Chapter Four: Perspectives on Digital Data Collections Policy

OVERVIEW

In this chapter we focus on the policy issues that arise from the complex and highly dynamic character of the digital data collections universe. First, we establish the context and the need for an evaluation of NSF strategy and policies for digital data collections. The remainder of the chapter describes specific policy issues that should be addressed. We conclude with a comparison of large instrument-based facilities to long-lived digital data collections.

NEED FOR AN EVALUATION OF NSF POLICIES

Digital data collections and their roles in the research and education enterprise have evolved. The NSF strategy and policies have not kept pace. It is timely for the Foundation to reconsider its overall strategy for supporting digital data collections, as well as the processes that would implement that strategy. That strategy needs to accommodate those policies that must be discipline-specific or data collection category-specific. For example, while NSF might require a data management plan for all proposals that will produce data for long-term preservation, the evaluation of the plan must take place at the appropriate disciplinary or programmatic level using criteria that are appropriate to the data type and standards that arise from the respective discipline or community. The needs of research must drive the determination of specific policies; however they need to be harmonized, removing any contradictions to better support the interdisciplinary world of today. We also recognize that in some cases, a specific NSF policy is not required and the agency should leave decisions to the appropriate communities to make in whatever forums they select.

NSF support and NSF policies for digital data collections have grown incrementally over the past several decades. And both the investment and the policies have grown piece-meal in programs for the individual disciplines. As a result there are some policies regarding data sharing and archiving (see Appendix C). We could not find parallel policies for all disciplines.

NSF has a history of funding collections maintained by outside organizations. How many can it support? And how should the finite resources that the NSF has for this category of investments be used to assure that the benefits accrue to the broadest range of communities supported by the Foundation, and that this category is in balance with investments in all other areas, particularly with principal investigator grants?

Regardless of the approach that the Foundation ultimately adopts, the task force members stressed that the NSF must make its funding intentions transparent. The nature of any funding agency's support for a digital data collection can have significant impact on investments made by the research and education community, as well as by other U.S. and international agencies. Researchers must feel confident that a collection is truly long-lived because the decisions to use a particular collection can have considerable impact on their time and resources. Making such a commitment requires training their colleagues, including students, to use the collection effectively and necessitates that they all have a coherent and accurate view of the data, their metadata description, and the conditions in which the collection was built and is maintained. In order for researchers to make a sound decision about using a collection it is essential that agency policy to support its collections be well developed, broadly disseminated, and strictly observed.

NSF has created over time a portfolio of digital data collections. Today, that portfolio is not managed in a coherent, coordinated way. As mentioned earlier, we could not easily ascertain the number of long-lived data collections supported. It is time to take stock, not just of the numbers, but also of the strategy and policies that will best apply the NSF investment in digital collections.

SPECIFIC POLICY ISSUES

The following section discusses a set of policy issues. The first several issues very clearly involve strategic decisions for the NSF. There are many issues that we do not discuss here, for example technical standards choices. These are decisions that the community acting in concert must make.

1. Proliferating Collections

There are two basic Federal agency approaches to funding digital data collections: maintain collections primarily "in-house" (as do NOAA and NASA) or fund collections that are maintained by external organizations (as does NSF and in some cases NIH). These can be considered in-agency and out-agency collections, respectively.

In situations where there are just a few digital collections, there are a limited number of managing organizations making community-proxy decisions and there are fewer standards candidates, especially compared to the number of standards that arise when there are many smaller, independently managed collections. The majority of the in-agency collections are resource or reference collections because of their scale and because they support multiple data gathering missions.

In contrast, NSF funds digital data collections in response to requests from the community, and, as a result, it is more difficult for the Foundation to exercise the discipline in planning that the in-agency collection agencies can. Currently,

the NSF funds some hundreds, perhaps thousands, of resource and reference collections (although the NSF was unable to provide a definitive count). Some proliferation may be very healthy. But how many independent data collections does each of the NSF user communities need to provide reliable preservation and access to the essential information and range of data types necessary for continued advancement of a field? Certainly, other agencies disagree with widespread proliferation of independent collections – based on their actions. The question deserves serious consideration. It is our first example of an NSF-wide question. What rationale determines the number of long-lived collections? The answer may be somewhat different for different disciplines, but it is not likely different by an order of magnitude. And as research becomes more interdisciplinary, policies (especially the choice of technical standards) need to be harmonized across multiple disciplines. As the number of independent collections grows, that harmonization becomes more difficult.

2. Community-Proxy Policy

Resource and reference collections must provide accessible, high-quality assurance regarding data elements in their holdings. The organization maintaining such digital collections necessarily takes on community-proxy functions, that is, they make choices on behalf of the current and future user community on issues such as collection access, collection structure, data curation technical standards and processes, ontology development, annotation, and peer review.

Currently, data collection organizations that perform community-proxy functions are granted that authority in largely informal ways. Assignment of authority from the community is often implicit rather than explicit. In essence, community-proxy organizations are implicitly authorized when they receive project funding. Because the NSF supports a multitude of resource and reference collections within a field, there may be multiple community-proxy organizations making uncoordinated, conflicting decisions.

In the standards area, this lack of coordination can be both costly and detrimental to ease of access for the future data users. Each data author may choose different structures and formats, set different standards, and determine different defaults for user interfaces and data search algorithms – just to name a few examples of community-proxy technical decisions. This proliferation of community-proxy decisions adds unneeded complexity for the users. Note that much of the complexity and conflicting decisions arise from the fact that NSF funds a diverse set of out-agency collections, thus empowering a multiplicity of decision makers.

One challenge in creating consistent community-proxy standards is that the costs associated with exercising community-proxy functions can be high, representing in some cases a majority portion of the budget of a collection.

In some cases, this cost is so high that the community-proxy function responsibilities are ignored or treated casually. It is appropriate to develop a framework for establishing and guiding the work of community-proxy organizations, one that recognizes the true costs and value of this effort.

3. Data Sunset and Data Movement

Terminating funding for a data set or an entire digital collection (sunsetting) is a more difficult choice when there are many external collections than when an agency maintains a limited set of internal collections over which it exerts total administrative control. Fortunately, collection sunsetting is a relatively unusual event. By contrast, the movement of data between collections is routine in the data collections universe.

For example, data collected in a continuing research project may initially be placed in one research collection and then transferred to another as project responsibilities, organization, or funding changes. Or fragmentary data initially retained in a research collection may be transferred to a resource or reference collection when the data set is judged to be complete, of broad interest, and appropriate for general distribution. This regular movement of data creates two problems: tracking and attribution/access rights. Tracking is a challenge because links to the data in publications, Web sites, etc. may become obsolete. Finding the data that were previously available may be difficult for those outside the immediate project team. Strategies for location-independent identification of data objects, such as Digital Object Identifiers and permanent Universal Resource Locators (URLs) need to be developed and broadly applied to address this problem.

Information on proper attribution and on access restrictions and permissions may also be difficult to obtain since the organization maintaining the transferred data may not be the original authors. Standards for required metadata elements providing data history, authorship, and access information are needed to address this problem.

Several groups are exploring how to achieve these ends for digital artifacts. One example can be found in the 'Commons Deed' concept of the Creative Commons project, which seeks to provide a "reasonable, flexible copyright in the face of increasingly restrictive default rules" for creative, digital works (see http://www.creativecommons.org). The digital preservation program of the Library of Congress (see http://www.digitalpreservation.gov/) recognizes that almost anyone can be a publisher of digital artifacts. The challenge is to determine how society will preserve this information and make it available to future generations; and how data collections will classify this information so that their patrons can find it. The interagency Digital Libraries program led by NSF (http://www.dli2.nsf.gov; http://www.dli2.nsf.gov/dlione/) seeks to advance means for collecting, storing,

and organizing digital information and making this information readily available. There are still other activities at NSF including the Digital Archiving and Long-Term Preservation program (http://www.nsf.gov/pubs/2004/nsf04592/nsf04592.htm) and the National Science, Technology, Engineering and Mathematics Education Digital Library program (http://www.ehr.nsf.gov/due/programs/nsdl/). These programs seek to take leadership roles in addressing the challenges faced by digital libraries and archives, including those arising from the movement of data among collections.

These are only a few of a broad number of exploratory activities within and without the research community that are grappling with the many issues related to the rise of digital data collections, the empowerment of the individual anywhere within the Web, and creative sharing opportunities made possible by the very low cost of computation and communications. The Foundation is supporting these explorations, even actively participating.

The unchecked proliferation of long-lived digital collections funded by the NSF, however, makes it imperative that the Foundation develop its own strategy that incorporates all these dimensions of policy and investment, in contrast to the current decentralized, multiplicity of strategies and policies, or lack of policies that exists in the Foundation today.

In summary, many of the issues involved in data movement are community issues. The NSF, through its support for activities that promote interactions, can help communities in resolving these issues. And as solutions arise in the various communities, NSF can be a catalyst for the coherent application of community decisions and community policies across collections that users access in concert.

4. Data Management Plans

In this report we have asserted that NSF should have a coherent and thoughtful digital data collection strategy. The same is true for the individual or teams of researchers who will author and curate data. They need to have a strategy for dealing with data from their inception to their demise, or at least the foreseeable future.

We define a data management plan to be a plan that describes the data that will be authored as well as how the data will be managed and made accessible throughout its lifetime. Such a plan should be an integral part of a research project. The first version of the plan should be determined and documented at the research proposal stage of a research project.

The contents of the data management plan should include:
- the types of data to be authored;
- the standards that would be applied for format, metadata content, etc.;

- provisions for archiving and preservation;
- access policies and provisions; and
- plans for eventual transition or termination of the data collection in the long-term future.

In effect, this would provide specific guidance to applicants (and reviewers) to meet the current requirements of the Grant Proposal Guide (NSF-04-2), which specifies that the project description of a proposal should include, where appropriate, "plans for preservation, documentation, and sharing of data".

Any research proposal should give evidence that data management was considered. For proposals that do not involve the creation of data requiring long-term preservation, a simple statement that such a plan is not required would suffice. The validity of this assertion could be evaluated by peer review. If inclusion of specific data management plans is appropriate, then peer review will evaluate what is proposed. Providing such a plan assures that reviewers can assay whether the proposed budget is adequate to support data collection activities if direct funding is proposed.

In reviewing cutting-edge and interdisciplinary data management plans, peer reviewers (who represent the community) would have the opportunity to recognize where standards are missing and needed, where they may be unnecessarily limiting or outdated, where standards may be made compatible across disciplines, etc. It is not the Foundation's responsibility to decide how data will be managed, but it is the Foundation's responsibility to assure that coherent and cost-effective plans are defined and executed.

5. Data Access/Release Policies

The overall Foundation philosophy regarding access to the results of research is embodied in the NSF Grant General Conditions (GC-1):

> *NSF expects significant findings from research and education activities it supports to be promptly submitted for publication, with authorship that accurately reflects the contributions of those involved. It expects investigators to share with other researchers, at no more than incremental costs and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work. It also encourages grantees to share software and inventions or otherwise act to make the innovations they embody widely useful and usable. Adjustments and, where essential, exceptions may be allowed to safeguard the rights of individuals and subjects, the validity of results, or the integrity of collections or to accommodate legitimate interests of investigators.*
> *(see* http://www.nsf.gov/home/grants/gc102.pdf*)*

A number of NSF divisions and programs have developed specific data access policy statements that are in keeping with this general philosophy but which also recognize discipline, community, or program-specific needs, limitations, and standards. Examples of such statements can be found in Appendix C.

Concerns about the existing set of NSF policy statements for data access and release include the following. First, there is no single site at which a member of the community can readily locate all applicable or relevant policy statements. Second, many programs lack an explicit statement of data access and release policy. Third, there is little coherence and consistency among the set of existing statements.

The absence of coherent, accessible, and transparent data access policies creates barriers to interdisciplinary research and to effective data collections management. Researchers working at the interface between disciplines can find themselves subject to conflicting data release policies and deposition requirements. Collections managers who work with multiple communities are often faced with differing rules for deposition, conflicting technical standards, and varying access restrictions. Development of a comprehensive set of policy statements for data access and release that provides for consistency and coherence across disciplines while meeting the distinct needs of individual disciplines and communities, that are transparent and readily accessible to the community, and that prevent unnecessary proliferation and duplication of standards could greatly facilitate progress in research, education, and collections management.

6. Digital Data Commons as a Means for Broadening Participation

Many individuals and even entire communities are limited in their opportunities to create and maintain digital data collections by lack of access to the necessary resources and expertise. As described above, digital data commons can be broadly enabling, allowing individuals (even entire communities) who are not information specialists to contribute actively to the data collections universe.

There is a question of how to fund such "commons" data spaces. Research proposal data management plans could provide an overt statement of need through researcher's preference for such common space, and of the need for indirect funding of such digital common spaces. The data management plan would provide factual statements that could be used to justify indirect funding for data archiving, rather than to have each proposal include direct line budget elements to fund data archiving. It has been proposed that with an indirect cost model, archiving and curation could be funded in whole or in part through an allowance in the institutional indirect costs. Requiring peer review of data management plans provides a kind of forum in which researchers can state the value for the indirect funding model for archiving of data. Workshop participants

urged that the NSB and NSF undertake an evaluation of the comparative merits of direct funding versus indirect funding for data collections infrastructure. The Board recognizes that the development of an enabling legal framework for "commons" data spaces is another significant challenge and looks to the development of community, interagency, and international partnerships to address this challenge.

7. Opportunities for Education, Training, and Workforce Development

Digital data collections are a remarkably empowering resource for research and education. Useful access to such collections enables scientists, students, and educators from across the full spectrum of institutional, cultural, and geographic settings to make innovative contributions at the cutting edge of the research and education enterprise. Providing for such access requires not only that the necessary infrastructure be available but also that training in the knowledge and skills required to use the collection infrastructure be broadly accessible at all levels and that a workforce of innovative data scientists be available to create cutting-edge collections technology.

There are two kinds of training. First, there is training to permit researchers who are domain experts to be able to access collections in sophisticated ways. Collection managers will routinely run seminars and courses to educate these relatively sophisticated users who need deep understanding of both content and metadata descriptions of content. Even this kind of training needs to be multidisciplinary in character and targeted to researchers with diverse backgrounds.

Second, digital data collections have a remarkable ability to provide meaningful access to information to all people. Digital data collections are accessible in a way that research activities often cannot be. So, strategic investments in data collections can provide one important means for addressing the general public, young children as well as adults. Making collections intelligible to the general public and providing for those who want education and training are a challenge to the data scientists who devise the interfaces and the training program. This community has a wide variety of skills and interests that they bring to the task.

Implementing both kinds of training programs requires adequate funding. We recognize that this need for education, training, and workforce development at all levels is not limited to data collections, but represents a more general need for all cyberinfrastructure, as was specifically stated in the report of the NSF Blue Ribbon Advisory Panel on Cyberinfrastructure (see http://www.cise.nsf.gov/sci/reports/CH2.pdf). These goals are also consonant with the NSF priority for investment in people and its priority for improving the productivity of researchers

and expanding opportunities for students. This is explicitly embodied in the Workforce for the 21st Century priority area defined in the NSF FY2005 budget proposal as follows:

> *This priority area aims to strengthen the nation's capacity to produce world-class scientists and engineers and a general workforce with the science, engineering, mathematics and technology skills to thrive in the 21st Century workplace. Funding will support innovations to integrate NSF's education investments at all levels, K-12 through postdoctoral level, as well as attract more U.S. students into science and engineering fields and broaden participation (see* http://www.nsf.gov/od/lpa/news/04/fsfy05priorityareas.htm*).*

Thus, effective use of the investment in digital data collections to enhance educational opportunities in a digital environment should be viewed as an important and integral component in the broader efforts of the Foundation to meet the unique needs of the 21st century workplace. A comprehensive strategy for investments in data collections is needed to ensure that the educational benefits of these investments accrue to all who are represented at NSF.

8. Duration of NSF Commitment to Support Long-Lived Digital Collections

The vast majority of NSF support carries with it no long-term commitment. Principal investigator grants have a duration of several years. Centers are typically funded for five years with a potential for an additional five years of funding. Long-lived digital data collections raise a new issue. They potentially can live in perpetuity. Indeed, as mentioned earlier, the value of a collection may increase with age.

It is timely for NSF to consider whether it should make very long-term commitments to a digital collection. This would be in sharp contrast to any commitment to the organization managing the collection. Periodic reviews – as are now performed – of the management organization help assure quality of that management. It is not infrequent that NSF, through a competitive process, changes the management organization. The Protein Data Bank provides one example of this. The current managing organization was not the founding management organization. Indeed, as the Board has seen some months ago, the issue of NSF commitment of support was entwined with the issue of the renewal of funding of the current managers. It is timely to consider whether commitment to the collection should be a separate decision from commitment to fund the current management organization and their immediate plans.

It was observed in the workshops that long-lived digital collections share some attributes with instrument-based facilities. So, we explore the larger issue of long-duration support by considering the similarities and differences between collections and large instrument-based facilities.

LONG-LIVED DIGITAL DATA COLLECTIONS AND LARGE FACILITIES

Workshop participants drew analogies between resource/reference collections and large facilities such as telescopes, ocean drilling ships and long-term ecological research projects. The parallels are significant. Digital data collections resemble large facility projects in terms of their extended lifetime; the need for stable, core support; the critical importance of effective project management in combination with domain expertise; the ability to energize and enable broad research and education communities; and the importance of partnerships, both national and international. Considering these similarities, it may be informative to consider NSF processes for managing large facilities as a way of better understanding the issues involved in developing policy to manage long-lived digital data collections.

The Foundation's facility evaluation and approval process is formal. The deputy director periodically convenes the Major Research Equipment and Facilities Construction (MREFC) panel to consider proposed facility projects, to discuss them in comparison with one another, and very importantly to discuss the best way to nurture rising projects that might deserve funding in the future. The deputy director reports to the National Science Board several times a year on the status of emerging facility projects.

The National Science Board Guidelines for the Evaluation of Large Facility Projects (NSB 02-191) include the following:

- need for the facility;
- opportunities for research that will be enabled;
- project readiness;
- budget estimates;
- degree to which the project would broadly serve the many disciplines supported by the Foundation;
- multiple projects for a single discipline, or for closely related disciplines, are ordered based on a judgment of the contribution that they will make toward the advancement of research in those related fields; community judgment is considered; and
- international and interagency commitments are considered in setting priorities among projects.

Similar guidelines may or may not be appropriate for establishing new resource and reference collections, but the example of large facilities demonstrates that a set of organized processes and well-documented criteria will be critical in nurturing, evaluating, and selecting proposals for long-lived digital data collections.

However, instrument-based facilities differ from long-lived digital data collections in significant ways. With instrument-based facilities, there are clear funding decisions occasioned by the mechanical or physical decline of the instrument

or by an improvement in technology that renders the instrument less valuable than an instrument based on newer technology. At an appropriate time, the community downgrades the priority of the instrument-based facility in favor of building a new facility to realize the promise of new instruments. Of course new instruments can be housed at the same location as old instruments, and are occasionally an upgrade of an old instrument. But, it is clear to the community of users that the new instrument is replacing something older. As a result there are forces that assure the curtailment of Foundation funding of one facility in favor of newer facilities.

Today, with long-lived digital data collections, there are few natural decision points at which a funding agency might engage the research community to discuss the future of the collection. There are no physical instruments to deteriorate, and well-designed collections can anticipate changes in technology, necessitating migration to a new generation of media. Furthermore, unlike instrument-based facilities, data collections tend to increase in value the longer they are in operation, attracting ever-expanding groups of data users as the amount of data they include increases and spans greater periods of time. So valuable do they become that the appearance of a new data collection in the same field does not necessarily diminish the desire of the community to maintain existing collections.

In the absence of circumstances that may lead agencies to reevaluate their funding, research communities may come to expect permanent – and permanently increasing – support for selected data collections. Given the extremely limited funds available to the Foundation and the exceedingly slow growth in the overall NSF budget over the last decade, the Foundation will not be able to meet this expectation.

Clarity in the commitment of NSF to a digital collection is important to researchers that depend upon a collection and need to be able to predict its future accessibility and stability. Such clarity is also key to forming stable, multi-agency and international partnerships to support collections that should, appropriately, operate on a global scale. Determining the length of the NSF's commitment to a digital data collection should be considered from two perspectives: the Foundation's commitment to keeping the data available and its commitment to a specific team managing the collection. In many cases, particularly in those of reference collections, this first commitment may be indefinite. As part of its policy for long-lived digital data collections, the NSF must decide the criteria used to determine whether a commitment is indefinite or not, it must develop protocols for seeking input, and it should develop a process by which this decision is periodically revisited.

The duration of the NSF commitment to the team managing a long-lived digital data collection should be limited and subject to appropriately frequent performance review. Under some circumstances, it may not be appropriate to solicit competitive proposals to manage the collection, but in all cases periodic peer review that includes user communities is appropriate. This review should include an assessment of management strategies, management's ability to adopt new technology, and the quality of access provided by different collection managers. A new kind of management competition and associated peer review mechanism may be needed to accomplish these aims.