Executive Summary

It is exceedingly rare that fundamentally new approaches to research and education arise. Information technology has ushered in such a fundamental change.   Digital data collections are at the heart of this change. They enable analysis at unprecedented levels of accuracy and sophistication and provide novel insights through innovative information integration. Through their very size and complexity, such digital collections provide new phenomena for study. At the same time, such collections are a powerful force for inclusion, removing barriers to participation at all ages and levels of education.

The long-lived digital data collections that are the subjects of this report are those that meet the following definitions.

- The term 'data' is used in this report to refer to any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc. Such data may be generated by various means including observation, computation, or experiment.
- The term 'collection' is used here to refer not only to stored data but also to the infrastructure, organizations, and individuals necessary to preserve access to the data.
- The digital collections that are the focus for this report are limited to those that can be accessed electronically, via the Internet for example.
- This report adopts the definition of 'long-lived' that is provided in the Open Archival Information System (OAIS) standards, namely a period of time long enough for there to be concern about the impacts of changing technology.

The digital data collections that fall within these definitions span a wide spectrum of activities from focused collections for an individual research project at one end to reference collections with global user populations and impact at the other. Along the continuum in between are intermediate level resource collections such as those derived from a specific facility or center.

The National Science Board (NSB, the Board) recognizes the growing importance of these digital data collections for research and education, their potential for broadening participation in research at all levels, the ever increasing National Science Foundation (NSF, the Foundation) investment in creating and maintaining the collections, and the rapid multiplication of collections with a potential for decades of curation. In response the Board formed the Long-lived Data Collections Task Force. The Board and the task force undertook an analysis of the policy issues relevant to long-lived digital data collections. This report provides the findings and recommendations arising from that analysis.

The primary purpose of this report is to frame the issues and to begin a broad discourse. Specifically, the NSB and NSF working together – with each fulfilling its respective responsibilities – need to take stock of the current NSF policies that lead to Foundation funding of a large number of data collections with an indeterminate lifetime and to ask what deliberate strategies will best serve the multiple research and education communities. The analysis of policy issues in Chapter Four and the specific recommendations in Chapter Five of this report provide a framework within which that shared goal can be pursued over the coming months. The broader discourse would be better served by interaction, cooperation, and coordination among the relevant agencies and communities at the national and international levels. Chapters Two and Three of this report, describing the fundamental elements of data collections and curation, provide a useful reference upon which interagency and international discussions can be undertaken. The Board recommends that the Office of Science and Technology Policy (OSTP) take the lead in initiating and coordinating these interagency and international discussions.

WORKSHOP FINDINGS

The Board task force held two workshops to hear the opinions of relevant communities. These workshops have shaped the Board's analysis of issues. The first workshop focused on the experience of the NSF and other Federal agencies with digital data collections. The second workshop provided a forum to gather the views of the NSF grantee community. The outcomes of these workshops can be summarized as follows:

- Long-lived digital data collections are powerful catalysts for progress and for democratization of science and education. Proper stewardship of research requires effective policy to maximize their potential.
- The need for digital collections is increasing rapidly, driven by the exponential increase in the volume of digital information. The number of different collections supported by the NSF is also increasing rapidly. There is a need to rationalize action and investment – in the communities and in the NSF.
- The National Science Board and the National Science Foundation are uniquely positioned to take leadership roles in developing a comprehensive strategy for long-lived digital data collections and translating this strategy into a consistent policy framework to govern such collections.
- Policies and strategies that are developed to facilitate the management, preservation, and sharing of digital data will have to fully embrace the essential heterogeneity in technical, scientific, and other features found across the spectrum of digital data collections.

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;
- Create policy for key issues consistent with the technical and financial strategy.

Develop a Clear Technical and Financial Strategy

Recommendation 1: The NSF should clarify its current investments in resource and reference digital data collections – the truly long-lived 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 support a digital data collection and the need to undertake frequent peer review of the management of a collection.

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 community policies and priorities are established and then updated in a timely way.

Create Policy for Key Issues Consistent with the Technical and Financial Strategy

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.

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.

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.

Recommendation 6: The NSF, working in partnership 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.

CONCLUSIONS

The weakness of NSF strategies and policies governing long-lived data collections is that they have been developed incrementally and have not been considered collectively. Given the proliferation of these collections, the complexity of managing them, and their cost, action is imperative. The National Science Board is concerned about the current situation. Prompt and effective action will ensure that researchers and educators derive even higher value from these collections. The communities that create and use the collections will have to be fully engaged in this process. Consensus within the communities will have to inform Foundation policy, investment, and action. The need to address these issues is urgent. The opportunities are substantial.