Chapter Five: Findings and Recommendations

WORKSHOP OUTCOMES

The general findings and conclusions developed by participants at both workshops held to discuss long-lived digital data collections can be summarized as follows:

- Digital data collections are powerful catalysts for progress and for democratization of the research and education enterprise. Proper stewardship requires effective support for these essential components of the digital research and education environment of the 21st century.
- The need for digital collections is increasing rapidly, driven by the continuing exponential increase in the volume of digital information. The number of different collections supported by the NSF is also increasing rapidly. This increase in number necessitates that NSF use strategies for managing its portfolio of out-agency collections that differ from those used by agencies with primarily in-agency collections. There is an urgent need to rationalize action – in the communities and in the NSF. Enlightened strategic planning and careful investment management are needed to ensure the continued health of the data dimension of the research and education enterprise.
- The National Science Board and the National Science Foundation are uniquely positioned to take leadership roles in developing comprehensive strategic policy and enabling the system of digital data collections, respectively. Because the Foundation does not maintain data collections internally, as do some other agencies, it has and is perceived to have a more objective position. This out-agency emphasis does not reduce the ability of NSF to take a broad international leadership role. Many, in fact most, of the policy issues are not specific to an agency or the collections that it supports; they are specific to the conduct of data-rich research. This unique position of the Board and the Foundation, in combination with the urgent needs, creates a responsibility to act.
- Policies and strategies developed to facilitate the management, preservation, and sharing of digital data will have to fully embrace the essential diversity in technical, scientific, and other features found across the spectrum of digital data collections. This diversity arises from many sources including differences in data and metadata content among the various disciplines; differences in user needs, expectations, and access procedures; and differences in the legal restrictions and requirements that may apply to a given data set. Thus, heterogeneity is an essential feature of the data collections universe that should be enabled and not constrained by policy.

RECOMMENDATIONS

The following recommendations call for clarifying and harmonizing NSF strategy, policies, processes, and budget for long-lived digital data collections. Because the issues are urgent and because undertaking broader discussions depends upon an understanding of the Foundation's objectives and capabilities, we look for a timely response to these recommendations from NSF. The Board anticipates that a broader dialog among other agencies in the U.S. and with international partners will be required. The Board recommends that the broader dialogue be undertaken with the highest priority in a coordinated interagency effort led by OSTP.

These recommendations are divided into two groups. They call for the NSF to:

- develop a clear technical and financial strategy; and
- create policy for key issues consistent with the technical and financial strategy.

Develop a Clear Technical and Financial Strategy

NSF support for long-lived data collections has evolved incrementally, and in slightly different forms, across the multiple disciplines that the Foundation supports. Given the proliferation of resource and reference collections and the costs associated with creating and maintaining them, it is imperative that the Foundation develop a comprehensive strategy – incorporating and integrating technical and financial considerations – for long-lived data collections and determine the steps necessary to anticipate future needs.

Recommendation 1: The NSF should clarify its current investments in resource and reference digital data collections and describe the processes that are, or could be, used to relate investments in collections across the Foundation to the corresponding investments in research and education that utilize the collections. In matters of strategy, policy, and implementation, the Foundation should distinguish between a truly long-term commitment that it may make to supporting a digital data collection and the need to undertake frequent, peer review of the management of a collection.

Clarification of current NSF investments in digital data collections should address the following questions:

- How is the current investment distributed between the costs of creation; maintenance and operations; technology updating, including migration to new media/systems; and provision of user access to collections?
- What is the current balance between the investment in data collections compared to the investment being made in the research that exploits collections? How is this balance currently evaluated, and how should it be evaluated in the future?

- Does the Foundation currently make a formal distinction between a long-term commitment to a data collection and a limited commitment to collection managers that is subject to frequent peer review? How many such long-term commitments does the Foundation have?

Recommendation 2: The NSF should develop an agency-wide umbrella strategy for supporting and advancing long-lived digital data collections. The strategy must meet two goals: it must provide an effective framework for planning and managing NSF investments in this area, and it must fully support the appropriate diversity of needs and practices among the various data collections and the communities that they serve. Working with the affected communities NSF should determine what policies are needed, including which should be defined by the Foundation and which should be defined through community processes. The Foundation should actively engage with the community to ensure that their policies and priorities are established and then updated in a timely way.

Where appropriate, elements of the strategy under the umbrella may be discipline-specific, and possibly even program-specific. But because research is increasingly interdisciplinary, the Foundation's overall digital data collections strategy and associated policies need to be coherent across disciplines.

Clarification of NSF's approach to long-lived digital data collections should address the following questions:

- At what level can it support research, resource, and reference data collections?
- How should support be distributed among research, resource, and reference collections in the various disciplines?
- Under what conditions should the NSF make a commitment to support a long-lived data collection, and what process should be used to decide to terminate that support?
- Should the length of time that the Foundation commits to fund a collection be longer than the duration of an award to a specific organization to manage that collection?
- When is the use of sole-source rather than competitive proposals appropriate for continuing/initiating support for a collection?
- What is NSF's responsibility to ensure that users from other disciplines will be able to access a long-lived data collection?
- Is there an unmet need for digital common spaces to enable data collections, particularly at the research level? Should the Foundation fund any digital commons and if so, how?
- Under what conditions are discipline-specific, even program-specific policies appropriate, and how do they fit into the overall Foundation strategy?

In developing agency-wide strategy, the NSF should review all issues related to long-lived digital data collections and determine which require NSF to develop a policy, carefully designating those for which policy should come from some other source. A listing of some of the central policy issues for consideration by NSF is provided in Chapter Four of this report.

The following considerations should guide the Foundation in developing policies for long-lived digital data collections. First, policies need to be clearly stated, and NSF review processes need to assure the Foundation that funded projects adhere to relevant policies. Second, policies should place the communities at the center, empowering them to identify their needs; to develop and implement standards, customs and norms; and to reach out to other communities to bridge disciplinary, geographical, organizational, and other barriers. Finally, mechanisms for policy development and implementation should provide for a continuing process undertaken in partnership with the community and responsive to changes in needs and opportunities.

Create Policy for Key Issues Consistent with Technical and Financial Strategy

Although the Foundation has formulated policy that affects long-lived digital data collections, this policy must be brought into conformity with the NSF's overall strategy for these collections. There are also a number of areas in which policy is lacking. This is the focus of the next four recommendations.

Recommendation 3: Many organizations that manage digital collections necessarily take on the responsibility for 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; technical standards and processes for data curation; ontology development; annotation; and peer review. The NSF should evaluate how responsibility for community–proxy functions is acquired and implemented by data managers and how these activities are supported.

The activities of the organization that manages a resource or reference collection often go well beyond the collection and distribution of data. These activities include curation, expert annotation, peer review, quality assessment and control, author attribution and credit, and standards development and implementation. These essential, community-proxy functions can provide a robust framework for the digital data environment. However, data managers can meet these responsibilities only if they have the full trust and endorsement of the communities that they serve as well as adequate funding to support the activities.

Development of policy by which collections acquire the authority to perform community-proxy functions should address questions in two categories. The first focuses on how the need for community-proxy functions and the qualifications of a data manager to perform those functions are evaluated:

- Do formal or informal mechanisms exist at NSF or elsewhere for evaluating what community-proxy functions are needed and for determining which collection managers are qualified to take on the corresponding responsibilities?
- Is competitive review used in making these evaluations or are these primarily sole-source situations in which a collection has 'grown into' the corresponding responsibilities?
- What criteria are used in carrying out such evaluations?
- How does NSF act to ensure that the community is involved in reaching decisions in an efficient and effective manner?
- Are new or additional mechanisms and/or criteria needed in evaluating these needs and qualifications? For example, would distinguishing more clearly between resource or reference collections facilitate the evaluation process?

The second category of questions focuses on the costs to support community-proxy functions:

- What are those costs and how are they currently supported?
- How is the performance of an organization performing community-proxy functions evaluated?
- How is the current investment in these activities distributed across the disciplinary areas represented at NSF?

Recommendation 4: The NSF should require that research proposals for activities that will generate digital data, especially long-lived data, should state such intentions in the proposal so that peer reviewers can evaluate a proposed data management plan.

The inclusion of a data management plan in a proposal would permit representatives of the relevant communities, through the peer review process, to comment on the degree to which the plan meets the standards, norms, and expectations of the community. Reviewers and NSF program officers would be able to determine if the proposed budget adequately supports the data management plan, and the Foundation would use the project's annual and final project reports to track the manager's effectiveness in implementing the data management plan.

Many proposals do not involve the creation of data that will ever be part of a long-lived digital data collection. It is sufficient that such a proposal simply state, "No data management plan is appropriate." The validity of such an assertion could be tested by peer review, ensuring that the community has a chance to comment on overlooked or underappreciated needs for data access and preservation.

Recommendation 5: The NSF should ensure that education and training in the use of digital collections are available and effectively delivered to broaden participation in digitally enabled research. The Foundation should evaluate in an integrated way the impact of the full portfolio of programs of outreach to students and citizens of all ages that are, or could be, implemented through digital data collections.

Advancing research and education through the use of digital data collections is new and has the potential to be remarkably empowering. The existence of collections creates opportunities for cutting-edge contributions from a broad diversity of scientists, students, and educators across the full spectrum of institutional and geographic settings. Achieving this potential requires that training in the knowledge and skills required to use the collection infrastructure be broadly accessible at all levels. The resource and reference collections should provide this kind of training and education. Such programs need to be multidisciplinary in character and targeted to a wide variety of user levels and interests. Implementing such programs requires adequate funding.

Efforts to optimize the use of data collections to enhance research and education activities should be undertaken in concert with other efforts directed at NSF goals for cyberinfrastructure (see the report of the NSF Blue Ribbon Advisory Panel on Cyberinfrastructure; http://www.cise.nsf.gov/sci/reports/CH2.pdf) and with those undertaken under the Workforce for the 21st Century priority area as defined in the NSF FY2005 budget proposal.

Recommendation 6: The Foundation, working with collection managers and the community at large, should act to develop and mature the career path for data scientists and to ensure that the research enterprise includes a sufficient number of high-quality data scientists.

Data scientists materially determine the quality of the data collections that now play a vital role in research. Their role is new, so it is crucial that the professional career of data scientist be defined and recognized so that it will attract the best and brightest. NSF should be proactive in advancing programs that educate and reward data scientists.

Creating a culture in which the innovative use of digital data is valued as both a research product and a resource can contribute significantly to this goal. The NSF can encourage career field development, but it will fall primarily to the leaders of the large resource and reference collections who can put in place a culture to enable these scientists to receive the recognition through publication, promotion, community exposure, respect, and remuneration.

In creating policy to ensure that a sufficient number of high quality data scientists is available, the Foundation should consider the following questions:

- What aspects of current NSF policy and investments promote recognition of the contributions of data scientists? What opportunities exist for improvements in this regard?
- How can NSF encourage and facilitate the efforts of the community at large to create a culture that is supportive of data scientists?

CONCLUSIONS

It is exceedingly rare that fundamental new approaches to research arise. Information technology has ushered in such a fundamental change. Digital data collections are at the heart of this change. The existence of a new data collection can effectively serve as new phenomena to study. Such phenomena are equally accessible to study at all levels – by teams of scientists or by an individual investigator with a computer and Internet access. In addition, digital data collections serve as an instrument for performing analysis with an accuracy that was not possible previously or, by combining information in new ways, from a perspective that was previously inaccessible. And data collections that are genuinely accessible by non-experts provide open windows into science and engineering that can be used at all ages and all levels of education. Full realization of the exciting opportunities created by digital data collections requires the development of policies and strategies that are robust, responsible, and responsive.

Because digital data collections have proliferated and increased in size incrementally, the NSF investment and its policies have been determined by incremental decisions. It is timely to evaluate all aspects of the data-rich research and education environment, especially the strategy and the policies of the NSF. The National Science Board has concern about the current situation, yet sees the immense opportunity that such collections enable. The next step in advancing digital research through long-lived data collections is for these recommendations to be acted upon.

In addition to the analysis described above, the NSB anticipates the need for discussions beyond NSF to be led by OSTP and to encompass the full spectrum of digital data collections and supporting agencies. These discussions should be designed to examine in both the national and global contexts how the investment, the policies, and inter-agency management can provide cost-effective, high-quality digital data collections. The need to address these issues is urgent. The opportunities are substantial.