Chapter Two: 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 Four and the recommendations in Chapter Five, 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

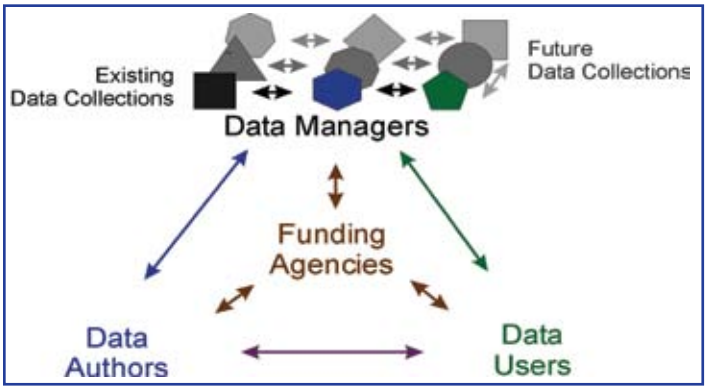
The actors in the digital data collections universe are both individuals and institutions. *Data users* include researchers, educators, administrators, students, and others who exploit information in data collections to pursue their research and education activities. *Data authors* are the individuals involved in research, education, or other activities that generate digital data that are subsequently deposited in a data collection. *Data managers* are the individuals and organizations responsible for database operation and maintenance. Note that 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 the figure. Arrows in the diagram represent the dynamic interactions and relationships among these functional entities and these are addressed in Chapter Three 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.



Structure of the Digital Data Collections Universe. Characteristics of the entities depicted in the figure are described in Chapter Two of the text. Relationships among these entities, represented by arrows in the diagram, are described in Chapter Three.

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 these 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, derivative data may be more readily usable by others. Thus, preservation of data in multiple forms may be warranted in many circumstances.

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 USGS National Center for Earth Resources Observation and Science (EROS), 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 (http://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.

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. Examples include D-Space, a joint initiative of the Massachusetts Institute of Technology and Hewlett-Packard (see http://dspace.org/), the CalTech Collection of Open Digital Archives (CODA; see http://library.caltech.edu/digital/), and the eScholarship program of the California Digital Library (see http://www.cdlib.org/programs/escholarship.html). These are examples of digital data commons – defined here as elements 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. As a result, a commons may also provide technologies and expertise to facilitate transitions between stages in the life cycle of a collection. 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.

[Blank Page]