Services for the Management and Preservation of Research Data

Executive Summary

The scientific and technical data generated in the course of research and scholarly activity lie at the very heart of the research enterprise. Such data is of immediate use to researchers as they conduct their investigations, and may prove to be a resource of continuing value to the researchers and to others over time. NIH encourages timely release and sharing of research results of its grantees, and requires retention of all data for three years following the final expenditure report. NSF requires a data management plan of its grantees. Overall, we have the following issues:

- The management of research data is ad hoc, potentially resulting in lost or unusable information.
- Federal grant agencies are noticing these issues and starting to impose requirements for data management.
- Costs are ongoing, but grant funding is not.

Data is not static; it begins either as observed data or as a human creation, most likely the product of collaboration. As a research project proceeds, that data is revised, annotated, formally described, indexed, disseminated, and published. Even after the end of a project, data selected for long-term preservation mus periodically be transformed to utilize evolving storage technologies and file format standards.

Planning is required to ensure the usefulness of data over time, and those plans require the existence of technical and curatorial services that are typically provided by the institution. These services are categorized by their phase in the data life cycle, either during a research project (when the researcher is the primary curator of the data), or after completion of the project (when the institution is the primary curator).

- Services required during a research project
    - File storage
    - Services for creation, ingest, description, annotation, indexing, revision, and publishing of data
    - Consultation to assist in the creation and implementation of data management plans
- Services required after the completion of a research project
    - Highly-reliable, "write once, read forever" storage services
    - Curation to assure long-term usefulness of the data

We propose an ongoing, UC Davis provided infrastructure that facilitates management the intellectual raw materials, interim results, and products of research in such a manner that the data not only enhances the immediate process of research, but also remains useful in the future. While we envision this as a UC Davis program, we view the operation of this program as a collection of strategic partnerships with other

universities, institutions, and commercial providers that are managed to deliver an optimal balance of service, risk, and cost.

We further propose, as a strawman, a twelve-month first-phase project to start delivering these services. The estimated cost of this first phase is $570,000, and focuses on the deployment of basic storage services and assurance that UCD's obligations to federal grant agencies are met. This includes:

- Consulting services to assist researchers with the creation and implementation of their data management plans.
- A UCD-hosted file storage service, available both as network mountable file structures (primarily for internal use), and as web-accessible repositories (primarily for external use).
- Provide preservation of files through the CDL's Merritt service.

Introduction

The scientific and technical data generated in the course of research and scholarly activity lie at the very heart of the research enterprise. Such data is of immediate use to researchers as they conduct their investigations, and may prove to be a resource of continuing value to the researchers and to others over time. NIH encourages timely release and sharing of research results of its grantees, and requires retention of all data for three years following the final expenditure report. NSF requires a data management plan of its grantees.

Overall, we face the following issues:

- The management of research data is ad hoc, potentially resulting in lost or unusable information.
- Federal grant agencies are noticing these issues and starting to impose requirements for data management.
- Costs are ongoing, but grant funding is not.

This white paper outlines a set of storage needs for research and proposes campus-level services to address those needs that are not easily addressed by individual researchers.

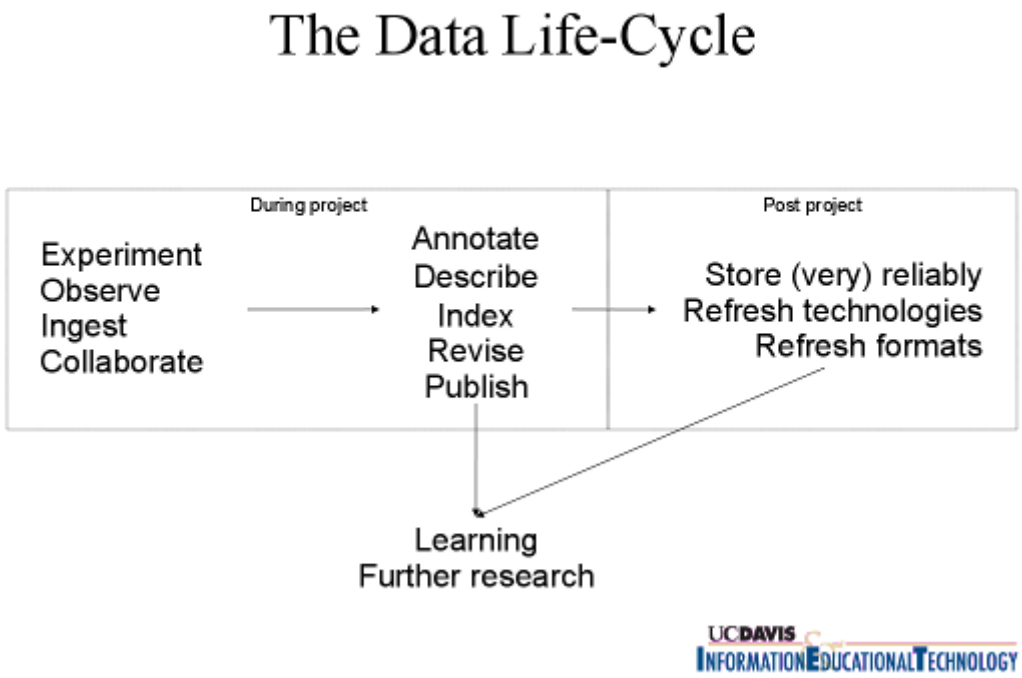
The Data Life Cycle

Research data is not static. It passes through various stages of refinement, selection, and transformation:

- Research projects collect and create data in various ways. Experimental and observational data is ingested into project data stores People collaborate to create papers, software, and other artifacts of the research process.
- Once this data has been incorporated into a project's data store, the process continues. Selected information is revised, described, annotated, indexed, and published.

- Finally, at the end of a project, information is preserved for posterity. Even in this stage, it must be transformed to accommodate to changing storage technologies and usable file formats.

The following diagram illustrates these processes:



Research Data Storage Needs

Storage needs can be classified according to the retention periods addressed by the services:

- The lifetime of a research project
- Long-Term Preservation

Services addressing storage over these time periods will have differing technical requirements for aspects like performance, replication, and backup. In general, the need for high performance decreases as the retention period increases, while the need for long-term survivability of the information increases. Also, the nature of backup is different for long-term preservation from the others, as the stored objects do not change over time.

The primary stakeholder for the preservation of data changes as the retention period increases. For research projects, the primary stakeholder is the researcher. For long-term preservation, however, the institution becomes the primary stakeholder.

Also, the need for curation services increases as the retention period increases. These services are typically provided by librarians and archivists, not IT professionals, so a comprehensive strategy must involve a partnership.

The following table summarizes these needs:

	Performance	Survivability	Curation	Primary Curator
Lifetime of a Research Project	High	High	High	Researcher
Long-Term Preservation	Moderate	Very High	Very High	Institution

Storage is also required for the lifetime of a computation, but that is well understood and well within the reach of individual researchers.

The serious issues facing UC Davis are data management requirements during a research project and for long-term preservation, how data is managed when the original funding and personnel that collected and created that data are no longer available. We will address long-term preservation first.

Long-Term Preservation

The typical progression at the end of funding for a research project is that the project participants perform an *ad hoc* selection of the project's data (experimental data, analyses, simulations, results, published papers, etc.) and preserve that data in whatever manner is available. It may be available for retrieval by others over the Internet, or it may be recorded on a stack of floppy disks in the closet, but the researchers know where it is. Over time, however, researchers move on to other interests, floppy disks become obsolete, and the data is lost if stewardship of that data has not been passed on to someone else.

Two complementary services, operating in tandem, are required to address this issue, one primarily technical, a storage service led by IET, and one primarily curatorial, led by the University Library:

- The storage service provides long-term, highly-reliable storage of the binary representation of research data
- The curatorial service provides long-term stewardship of the stored information

In the digital world, providing one of these services without the other will not result in long-term preservation of usable data. Providing only the technical service will result in bits that are stored forever, but that cannot be found and are in obsolete formats if they are. Providing only the curatorial service does not enable storage in the first place.

The Storage Service

The storage service creates one-time replicated copies of files that have been submitted and provides network access. The following are included in the storage service:

- A unique and persistent identifier for each submitted file to enable retrieval
- Integrity checks to detect and correct corrupted copies of the data
- Administrative metadata to identify stewardship of the data
- Migration to new storage technologies over time

This is a "write once, read forever" storage service. New versions of files can be submitted to the service, but old versions are never updated or deleted.

The Curatorial Service

The curatorial service utilizes the storage service to manage collections of information that are made up of the stored files. The following are included:

- An inventory service, including administrative and descriptive metadata about the files
- Indexing and search services to enable discovery of the information
- Format transformation over time to ensure the information is usable by then-current technology

Data Management during a Research Project

The challenge for data management during a research project is to ensure that the data created by the project can be made useful to others, particularly after the lifetime of the project. Technically, this means the use of standard file formats and access methods. From a curatorial point of view, this means data management practices that ensure creation of descriptive metadata, indexing, dissemination, publishing, *etc*.

There are two services that are needed:

- Consultation for researchers to assist in the creation and implementation of data management plans.
- A file storage service with associated tools for the creation, ingest, description, annotation, indexing, revision, and publishing of data. There are two important characteristics for these tools:
    - the tools should support collaborative management of the data, and
    - the tools should facilitate good data management practice and preservable data.

What Are Others Doing?

Princeton

Princeton has established DataSpace (http://dataspace.princeton.edu/jspui/), "...a digital repository meant for both archiving and publicly disseminating digital data which are the result of research, academic, or administrative work performed by members of the Princeton University community." Content is managed by communities of interest, such as Civil and Environmental Engineering and the Woodrow Wilson School of Public and International Affairs. Princeton's DataSpace uses a "Pay Once, Store Endlessly (POSE)" pricing model that preserves data for an initial submission fee of $0.006/MB.

UCSD

UCSD is beginning to provide data management and preservation services as part of their comprehensive Blueprint for the Digital University (April, 2009) (http://research.ucsd.edu/documents/rcidt/RCIDTReportFinal2009.pdf).

UC Curation Center (UC3)

The California Digital Library's UC Curation Center (UC3) (http://www,cdlib.org/services/uc3/) provides preservation storage services (Merritt - http://www.cdlib.org/services/uc3/merritt/index.html), using technical services provided by SDSC; the price is $1,090/TB/year. Curation "micro-services" are available and can be used as building blocks for a data management implementation. UC3 also provides consulting services to help researchers with the creation and implementation of their data management plans.

Discipline-Specific Repositories

Certain academic disciplines have created repositories for data that is of general interest to researchers in the field. For example, The Incorporated Research Institutions for Seismology (IRIS) Data Management Center (http://www.iris.edu/dms/dmc/) "...is a consortium of over 100 US universities dedicated to the operation of science facilities for the acquisition, management, and distribution of seismological data." IRIS is funded by the NSF, and seismology researchers can request IRIS resources as part of their grant proposals. Long-term preservation is not explicitly addressed, although IRIS has existed since 1984.

A Proposed Beginning

The Vision

An infrastructure that leverages both local and remote resources to facilitate management of the intellectual raw materials, interim results, and products of research in such a manner that the data not only enhances the immediate process of research, but also remains useful in the future.

Phase I Goals

In the short term (12 months), ensure that obligations to federal grant agencies are met and deploy basic storage services. As a strawman, this would include:

- Consulting services to assist researchers with the creation and implementation of their data management plans.
- A UCD-hosted file storage service, available both as network mountable file structures (primarily for internal use), and as web-accessible repositories (primarily for external use).
- Provide preservation of files through the CDL's Merritt service.

The following table contains estimated costs for such a strawman implementation. The estimates for Storage and Merritt Preservation Fees scale approximately linearly with the amount of storage and proportion of that storage that is preserved.

Estimated Costs for Phase I	
Storage (250 TB @ $500/TB)	$125,000
Server Hardware	$50,000
System Administration (1 FTE)	$100,000
Software Development / Acquisition (1 FTE)	$120,000
Service Management (0.5 FTE)	$60,000
Consulting Services (0.5 FTE)	$60,000
Merritt Preservation Fees (20% preservation - 50 TB @ $1090/TB/year)	$55,000
Total	**$570,000**

Longer-Term Goals

The following tasks will be required to complete the vision stated above:

- Addition of layered services on the file storage service to facilitate ingest, collaboration, indexing, dissemination, publication, etc.
- Implementation of a pricing model for the services, including a "Pay once, store endlessly (POSE)" model for the preservation service.
- Identify funding mechanisms faculty can tap to pay for these services.

This is only as we understand that vision today, however; the needs for data management services will evolve over time. Researcher input and collaboration with peer institutions will be needed to chart UCD's course forward. There are also other collaboration services that are important to modern research, although they are less associated with data management. These include:

- Video and audio group communication services
- Researcher information systems that assist in the discovery of potential collaborators

Appendix: NSF Data Management Requirements

Excerpt from NSF 11-1 January 2011 - Chapter II - Proposal Preparation Instructions

Plans for data management and sharing of the products of research. Proposals must include a supplementary document of no more than two pages labeled "Data Management Plan". This supplement should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results (see AAG Chapter VI.D.4), and may include:

1. the types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;
2. the standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies);
3. policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;
4. policies and provisions for re-use, re-distribution, and the production of derivatives; and
5. plans for archiving data, samples, and other research products, and for preservation of access to them.

Data management requirements and plans specific to the Directorate, Office, Division, Program, or other NSF unit, relevant to a proposal are available at: http://www.nsf.gov/bfa/dias/policy/dmp.jsp. If guidance specific to the program is not available, then the requirements established in this section apply.

Simultaneously submitted collaborative proposals and proposals that include subawards are a single unified project and should include only one supplemental combined Data Management Plan, regardless of the number of non-lead collaborative proposals or subawards included. Fastlane will not permit submission of a proposal that is missing a Data Management Plan. Proposals for supplementary support to an existing award are not required to include a Data Management Plan.

A valid Data Management Plan may include only the statement that no detailed plan is needed, as long as the statement is accompanied by a clear justification. Proposers who feel that the plan cannot fit within the supplement limit of two pages may use part of the 15-page Project Description for additional data management information. Proposers are advised that the Data Management Plan may not be used to circumvent the 15-page Project Description limitation. The Data Management Plan will be reviewed as an integral part of the proposal, coming under Intellectual Merit or Broader Impacts or both, as appropriate for the scientific community of relevance.

Appendix: NIH Data Management Policy

Excerpt from "NIH Grants Policy Statement (12/03) Part II: Terms and Conditions of NIH Grant Awards"

Sharing of Research Data

NIH believes that data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health. NIH endorses the sharing of final research data to serve these and other important scientific goals and expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers. "Timely release and sharing" is defined as no later than the acceptance for publication of the main findings from the final data set. Effective with the October 1, 2003 receipt date, investigators submitting an NIH application seeking $500,000 or more in direct costs in any single budget period are expected to include a plan for data sharing or state why data sharing is not possible.

NIH recognizes that data sharing may be complicated or limited, in some cases, by organizational policies, local IRB rules, and local, State and Federal laws and regulations, including the "Privacy Rule" (See "Public Policy Requirements and Objectives—Requirements Affecting the Rights and Welfare of Individuals as Research Subjects, Patients, or Recipients of Services—Confidentiality—Standards for Privacy of Individually Identifiable Health Information"). The rights and privacy of individuals who participate in NIH-sponsored research must be protected at all times. Thus, data intended for broader use should be free of identifiers that would permit linkages to individual research participants and variables that could lead to deductive disclosure of the identity of individual subjects.