**NSF’s Data Management Plan Requirements**

**Kansas State University Guidance for Sharing and Archiving Data**

**Background**

 Recognizing the value of data, as well as its results, the NSF aims to allow scientists to have access to data funded through NSF efforts as they currently have access to other products. As a result , starting January 18, 2011, NSF now requires a data management plan to be included with every proposal. NSF guidance (GPG, II.C.2.j) suggests that these plans address the following:

* the **types of data**, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;
* 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);
* **policies for access and sharing** including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;
* policies and provisions **for re-use, re-distribution, and the production** of derivatives; and
* **plans for archiving data**, samples, and other research products, and for preservation of access to them.

The most troublesome aspects of these requirements are the access/sharing and archiving of data particularly after the research grant is complete To aid researchers in this aspect of their data management planning, the Office of Research and Sponsored Programs assembled a work group to develop this guidance document which specifically addresses these issues. For further information on developing the plan itself, see: <http://www.k-state.edu/research/resources/writing/Data%20Mgmt/NSF%20Data%20Mgmt/index.htm> and

<http://guides.lib.k-state.edu/content.php?pid=298773&sid=2450849>.

**NSF Requirements**

The NSF is deliberately vague regarding the access/sharing and archiving requirements, only indicating they will be shaped over time in a field-specific manner. The NSF also does not specify whether local, regional, or national data stores should be used, what the data formats should be, how the data should be accessed, what data will be made available, and what the timeline for availability should be. All of these issues are considered to be domain-specific. That being said, the NSF has only two immutable requirements at this time:

1. The data specified in the data management plan must be stored, and
2. It must be made available at no more than “incremental” cost.

It is assumed that as standards evolve, the expectations will evolve to include: 1) searchability requirements for the data (i.e., does the data have easily searchable metadata to locate the data products) and 2) is the data, once located, readily available through the Internet. This data may be in human-accessible forms such as HTML, or electronically-oriented formats such as IRODS.

NSF is very serious about the dissemination and sharing of research data. As the agency’s approach to data management plans evolve, it is likely that data storage on personal computers with access via the PI will no longer be acceptable solutions. Below, current K-State sharing and archiving options available to researchers are outlined as well as a strategy for the future.

**K-State Libraries**

The Libraries can assist researchers in developing their data management plans for grant proposals. With the new NSF requirement, researchers and librarians can collaborate during the grant-writing process to plan for the curation of research data before the data have been generated. Together, they can determine the researchers’ needs for sharing, distributing, and managing research results, as well as identify appropriate metadata standards and policies for access and use. A newly established position of data services librarian (Meagan Duever, mjduever@ksu.edu ) is available to assist researchers developing data management plans for grant applications and will work with campus partners to develop a broader infrastructure to support data curation at Kansas State University.

One possible data solution, particularly for smaller data sets, is the K-State Research Exchange (K-REx). Operated by the Libraries, it is the institutional repository and digital archive of scholarly works produced by K-State faculty and students. K-REx accommodates most data files and data sets and includes searchable metadata. It also provides a permanent URL, long-term access and links to publications that result from the research. Works in K-REx are openly accessible to the public. In some cases, access can be restricted to certain groups for a particular time. See information about depositing at <http://krex.k-state.edu/dspace/>

**Constraints and Constants**

Data produced through scientific research tends to vary broadly in two major axis – size and activity – which dramatically alter the scale and practicality of various approaches to a data storage strategy. It is generally felt that 90% or more of the archived data will seldom or never be accessed. We assume large datasets (terabytes to petabytes) will also tend to be impractical to transmit dynamically for the short term, instead requiring the shipment of physical media. Subsets of large datasets, however, may be able to be made instantly accessible. We also assume that web-enabling curated data will take nontrivial resources to do well, and may not be considered worthwhile for some datasets.

We can organize these characteristics into a quadrant indicating an approach to each class of data.

|  |  |  |
| --- | --- | --- |
|  | Highly active | Largely inactive |
| Small data (KB to GB) | Make searchable and available online  | Make available on request |
| Big data (GB to TB) | Make searchable online and available on request | Make available on request |

So for the various data classifications above, individual access strategies are needed as is outlined in the bullets below:

* Small/active – make both the metadata and the datasets themselves readily available over the web in human- and computer-oriented forms.
* Small/inactive – make the data available on request through the individual researcher(s), conserving curator and researcher time while meeting NSF requirements.
* Large/active – make the metadata easily searched and discoverable, but send the data only on request on physical media (to avoid overwhelming network connections), and/or make subsets available through web services.
* Large/inactive - make the data available on request through the individual researcher(s), conserving curator and researcher time while meeting NSF requirements.

Likely all data will be stored in a central professionally-managed repository and any non-web-enabled metadata approach will be deprecated in the long term. Ultimately, where a dataset resides will depend on the judgment of the researchers involved

 As an institutional repository, K-REx can readily accommodate small data (active and inactive) and selected large data sets. However, data and data not stored in typical file-based structures such as enterprise geodatabases (Oracle, SQL Server, etc.), large scale file geodatabases (ESRI ArcGIS), or imagery and LiDAR (Light Distance and Ranging) datasets present unique challenges in storage and availability. The best option for these datasets involves the central storage in enterprise geodatabases connected to fast disks such as SANs. They could be made accessible by both researchers and supercomputing clusters via fiber or, at a minimum, 1 GB connections. This approach would also avoid problems of upload when archiving the data. The ability to serve these resources efficiently and effectively (or work across project teams with multi-user access and versioning) requires storage in an enterprise level geodatabase such as Oracle or SQL Server. Clients would connect via a client side ArcSDE (spatial database engine) connection. Implementation of an enterprise database and deployment of site license software and instructions can easily allow researchers to securely connect to the database and upload data resources. Additionally, the enterprise geodatabase(s) can efficiently be made web available using tools such as ArcServer, ArcEngine, and/or ArcMobile.

The ArcGIS Portal Toolkit can be used to create a metadata catalog of all data stores in the central repository and to deliver a browser based search site enabling searches by key words or by variables and spatial location. The Portal toolkit could potentially be linked to the K-REX database allowing spatial searches of data stored in K-Rex. To be effective and efficient, specialized staff members with geospatial database tuning and management training are essential as are map caches of data.

Map caches index large data sets and create multiple versions of the data at lower resolutions to speed the search process and delivery via the web. However, creating map caches is a processor intensive operation (suitable for the BeoCAT cluster) that typically a doubles the storage requirement for the actual dataset. At present K-State has the required software via the ESRI Site License, but lacks the central storage space, enterprise geodatabase instance(s), and qualified staff to build, manage, maintain, secure and serve enterprise geodatabase stores. Resolution of this issue is will take 1 to 2 years and is part of Phase 2 in the phased implementation strategy proposed below.

**General Thoughts**

The plethora of current data repositories scattered throughout servers, desktops, USB keys, DVDs, and external hard drives represents a compromise between economy (cheap), administrative overhead (minimal), security (minimal), and longevity (problematic). This approach is unlikely to be viewed favorably by NSF review panels even in the near term, leaving KSU at a significant competitive disadvantage for research dollars. We must address these issues quickly, providing a solution that is professionally managed, has reasonable assurances of data availability over time, and is both financially and administratively economical .

A significant number of research efforts will be better served by national disciplinary repositories – e.g., carbon flux tower data from the Konza could be included in the Ameriflux system – and thus will not need a KSU system (and conversely, should not be required to use it). However, not all projects have national repositories, and enough are likely to have significant data storage requirements that a petabyte-class storage architecture is likely to be required in the medium term (3-5 years).

We also note that the NSF specifically has said archival costs can be included in proposal budgets. On the other hand, any data storage approach needs to provide continuity (indicating an official university branch rather than an individual researcher’s resources are desirable), economy (clearly data storage funding detracts from the actual research activities), and researcher funding varies over time. Typical data retention times will vary by discipline, but range from three to ten years or more beyond the grant period. Thus, storage costs need to be covered during a typical 3-year grant but will extend beyond the end of the grant. Costs will be met through a combination of grant and university resources, which can be negotiated on an individual basis.

**Suggested Strategy**

We suggest a two-phased strategy to provide a coherent, university-wide solution to these issues. For the short term, we suggest economical storage be available to the researchers which is administered by Central IT Staff (CTS) but managed by the individual researcher (or project), in addition to existing solutions such as K-REx. For the longer term, we suggest a unified web-enabled data store for projects where a discipline-specific solution is not available. The details for each phase are outlined below.

**Phase 1** (2-3 months implementation for 2-50 TB total, active/inactive): CTS agrees to make available storage to research projects at the cost of the media (ongoing costs are absorbed by KSU), and further agrees to keep this capacity and contents available to the project for a minimum of five years. The PIs sign an agreement to comply with all relevant laws (HIPAA, FERPA, etc.), and handle any requests for the data. Any data requests beyond the capability of the PI to deal with (such as very large data transfers) are handled by CTS at a nominal cost. Data security is the responsibility of the PI. (We can model this approach after what is being done for the Oklahoma universities by the OU Super Computing Center for Education & Research (OSCER).) K-REx and other existing solutions are used for small/active datasets.

**Phase 2** (1-2 years implementation for 100TB – 1PB total, active/inactive): We add on the capability to provide easy online search capabilities on metadata for data stored within our system (active or inactive), and provide online access where practical to the data itself (small and potentially large datasets). This structure will allow us to capitalize on NSF research projects such as the NEON data repository and the DataONE project for research data access, as well as efforts such as those at the University of Missouri (<http://libraryguides.missouri.edu/content.php?pid=194910&sid=1633229>), and Purdue University (<http://www4.lib.purdue.edu/dcp/overview> . We will also add capabilities for non-tabular data such as GIS. Because the metadata and much of the data is now web accessible, the PIs and metadata managers will collaborate to provide a security model for data access which meets discipline and grant requirements.

In general, we suggest that the Library, through K-REx and other resources that become available, will be responsible for access control policy, metadata management, and storage of datasets were practical. We suggest central IT be responsible for providing and managing storage for user-managed datasets, access control for these, and collaborate with the Library to implement any joint security or access control needs.

**Information for PIs**

PIs in the process of writing grants or producing data from existing projects may contact:

* The CTS Data Management lead at (Chris Casey-- 532-3517 or csc@ksu.edu) to discuss needs and pricing for storage, and
* The Office of Sponsored Research (Mary Lou Marino--532-3211 or mlmarino@ksu.edu) for sample data management plans and text for inclusion in NSF-sponsored research proposals.
* The Libraries (Meagan Duever--532-7447 or mjduever@ksu.edu) for consultation on developing a data management plan (Note - we do not yet have a phone number or email address for the newly hired data services librarian who start work December 1st.)
* The K-State Research Exchange (Marty Courtois--532-4428 or courtois@ksu.edu) to archive data sets in K-Rex
* Outside archival and sharing resources:
	+ <http://oad.simmons.edu/oadwiki/Data_repositories>
	+ <http://www.lib.umn.edu/datamanagement/datacenters>