DRAFT

NSB-05-40
March 30, 2005

DRAFT

Report of the National Science Board

Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century

CONTENTS

APPENDICES

A. Task Force Charter
B. Sources of Additional Information
C. Current Policies on Data Sharing and Archiving
D. Digital Data Collections by Categories

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 IV and the specific recommendations in Chapter V of this report provide a framework within which that shared goal can be pursued over the coming months. The broader discourse will involve interaction, cooperation, and coordination among the relevant agencies and communities at the national and international levels. Chapters II and III

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.

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 in order 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 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. The Board anticipates that a broader dialog with other agencies in the U.S. and with international partners will eventually be required, before all the issues are resolved. Since the issues are urgent and since undertaking broader discussions depends upon a clear understanding of the Foundation's objectives and capabilities, we look for a timely response to these recommendations from NSF.

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.

CONCLUSION

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.

I. 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 this data in large public databases found on the Internet.

New analytic 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

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 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. (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.

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 IV and the specific recommendations in Chapter V 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 II and III 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.
- *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*, 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*, 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.

II. THE ELEMENTS OF THE DIGITAL DATA COLLECTIONS UNIVERSE

OVERVIEW

Developing a policy to ensure that researchers and educators derive the maximum value from digital data collections consistent with legal and technological constraints is a difficult undertaking. The issues involved are extraordinary in their range and complexity. Addressing them requires a precise understanding of the elements of the data collections universe. To provide a common ground for discussion and to prepare the reader for the policy discussion in Chapter IV and the recommendations in Chapter V, the task force has prepared some core definitions to ensure that the participants have a shared vocabulary.

To begin with, the phrase *data collections universe* is used throughout this report to refer to the system of digital data, data collections, related software, hardware and communications links, data authors, managers, users, data scientists and supporting agencies and research centers that allow the collection, curation, analysis, distribution and preservation of digital data in the current research and education environment.

INDIVIDUALS AND INSTITUTIONS

The actors in the digital data collections universe are both individuals and institutions. *Data users* include researchers, educators, administrators, students, and others who exploit information in data collections to pursue their research and education activities. *Data authors* are the individuals involved in gathering data that are subsequently deposited in a data collection. *Data managers* are the individuals and organizations responsible for database operation and maintenance. Note that archiving – the process of depositing data in a collection – is often a shared responsibility of data authors and managers. Although the sharing of responsibilities varies among data collections, authors are often responsible for authorizing archiving of data and for providing required information in a usable format; managers are often responsible for ensuring that depositions are of a content and format appropriate for the collection.

Among the members of a data management organization are the *data scientists*, the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, and others, who are crucial to the successful management of a digital data collection. The intellectual contributions of data scientists are key drivers for progress in the information sciences/data collections field. The career path for data scientists is not yet mature. The mechanisms to recognize their contributions are not fully in place.

The terms *data authors*, *data managers*, *data scientists*, and *data users* reflect functional categories. A single person may at varying times act as a data user, manager, data scientist, or author. For instance, a data user who undertakes new research may quickly become a data

author or an experienced data author who creates a new research collection may become a data manager.

The term *funding agencies* is used to refer to all of the entities – local, national, and international; government, non-profit, and for-profit entities – that provide financial support for data production, archiving, management and use. This term includes agencies that primarily support data collections that reside within research and education organizations (as is typical for collections funded by NSF), and those that support collections that reside within the funding agency. The central role of the funding agencies was a common thread through many of the workshop discussions.

The structure of the digital data collections universe, building on the elements discussed above, is illustrated in figure 1. Arrows in the diagram represent the dynamic interactions and relationships among these functional entities and these are addressed in Chapter III of the report. The reason for the use of multiple icons representing data collections will become clear later. The arrows that relate the collections represent the orchestrated use of multiple data collections by a user on a single project. There are deep technical issues arising from the need and desire to use multiple collections in concert.

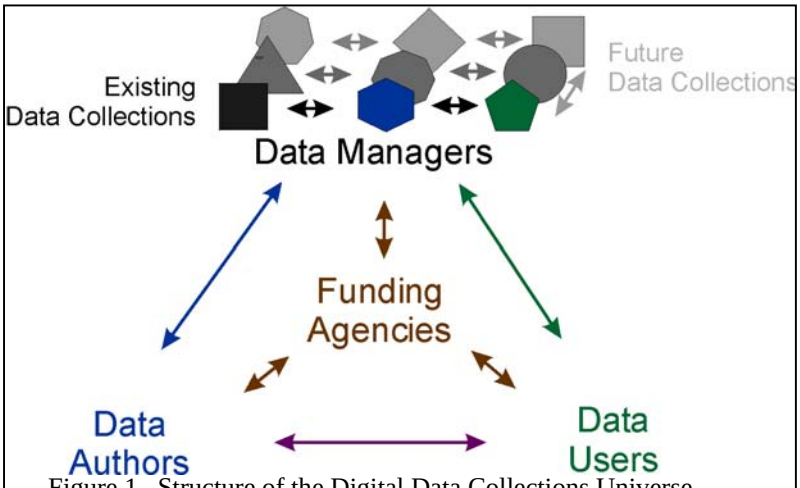

Figure 1. Structure of the Digital Data Collections Universe
Characteristics of the entities depicted in the figure are described in Section II of the text. Relationships among these entities, represented by arrows in the diagram, are described in Chapter III.

DATA

Digital data are the currency of the data collection universe, which, like currency in the financial realm, comes in many different forms. These differences include the nature of the data, their reproducibility, and the level of processing to which they have been subjected. Each of these differences has important policy implications.

First, the nature of data in a collection may be diverse, including numbers, images, video or audio streams, software and software versioning information, algorithms, equations, animations, or models/simulations. This essential heterogeneity, and the issues it raises, was stressed during the presentations of the workshop participants, who emphasized that a "one-size-fits-all" approach to policy development is inadequate. They argued that robust policies that not only recognize, but also effectively support, various kinds of data are required.

Data can also be distinguished by their origins – whether they are observational, computational, or experimental. This distinction is crucial to choices made for archiving and preservation. Observational data, such as direct observations of ocean temperature on a

specific date, the attitude of voters before an election, or photographs of a supernova are historical records that cannot be recollected. Thus, these observational data are usually archived indefinitely.

A different set of considerations applies to computational data, such as the results from executing a computer model or simulation. If comprehensive information about the model (including a full description of the hardware, software, and input data) is available, preservation in a long-term repository may not be necessary because the data can be reproduced. Thus, although the outputs of a model may not need to be preserved, archiving of the model itself and of a robust metadata set may be essential.

Experimental data such as measurements of patterns of gene expression, chemical reaction rates, or engine performance present a more complex picture. In principle, data from experiments that can be accurately reproduced need not be stored indefinitely. In practice, however, it may not be possible to reproduce precisely all of the experimental conditions, particularly where some conditions and experimental variables may not be known and when the costs of reproducing the experiment are prohibitive. In theses instances, long-term preservation of the data is warranted. Thus, considerations of cost and reproducibility are key in considering policies for preservation of experimental data.

Finally, processing and curatorial activities generate derivative data. Initially, data may be gathered in raw form, for instance as a digital signal generated by an instrument or sensor. These raw data are frequently subject to subsequent stages of refinement and analysis, depending on the research objectives. There may be a succession of versions. While the raw data may be the most complete form of data, in many cases their value is limited. Raw data, for example, may contain faulty information due to equipment anomalies or may require extensive processing to be usable.

The experimental process is the origin of another distinction, in this case between the intermediate data gathered during preliminary investigations and final data. Researchers may often conduct variations of an experiment or collect data under a variety of circumstances and report only the results they think are the most interesting. Selected final data are routinely included in data collections, but quite often the intermediate data are either not archived or are inaccessible to other researchers. There is, however, the growing realization that intermediate data may be of use to other researchers. And this gives rise to cost/value tradeoffs.

To make data usable, it is necessary to preserve adequate documentation relating to the content, structure, context, and source (e.g., experimental parameters and environmental conditions) of the data collection – collectively called *metadata*. Ideally, the metadata are a record of everything that might be of interest to another researcher. For computational data, for instance, preservation of data models and specific software is as important as the preservation of data they generate. Similarly, for observational and laboratory data, hardware and instrument specifications and other contextual information are critical. Metadata is crucial to assuring that the data element is useful in the future. The use of metadata and their accuracy have increased over the past several decades.

DIGITAL DATA COLLECTIONS

We use the term *data collections*, rather than the more restrictive term *databases*, because any policy discussion must include the full range of elements that impact the management of digital data collections and our investment in them. Throughout the report, *data collection* will refer to not only a database or group of databases, but also to the infrastructure, organization and individuals essential to managing the collection.

Data collections fall into one of three functional categories (examples of data collections in each of these categories are provided in Appendix D). Each of these three types of digital data collections raises unique issues for policy makers.

- *Research data collections* are the products of one or more focused research projects and typically contain data that are subject to limited processing or curation. They may or may not conform to community standards, such as standards for file formats, metadata structure, and content access policies. Quite often, applicable standards may be nonexistent or rudimentary because the data types are novel and the size of the user community small. Research collections may vary greatly in size but are intended to serve a specific group, often limited to immediate participants. There may be no intention to preserve the collection beyond the end of a project. One reason for this is funding. These collections are supported by relatively small budgets, often through research grants funding a specific project.
- *Resource or community data collections* serve a single science or engineering community. These digital collections often establish community-level standards either by selecting from among preexisting standards or by bringing the community together to develop new standards where they are absent or inadequate. The budgets for resource or community data collections are intermediate in size and generally are provided through direct funding from agencies. Because of changes in agency priorities, it is often difficult to anticipate how long a resource or community data collection will be maintained.
- *Reference data collections* are intended to serve large segments of the scientific and education community. Characteristic features of this category of digital collections are a broad scope and a diverse set of user communities including scientists, students, and educators from a wide variety of disciplinary, institutional, and geographical settings. In these circumstances, conformance to robust, well-established, and comprehensive standards is essential, and the selection of standards by reference collections often has the effect of creating a universal standard. Budgets supporting reference collections are often large, reflecting the scope of the collection and breadth of impact. Typically, the budgets come from multiple sources and are in the form of direct, long-term support, and the expectation is that these collections will be maintained indefinitely.

Note that digital collections in each of these three categories can be housed in a single physical location or they may be virtual, housed in a set of physical locations and linked together electronically to create a single, coherent collection. The distinction between centralized and distributed collections can have important implications for developing policy for funding and for ensuring their persistence and longevity.

Data collections may also differ because of the unique policies, goals, and structure of their funding agencies. Collections created and maintained by government data centers such as the Earth Resources Observation Systems (EROS) Data Center, data federations such as the Mammal Networked Information System (MaNIS), and university consortia such as the University Corporation for Atmospheric Research (UCAR) each pose unique challenges for policy makers.

EXAMPLE OF THE EVOLUTION OF A COLLECTION: THE PROTEIN DATA BANK

It is informative to review the history of a collection in order to illustrate the dynamic nature of data collections as well as the complexity of issues that are characteristic of the data collections universe. The history of the Protein Data Bank (www.pdb.org) highlights the difficulty of devising policy for long-lived data collections, namely addressing the evolution of the collection over time. The Protein Data Bank was launched in 1971 as a digital collection with fewer than a dozen files that described experimentally determined, three-dimensional structures of certain biological macromolecules. It was a research-level collection at its inception. Today, the collection is considered the premier, authoritative source for experimental structural information on biological macromolecules. More than 2,700 structures were deposited in the collection during the first six months of 2004 alone. The primary site and its seven mirror sites worldwide serve an average of more than 130,000 file downloads per day. In summary, the Protein Data Bank has been transformed from a research collection into a global, reference collection of the first rank.

The evolution of the Protein Data Bank is not simply a matter of size. Responsibilities of those managing the collection changed from simply providing a reliable archive to providing a robust set of community-proxy services that includes community-based standards development and implementation, quality assessment and control, expert annotation, and linkage to related resources. With this increase in responsibilities came a need for increased funds. The collection was originally launched at Brookhaven National Laboratory with support from the Department of Energy. The first extramural support was requested from the NSF in 1974 through an unsolicited research proposal. Today, the Protein Data Bank is supported by a coalition of eight Federal agencies along with multiple international partners.

The evolution of the Protein Data Bank is illustrative of a common feature of the data collections universe: the needs and responsibilities of data authors, managers, and users as well as those of the funding agencies can change over time with changes in research priorities and the appearance of new research techniques and questions. In the past, this process has been managed at the level of the discipline or community (and at the corresponding NSF program level). However, given the substantial cost of creating data collections and managing their growth and evolution, this approach is no longer adequate.

LONG-LIVED DIGITAL DATA COLLECTIONS

The meaning of *long-lived* or *long-term* in reference to digital collections has been defined as follows in the Open Archival Information System (OAIS) standards of the Consultative

Committee for Space Data Systems (CCSDS) of the Organization for Standardization (ISO) (see http://www.ccsds.org/CCSDS/documents/650x0b1.pdf):

> *A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing user community, on the information being held in a repository. This period extends into the indefinite future.*

The OAIS definition is technology driven in that it states that the defining characteristic of a long-lived collection is the migration of data content across multiple generations of technological media. The cost and technical difficulty of sustaining a large data collection as it transitions from one recording media to another is formidable.

This report focuses on those digital data collections that are long-lived according to this OAIS definition. Essentially all reference and most resource data collections fall under this definition. Many research collections are intended to be short-lived and do not. However, there are important exceptions. These include research collections that have enduring value to continuing projects and therefore must be maintained over a long period. Also, the community may recognize certain research collections as worthy of preservation. These research collections may then become (or be subsumed by) resource or reference collections. Thus, this report considers policy issues relevant to long-lived digital data collections at the research, resource and reference levels.

DIGITAL DATA COMMON SPACES

Not all researchers have equal access to the resources and expertise necessary to create and operate a digital data collection. The need is especially apparent at the level of an individual investigator developing a research collection. However, reliable and continuing access to the necessary resources and expertise presents a significant barrier to many communities seeking to establish resource or reference level collections. Today, there are several efforts to provide broad access to the hardware, software, connectivity, and expertise necessary to support data collections at all levels. For example, the Massachusetts Institute of Technology and Hewlett-Packard have developed D-Space (see http://dspace.org/). This is an example of a digital data commons – defined here as an element of infrastructure, much as a university library or a campus core facility for DNA sequencing would be considered as infrastructure. The data commons consists of the cyberinfrastructure for data preservation, retrieval and analysis, robust communications links for global access, and data scientists who direct the facility and can act as consultants and collaborators to the researchers served by the facility. A data commons may simultaneously support many short-term and long-lived collections, including multiple instances of research, resource and reference collections. A commons can be broadly enabling, allowing individual investigators who are not information specialists to launch and maintain digital data collections.

CONCLUSIONS

The digital data collections universe is complex, involving many participants using many types of data for many different purposes. In recent years, the research community has witnessed the rise of a multitude of collections that are robust and flexible, while allowing for

heterogeneous data types and associated metadata, allowing them to meet the wide range of needs, customs, and expectations that are found among the communities of data authors and users. To be effective in supporting data collections and enabling research in a digital environment, informed policy must build on these examples to enable all of the elements of the data collection universe.

III. ROLES AND RESPONSIBILITIES OF INDIVIDUALS AND INSTITUTIONS

SHARED GOALS AND RESPONSIBILITIES

Sound policy development and implementation rest on the recognition of the roles and responsibilities of those who play an active part in the digital data collection universe—the data users, authors, managers, and funding agencies. One of the goals of policy is to ensure that these roles and responsibilities are clearly defined and properly fulfilled. In pursuing their respective interests in data collections, each actor in the data collection universe has a distinct set of responsibilities, which are outlined in the paragraphs that follow. In addition to their separate responsibilities, the groups must also act collectively to pursue some of the higher-level goals important to the entire fields. Examples of such goals are the following:

- ensure that all legal obligations and community expectations for protecting privacy, security, and intellectual property are fully met;
- participate in the development of community standards for data collection, deposition, use, maintenance, and migration;
- work towards interoperability between communities and encourage cross-disciplinary data integration;
- ensure that community decisions about data collections take into account the needs of users outside the community;
- encourage free and open access wherever feasible; and
- provide incentives, rewards, and recognition for scientists who share and archive data.

An important policy consideration is the creation of opportunities and mechanisms by which all of the groups can work together in addressing universal goals.

DATA AUTHORS

The interests of the data authors – the scientists, educators, students, and others involved in research that produces digital data – lie in ensuring that they enjoy the benefits of their own work, including gaining appropriate credit and recognition, and that their results can be broadly disseminated and safely archived. In pursuing these interests, the data authors have the following responsibilities:

- conform to community standards for recording data and metadata that adequately describe the context and quality of the data and help others find and use the data;
- allow free and open access to data consistent with accepted standards for proper attribution and credit, subject to fair opportunity to exploit the results of one's own research and appropriate legal standards for protecting security, privacy and intellectual property rights;
- conform to community standards for the type, quality, and content of data, including associated metadata, for deposition in relevant data collections;
- meet the requirements for data management specified in grants, contracts, and cooperative agreements with funding agencies; and

- develop and continuously refine a data management plan that describes the intended duration and migration path of the data.

Robust, comprehensive, and broadly endorsed and disseminated community standards are crucial to the ability of authors to meet these responsibilities. Thus, active support for the development of community standards is an important policy goal.

DATA MANAGERS

Data managers – the organizations and data scientists responsible for database operation and maintenance – have the responsibility to:

- be a reliable and competent partner in data archiving and preservation, while maintaining open and effective communication with the served community;
- participate in the development of community standards including format, content (including metadata), and quality assessment and control;
- ensure that the community standards referenced above are universally applied to data submissions and that updated standards are reflected back into the data in a timely way;
- provide for the integrity, reliability, and preservation of the collection by developing and implementing plans for backup, migration, maintenance, and all aspects of change control;
- implement community standards through processes such as curation, annotation, technical standards development and implementation, quality analysis, and peer-review (some of these functions, defined in this report as *community-proxy functions*, apply primarily to resource and reference collections and may not apply to many research collections);
- provide for the security of the collection;
- provide mechanisms for limiting access to protect property rights, confidentiality, privacy, and to enable other restrictions as necessary or appropriate;
- encourage data deposition by authors by making it as easy as possible to submit data; and
- provide appropriate contextual information including cross-references to other data sources.

To be successful, the data manager must gain the trust of the community that the collection serves. Thus, collections policy should emphasize the role of the community in working with data managers.

DATA SCIENTISTS

The interests of data scientists – the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, and others, who are crucial to the successful management of a digital data collection – lie in having their

creativity and intellectual contributions fully recognized. In pursuing these interests, they have the responsibility to:

- conduct creative inquiry and analysis;
- enhance through consultation, collaboration, and coordination the ability of others to conduct research and education using digital data collections;
- be at the forefront in developing innovative concepts in database technology and information sciences and applying these in the fields of science and education relevant to the collection;
- implement best practices and technology;
- serve as a mentor to beginning or transitioning investigators, students and others interested in pursuing data science; and
- design and implement education and outreach programs that make the benefits of data collections and digital information science available to the broadest possible range of researchers, educators, students, and the general public.

Almost all long-lived digital data collections contain data that are materially different: text, electro-optical images, x-ray images, spatial coordinates, topographical maps, acoustic returns, and hyper-spectral images. In some cases, it has been the data scientist who has determined how to register one category of representation against another and how to cross-check and combine the metadata to ensure accurate feature registration. Likewise, there have been cases of data scientists developing a model that permits representation of behavior at very different levels to be integrated. Research insights can arise from the deep understanding of the data scientist of the fundamental nature of the representation. Such insights complement the insights of the domain expert. As a result, data scientists sometimes are primary contributors to research progress. Their contribution should be documented and recognized. One means for recognition is through publication, i.e. refereed papers in which they are among the leading authors.

DATA USERS

The interests of data users – construed here to include the larger scientific and education communities, including their representative professional and scientific communities – lie in having ready access to data sets that are searchable, robust, well defined, and well documented. In pursuing these interests, data users have the responsibility to:

- adhere to appropriate standards for attribution and credit in the use of data generated by others and observe appropriate limits on redistribution;
- report significant errors to data managers or authors as appropriate;
- provide primary input to decisions on what data are valuable to archive (for instance, raw versus processed data) and for how long;
- reach consensus on data center needs/structure for their user community and evaluate the quality of the available centers; and
- respect restrictions on use, such as copyright and no-derivatives, placed on data sets.

Meeting responsibilities for attribution and for respecting restrictions on use requires that the relevant information be readily available to the user. Thus, an important policy consideration is the development of metadata systems that provide authorship, versioning, modification,

licensing, and other relevant information. The system of digital licensing being developed by Creative Commons (see www.creativecommons.org) provides an example in this regard.

FUNDING AGENCIES

Much of the data currently being collected are 'born-digital' and lack any analog counterpart. Additional data are being converted to digital form and, in the process, are often dissociated from their analog representation. The digital data, and the investments made in gathering them, could be lost unless a robust preservation plan is created for digital data. This is the role and responsibility of NSF and other funding agencies, working in concert with data authors, managers, and users to:

- create a culture in which digital data receives the same consideration as data published in print form so that an author's contribution is judged by the insights, creativity, and significance of the analysis and not by the media in which the data are created and stored. Compiling, editing, and publishing data in a data collection should be seen as a fundamental research responsibility. The emphasis on preservation (and the development of a stable preservation infrastructure) would be the equivalent to that now attached to the preservation of data in printed form.
- catalyze the creation of an accessible digital commons for research and education that provides the foundation for launching, operating, and preserving research, resource, and reference collections;
- support interactions within and between communities to allow the development of robust community standards for digital data and interoperability and facilitate the development of community norms, customs, and expectations for digital research; and
- enable the broadest possible access to the digital research environment by ensuring that both the physical resources and the necessary training are broadly available. Provide the oversight to ensure that this training supports the development of the expert workforce and scientific leadership required for innovative digital discovery through digital data systems and collections.

The Foundation is in a unique position to act because of the fundamental support it provides for the research and education enterprise, its history of leadership in the area of digital data and research, and the breadth of disciplinary representation and participation found across the Foundation. Because digital data collections have become indispensable to advances in research and education, the task force believes that urgent action, involving transformative, rather than incremental, change is required.

DATA QUALITY ACT

Federal agencies have responsibilities under the so-called 'Data Quality Act' (Public Law 106-554; H.R. 5658, Sec. 515). In accordance with the Act, the Office of Management and Budget (OMB) has issued guidelines that "provide policy and procedural guidance to Federal agencies for ensuring and maximizing the quality, objectivity, utility, and integrity of information ... disseminated by Federal agencies" (see http://www.whitehouse.gov/omb/fedreg/reproducible2.pdf). These guidelines apply to information whose collection and dissemination to the public is initiated or sponsored by a Federal agency. NSF examples include the annual Science and Engineering Indicators report

and certain other publications produced by the NSF Division of Science and Engineering Statistics.

Importantly, the OMB guidelines do not apply to information disseminated by a Federal grantee or contractor or Federally employed scientist when he or she publishes and communicates research findings in the same manner as academic colleagues, or decides whether to disseminate research results or other data and what information will be included in the dissemination. Thus, the guidelines do not apply to information disseminated by NSF-funded grantees as outlined in the NSF Information Quality Guidelines (see http://www.nsf.gov/policies/nsfinfoqual.pdf):

> *NSF grantees are wholly responsible for conducting their project activities and preparing the results for publication or other distribution. NSF promotes data sharing by its grantees through its data sharing policy and by data archiving by its grantees. NSF does not create, endorse, or approve such data or research materials, nor does the agency assume responsibility for their accuracy.*

As the Foundation develops policy and strategy for long-lived digital data, it is essential that the traditional distinction between NSF initiated and disseminated data, and data created, maintained, and shared by its grantees be maintained.

IV. 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? 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 (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 polices, 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.

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.

V. 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. The Board anticipates that a larger dialog with other agencies in the U.S. and with international partners will eventually be required before all the issues are resolved. Since the issues are urgent and since undertaking

broader discussions depends on a clear understanding of the Foundation's objectives and capabilities, we look for a timely response to these recommendations from NSF.

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

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 IV 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 broadening the scope of its discussions beyond NSF to include the full spectrum of data collections and supporting agencies, both national and international. These discussions would be informed by the response from NSF and could be designed to examine in both the national and global contexts how the investment, the policies, and the management of in-agency and out-agency collections works to the end objective of cost-effective, high-quality research and education in data-rich environments. The need to address these issues is urgent. The opportunities are substantial.

APPENDIX A
TASK FORCE CHARTER

NSB 04-19
February 5, 2004

CHARTER FOR THE LONG-LIVED DATA COLLECTION TASK FORCE

Data collections, particularly digital data collections for research, have been increasing in number and size over the past couple of decades. They range from small single investigator collections to very large collections whose content is derived from instruments housed in large facilities. Over this same period, the National Science Foundation (NSF) obligations for support for both data collection and curation has been increasing.

The Foundation differs from agencies, such as NASA, NOAA, and the Department of Energy. Most frequently, their strategy is for the agency to own and manage the collections. As a consequence, they own and manage many fewer independent collections than the NSF supports. With ownership comes agency control for data format standards and access policies. The Foundation does not typically maintain data collections itself. It is individual researchers, consortia, and organizations that develop and maintain large facilities that manage the collections. This has resulted in a proliferation of data collections, large and small, across all disciplines. There is divergence in formats, access policies, and in quality.

It is timely to consider the policy ramifications of this rapid growth of data collections in the NSF supported community. This National Science Board task force will address the policy issues directly relevant to the NSF's style of data collection support. These policy issues and questions include:

- When, why, and for how long the NSF will fund data collections that are or appear to be very long-lived (decades)?
- Are there conditions under which it is appropriate for NSF to maintain a data collection intramurally, as most other agencies routinely do?
- What responsibility does NSF have to assure quality of the collections that it supports?
- What part, if any, should NSF play in creating and enforcing community technical standards, for example, for the use and form of metadata?
- How does NSF assure that data is accessible to a broad, diverse, and interdisciplinary community?
- Should the budget for collection curation be made more visible, or remain (more or less) integral to the many different research budgets?
- Is there a desired balance between expenditure on collection and curation?
- Since digital media are impermanent, migration to new media are crucial if a collection is to persist. Under what conditions should NSF support such migration?

- What policies should guide relationships with national, international, and private agencies and organizations to cooperatively support data curation?

The objective of this National Science Board task force is to delineate the policy issues relevant to the National Science Foundation and its style and culture of supporting the collection and curation of research data. For those issues where guidance to the Foundation is appropriate, the task force should make recommendations for the National Science Board and the community to consider.

APPENDIX B
SOURCES OF ADDITIONAL INFORMATION

Blue-Ribbon Advisory Panel on Cyberinfrastructure (Atkins, Daniel E. Chair). 2003. *Revolutionizing Science and Engineering through Cyberinfrastructure.* Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. January 2003. Available online at http://www.cise.nsf.gov/sci/reports/atkins.pdf

This report of an NSF advisory panel provides an evaluation of current major investments in cyberinfrastructure and its use, recommends new areas of emphasis relevant to cyberinfrastructure, and proposes an implementation plan for pursuing the recommendations.

Consultative Committee for Space Data Systems. 2002. *Recommendation for Space Data System Standards: Reference Model for an Open Archival Information System (OAIS*). CCSDS 650.0-B-1, Available online at http://www.ccsds.org/CCSDS/documents/650x0b1.pdf.

This document was produced by the Consultative Committee for Space Data Systems (CCSDS) of the Organization for Standardization (ISO) and provides a reference model for archival systems that serve to preserve and maintain long-term access to digital information.

Dublin Core Metadata Initiative Web site. http://dublincore.org/

The Dublin Core Metadata Initiative (DCMI) is an organization dedicated to promoting the widespread adoption of interoperable metadata standards and developing specialized metadata vocabularies for describing resources that enable more intelligent information discovery systems.

Hodge, Gail and Bonnie Carroll. "Digital Archiving: The State of the Art, the State of the Practice." April 1999 http://www.icsti.org/icsti/Dig_Archiving_Report_1999.pdf

This report, sponsored by the International Council for Scientific and Technical Information's Information Policy Committee and CENDI, provides information on the state-of-the-art and practice in digital electronic archiving in terms of policy, models, and best practices, with an emphasis on cutting-edge approaches.

Hodge, Gail, and Evelyn Frangakis. 2004. Digital Preservation and Permanent Access to Scientific Information: The State of the Practice. CENDI US Federal Information Managers Group. CENDI 2004-3. Available online at: http://www.icsti.org/icsti/icsti_reports.html

This report, by the International Council for Scientific and Technical Information (ICSTI) and the CENDI US Federal Information Managers Group, focuses on operational digital preservation systems specifically in science and technology (S&T). It considers the wide range of digital objects of interest to S&T, including e-journals, technical reports, e-records, project documents, and scientific data.

Inter-university Consortium for Political and Social Research. 2002. Guide to Social Science Data Preparation and Archiving. Available online at http://www.icpsr.umich.edu/ACCESS/dpm.html

This guide is intended to help researchers document their datasets and prepare them for archiving. It describes in detail the processes involved in data creation and management, and in preparing materials for deposit in ICPSR. The project was supported by the Robert Wood Johnson Foundation.

Kurtzman, Howard S., Russell M. Church, and Jonathon D. Crystal. 2002. *Data Archiving for Animal Cognition Research: Report of an NIMH Workshop.* Workshop report available online: http://www.brown.edu/Departments/Psychology/anicog/dar-25jul02.pdf. Also published in *Animal Learning & Behavior 30* (4), 405-412.

The workshop report provides a set of conclusions and recommendations concerning: (A) the impact of data archiving on research; (B) how to incorporate data archiving into research practice; (C) contents of data archives; (D) technical and archival standards; and (E) organizational, financing, and policy issues.

Lord, Philip, and Alison Macdonald. 2003. *e-Science Curation Report--Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision.* Prepared for the JISC Committee for the Support of Research (JCSR). Twickenham,U.K. The Digital Archiving Consultancy Limited. Available online at: http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

The study examines the current provision and future needs of curation of primary research data in the UK, particularly within the e-Science context.

Lyman, Peter and Hal R. Varian, "How Much Information", 2003. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003.

This online study is an attempt to estimate how much new information is created each year. It covers information distributed in four storage media – print, film, magnetic, and optical – and seen or heard in four information flows – telephone, radio and TV, and the Internet.

MacDonald, Alison, and Philip Lord. 2003. *Digital Data Curation Task Force: Report of the Task Force Strategy Discussion Day, Tuesday, 26th November 2002.* Twickenham, U.K.: The Digital Archiving Consultancy. January 2003. Available online at: http://www.jisc.ac.uk/uploaded_documents/CurationTaskForceFinal1.pdf

This is report summarizes the meeting of a United Kingdom task force organized under the auspices of the Joint Information Systems Committee's Committee for the Support of Research. The task force's goal was to define and structure a strategy for the "curation" of primary research data in the UK.

National Research Council. 1995. *Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources.* Washington: National Academy Press.

This report, under the auspices of the Commission on Physical Sciences, Mathematics, and Applications, was initiated at the request of the National Archives and Records Administration (NARA) and the National Oceanic and Atmospheric Administration (NOAA). It defines a set of goals, principles, and priorities, and generic retention criteria for archiving of physical science data.

National Research Council. 2003. *The Role of Scientific and Technical Data and Information in the Public Domain: Proceedings of a Symposium.* Board on International Scientific Organizations. Washington: National Academy Press. Available online at http://www.nap.edu/books/030908850X/html/

This symposium report covers 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.

National Science Board. 2003. *Science and Engineering Infrastructure Report for the 21st Century: -The Role of the National Science Foundation*. Arlington, VA: National Science Foundation. (NSB-02-190). February 8, 2003. Available online at: http://www.nsf.gov/nsb/documents/2002/nsb02190/nsb02190.pdf

This report presents the findings and recommendations developed by the Task Force on Science and Engineering Infrastructure of the National Science Board Committee on Programs and Plans. The task force assessed the current state of U.S. S&E academic research infrastructure, examined its role in enabling S&E advances, and identified requirements for a future infrastructure capability.

National Science Foundation and the Library of Congress. *It's About Time: Research Challenges in Digital Archiving and Long-term Preservation.* Final Report ,Workshop On Research Challenges In Digital Archiving And Long-Term Preservation, held April 12-13, 2002. Sponsored by the National Science Foundation, Digital Government Program and Digital Libraries Program, Directorate for Computing and Information Sciences and Engineering and the Library of Congress National Digital Information Infrastructure and Preservation Program. August 2003. Available online at http://www.digitalpreservation.gov/repor/NSF_LC_Final_Report.pdf

This workshop report provides a research agenda to address key technological and computer and information sciences challenges in digital archiving and preservation. In addition, a broader discussion of issues relevant to a national digital preservation program, including archival architecture and property rights considerations, technical challenges, and potential roles of institutional and agency players can be found at the website of the National Digital Information Infrastructure Preservation Program: http://www.digitalpreservation.gov/.

The Wellcome Trust. 2003. *Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility*. Report of a meeting organized by the Wellcome Trust, held on 14–15 January 2003, Fort Lauderdale, USA. Available online at: http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf

The report discusses how pre-publication data release can promote the best interests of science and help to maximize the public benefit to be gained from research. It delineates responsibilities of funding agencies, (data) resource providers, and resource users.

APPENDIX C
CURRENT POLICIES ON DATA SHARING AND ARCHIVING

There are a variety of current policies in place at NSF and in other agencies that vary considerably in their scope and in their provisions. There are also a variety of community standards, some set by professional societies, some set by journals, and some established through community meetings.

This Appendix provides examples of existing policies, illustrates how policies can differ across the NSF and across agencies, and identifies areas where there may be a lack of adequate policy or a lack of appropriate consistency across different policies.

EXAMPLES OF DATA POLICIES

NSF POLICIES

This section includes examples of NSF policies, including NSF's general conditions for grants as well as the data policies of several specific programs.

NATIONAL SCIENCE FOUNDATION GRANT GENERAL CONDITIONS

NSF's Grant General Conditions include the following:

> Article 36. Sharing of Findings, Data, and Other Research Products
>
> a. 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 cost 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 awardees to share software and inventions or otherwise act to make the innovations they embody widely useful and usable.
>
> b. 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.

These conditions are quite general, and do not address archiving of data, the duration of data collections, or requirements for providing metadata or finding aids.

DIVISION OF ENVIRONMENTAL BIOLOGY

The Division of Environmental Biology, within the Directorate for Biological Sciences, follows general NSF policy and has developed the following statement for program announcements.

> Proposals submitted to all programs in DEB must adhere to the general NSF policy on data sharing as described in the Grant Proposal Guide… Thus, proposals should describe plans for specimen and information management and sharing, including where data and metadata, will be stored and maintained, and the likely schedule for release. These plans will be considered as part of the review process.
> http://www.nsf.gov/bio/deb/

Division of Ocean Sciences

The Division of Ocean Sciences, within the Geosciences Directorate has a long-standing and detailed policy for oceanographic data. Excerpts from the policy statement are provided below.

> POLICY FOR OCEANOGRAPHIC DATA, NSF 94-126
>
> Ocean data collected under Federal sponsorship and identified as appropriate for submission to a national data center are to be made available within a reasonable time as described below.
>
> Principal investigators are required to submit all environmental data collected to the designated national data centers as soon as possible, but no later than two (2) years after the data are collected. Inventories of all marine environmental data collected should be submitted to the designated national data centers within sixty (60) days after the observational period/cruise…
>
> Data sets identified for submission to the national data centers must be submitted to the designated center within two (2) years after the observational period. This period may be extended under exceptional circumstances by agreement between the principal investigator and NSF. Data produced by long-term (multi-year) projects are to be submitted annually…
>
> NOAA's National Environmental Satellite Data and Information Service staff and program representatives from funding agencies will identify the data sets that are likely to be of high utility and will require their principal investigators to submit these data and related information to the designated center.
>
> Funding agencies will apply this policy to their internal ocean data collection and research programs and to their contractors and grantees and will establish procedures to enforce this policy.

> A list of oceanographic data types and the centers designated to receive them are the following….:
>
> Data are to be submitted according to formats and via the media designated by the pertinent national data center.
>
> Principal investigators and ship-operating institutions are also responsible for meeting all legal requirements for submission of data and research results, which are imposed by foreign governments as a condition of that government's granting research clearances….

The full policy is available at: http://www.nsf.gov/pubs/stis1994/nsf94126/nsf94126.html

DIVISION OF BEHAVIORAL AND COGNITIVE SCIENCES (BES)

NSF's Division of Behavioral and Cognitive Sciences (BCS), within the Directorate for Social, Behavioral and Economic Sciences, has a data policy that recognizes the diversity of types of data handled by the division. Excerpts from this policy follow:

> BCS supports a wide range of disciplines. The nature of the data, the way they are collected, analyzed, and stored, and the pace at which this reasonably occurs varies widely. There are different storage facilities and different access requirements for, e.g., archaeological data, specimens from physical anthropology, large-scale survey data, oral interviews with scientists and other subjects, data generated by experimental research, and field records of tribal ceremonies. Where appropriate and possible, grantees from all fields will develop and submit specific plans to share materials collected with NSF support. These plans should cover how and where these materials will be stored, at reasonable cost, and how access will be provided to other researchers, generally at their cost.
>
> This policy explicitly recognizes that many complexities arise across the range of data collection supported by BCS programs, and that unusual circumstances may require modifications or even full exemptions. For example, human subjects protection requires removing identifiers, which may be prohibitively expensive or render the data meaningless in research that relies heavily on extensive in-depth interviews. Intellectual property rights may be at risk in some forms of data collection. The policy is intended to be flexible enough to accommodate the variety of scientific enterprises that constitute BCS programs. No comprehensive set of rules is possible, but the procedures indicated below are designed to provide guidance for broad categories of data collection.
>
> Experimental Research
> In experimental research, individuals, be they people, animals, or objects, are subjected to preplanned conditions and their responses tabulated in some fashion. Investigators

should plan to make these tabulated data available to other investigators requesting them. In addition, complete information on how an experiment was conducted and any unusual stimulus materials should be made available, so that failures to replicate will not turn out to depend on one scientist's incomplete understanding of another's procedure.

Mathematical and Computer Models
Often in the course of conducting research, investigators develop mathematical and computer models, either as an innovative aid in the analysis of data or as a theoretical statement about the processes involved in generating some classes of data. Investigators should plan to make these models available to others wanting to apply them to other data sets or experimental situations…

Object Based Research
Some research supported by BCS is based on objects such as archaeological specimens or fossil remains. In these instances data consist of the objects themselves, contextual information such as geological sections and finally quantitative and qualitative descriptions of the materials. Because these physical objects rarely become the property of the investigator but belong to a host nation or cultural group, scientists often do not control access to them. This situation is further complicated by the fact that description of materials often must proceed slowly and may take several years to complete. However, it is still incumbent upon the investigator to make primary and contextual information available as rapidly as possible to permit other scientists to examine them and draw their own conclusions.

Qualitative Information
The kinds of qualitative information collected in research projects supported by BCS can range from microfilms and other copies of very old documents to oral interviews and video tapes about historical events in science or about contemporary technological controversies. They can consist of ethnographic or linguistic field notes or recordings or transcriptions, or hand written records of open-ended interviews. Investigators should consider whether and how they can develop special arrangements to keep or store these materials so that others can use them. If it is appropriate for other researchers to have access to them, the investigators should specify a time at which they will be made generally available, in an appropriate form and at a reasonable cost.

Quantitative Social and Economic Data Sets
For appropriate data sets, researchers should be prepared to place their data in fully cleaned and documented form in a data archive or library within one year after the expiration of an award. Before an award is made, investigators will be asked to specify in writing where they plan to deposit their data set(s)…

The full policy is available at http://www.nsf.gov/sbe/bcs/common/archive.htm

OTHER AGENCY AND INTERAGENCY DATA POLICIES

U.S. GLOBAL CHANGE RESEARCH PROGRAM

The interagency U.S. Global Change Research Program has a high level data policy that provides guidelines for more specific policies by participating agencies. Excerpts follow.

> The U.S. Global Change Research Program requires an early and continuing commitment to the establishment, maintenance, validation, description, accessibility, and distribution of high-quality, long-term data sets. Full and open sharing of the full suite of global data sets for all global change researchers is a fundamental objective.
>
> Preservation of all data needed for long-term global change research is required. For each and every global change data parameter, there should be at least one explicitly designated archive. Procedures and criteria for setting priorities for data acquisition, retention, and purging should be developed by participating agencies, both nationally and internationally. A clearinghouse process should be established to prevent the purging and loss of important data sets.
>
> Data archives must include easily accessible information about the data holdings, including quality assessments, supporting ancillary information, and guidance and aids for locating and obtaining the data.
>
> National and international standards should be used to the greatest extent possible for media and for processing and communication of global data sets.
>
> Data should be provided at the lowest possible cost to global change researchers in the interest of full and open access to data. This cost should, as a first principle, be no more than the marginal cost of filling a specific user request. Agencies should act to streamline administrative arrangements for exchanging data among researchers.
>
> For those programs in which selected principal investigators have initial periods of exclusive data use, data should be made openly available as soon as they become widely useful. In each case the funding agency should explicitly define the duration of any exclusive use period.

There are more details at http://www.globalchange.gov/policies/diwg/diwg-guidelines.html

NOAA COASTAL OCEAN PROGRAM (COP) DATA POLICY

Many of the programs and agencies involved in observational earth science data have data policies that are generally similar. The National Oceanic and Atmospheric Administration's Coastal Ocean Program is one example. Excerpts from its policies include:

The COP Data Policy promotes: (1) full and open sharing of data and other products of COP-sponsored research by all COP researchers; (2) entitling the investigator who collects the data to the fundamental benefits of the collected data set, derived models, etc.; (3) selection of methods and equipment to ensure sufficient accuracy and precision to meet the project requirements for inter-comparisons and syntheses; (4) preservation of all data collected under COP sponsorship, including derived models, in an easily accessible archive with transfer ultimately to a permanent archive at a National Data Center…

COP encourages the no-cost, open, voluntary and ethical exchange of data or other COP-related information among investigators. Publication of descriptive or interpretive results immediately and directly from the data is the privilege and responsibility of the investigators who collect the data. Prior to submission to a permanent data archive at a National Data Center, publication or presentation of any data derived by a co-participating investigator requires the permission of the scientist originating the data. Any scientist making substantial use of a data set should anticipate that the data collectors will be co-authors of published results. Originating investigators may not unreasonably impede use or publication of archived data, models, or model application.

Methods and equipment used to take measurements and collect samples must be of sufficient accuracy and precision to yield data with quality adequate to meet the objectives of the COP field projects, associated modeling efforts, and larger-scale synthesis…

A data archive system will be established by each COP-sponsored project within six (6) months of the project start date for temporary repository of the data prior to their submittal to a permanent archive. The data archive system must facilitate the exchange of data and insure the long-term existence of the data set. The COP Project Manager (or a designated project Data Manager) will ensure the following data archive system conditions are met:

- data integrity and appropriate metadata are maintained;
- all users are provided access in a timely manner;
- and the data are transferred to a designated National Data Center (e.g., National Oceanographic Data Center) within two (2) years from the time of initial observations.

…The submitted data will include the actual measurements and supporting descriptive information (i.e., metadata) sufficient to permit its effective use by researchers not familiar with the original project or the particular instrument making the measurements. The NOAA/Federal Geographic Data Committee Metadata Standard Format shall be used to describe the data.…

This policy also encourages the project archive to include selected models, and model products or results. Measurements which do not involve manual analysis should be submitted to the project archive within six (6) months. All measurements, including metadata, should be submitted to a National Data Center for permanent archive…

> Unclassified and/or unrestricted environmental data and information produced, sponsored, collected, or obtained by NOAA/COP are public property. It is NOAA policy to make environmental data and information available under NOAA's stewardship based on exchange, loan, cost of dissemination, or at no cost in the interest of full and open access to data.

The full policy is available at http://www.cop.noaa.gov/../Grants/datapolicy.PDF Other examples of earth-science policies that are generally similar in scope and terms are NASA's Global Change program (available at http://www.globalchange.gov/policies/agency/nasa.html and the Office of Naval Research's Ocean, Atmosphere, and Space Science and Technology Department, available at http://www.onr.navy.mil/../02/docs/tcpsod.pdf.. These policies are quite specific about what data and metadata must be provided, the timing of providing this data, and the data centers in which the data needs to be archived.

NATIONAL INSTITUTES OF HEALTH

The National Institutes of Health (NIH) has a relatively recent (2003) data sharing policy. This applies NIH-wide, but currently applies only to large grants. These policies apply to data sharing, but do not address long-term archiving. Excerpts from the policy include:

> Data should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data. To facilitate data sharing, investigators submitting a research application requesting $500,000 or more of direct costs in any single year to NIH on or after October 1, 2003 are expected to include a plan for sharing final research data for research purposes, or state why data sharing is not possible.
>
> The NIH policy on data sharing applies:
> - To the sharing of final research data for research purposes.
> - To basic research, clinical studies, surveys, and other types of research supported by NIH. It applies to research that involves human subjects and laboratory research that does not involve human subjects. It is especially important to share unique data that cannot be readily replicated.
> - To applicants seeking $500,000 or more in direct costs in any year of the proposed project period through grants, cooperative agreements, or contracts.
> - To research applications submitted beginning October 1, 2003.
>
> Final research data are recorded factual material commonly accepted in the scientific community as necessary to document, support, and validate research findings. This does not mean summary statistics or tables; rather, it means the data on which summary statistics and tables are based… For most studies, final research data will be a computerized dataset.

Given the breadth and variety of science that NIH supports, neither the precise content for the data documentation, nor the formatting, presentation, or transport mode for data is stipulated.

… if an application describes a data-sharing plan, NIH expects that plan to be enacted. In the final progress report, if not sooner, the grantee should note what steps have been taken with respect to the data-sharing plan. In the case of noncompliance (depending on its severity and duration) NIH can take various actions to protect the Federal Government's interests. In some instances, for example, NIH may make data sharing an explicit term and condition of subsequent awards.

Grantees should note that, under the NIH Grants Policy Statement, they are required to keep the data for 3 years following closeout of a grant or contract agreement… the grantee institution may have additional policies and procedures regarding the custody, distribution, and required retention period for data produced under research awards.

…NIH expects the timely release and sharing of data to be no later than the acceptance for publication of the main findings from the final dataset.

NIH recognizes that the investigators who collected the data have a legitimate interest in benefiting from their investment of time and effort. NIH continues to expect that the initial investigators may benefit from first and continuing use but not from prolonged exclusive use.

The rights and privacy of human subjects who participate in NIH-sponsored research must be protected at all times. It is the responsibility of the investigators, their Institutional Review Board (IRB), and their institution to protect the rights of subjects and the confidentiality of the data. Investigators may use different methods to reduce the risk of subject identification…

If research participants are promised that their data will not be shared with other researchers, the application should explain the reasons for such promises. Such promises should not be made routinely and without adequate justification.

For the most part, it is not appropriate for the initial investigator to place limits on the research questions or methods other investigators might pursue with the data. It is also not appropriate for the investigator who produced the data to require coauthorship as a condition for sharing the data.

…under the Small Business Act, SBIR grantees may withhold their data for 4 years after the end of the award. Issues related to proprietary data also can arise when cofunding is provided by the private sector (e.g., the pharmaceutical or biotechnology industries) with corresponding constraints on public disclosure. NIH recognizes the need to protect patentable and other proprietary data. Any restrictions on data sharing due to cofunding arrangements should be discussed in the data-sharing plan section of an application and will be considered by program staff.

There are many ways to share data.

- Under the auspices of the PI
- Data archive
- Data enclave
- Mixed mode sharing.

Investigators will need to determine which method of data sharing is best for their particular dataset.

Regardless of the mechanism used to share data, each dataset will require documentation. (Some fields refer to data documentation by other terms, such as metadata or codebooks). The precise content of documentation will vary by scientific area, study design, the type of data collected, and characteristics of the dataset.

It is appropriate for scientific authors to acknowledge the source of data upon which their manuscript is based. Many investigators include this information in the methods and/or reference sections of their manuscripts.

NIH recognizes that it takes time and money to prepare data for sharing. Thus, applicants can request funds for data sharing and archiving in their grant application.

The full policy and implementation guidance is available at
http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

PUBLICATIONS

In addition to government agencies, some publications have policies that affect the sharing and archiving of data.

SCIENCE

Science's policy is as follows:

Any reasonable request for materials, methods, or data necessary to verify the conclusions of the experiments reported must be honored.

Before publication, large data sets, including protein or DNA sequences and crystallographic coordinates, must be deposited in an approved database and an accession number provided for inclusion in the published paper, under the database deposition policy outlined below.

Database Deposition Policy
Science supports the efforts of databases that aggregate published data for the use of the scientific community. Therefore, before publication, large data sets (including microarray data, protein or DNA sequences, and atomic coordinates or electron

microscopy maps for macromolecular structures) must be deposited in an approved database and an accession number provided for inclusion in the published paper.

Macromolecular structure data. Atomic coordinates and structure factor files from x-ray structural studies or an ensemble of atomic coordinates from NMR structural studies must be deposited and released at the time of publication. Three-dimensional maps derived by electron microscopy and coordinate data derived from these maps must also be deposited. Approved databases are the Worldwide Protein Data Bank [through the Research Collaboratory for Structural Bioinformatics, Macromolecular Structure Database (MSD EMBL-EBI), or Protein Data Bank Japan], BioMag Res Bank, and Electron Microscopy Data Bank (MSD-EBI).

DNA and protein sequences. Approved databases are GenBank or other members of the International Nucleotide Sequence Database Collaboration (EMBL or DDBJ) and SWISS-PROT.

Microarray data. Data should be presented in MIAME-compliant standard format. Approved databases are Gene Expression Omnibus and ArrayExpress.

Large data sets with no appropriate approved repository must be housed as supporting online material at Science, or when this is not possible, on the author's Web site, provided a copy of the data is held in escrow at Science to ensure availability to readers.

For more information, see the *Science* Web site, http://www.sciencemag.org/feature/contribinfo/prep/gen_info.shtml#datadep

Nature has generally similar policies, available at http://www.nature.com/nature/submit/policies/index.html#6

AMERICAN GEOPHYSICAL UNION

The American Geophysical Union (AGU) has an extensive set of policies that govern (1) citations of publicly available data sets in regular AGU journal papers; (2) long-term access to small supporting data sets and graphics files that are published concurrently with, and are an electronic component of, some AGU journal papers; and (3) a special class of data and analysis papers that are offered in some AGU journals. Excerpts from these policies are as follows:

Citing Data in Regular AGU Journal Papers

1. Data sets cited in AGU publications must meet the same type of standards for public access and long-term availability as are applied to citations to the scientific literature.

Thus data cited in AGU publications must be permanently archived in a data center or centers that meet the following conditions:
a) are open to scientists throughout the world.
b) are committed to archiving data sets indefinitely.
c) provide services at reasonable costs.

The World and National data centers meet these criteria. Other data centers, though chartered for specific lengths of time, may also be acceptable as an archive for this material if there is a commitment to migrating data to a permanent archive when the center ceases operation. Citing data sets available through these alternative centers is subject to approval by AGU.

2. Data sets that are available only from the author, through miscellaneous public network services, or academic, government or commercial institutions not chartered specifically for archiving data, may not be cited in AGU publications. This type of data set availability is judged to be equivalent to material in the gray literature. If such data sets are essential to the paper and authors should treat their mention just as they would a personal communication. These mentions will appear in the body of the paper but not in the reference list.

3. To assist scientists in accessing the data sets, authors are encouraged to include a brief data section in their papers. This section should contain the key information needed to obtain the data set being cited.

4. Data sets that meet the requirements stated in paragraph 1 above can be included in the reference list of an article in an AGU publication. The format for the reference will be specified in AGU's guide for contributors. The following elements must be included in the reference: author(s), title of data set, access number or code, data center, location including city, state, and country, and date.

Data Papers

1. Editors are free to establish a category of articles that are primarily designed to discuss the acquisition, preparation, and use of key data sets. The requirements for the substance of these articles and their lengths will be determined by the editor.

2. Data sets discussed in data papers published in AGU books and journals must be publicly available and accessible to the scientific community indefinitely. Authors of such papers are required to deposit their data sets in a data center that meets the criteria discussed above. In the event that an appropriate data center cannot be found by the author, AGU will take an active role in recommending the acceptance of the data by a suitable data center. AGU will provide temporary storage services, for a fee, and will facilitate the migration of the data sets to an approved center as soon as practical. (Also see section below on AGU's role in archiving data.)

3. Data sets that are the basis of data papers are subject to review. A sample of these data sufficient for the review process must be supplied with the submission of the paper. The reviewer is expected to comment on the data as if they were an integral part of the paper and on their usability.

4. Data sets for data papers must include a descriptive section that provides the user with key information about the collection, preparation and use of the data set. (This section is sometimes called the "metadata.") The format and content of this section will be specified in AGU's guide to contributors.

5. At the time of submission, authors must supply complete information about the archiving of the data sets. To avoid possible delays in the publication of the data paper, authors should consult with the data center(s) before submitting the paper to AGU. If the data sets have been archived before the paper is submitted, information on accessing them must be supplied to the reviewers.

6. The data sets will be listed in AGU's electronic index to publications (EASI). The citation in the index will include sufficient information for locating the data set.

Characteristics of Data Archive to be Maintained by AGU

1. Permanent archive: AGU makes a commitment to maintain and provide long-term access to the data sets.

2. Platform independent: The format of such data sets and graphics files shall be platform-neutral to allow the widest possible availability.

3. Future portability: Formats for archiving data and graphics files must be in a generic, preferably non-proprietary format consistent with conversion to future open standards if necessary.

4. Ease of management: Files shall not require significant pre-processing or reformatting for administrators in order to archive the data.

5. Usability: Compression techniques used for data sets should be available on multiple platforms, such as zip utility.

6. Flexibility: The guidelines and their recommended standards should be sufficiently flexible to allow for future incorporation of technology advances, and to allow for future user input gained from practical experience.

AGU's Role in Archiving Data

It is AGU’s intent to ensure the continuity of archived data sets by providing long-term access to small supporting data sets and graphic files that are an electronic component of and other supplemental materials that are published concurrently with AGU journal

articles; entering into agreements with data centers to acquire archived data sets should the center no longer offer this service; providing temporary storage if needed for these archived data sets until a new storage center is found; and maintaining a catalog of data papers which provides the current location of data sets.

1. AGU does not expect to archive data sets subject to this policy, except on a for-fee basis and for sets of a small size. In general AGU expects data to be deposited with and maintained by facilities that are specifically chartered for that purpose. AGU will work with these facilities as described below.

2. AGU will work with data centers to help advertise their services and to help inform authors about the formats and standards established by the data centers. This information will be provided in order to assist authors in finding an approved archive for their data sets.

3. AGU will take an active role in helping to expand the scope of data centers if authors have been turned down because the subject of the data sets does not fit the charter of existing data centers.

4. It is not AGU's intention to serve as an archive for large data sets that should be housed in data centers. Nor do we expect to take on the responsibilities of handling such data sets even temporarily unless they are an electronic component of a regular AGU journal paper.

5. It is AGU's intent to ensure the continuity of archiving of data sets in the data papers. Thus, AGU will attempt to enter into agreements with data centers to acquire archived data sets should the center decide to cease storing them. AGU will provide temporary storage services while another approved center is found. To meet the continuity objective, AGU will maintain a catalog of data papers and the location of current storage.

6. AGU maintains a deposit service for supplementary material of different types in order to provide long-term access to small supporting data sets and graphics files that are published concurrently with, and are an electronic component of, some AGU journal articles. Procedures related to this service are discussed in "Guidelines for AGU Electronic Supplemental Data Set Archive."

These policies are available at http://www.agu.org/pubs/data_policy.html.

ANALYSIS OF DATA POLICIES

The examples of data policies provided here illustrate that there is a wide range in the scope, specificity, and terms of data policies within NSF, across Federal agencies, and across scientific communities. Some observations about these policies are as follows.

- Overall NSF policy is quite general, and does not address requirements for archiving (or sunsetting) data, requirements for metadata, or enforcement of policy.
- Some NSF programs have detailed data policies; others do not.
- Policies vary considerably in whether or not they require archiving of data or just sharing.
- Data policies are well established and stable for observational earth science data. This may arise in part because of the existence of a well-established system of world data centers that provide archives for data.
- Data policies are newer and evolving in the life sciences. Publication policies have an important influence on data practices in these fields. NIH policy is a recent addition to this field.
- Human subjects provisions and proprietary data concerns are major elements of data policies in the life and social sciences.

APPENDIX D
DIGITAL DATA COLLECTIONS BY CATEGORIES

INTRODUCTION

Digital data collections vary greatly in size, scope, usage, planned duration, and other dimensions. We distinguish between three functional categories of data collections: (1) *research database collections*, which are specific to a single investigator or research project; (2) r*esource or community database collections*, which are intermediate in duration, standardization, and community of users*;* and (3) r*eference collections*, which are managed for long-term use by many users. The following sections provide descriptions and examples of each of these types of digital data collections.

It should be noted that there not always clear distinctions between these categories: data collections for large research projects overlap with community database collections, and many community data collections transition to become reference data collections. These categories are based on functional attributes of the collection rather than location or size of the data set, and some data centers support all three kinds of collections.

RESEARCH DATABASE COLLECTIONS

DESCRIPTION

Research database collections are the products of one or a few focused research projects. The collections may vary greatly in size, but are intended to serve a specific group, often limited to immediate participants. These collections have relatively small budgets and may be supported directly or indirectly, often through the research grants supporting the project that they serve. Funding is assured for only a short period of time. They typically contain data that is subject to limited processing or curation, and may or may not conform to community standards (e.g. standards for file formats, metadata structure and content, access policies, etc). Often, applicable standards may be limited or rudimentary as the data types may be novel and the size of the user community may be small. The collection may not be intended to persist beyond the end of the project. Some research collections are accessible to the public through the Web, but many are not, and many of the Web links to research collections are ephemeral.

EXAMPLES

There are many thousands of research databases, and they are highly variable in size, number of users, consistency of data and metadata format, duration, and other attributes.

In the Earth Sciences, many research data sets result from field-based research projects. Examples of data sets available on the web from recent field programs can be found at http://www.atd.ucar.edu/atd_data.html. A specific example is the data collection from the Fluxes Over Snow Surfaces (FLOSS) project, which is studying the surface meteorology of snow-covered rangeland in Colorado. This collection includes data from a wide variety of project measurement instruments. http://www.atd.ucar.edu/rtf/projects/FLOSS/. Many research databases in the earth sciences use well-established file format and structures that conform to the requirements of major data systems funded by NSF or other agencies, such NOAA or NASA.

An example of a biology research data collection is the Ares Lab yeast intron database. This site contains information and analyses about many specific segments of the genome of the yeast *Saccharomyces cerevisiae.* It was created and managed by a group that includes biologists and bioinformatics specialists. It is available at http://www.cse.ucsc.edu/research/compbio/yeast_introns.html

In economics, some research data collections result from laboratory experiments. An example is NSF-funded research at the University of Virginia and collaborating colleges that collects data via online game-like programs. The project website contains computer programs and a data base of experimental results that can be further analyzed. Examples of these can be found in links from http://www.people.virginia.edu/~cah2k/research.html. Many other empirical economics projects create new datasets based on the compilation and analsysis of economic, industrial, and behavior data. In many cases the project data collections are not available on the web, but may be available to other researchers from the author.

RESOURCE OR COMMUNITY DATA COLLECTIONS

DESCRIPTION

Resource or community data collections serve a specific science and engineering community. They are typically between research and reference data collections in size, scale, funding, community of users, and duration. They typically conform to community standards, where such standards exist. Often these digital collections can play key roles in bringing communities together to develop appropriate standards where a need exists. In many cases community database collections migrate to reference collections. In some fields, such as biology, resource data collections are often separate, directly funded projects. In other areas, such as the earth and environmental sciences, resource database collections are often managed under the umbrella of a data center that also supports research and reference databases.

EXAMPLES

Examples of resource data collections in the biological sciences include:

- The Arabidopsis Information Resource (TAIR) http://www.arabidopsis.org/ is managed by an organization that involves 20 developers (programmers and curators) and serves about 13,000 registered users and 5,000 laboratories. In early 2004, the collection contained around 3 gigabits of actual data and 16 gigabits for indexes for searching and analyzing data. The data is available to the public. Its continued availability depends on the duration of the project.

- PlasmoDB is a community data collection for the study of genomics of the malaria parasite *Plasmodium.* Researchers can view genomic data, obtain detailed information about individual genes, and access tools to facilitate analysis. http://www.plasmodb.org/bdbs.shtml.

- The Maize Genetics and Genomics Database (MaizeGDB) provides a similar set of databases and tools for maize research. MaizeGDB is funded by a cooperative

agreement through the USDA Agricultural Research Service. http://www.maizegdb.org/.

- The Canopy Database Project supports data acquisition, management, analysis and exchange relating to forest canopy studies at all stages of the research process. It develops informatics tools, document and publishes datasets that demonstrate use of these tools, characterizes fundamental structures of the forest canopy, and relates those structures to functional characterizations for retrospective, comparative, and integrative studies. http://canopy.evergreen.edu/home.asp

An example of a community database in the physical sciences is the LIGO Scientific Collaboration (LSC), which is a community resource for organizing technical and scientific research in the Laser Interferometer Gravitational Wave Observatory (LIGO). Around 500 scientists are involved in the collaboration. Access to the data is available only to members of the LSC, but the LSC is open to all scientists who apply and who propose an acceptable research plan – no groups have been rejected. LIGO data is characterized by very small signals buried in large amounts of instrument noise, and data is analyzed by internal teams consisting of instrument experts teamed with analysis experts. http://ligo.org.

In the earth and space sciences, many resource databases are housed within larger data centers that contain a combination of research, resource, and reference databases. For example the University Corporation for Atmospheric Research (UCAR), which is jointly funded by NSF and NOAA, operates the Joint Office of Science Support, which provides scientific, technical, and administrative support services to help the research community plan, organize, and implement research programs and associated field projects. Its CODIAC data management system offers scientists access to research and operational geophysical data. It maintains data archives and provides data support for current projects and field programs, including aircraft data, ground radars, and satellite photos. http://www.ofps.ucar.edu/codiac/.

NASA's Earth Science Enterprise (ESE) has ten discipline-specific data centers, known as Distributed Active Archive Centers (DAACs) that process, archive, document, and distribute data from NASA's Earth observing satellites and field measurement programs. Each data center has its own data-delivery methods and data-analysis tools. Most contain a combination of resource and reference data collections. Data can be accessed through http://nasadaacs.eos.nasa.gov/search.html. Examples of these distributed active archives include:

- The Alaska Satellite Facility (ASF) DAAC at the University of Alaska, Fairbanks, operates under contract to NASA to acquire, process, archive, and distribute satellite Synthetic Aperture Radar (SAR) data for the U.S. government and research communities. The ASF DAAC archives both restricted and unrestricted data. Restricted data is available only to registered and approved users while unrestricted data is available to the general public. http://www.asf.alaska.edu/.

- The DAAC at Goddard Space Flight Center manages data related to the upper atmosphere, atmospheric dynamics, global precipitation, global biosphere, ocean biology, ocean dynamics, solar irradiance. http://daac.gsfc.nasa.gov/www/.

Another resource data collection is the Ocean Drilling Program database managed at Texas A&M University. The Ocean Drilling Program is supported by NSF and 22 international partners. It contains data relating to decades of ocean drilling. http://www-odp.tamu.edu/database/

REFERENCE COLLECTIONS

DESCRIPTION

Reference collections are intended to serve large segments of the general scientific and education community. Conformance to robust and comprehensive standards is essential to provide the diverse user access and impact that are the mission of these collections. Adoption of standards by reference collections often 'sets the bar' for a large segment of the community, effectively creating a 'universal' standard. Budgets are often large, reflecting the scope of the collection and breadth of impact, and are typically provided by long term, direct support from one or more funding sources.

EXAMPLES

Examples of biological reference data collections include:

- The Protein Data Bank, which serves as the authoritative, international repository for macromolecular structure information. This collection was first created more than 30 years ago and its activities are currently supported by a coalition of eight US agencies. (http://www.pdb.org)

- Uniprot - the Universal Protein Resource, is the world's most comprehensive catalog of information on proteins. The UniProt Archive (UniParc) is a comprehensive repository, reflecting the history of all protein sequences. The UniProt Consortium is comprised of the U.K.-based European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and the U.S.-based Protein Information Resource (PIR). UniProt is supported, in part, by the National Institutes of Health and by the European Union. http://www.pir.uniprot.org/

Examples of space science reference data collections include:

- The SIMBAD astronomical database housed at the Centre de Données Astronomiques de Strasbourg in France. It provides basic data, cross-identifications and bibliography for astronomical objects outside the solar system. On October 1, 2004, Simbad contained over 3 million objects, 8.7 million identifiers, and nearly 15,000 bibliographical references. http://simbad.u-strasbg.fr/Simbad

- The National Space Science Data Center serves as the permanent archive for NASA space science mission data, and includes data on astronomy and astrophysics, solar and space plasma physics, and planetary and lunar science. NSSDC archives about 20 TB of digital data from about 420 mostly-NASA space science spacecraft, of which the most current 3 TB are electronically accessible. In addition to serving as the permanent archive, NSSDC also serves as NASA's primary active archive for space physics

mission data and for long-wavelength data (IR, etc.) from selected NASA astrophysics missions. It provides access to several geophysical models and to data from some non-NASA mission data. NSSDC also supports several public-interest web-based services that provide, for examples photo images of interest to the public. http://nssdc.gsfc.nasa.gov/

An example of a physical sciences reference data collection is the Physical Reference Data at the National Institutes of Standards and Technology. This collection contains high quality reference data on physical constants, atomic and molecular data, spectroscopy, and other areas. http://physics.nist.gov/PhysRefData/contents.html.

Examples of geoscience reference data collections include the reference datasets managed by the National Center for Atmospheric Research. These includes hundreds of atmospheric, oceanographic, and geophysical datasets. As noted previously, some of these are research or community datasets, but evolve to become reference datasets over time. These can be accessed through http://dss.ucar.edu/. A specific example of a reference dataset at NCAR is the Re-analysis project which was carried out jointly with the European Center for Medium Range Forecasting. This project used the latest atmospheric global models and previously collected data (decades back in time) to derive past atmospheric circulation patterns. These are essential data sets for understanding how the atmosphere is changing and how well the simulation models can re-create the "observed" atmosphere. These data are accessible at http://dss.ucar.edu/pub/reanalyses.html

Examples of social science reference data collections include:

- SEDAC, the Socioeconomic Data and Applications Center, which is one of the Distributed Active Archive Centers (DAACs) in the Earth Observing System Data and Information System (EOSDIS) of the U.S. National Aeronautics and Space Administration. SEDAC focuses on human interactions in the environment. Its mission is to develop and operate applications that support the integration of socioeconomic and Earth science data and to serve as an "Information Gateway" between the Earth and social sciences. http://sedac.ciesin.columbia.edu/data.html

- The reference datasets from the Panel Study of Income Dynamics (PSID) conducted at the Survey Research Center, Institute for Social Research, University of Michigan. PSID, begun in 1968, is a longitudinal study of a representative sample of U.S. individuals (men, women, and children) and the family units in which they reside. The sample size has grown from 4,800 families in 1968 to more than 7,000 families in 2001. At the end of 2003, PSID had collected information about more than 65,000 individuals spanning as much as 36 years of their lives. In the last five years, more than 290 journal articles and 70 Ph.D. dissertations were based on the PSID. PSID datasets include public release data files that have been processed and edited, and are available to all users. Other PSID datasets are still undergoing active processing and revision by the project team and others, and would be considered to be research or community datasets. http://psidonline.isr.umich.edu/