Chapter One: Introduction

Long-lived digital data collections are increasingly crucial to research and education in science and engineering. A number of well-known factors have contributed to this phenomenon. Powerful and increasingly affordable sensors, processors, and automated equipment (for example, digital remote sensing, gene sequencers, micro arrays, and automated physical behavior simulations) have produced a proliferation of data in digital form. Reductions in storage costs have made it cost-effective to create and maintain large databases. And the existence of the Internet and other computer-based communications have made it easier to share data. As a result, researchers in such fields as genomics, climate modeling, and demographic studies increasingly conduct research using data originally generated by others and frequently access these data in large public databases found on the Internet.

The long-lived digital data collections that fall within the scope 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 throughout 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. (see http://public.ccsds.org/documents/650x0b1.pdf).

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. Appendix D provides a listing of examples to illustrate this spectrum of activities.

New analytical techniques, access technologies, and organizational arrangements are being developed to exploit these digital collections in innovative ways. In some cases, new analytical tools are developed that perform better and more extensive analyses than could be completed at the time when data were collected. Often analysis depends not just on the sensed or computer-generated data, but upon the metadata that characterizes the environment and the sensing instrument. As a result of these innovative approaches, data collections often have value beyond that envisioned when the collection was started.

Data collections provide more than an increase in the efficiency and accuracy of research; they enable new research opportunities. They do this in two quite different ways. First, digital data collections provide a foundation for using automated analytical tools, giving researchers the ability to develop descriptions of phenomena that could not be created in any other way. While this is true for science that studies natural physical processes, it is particularly enabling for the social scientists.

Second, digital data collections give researchers access to data from a variety of sources and enable them to integrate data across fields. The relative ease of sharing digital data – compared to data recorded on paper – allows researchers, students, and educators from different disciplines, institutions, and geographical locations to contribute to the research enterprise. It democratizes research by providing the opportunity for all who have access to these data collections to make a contribution.

Recognizing the growing importance of these digital data collections for research and education, their potential for broadening participation, and the vast sums invested in creating and maintaining them, the National Science Board formed the Long-lived Data Collections Task Force. The Board charged the task force with identifying the policy issues relevant to long-lived data collections and making recommendations for consideration by the Board and the community (see Appendix A for the task force charter).

As a first step in informing analyses of these issues, the Board and its task force held two workshops with the goal of identifying key policy issues for further consideration. The first workshop, held on November 18, 2003, focused on the experiences of NSF programs and other Federal agencies with long-lived data collections. Participants agreed to a considerable extent on the main policy issues, even though there is one stark difference between NSF and many other agencies: the vast majority of long-lived data collections supported by the NSF are managed by external research organizations, while other agencies, such as the National Aeronautics and Space Administration (NASA) and the National Oceanographic and Atmospheric Administration (NOAA) focus more heavily on archiving and curating many such data collections themselves. The second workshop, held on March 23, 2004, focused on the experience of the NSF grantee community.

This report summarizes the discussions and recommendations made at these two workshops, supplemented by the findings of other researchers who have examined these issues in detail (see Appendix B for a short bibliography of relevant studies). At both workshops, participants emphasized that policy development must be guided by a clear understanding of the unique features of the “data collection universe” – the system of data collectors, users, managers, and funding agencies central to the research and education activities that involve digital data collections. Accordingly, the **second** and **third** chapters of the report outline the complex structure of the digital data collections universe and the responsibilities of the individuals and institutions that play a role in creating and maintaining the collections that are in it.

The **fourth** chapter builds on this framework to highlight what the task force believes to be the key considerations when formulating policy and strategy for long-lived data collections, focusing on issues that are germane to the NSF.

The **fifth** and final chapter of the report summarizes the workshop outcomes and provides recommendations. In keeping with the charge to the task force, these recommendations focus specifically on “the policy issues relevant to the National Science Foundation and its style and culture of supporting the collection and curation of research data.”

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 will require substantial 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 the data collections universe and the relationships among its constituents, are intended to provide a useful reference upon which to begin broader interagency and international discussions.

SOURCES FOR ADDITIONAL INFORMATION

There have been a series of studies of data collections that can provide an excellent starting point for action on the task force recommendations (see Appendix B for citations).

- *The National Digital Information Infrastructure Preservation Program*, led by the Library of Congress working closely with other Federal partners, seeks to address a number of issues, including archival architecture and property rights considerations, technical challenges, and potential roles of institutional and agency participants.
- *It's About Time: Research Challenges in Digital Archiving and Long-Term Preservation*, the report of a workshop jointly sponsored by NSF and the Library of Congress, provides a research agenda to address key technological and computer and information sciences challenges in digital archiving and preservation.
- *The Role of Scientific and Technical Data and Information in the Public Domain: Proceedings of a Symposium*, the report of a recent National Research Council symposium, reviews the legal, technical and policy challenges in establishing an effective balance between the benefits of open access and the need for proper protection of intellectual property rights.
- *How Much Information? 2003*, a report from the School of Information Management and Systems of the University of California, Berkeley, provides a compendium of information on the increasing complexity of digital information types and the global expansion in digital information flux.
- *Revolutionizing Science and Engineering through Cyberinfrastructure*, the report of the NSF Blue-Ribbon Advisory Panel on Cyberinfrastructure, describes the opportunities that exist for creating new research environments through cyberinfrastructure, including the important role of digital data collections.
- *Science and Engineering Infrastructure for the 21st Century: The Role of the National Science Foundation* (NSB-02-190), prepared by the National Science Board, provides an analysis of academic research infrastructure, including current status and anticipated needs, and provides a discussion of data collections in the context of infrastructure needs.