will be an operational platform that will host the Collective; an initial set of datasets ingested into the Collective; experience gained by the

[27] https://www.dataone.org/software-tools/metacat
[28] https://www.dataone.org/
[29] https://osf.io/
[30] https://figshare.com/
[31] https://islandora.ca/
[32] http://www.dspace.org/

Library in operating the platform; and awareness by pilot subjects of the Collective's existence and value-adding features.

In addition to selecting a platform, the pilot project will:

- Configure and customize the platform installation for UCSB's purposes.
- Put in place a high availability production hosting environment and separate staging environment.
- Arrange external backup of the platform system and database.
- Integrate the platform's authentication and user identity system with UCSB's single sign-on and identity systems.
- Arrange administrative access and begin to develop administrative oversight procedures.
- Develop and document policy governing the purpose, scope, restrictions, limitations, and appropriate uses of the Collective. The development of this policy will be coordinated with the Library's "Task Force on Collection Development Policy for Faculty Data," recently convened as of this writing, and which is addressing accession of research data into the Library's own collections. We anticipate that the Collective's policy will fit into the Library's larger policy.
- Create user documentation. Additionally, create boilerplate language researchers can insert into data management plans to propose use of the Collective on their projects.
- Train Library staff (curators, but also subject librarians and department liaisons) in use and operation of the Collective. Develop support and help desk procedures.
- Develop outreach materials and presentations, and begin departmental outreach efforts.

Storage architecture

Regarding storage of research data, the high level goal of the pilot project is to gain enough experience to start assessing 1) the cost of local data curation and 2) the adoption and growth rates of a local solution by campus researchers. But it is also a goal to establish the foundation of a production system that will continue beyond the pilot phase, and thus it will be necessary to assemble a longer-lived, operational storage system. In doing so, the project will develop a storage architecture that is both cost-effective and sufficiently robust to serve as a preservation-supportive platform.

The proposal is to develop a local storage repository with an initial net capacity of 250TB. The repository will consist of three storage arrays, managed and mirrored using ZFS[33], at least one of which will be physically located in the North Hall data center and at least one of which will not. A fourth copy of Collective data will be maintained at Box.com using UCSB's institutional account, synchronized using software developed by this project (the software will be reusable by ADRL and other campus repositories for the same purpose). The capital cost of the local storage component is estimated at $78,000 and will be procured, installed, and initially managed by ERI. ERI has agreed to manage the storage *gratis* for the duration of the pilot project. The remainder of this section discusses this decision.

Storage is at once the easiest part of assembling a repository and the hardest. It is the easiest in the sense that the approaches, technologies, and vendors are all well-known and well-understood; the question is not *whether* a storage system can be built, but only at what cost and with which features and tradeoffs. But storage is the hardest part in that it represents the greatest single capital cost.

The proposed initial size of the Collective is 250TB. In the aforementioned survey of ERI and MSI principal investigators, which received a 40% response rate, respondents identified 500TB of data they would like to see archived, though this number must be tempered by the reality that, in practice and when confronted with repository deposit requirements, not all researchers will actually participate. A closer analysis of dataset sizes reveals that 250TB would be sufficient to accommodate up to 30 researchers and their datasets, including at least one large dataset on the order of 100TB. We believe an initial capacity of 250TB will be sufficient to be able to evaluate the Collective over the course of the pilot project and to subsequently support opening the Collective up to campus as a whole. Furthermore, in the vendor quotes received to date, 250TB corresponds to standard storage rack and incremental expansion sizes.

The performance of the storage system must be such that it supports active use of stored data by researchers in existing computational environments.

Beyond capacity and performance, the major requirement of the storage system is that it protect against loss, which is a critical concern for any preservation-supportive repository. What constitutes sufficient protection is not well-defined, however. Beyond the obvious dictum that "more copies are better," and despite formalized threat models and risk analysis[34], there are no formally recognized standards or minimum requirements in this area. The closest to a recognized standard is perhaps the so-called "3-2-1 rule," which is a shorthand way of saying one should maintain at least 3 copies of data, on at least 2 different types of media or systems, at least 1 of which is physically separated[35]. This rule originated in the context of personal

[33] https://en.wikipedia.org/wiki/ZFS
[34] David Rosenthal (2014). What could go wrong? DSHR's Blog. http://blog.dshr.org/2014/04/what-could-possibly-go-wrong.html
[35] Paul Ruggiero and Matthew A. Heckathorn. Data Backup Options. US-CERT. https://www.us-cert.gov/sites/default/files/publications/data_backup_options.pdf

archiving, but its intentions carry over and it can and has been applied to large-scale repository architectures as well. The purpose of having multiple, distinct copies is to avoid loss of a single copy, of course, and the purpose of having at least 3 copies is to avoid being in the delicate situation of having only one copy remaining if the only other copy is lost. The purpose of employing 2 media types (or system types, or technologies) is to reduce the risk of correlated system failures, that is, simultaneous failures that are due to using the same underlying technology. And the purpose of having at least one physically remote copy is to reduce the risk of physically correlated failures, i.e., simultaneous failures that arise due to the copies being in the same physical space. A further risk to reduce is that of operator-correlated losses, i.e., simultaneous losses due to human factors ranging from the inevitable mistakes to rogue system administrators. However, mitigating against this last threat is beyond the scope of this pilot project, for it would require yet more copies and a distributed, protocol-based storage architecture such as LOCKSS[36].

Because there is no standard for reducing risk of loss, many approaches have been taken that differ both quantitatively and qualitatively, making direct comparisons difficult. To give a sense of the breadth of contemporary solutions, listed below are some relevant approaches and their associated costs. In this discussion storage costs are expressed in dollars per terabyte per year. In practice, hardware is typically paid for once and then maintained over its warranty period, but annualizing the purchase price allows for more uniform comparisons. Also, we are careful here to distinguish single terabytes from replicated terabytes. Replication is omnipresent in storage systems; even a single physical disk drive employs some redundancy internally. For the purposes of this discussion, however, we will consider storage to be replicated only if it is replicated in the sense of the 3-2-1 rule, that is, replicated across different systems and/or technologies and/or physical locations.

- CDL freely provides the Merritt repository, and the Dash interface to Merritt, charging only for the storage used on a cost recovery basis. The storage is highly replicated and its current rate is $650/TB/year. CDL has done much work in cost modeling and in examining the space of solutions, particularly those offered within the UC system. Stephen Abrams, Merritt's storage architect, explains its approach and costing as follows [37]:

  "The customer cost for using Merritt (and now, Dash) has always been based on our recovery cost for provisioning the consumed storage capacity. In the old days when we were running all of our own computing equipment, we went through a pretty involved process to determine a plausible number of the full economic cost of the storage, including data center hosting, network charges, some portion of storage administrators' time, etc. Since then, we first moved to relying on SDSC's private cloud service[38], offering 3 replicas on independent arrays in a single datacenter, then added a similar

[36] https://www.lockss.org/
[37] Stephen Abrams, personal communication.
[38] https://www.sdsc.edu/services/it/cloud.html

private cloud service at UCLA[39], offering 2 replicas in independent datacenters, which we recently deprecated in favor of AWS S3 (for bright content) and Glacier (for dark content), whose SLA's imply a minimum of 3 replicas.

SDSC is based on the OpenStack Swift API[40] and has a simple cost model: $32.16/TB/mo = $390/TB/yr. UCLA is based on ZFS-mirrored arrays that are exposed as an NFS share, and also has a simple cost model: $250/TB/yr. We then added on a $10 'contingency' fee (to build up a modest surplus in readiness for some future migration or unexpected mitigation activity) to arrive at our current $650/TB/yr price point.

Even though we have now replaced UCLA with S3 and Glacier, we are still continuing with the same price for the time being, although our modeling suggests that we should expect a reduction in our overall costs of around 20%, which, if realized, we would pass along as a reduced price point (~ $520), probably in the next fiscal year."

- Supporting the Alexandria Digital Research Library (ADRL), the Library manages a single 180TB NetApp RAID-DP system housed in the North Hall data center. In a recent proposal to NSF, expanding the capacity of this facility was costed at approximately $120/TB/year.

- The University of Colorado's PetaLibrary[41] is "a National Science Foundation-subsidized service for the storage, archival, and sharing of research data." It has taken a different approach, and instead of offering one storage solution, offers a variety of storage options at different price points selectable by the researcher. A single copy of data can be stored for $65/TB/year, with additional copies (both online copies and tape backups) available at additional cost.

- Aristotle[42] is a 5-year, NSF-funded, multi-institution project to create a "federated cloud for academic research," allowing institutions to pool, and researchers to borrow, storage resources. UCSB is one participant in the federation, and the Letters & Science Information Technology (LSIT) group has built a node that operates on a condo principle [43] with access provided by CEPH[44], which supports both Amazon S3 API and Posix filesystem access. While a research project is not necessarily the best fit for a longer-term data repository, we note that novel storage architectures such as Aristotle

[39] https://idre.ucla.edu/cass
[40] https://docs.openstack.org/developer/swift/
[41] https://www.rc.colorado.edu/resources/storage/petalibrary
[42] https://aristotle.ucsb.edu/
[43] http://csc.cnsi.ucsb.edu/overview/condo-clusters
[44] http://docs.ceph.com/docs/hammer/

and Berkeley's OceanStore[45], and peer-to-peer data sharing and publishing technologies such as Dat[46], are highly relevant in this space and merit further investigation.

This proposal is to leverage the considerable expertise of UCSB's computing centers. The Life Sciences Computing Group (LSCG) recently assembled a 1.6PB storage solution at a capital cost of $18/TB/year, and the hardware's performance and reliability have reportedly been positive to date. A smaller system that does not amortize certain costs quite as effectively would cost $21/TB/year, with further expansion in 250TB increments up to 1PB for only $14/TB/year. The storage systems would be managed and mirrored with ZFS, which local system administrators have many years of experience with and have expressed much faith in.

ERI has offered to take the lead on procuring, installing, and managing the storage, and both ERI and LSCG IT staff have offered to donate their time for this pilot project. Should ERI continue to host Collective storage following the pilot phase, its nominal recharge rate is $120/TB/year (single copy), or $360/TB/year for its standard "primary copy plus two backup copies" offering.

In addition to on-campus storage, it will be highly desirable to maintain a copy of Collective data physically separate from the UCSB campus, preferably distant enough that it would not be affected by natural disasters that might befall the Santa Barbara environment. Given 3 copies on campus, an additional copy may be relied on for disaster recovery only, i.e., it need not be immediately online. UCSB already has a fixed-cost, unlimited-storage institutional account with Box.com, a cloud storage provider, and thus it would be most cost effective to use Box.com for this purpose. However, Box.com is not without its risks. There have been anecdotal reports of difficulties in syncing data. And it is unknown if Box.com will function effectively at the 10TB to 100TB scales. (While cloud providers may offer unlimited storage, they rely on the fact that most users will not store very much data in practice, and too, they have additional means, e.g., throttling, to make it difficult to take full advantage of advertised capacities.) The pilot project will evaluate the viability of using Box.com, will evaluate different upload tools and synchronization techniques, and will create approaches to address known Box.com restrictions such as caps on individual file sizes and disallowed characters in filenames.

Should Box.com not work out, the pilot project will pursue other strategies for storing data off-campus. One fallback strategy is to look at other, lower-cost cloud storage solutions, Amazon Glacier in particular. Another is to arrange a mutual backup arrangement with a data center on another UC campus. Local system administrators have related that they have friendly working relationships with counterparts on other campuses, so such an arrangement may be possible. And this type of collaboration would represent a concrete way of leveraging the fact that UCSB is not a standalone campus, but a member of a larger system. It is noteworthy that, along these same lines of thought, CDL is re-evaluating its storage architecture for Merritt,

[45] https://oceanstore.cs.berkeley.edu/
[46] https://datproject.org/

looking to replace its centrally-managed solution with a collaborative model of campus-contributed storage pools.

All storage solutions require refreshing and reevaluation over time, and this one is no different. Any solution should be considered temporary, to be replaced by a newer technology. The hardware purchased for this pilot project would carry a 5 year warranty, and that would define its lifetime.

Staffing and resources

The pilot project will be led by Greg Janée, Director of the Data Curation Program.

Evaluating and setting up the Collective platform, and developing a cloud backup mechanism, will be performed by Greg and other Program members, and will be fit into the normal operations of the Program. The Data Curation Program and, more broadly, the Library, will continue to support and operate the Collective following the conclusion of the pilot project.

Procurement and installation of storage will be led by Michael Colee, Director, ERI Computing, who will consult with the Life Sciences Computing Group (LSCG) and other computing groups as necessary. ERI has agreed to cover the cost of Michael's time for this project. ERI and LSCG will track the staff time expended on the pilot project to inform subsequent discussion of data curation costs. Additionally, if desired, Michael will train Library system administrators in the administration of ZFS-based storage systems.

The services of a graphic designer may be required to develop graphical elements for the Collective online system and documentation and outreach materials.

Training time will be required for Library data curators, subject librarians, and other liaisons to campus departments.

Project costs will consist primarily of a capital expenditure of approximately $78,000 to purchase an initial 250TB (net) storage. Relatively minor costs ($500) include local VM hosting of the Collective platform and an SSL certificate.

Over the long term, following the end of the pilot project, additional storage purchases will be required as storage needs increase and as storage systems age out of their 5-year warranty periods.

Timeline

The following is a 13 month plan. The storage and backup layer will be in place within 7 months. The other parts of the project fall into three rough phases: a 2 month setup phase; a 6 month evaluation phase; and a 5 month implementation and outreach phase.

Month:	1	2	3	4	5	6	7	8	9	10	11	12	13
Storage													
Purchase and install storage	■	■	■										
Create, evaluate Box.com backup mechanism				■	■	■	■						
Platform evaluation													
Identify test platform(s)	■												
Identify initial test researchers, datasets	■	■											
First round of evaluations			■	■	■								
Identify additional test datasets, use cases					■								
Second round of evaluations						■	■						
Decide on platform								■					
Production implementation													
Set up, customize production environment, system									■	■			
Integrate with campus single sign on										■	■		
Develop and approve policy									■	■	■		
Create documentation and DMP language											■	■	■
Develop support procedures												■	■
Outreach													
Train Library staff											■	■	■
Develop departmental presentation												■	■
Announce availability													■