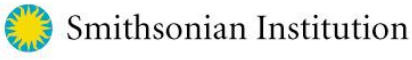
Sharing Smithsonian Digital Scientific Research Data from Biology

March 2011



Office of Policy and Analysis

Washington, DC 20013

Contents

Preface

Rapid progress in information technologies over the last few decades offers new opportunities to maximize the benefits of Smithsonian scientific research, not only by providing it with increased visibility, but also by facilitating broader use of the vast amounts of data that Smithsonian scientists have gathered over the years. In a growing number of cases, drawing together data gathered in different times, places, and scientific fields offers the best hope for deepening our insight into major concerns about the world's environment, climate, and biodiversity.

Recognizing this, the Office of the Under Secretary for Science recently asked the Office of Policy and Analysis (OP&A) to examine how and to what extent the Smithsonian makes its digital scientific data available to internal and external users, and how it might extend access to these data. This request was prescient—soon afterward, the Smithsonian adopted the 2010-2015 strategic plan, which makes broadening access to Smithsonian resources, including scientific data, a priority.

This report provides an overview of the issues, challenges, and opportunities that the Smithsonian and the wider scientific community face as they work to increase access to and use of the growing volume of digital data produced by the world's researchers. Based on the findings, the report presents conclusions and provides recommendations on how the Smithsonian can better share its wealth of such data. A great challenge in writing this report was the very rapid rate at which the fields of data management and sharing are advancing, and the need for recommendations to take into account this very fluid situation.

This broad, comprehensive report was driven by two highly capable senior analysts, James Smith and Whitney Watriss, who conducted most of the research and interviews for the study, analyzed the resulting information, wrote the bulk of main report, formulated the conclusions and recommendations, and worked many hours to edit, polish, and review the final draft. They were assisted by Ioana Munteanu, an OP&A staff member who undertook research on the use of social media to share data. All three were very ably assisted by Sarah Block, an OP&A researcher.

I also wish to acknowledge the assistance of four interns: Damaris Altomerianos and Grace Hart, who researched and wrote appendices; William Hix, who conducted initial interviews for the project when it was still in its incipient stages; and Daniel Garcia, who carefully read and reviewed a late draft.

Finally, I owe a debt to the reviewers who provided invaluable comments on drafts of the report or parts of it, including Riccardo Ferrante, Tom Garnett, Anson Hines, Leonard Hirsh, Lori Beth Magruder, Steve Paton, Pamela Smith, Thornton Staples, George Vandyke, Gunter Waibel, Anna Weitzman, and Donald Weller. I also wish to thank Eva Pell, the Under Secretary for Science, who patiently waited for OP&A to finish this project. Finally, Scott Miller, Deputy Under Secretary for Collections and Interdisciplinary Support, who originally encouraged us to pursue this project, deserves special mention for serving as a contributor, listener, and reviewer. I am grateful for his insights and advice.

Carole M.P. Neves, PhD
Director, Smithsonian Office
of Policy and Analysis

Acronyms

APA	Alliance for Permanent Access (EU)
ARKS	Animal Records Keeping System (ISIS software)
BHL	Biodiversity Heritage Library
CBOL	Consortium for the Barcode of Life
CIO	Chief Information Officer
CIS	Collections information system
CNI	Coalition for Networked Information
CODATA	Committee on Data for Science and Technology (ICSU)
CTFS	Center for Tropical Forest Science (Smithsonian Tropical Research Institute, Smithsonian Institution)
DAMS	Digital Asset Management System (Smithsonian Institution)
DataONE	Data Observation Network for Earth
DTIC	Defense Technical Information Center
EDAN	Enterprise Digital Asset Network (Smithsonian Institution)
EML	Ecological Metadata Language
EMu	Electronic Museum (CIS software)
EOL	Encyclopedia of Life
EPA	U.S. Environmental Protection Agency
ESA	Ecological Society of America
ESIP	Federation of Earth Science Information Partners
EU	European Union
GBIF	Global Biodiversity Information Facility
GCMD	Global Change Master Directory (National Aeronautics and Space Administration)
GEO/GEOSS	Group on Earth Observations / Global Earth Observations Systems of Systems
GIS	Geographic Information System
ICSU	International Council for Science
INOTAXA	Integrated Open Taxonomic Access project
ISIS	International Species Information System

IT	Information technology
ITIS	Integrated Taxonomic Information System
ITO	Information Technology Office (National Museum of Natural History, Smithsonian Institution)
IUBS	International Union for Biological Sciences
IWGDD	Interagency Working Group on Digital Data
JISC	Joint Information Systems Committee (United Kingdom)
LOC	Library of Congress
LTER	Long Term Ecological Research Network
NARA	National Archives and Records Administration
NASA	National Aeronautics and Space Administration
NBIC	National Ballast Water Information Clearinghouse (Smithsonian Environmental Research Center)
NBII	National Biological Information Infrastructure
NCEAS	National Center for Ecological Analysis and Synthesis (University of California at Santa Barbara)
NEMESIS	National Exotic Marine and Estuarine Species Information System (Smithsonian Environmental Research Center)
NEON	National Ecological Observation Network
NGO	Non-governmental organization
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NMNH	National Museum of Natural History (Smithsonian Institution)
NOAA	National Oceanic and Atmospheric Administration (U.S. Department of Commerce)
NSF	National Science Foundation
NSTC	National Science and Technology Council
NZP	National Zoological Park (Smithsonian Institution)
OAIS	Open Archival Information System
OCIO	Office of the Chief Information Officer (Smithsonian Institution)
OECD	Organization for Economic Cooperation and Development
OMB	Office of Management and Budget

OP&A	Office of Policy and Analysis (Smithsonian Institution)
OSTP	Office of Science and Technology Policy
OUSS	Office of the Under Secretary for Science (Smithsonian Institution)
PI	Principal investigator
SAO	Smithsonian Astrophysical Observatory (Smithsonian Institution)
SCBI	Smithsonian Conservation Biology Institute (National Zoological Park, Smithsonian Institution)
SD	Smithsonian Directive
SERC	Smithsonian Environmental Research Center (Smithsonian Institution)
SIA	Smithsonian Institution Archives
SIGEO	Smithsonian Institution Global Earth Observatory (Smithsonian Tropical Research Institute, Smithsonian Institution)
SIGIS	Smithsonian Institution Geographic Information System
SIL	Smithsonian Institution Libraries
SITP	Smithsonian Information Technology Plan
SPARKS	Single Population Analysis Records Keeping System (ISIS software)
SRO	Smithsonian Research Online
STRI	Smithsonian Tropical Research Institute
TDR	Trusted digital repository
TDWG	Biodiversity Information Standards
UK	United Kingdom
UN	United Nations
UNEP	United Nations Environment Program
UNESCO	United Nations Educational, Scientific, and Cultural Organization
USGEO	U.S. Group on Earth Observations (U.S. partner of international GEO/GEOSS initiative)
USGS	U.S. Geological Survey (U.S. Department of the Interior)
WDS	World Data System
ZIMS	Zoological Information Management System (ISIS software)

Executive Summary

Purpose and Methodology of the Study

At the request of the Office of the Under Secretary for Science (OUSS), the Smithsonian Office of Policy and Analysis (OP&A) conducted a study on how the Smithsonian can improve the sharing of its digital biology data, both internally and externally.[1] In keeping with the focus on biology, the study looked at data management and sharing at four Smithsonian research units: the National Museum of Natural History (NMNH); National Zoological Park (NZP) and its Smithsonian Conservation Biology Institute (SCBI); Smithsonian Environmental Research Center (SERC); and Smithsonian Tropical Research Institute (STRI). It also looked at several pan-Institutional units that support data management and sharing: the Office of the Chief Information Officer (OCIO), Smithsonian Institution Archives (SIA), and Smithsonian Institution Libraries (SIL).

In carrying out the study, the OP&A study team reviewed external literature and internal Smithsonian documents on digital scientific data sharing and management; interviewed 47 Smithsonian and 29 external experts; and attended two conferences held by leading organizations working on data sharing and management. After analyzing the information collected from these sources, the study team developed conclusions and recommendations. The final report benefitted greatly from the comments of reviewers.

This report reflects a snapshot in time. In the fast-moving field of digital data management, some details will have become outdated by the time this report is issued. The examples of data management and sharing initiatives included in the report are by no means exhaustive; they are simply a representative selection of those that came to the study team's attention.

Background

The 21st century has seen rapid growth of interest in using digital scientific data across research teams, disciplines and fields of research, organizations, and nations. This

[1] As used in this report, "data" refers to digital scientific data, including but not limited to observational research and collections data. "Data management" refers to processes undertaken to facilitate data discovery (finding data through online channels), access (retrieving them), and use (actually working with them for research purposes other than those for which they were originally collected), for as long as the data are deemed to be of value. "Biology" includes the various sub-disciplines of traditional biological science, as well as environmental science.

interest has been driven by a number of factors, most importantly: global environmental challenges whose analysis and mitigation require drawing on and integrating data from many different sources; the burgeoning volume of biology data being generated; the escalating costs of managing and sharing data, which exceed the resources of individual organizations and even individual nations; growing concern about the loss of so-called "legacy data" (data that are at risk of becoming unusable due to inadequate data management and inattention to long-term preservation); and the ethical imperatives of justifying public investments in biology research and sharing the fruits of such research across all nations.

In response to these forces, research organizations, governments, and multinational entities are looking systematically at processes, technologies, standards, and infrastructure for effectively managing and sharing data. Many of these efforts involve collaborations at the organizational, national, and international levels. However, such efforts have been fragmented and widespread. Easy access to a wide range of usable biology data remains an elusive ideal. Achieving this ideal will require expensive technological infrastructure; new administrative policies, structures, and workflows; specialized teams of personnel that combine domain research, information-technology (IT), and information-management skills; and, ultimately, coordination and collaboration across organizations and governments.

It will also require changes in the biology research culture, which for the most part adheres to a traditional small-science approach to data management in which individual research teams see their data as proprietary and pay little attention to the data management necessary to facilitate their use by others or their long-term preservation. Professional disincentives to (and a parallel lack of incentives for) data management and sharing reinforce this norm; these include the emphasis on publishing peer-reviewed articles and the absence of professional credit for producing, curating, and sharing valuable data sets per se. One result has been the creation of a great many disconnected data sets at a great many locations in myriad forms. These can be very difficult to discover, access, and use, and the already-enormous volume of such data is growing exponentially. Some are already past the point of use and preservation, and many more are at risk of permanent loss.

Conclusions

Conclusion: The small-science approach to the management and sharing of digital biology research data is anachronistic. It is at variance with the growing emphasis both in U.S. policy and around the world on open access to scientific data, and it may put the results of important research investments at risk and impede long-term access to valuable scientific resources.

To meet growing expectations for access to scientific data, the Smithsonian needs to carry out systematic and thorough management of its biology data throughout their lifecycle. While the study team encountered a number of very noteworthy and important initiatives at the Smithsonian to further systematic data management and sharing, it also found that an Institutional strategy to guide progress in these areas is lacking. As a result, discovery, access, and use of Smithsonian biology data are constrained, and an ever-increasing volume of legacy data is at risk.

Going forward, the small-science approach to data management that has been the norm for Smithsonian biological science will have to be modified. While science that involves small research teams will always have a place in the scientific enterprise, the loose data-management norms associated with small science will have to give way to more standardized, systematic methods.

Conclusion: To make its digital biology data easily discoverable, accessible, and usable by internal and external users, the Smithsonian needs to unequivocally articulate a policy of open access and systematically establish the capacity and tools to implement that policy.

The study team believes that the Smithsonian, as a largely taxpayer-funded entity, has an obligation to provide open access to its biology data, subject to reasonable proprietary waiting periods and other justifiable exceptions. Both the 2010-15 Smithsonian Strategic Plan and the Digitization Strategic Plan specifically call for increased sharing of digital data, although neither uses the term "open access." A starting point would be the promulgation of a policy of open access to Smithsonian data. That step must be reinforced by making open access a fundamental operating principle of the Smithsonian's research enterprise and establishing external usability as a primary consideration in decisions regarding data-management processes, standards, infrastructure, and technology.

Conclusion: Sharing of Smithsonian biology data requires fundamental changes in current data-management and -dissemination practices.

Data sharing requires that the Smithsonian undertake proactive efforts to manage its biology data in a manner that makes them easily discoverable, accessible, and usable. The current small-science, seat-of-the-pants, fragmented, and mostly unit- or department-based approach will have to give way to a set of core Institution-wide principles and standards. These will have to be framed with careful attention to the distinctive needs of different sub-disciplines and research areas, and will have to allow some flexibility in application so as to accommodate particularly innovative or unique research.

In moving forward, it is important that Smithsonian IT staff, information-science personnel, and domain scientists work closely as a team to develop a supportive environment for researchers that offers them a variety of appropriate, continuously-upgraded tools, systems, and services to facilitate their role in data management and sharing. Internal Smithsonian efforts will need to draw on and coordinate with relevant experts and organizations in the external environment.

The following are basic elements critical to increased data sharing, as well as effective long-term preservation where the data merit such treatment. These could usefully be identified in a Smithsonian-wide policy:

- Data-management and -sharing standards for the entire lifecycle of various kinds of digital data.

- Criteria for deciding the appropriate level of management and preservation for specific data, established in consultation with the internal Smithsonian research community and external organizations.

- Compliance of Smithsonian researchers with relevant internal and external data-management standards and requirements.

- A central record (ideally including the information necessary for discovery, access, and use) of the Smithsonian's digital biology data holdings, and a system for regularly updating the information.

- Smithsonian tools that facilitate discovery and access—including a single point of entry to information on Smithsonian data, ideally with links to these data. The Smithsonian can also provide discovery and access through tools maintained by external organizations, including the U.S. federal government's data.gov and Science.gov portals.

- A trusted digital repository (TDR) for long-term storage, stewardship, and access to Smithsonian (and possibly other research organizations') scientific data. A Smithsonian TDR could serve as a repository for the work of the Institution's own researchers, as well as a central national or global repository for particular types of data. Decisions about the appropriate role for a Smithsonian TDR are best undertaken with extensive input from internal researchers, external research organizations, and other external organizations working on long-term data preservation issues.

Conclusion: Systematically and immediately addressing the risk of legacy data loss and preventing further growth in the backlog of Smithsonian legacy data are high priorities.

The starting point for protecting existing legacy data is to place them in secure storage in their present form until they can be managed properly.

Two steps to minimize the further growth of legacy data are to

- Ensure that the data of researchers soon to retire, and projects soon to end (or recently ended), receive near-term basic data management and are transferred to secure storage; and

- Require that scientists with ongoing projects list their data in the central record of Smithsonian holdings and routinely back them up on a stable medium.

Conclusion: It is important that the central administration be proactive in reaching out to Smithsonian biology researchers to raise awareness of the value of managing and sharing digital data, alert them to the support available to facilitate those tasks, and obtain buy-in among research staff for enhanced data management.

Researchers currently have virtually no incentives, and many disincentives, for engaging in data management beyond the minimum required for their own analytic purposes. The Smithsonian needs a program to get buy-in from its scientists for better data management. Of particular importance to such an effort are:

- Inclusion of professional credit for effective data management in performance evaluations;

- Providing researchers with access to support personnel and tools that facilitate data management and minimize the time researchers need to devote to it; and

- Access to funds earmarked for data management and additional data-management support staff.

Conclusion: Meeting the growing challenges of digital data management and sharing at the Smithsonian will require additional resources.

The growing demands for data management and sharing will require not just increased levels of funding for assets such as infrastructure and specialized staff, but also better use of existing Institutional resources. Leveraging resources through partnerships and participation in collaborative initiatives is one important avenue to follow; Smithsonian data-management and -sharing efforts are still too often undertaken in relative isolation from external organizations. The Institution might also pursue a budget line item for digital data biology management and sharing, given that such data are core national assets that can be critical to the formulation and implementation of federal science policy in numerous areas. The Smithsonian will need to pursue a combination of these and other strategies, such as more efficient use of internal resources; shifting funds from lower-priority functions to data management and sharing; increasing grant overhead rates and allocating part of them to data management; and potentially offering fee-based services (for example, for preserving data from other organizations in the Smithsonian TDR).

Additional data-management support staff are an especially critical resource need. Effective data management and sharing require a cadre of research support personnel who combine IT, information-science, and domain-science expertise. The allocation of such personnel between research and central support units remains an open question. In some cases, specialized expertise clearly needs to reside at the unit level; in others, a central corps of specialists available on an as-needed basis may be a more efficient arrangement.

Conclusion: Significant collaboration with external organizations will need to be part of the Smithsonian's approach to managing and sharing its data.

The OP&A study team did not think that the Smithsonian has systematically explored and taken advantage of opportunities to engage with external organizations working on similar data-management and -sharing challenges. As a result, it is missing out on opportunities to leverage resources, share expertise,

form collaborations, stay on top of rapidly-changing developments, and learn valuable lessons from the successes and failures of other organizations.

Greater participation by Smithsonian researchers and data-management personnel in external forums will likely require formal inclusion of expectations for such participation in staff position descriptions, credit for it in performance evaluations, and increased resources for staff travel. The Smithsonian might also identify specific areas where it is especially suited to assume a leadership role. One such area might be usability of data across disciplines, given how many are represented within the Institution. Another might be data publishing, which requires a system for peer-reviewing data sets, rules for citations, and a means of tracking data re-use. The Smithsonian has the stature to demonstrate how data publishing might work and be a catalyst for broad acceptance.

Conclusion: The Smithsonian will need to put in place an organizational structure, with clear roles and responsibilities at the levels of the central administration and research units, to ensure coordinated implementation of sound data-management and -sharing standards, systems, and practices.

The absence of an overarching Institutional strategy and framework for data management and sharing has contributed to fragmented and often opportunistic efforts in this area. A critical organizational question that needs to be answered concerns the appropriate roles that the biology units, OCIO, SIA, SIL, and other central support offices should have in data management and sharing. A single Institutional focal point is needed to mediate among and manage the needs of the various research units and the central administration. OUSS seems well-positioned to take on this role, but will need additional resources to be effective.

Recommendations

The OP&A study team has three overarching recommendations:

1. **The Smithsonian should unequivocally commit to a policy of open access to its digital biology data,[2] subject to reasonable restrictions, including an initial embargo period to allow researchers to publish. Data sharing and the systematic underlying data management needed to support it should be fully integrated into the Institution's biology research enterprise, and external**

[2] The OP&A study team limits its recommendations to digital biology data, which were the topic of the study. It presumes, however, that any policy directive would likely encompass digital data from all science (and possibly social science) research conducted at the Smithsonian.

usability should be a primary consideration in decisions regarding data-management processes, standards, infrastructure, and technology.

2. **The Smithsonian should establish the capacity and tools to make its digital biology data easily discoverable, accessible, and usable by present and future users, internal and external.**

3. **The Smithsonian should engage more fully and systematically with external organizations working to advance data sharing and management, taking on a leadership role in areas where it has particular expertise and resources or where it is in the Institution's strategic interests.**

To accomplish these three core recommendations:

4. **OUSS should convene a working group to (1) develop a plan of action for managing and sharing digital biology data and (2) draft a policy to govern biology data management and sharing.**

    - The **working group** should include representatives from the Smithsonian biology research units, OCIO, SIL, SIA, the Office of Human Resources, central management, and other relevant personnel.

    - The working group should draft a **plan of action** that addresses the issues enumerated below, with the option of splitting the work into two parts so that priority issues such as the potential loss of legacy data can be addressed in the near term.

    - Based on the plan of action, the working group should draft a **policy for digital biology data management and sharing**. This policy could be integrated into existing Smithsonian Directives, included in the proposed SDs 609 and 610, or issued as a stand-alone Directive.

    - The plan of action referenced above should address the following:

        - *Definition of open access*—including when data would be made available, guidelines for restricting access, and processes for decision making on access.

        - *Provision of a status for Smithsonian digital biology data comparable to that of the National Collections covered in SD 600*.

- *Core requirements for digital biology data management over their lifecycle to facilitate discovery, access, and use.* Such requirements might include common Smithsonian-wide data management standards, specifications for metadata (defined in the context of particular fields or types of research), and acceptable formats.

- *Infrastructure needed to support digital biology data sharing*—including a data portal and a TDR for long-term preservation.

- *Near-term secure storage of the Smithsonian's digital biology legacy data*—including storage as-is in a secure repository until the data can be assessed for future value and subsequently curated or disposed of as appropriate.

- *Development and maintenance of a record of the Smithsonian's digital biology data holdings*—including adequate metadata to support discovery, access, and use.

- *Measures to minimize the continued growth of legacy data*—including a requirement that researchers prepare a data management plan as part of project design; list the data they are collecting in a central record and update their status over the lifetime of the project; and ensure that their data meet at least the basic criteria for discovery, access, and use before being transferred to a secure central location.

- *Criteria for determining the appropriate level of data management for specific data sets.*

- *A central biodiversity informatics/data-management support capability for the Smithsonian biology research community.*

- *Professional incentives for biology researchers to engage in data management and sharing, and provision of tools and support services to assist them in these efforts:*

    * Include in staff position descriptions expectations for data management and sharing and for participation in forums on data management, and provide formal credit in performance evaluations

for data publishing and citations, and for participation in external and internal data-management and -sharing initiatives.
    * Require that Smithsonian scientists, prior to leaving the Institution, consult with a responsible party to determine what should happen to their data, and engage with support personnel to carry out whatever data management is necessary to prepare these data for transfer for long-term preservation if appropriate.
    * Provide services and tools to minimize the time researchers need to spend on data management and sharing.
    * Raise researchers' awareness of the importance of data management, long-term preservation, and sharing, and of the personal and societal benefits.

- *Systems and tools for easy discovery, access, and use of Smithsonian data by internal and external users*.

- *Workforce requirements and deployments at the central and unit levels*—including consideration of the appropriate numbers and types of support personnel for major data-management and -sharing tasks, and the appropriate balance between researchers and research support staff.

- *Increased Smithsonian participation in relevant external initiatives and forums*—including taking on a leadership role in appropriate areas of Smithsonian strength and strategic interests.

- *Strategies to increase or leverage funding for data management and sharing:*

    * Pursuing federal allocations specifically for data management and sharing, including a line item in the Smithsonian's federal budget, particularly in the context of assuming responsibility for appropriate parts of federal cyberinfrastructure for scientific data.
    * Increasing and reallocating overhead allowance rates on scientific grants to reflect growing requirements and costs for data management and sharing.
    * Leveraging resources through partnerships and participation in collaborative initiatives.
    * Making more efficient use of internal resources, and shifting funds from lower-priority functions to data management and sharing.

    * Providing services for fees.
    * Coordinating with the National Science Foundation (NSF), Library of Congress (LOC), National Archives and Records Administration (NARA), and other major national scientific and library/archival organizations to raise awareness in Congress, the Office of Management and Budget (OMB), and the public about the wider societal benefits of data sharing (particularly with respect to addressing global environmental challenges), and the importance of federal investment to defray the costs of a national data-sharing infrastructure.
    * Participating actively in forums that discuss federal investment in data management and sharing.

- *Design of an organizational structure to support data management and sharing at all levels:*

    * Definition of roles and responsibilities for Smithsonian central support offices (particularly OUSS, OCIO, SIL, and SIA) and research units—including which unit(s) have primarily responsibility for implementing specific parts of the plan of action.
    * Development of communication and coordination mechanisms to leverage relevant resources across units, ensure smooth internal collaboration, and disseminate lessons learned across the Institution.

- *Identification of the highest near-term priorities:*

    * Prevention of further loss of unmanaged legacy data.
    * Identification of national and global initiatives in which the Smithsonian should participate.
    * Explicit inclusion of participation in such initiatives in staff job descriptions, and/or professional credit for participation.
    * Provision of funds for travel and other support for such participation.
    * Development of a system for keeping the Smithsonian abreast of relevant external developments on an ongoing basis, and identifying promising opportunities for leveraging resources through new collaborations.
    * Engagement with data.gov and Science.gov.

5. **The Smithsonian should issue a digital biology data-management and -sharing policy, based on the draft policy of the working group.**

6. **The Smithsonian should begin implementation of the near-term priorities identified in the plan of action as soon as possible following receipt of the working group's recommendations.**

Introduction

Purpose and Scope of the Study

The Smithsonian Office of the Under Secretary for Science (OUSS) asked the Smithsonian Office of Policy and Analysis (OP&A) to conduct a study of how the Institution could improve access to its scientific research data and results (for example, publications, research tools, and software). As the OP&A study team began the work, it became clear that access to the results of Smithsonian scientific research was not in most cases the primary issue—most Smithsonian scientists, for example, regularly publish peer-reviewed articles and, to a lesser extent, engage in efforts to inform the general public and decision makers about the substance of their work. Rather, the larger issue was that the *digital data* underlying the scientific results were not adequately accessible to or usable by the wider research community and other professionals such as policy and decision makers. It was decided that the study would focus on access to these data. Thus, unless otherwise stated, when this report uses the term "data," it refers to *digital scientific research data.*

It also became evident that the study should focus on research in the sciences associated with the Grand Challenge of "Understanding and Sustaining a Biodiverse Planet" in the 2010-2015 Smithsonian Strategic Plan. While for convenience this report calls these sciences **biology**, the term refers not only to biology[1] proper (for example, systematic biology, ecology, marine biology, botany, and reproductive science), but also to environmental science.[2] A primary reason for this focus is that, in general, data-sharing and data-management practices in biology are less advanced than in areas of science associated with the Grand Challenge of "Unlocking the Mysteries of the Universe," such as astrophysics, planetary science, and astronomy.[3] That said, interviewees pointed out

[1] Harley, et al. (2020) offer the following definition:

> *Biology, broadly defined, is the scientific study of life and living organisms. … The field can be clustered into two primary academic divisions: the "bench" sciences, which encompass molecular and cell biology (MCB, e.g., genomics, neurobiology, microbiology, developmental biology, biochemistry, immunology, and biotechnology), and the "field" sciences, comprising organismal biology (OB, e.g., marine biology, ecology, zoology, and evolutionary biology).*

[2] Environmental science is usually understood as an interdisciplinary union of the physical, chemical, and biological sciences, employed to study interactions between habitats and living organisms.

[3] While interviewees from the Smithsonian Astrophysical Observatory (SAO) indicated that the treatment of data at that unit is variable, there was general agreement that most research areas there are farther along the road to consistent data management than most biology research areas at the Smithsonian.

that several non-biology disciplines at the Smithsonian (such as mineral sciences,[4] anthropology, and paleobiology) suffer from some of the same data-management and -sharing problems that biology does. Thus, many of the issues discussed in this report are relevant to these fields as well.

As discussed in more detail below, much research in biology, both at the Smithsonian and in the wider world, continues to be **small science**, in contrast to the **big science** that characterizes fields such as astronomy, particle physics, and seismology. The difference between "small" and "big" science has been defined in several ways, but the distinction most relevant to this report relates to digital data-management norms. In this report, "small science" refers to fields in which digital data-management practices tend to be driven by the needs of individual, often small-scale, research projects, rather than by standards common to a whole field. Big science fields, by contrast, have standards of data management that are widely accepted and used by researchers.[5] While some fields of biology already function more like big science, and while the overall picture is changing in response to the forces discussed in this report, interviewees and written sources generally supported this broad-brush characterization of biology as small science, at least relative to the physical sciences.

The term **field** itself requires clarification. In many areas of biology, it is difficult to draw clear lines among labels such as "discipline," "research program," "sub-discipline," "community of practice," and so on. This report uses "field" as a term of convenience for any level of scientific practice—sub-disciplinary, disciplinary, interdisciplinary—in which there is a sense of shared professional identity backed by professional communication and exchange.

Given the boundaries of the study, the Smithsonian science units of primary interest were the following:

- National Museum of Natural History (NMNH);

[4] A notable exception is the Global Volcanism Program of the Department of Mineral Sciences at the National Museum of Natural History. According to the Program's website (http://www.volcano.si.edu/info/about/about_gvp.cfm), its electronically accessible databases are a "foundation for all statistical statements concerning locations, frequencies, and magnitudes of Earth's volcanic eruptions during the last 10,000 years." In the early stages of an eruption anywhere in the world, the Program serves as an international clearinghouse for reports, data, and imagery, drawing on information from contributors who make up the Global Volcanism Network.

[5] Development of standards in big-science fields has historically been driven by the need for researchers in these fields to share a limited supply of very expensive, complex equipment such as high-performance earth-and space-based telescopes, particle accelerators, and globally-coordinated networks of sensors.

- National Zoological Park (NZP) and its Smithsonian Conservation Biology Institute (SCBI);

- Smithsonian Environmental Research Center (SERC); and

- Smithsonian Tropical Research Institute (STRI).[6]

Two points about this study bear emphasizing at the outset. First, the study team's charge was to identify ways to advance the sharing of Smithsonian scientific data, not to pass judgment on the state of data sharing at the Institution, whether absolutely or relative to that at other organizations. This required identifying both successes on which future efforts can build, and barriers that inhibit effective data sharing. The successes discussed in this report are those that came to the study team's attention, and do not constitute an exhaustive list of effective initiatives underway at the Smithsonian; thus, a failure to mention a particular effort is not intended as a negative judgment. This study also describes some important initiatives being carried out in other organizations, including several federal agencies. These are intended to point out the array of external knowledge, resources, and models from which the Smithsonian can benefit, and should not be interpreted as implied standards of practice. It is not possible to directly compare current efforts at the Smithsonian with, for example, those of federal agencies that are far larger and better funded, many of which have already adopted policies that explicitly promote and support systematic data management and sharing.

Second, the members of the OP&A study team are not scientists, information technology (IT) experts, or information-science professionals. Their exposure to this subject has for

[6] The Smithsonian Horticultural Services Division (HSD) maintains data on its orchid collections that could conceivably fall under the rubric of this study. However, because HSD does not conduct biology research (although some of its staff do collaborate with scientists at other units), its data-sharing and -management issues are not addressed here. Three Smithsonian non-biology science research units were similarly excluded:

- SAO, which focuses on astronomy and astrophysics, with the goal of understanding the basic physical processes that determine the nature and evolution of the universe;
- The Museum Conservation Institute (MCI), which conducts "in-depth studies of artistic, anthropological and historic objects using state-of-the-art analytical techniques to understand their provenance, composition and cultural context, and to improve our conservation techniques" (http://www.smithsonian.org/ResearchCenters/Museum-Conservation-Institute); and
- The Center for Earth and Planetary Studies (CEPS) at the National Air and Space Museum (NASM), which "performs original research and outreach activities on topics covering planetary science, terrestrial geophysics, and the remote sensing of environmental change" (http://www.nasm.si.edu/ceps/).

The study team cannot comment on the state of digital data sharing and management at those units.

the most part been confined to the interviews, literature review, and conferences that were involved in this study, although they were able to call as needed on relevant specialists at the Smithsonian and beyond to help them navigate the complicated world of scientific data sharing. Although the disadvantages of taking on a complex subject without baseline experience are obvious, one important advantage—especially given the lack of agreement among sources on a number of points of interest—is that the study team did not approach the project with any professional preconceptions, interests, biases, or axes to grind.

Methodology

The OP&A study team began by reviewing the literature on the sharing of scientific digital data in the United States and around the world, as well as reviewing relevant internal Smithsonian documents (see Appendix A, Bibliography). Topics of interest included policies related to data sharing; the state of data sharing in general and the factors that have influenced it; key players working to promote data sharing; and the directions in which data sharing in biology are moving and why.

OP&A next interviewed 29 experts at external organizations and 47 Smithsonian staff involved with policy and technical issues pertaining to data sharing and management (see Appendix B, Organizations Interviewed for the Study). Questions were intended to probe interviewees' insights both on the key issues of data sharing and data management, and on how the Smithsonian might expand access to its data.

The study team also attended two conferences sponsored by organizations that are among the foremost players in advancing scientific data sharing: (1) the Group on Earth Observations and its associated Global Earth Observation System of Systems (GEO/GEOSS), November 2009; and (2) the Federation of Earth Science Information Partners (ESIP), January 2010.

Following the data collection, the study team analyzed the data it had collected and developed conclusions and recommendations on how the Smithsonian could open up its digital scientific research data to more users. A draft report was sent to several reviewers to correct for errors in fact and get feedback on the conclusions and recommendations. The report was revised after reviewing the comments.

This report provides an overview of the issues, challenges, and opportunities facing the Smithsonian as it moves forward with digital scientific data sharing and management. It should be noted that most of the details and specific examples discussed in this report are

snapshots from a particular point in time. Some details will have become outdated by the time this report was finalized.

Terms and Definitions

There is considerable inconsistency and imprecision in external and internal usage of key terms and concepts related to data management and sharing. Thus, it is important at the outset to explain how key terms are to be understood *in the context of this report.* Less-central terms are defined in footnotes or parenthetical text when first used.

The most fundamental term used in this report, and the one whose meaning varies substantially in external usage, is **data**. The primary focus of this study is, as noted, on digital scientific research data—observational information collected for the purpose of answering scientific research questions.[7] Digital scientific research data can be divided into two broad categories, both of which are relevant to data sharing:[8]

- Raw (or primary) research data: individual observations, measurements, images, and so on—either as originally collected or after being checked, cleaned, corrected (e.g., for instrumentation error), and organized;[9] and

- Derivative (or derived) research data: refined data derived from raw research data through some form of manipulation, transformation, or abstraction (e.g., statistical analysis or modeling; and enhancement of images, videos, and audio files).

Information (images and documentation) contained in natural history collections databases (of specimens, tissue samples, fossils, and so on) is explicitly included in the definition of "data" used in this report, whether this information was collected for specific research purposes or as part of a general digitization effort. In addition, the following types of information are treated as data here:

[7] "Observational" information is used broadly in this context—not just as information literally measured or observed by researchers, but as information collected remotely via technology. Thus, it includes earth observation satellite data, automated sensor data, and so on.

[8] Raw data, for example, may be of interest to researchers attempting to verify or replicate a published result. Derived data may be of more interest to researchers attempting to synthesize a range of previous research conducted by multiple research teams at disparate locations and times.

[9] In some cases (for example, satellite imagery), data originally exist as bitstreams that are not themselves amenable to scientific analysis without processing, typically by engineers or IT personnel rather than by scientists per se. Where a distinction needs to be drawn between such bitstreams and other forms of data, this will be made clear in the text.

- Textual, video, audio, and photographic documentation of field and lab work (as distinguished from textual, video, audio, and photographic assets that serve as raw research data);
- Records of zookeepers and veterinary staff; and
- Unpublished reports and analyses.

For the most part, the focus in this report is squarely on the management and sharing of *digital* data. Indeed, many of the major issues addressed in this report would not exist in the absence of recent advances in digital technologies. However, for convenience, we sometimes refer to non-digital scientific information as "data"—for example, "legacy data" in pen-and-paper format. When non-digital information is at issue, this will be clear in context.

A **data set** is a collection of related research data. **Metadata** are descriptive documentation associated with research data sets, or "data about data." (Descriptions of the methodologies used in collecting the data are, for example, an important type of metadata.) The dividing line between data and metadata can be unclear, but for the most part this did not emerge as a concern in this study. It only needs to be noted that when data management and data sharing are discussed, the management and sharing of the associated metadata are covered as well.

For the most part, as used here the term "data" does *not* address formal publications, conference presentations, and similar expositions of research analyses and results.[10] (As noted, the professional communications and dissemination channels for the *results* of scientific analysis are well-established—although admittedly in flux, because of the impact of the web on traditional models of scientific publishing and dissemination—and Smithsonian scientists are for the most part fully integrated into these channels.) An exception is the category of "legacy publications" or "legacy literature," which refers to hard-copy published literature produced in the pre-digital era. It is not possible to draw a

[10] Although the study team excluded these dissemination channels, it is important to note that both the literature and interviewees frequently commented on the need for publications to do a better job linking back to the underlying data. Helly (2003), for example, noted that

> *Interpretations and figures [in publications] based on data are widely published and archived in libraries, while most of the primary data are confined to research files of investigators. These private archives, however, do not provide sufficient access for future research that might result in a reinterpretation of the data.*

Helly recommended publishing the underlying data in a "science data network" simultaneously with the publication of the paper.

precise historical line to separate legacy literature from modern scientific literature. However, the former can be thought of, in functional terms, as the yellowing journals in the collections of libraries and natural history museums that, absent active digitization initiatives, will remain unavailable in digital form.

The fundamental subject of this report is **data sharing,** which refers to making data available to people beyond the original collectors through *digital* dissemination channels. Successful data sharing requires effective **data management,** a term used broadly in this study as shorthand for the whole set of policies, practices, and procedures that govern the processing, documentation, dissemination, and storage of data throughout its lifecycle.

Three basic requirements for successful data sharing are that the data be **discoverable, accessible,** and **usable:**

- "Discoverable" means data can be located and initially assessed for relevance by potential users, typically through an online search. A critical requirement for discoverability is that the data are adequately documented with summary metadata available through online search channels.

- "Accessible" means data can be retrieved by potential users. A fundamental requirement for accessibility is that those who control the data are willing and able to release them. Accessibility also requires channels for conveying the data to potential users, typically through the internet.[11]

- "Usable" means that data both work in a *technical* sense, and are intelligible in a *scientific* sense. Here, the technical side of usability is defined as **interoperability**. In some contexts, "interoperability" refers exclusively to the ability of IT formats and components to work together.[12] However, the term can also be applied to the data themselves, and in the biodiversity community it is probably more common to see it applied to data than to technology. In this sense, "interoperability" refers to the ability to bring together data gathered

[11] Note that data can be discoverable without necessarily being accessible. For example, a search may identify the existence of a given data set, but fail to provide a direct link to it or even information on who to contact for access.

[12] Data are interoperable in this purely technological sense if people other than the originator can read and manipulate them with the standard technologies and software of the community of potential users—or, if such standards do not exist, with technologies and software in general use.

independently to form larger data sets.[13] The scientific side of usability refers to the ability of specialists familiar with data and metadata practices in the relevant field to correctly interpret the data.[14] **Data processing** refers to the steps taken to make data usable.

Data stewardship refers to the *long-term* (lifecycle) management of data for as long as they are judged to have value. **Data preservation** refers to processes of data stewardship (such as migration across technology platforms, bit checks, version control documentation, and so on) that are undertaken to ensure the survival, integrity, accessibility, usability, and reliability of data over their lifecycles.

In this report, **data curation** is used broadly to encompass the activities that scientists, research organizations, data centers, and others undertake to make data discoverable, accessible, and usable and to preserve these characteristics over time.[15] Thus, curation can include, but is not limited to, activities such as:

- Applying standard ontologies, nomenclatures, and metadata to data;
- Storing data on externally accessible technology platforms, with the appropriate formats, hardware, software, and middleware;
- Providing user tools that allow researchers to retrieve, visualize, and manipulate data more easily;
- Migrating data to new platforms as technologies evolve;

[13] A number of interviewees referenced the difficulties of combining digital data sets from different researchers even within the same discipline, let alone across different disciplines. Problems with interoperability in this sense are becoming more of an issue as interest grows in bringing together scattered data to discover inherent properties and patterns, and to facilitate research in new fields. One significant obstacle to interoperability not addressed in this report but cited by interviewees (one of whom considered it to be the single biggest obstacle to data usability) is the wide variation in data collection protocols—such as units of measures, timeframes, geographic scope, and resolution.

[14] Correct interpretation depends on knowing how the data were collected and recorded (for example, what units of measure were used and whether the data fields are labeled using standard nomenclatures), and how the data were documented with metadata (for example, the time, place, and context of data collection, and the equipment, calibrations, and protocols used).

[15] This definition is consistent with other examples of external usage. For example, the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (2010) argues that data curation starts at the moment of data collection (p. 53). However, some authors define data curation as a subset of activities under data stewardship, and thus more narrowly concerned with the preservation of data judged to have long-term value (Berman and Moore 2006).

- Undertaking quality-assurance and quality-control processes, both initially and over time (including validation of provenance and authenticity; documenting changes over time to provide version control; and checking for bit decay[16] and fidelity of migrated data);

- Maintaining links to published materials and annotations;

- Instituting and executing processes to ensure the reliable, long-term preservation of and access to data deemed of sufficient value.

Three general professional communities are referred to in this report—(**domain) science**, **information technology (IT)**, and **information science**:

- (Domain) science refers to the community of natural scientists. Historically, domain scientists have been concerned primarily with the collection and analysis of data to address specific research questions, and with presenting their findings through publications and other means of dissemination.

- The IT community is comprised of experts in the creation and administration of technologies[17] and technology systems[18] for storing, processing, preserving, and communicating digital data.[19] These systems encompass:

    - Hardware;

    - Technology tools (user tools and interfaces);

    - Protocols (software, middleware[20], programs, data processing algorithms); and

    - Systems architectures.

    Often, IT professionals, whether practical or theoretical in orientation, are not concerned with the specifics of the data stored, processed, or communicated on

[16] "Bit decay" refers to the loss of data associated with the degradation of storage media over time; it is also known as "bit rot."
[17] For example, automated data collection equipment, data processing hardware, data storage technologies, and communications networks.
[18] For example, operating systems, software and middleware, protocols, and systems architectures.
[19] The term as used here explicitly excludes professionals whose expertise pertains to the creation and administration of data-collection technologies (for example, space- or earth-based environmental sensors).
[20] "Middleware" is the generic term for computer software that allows individual software components or applications running on one or more machines to interact with one another. That is, middleware mediates between different software programs and operating systems.

the systems for which they are responsible, although some inevitably get immersed in the nuances of the discipline for which they create systems. They may, however, be called upon to customize parts of a system to meet the data needs of a particular field.

- The information-science community refers to a category of multidisciplinary professionals that includes, but is not limited to, library professionals, archivists, and some types of computer scientists. More generally, it includes interdisciplinary professionals with expertise in the principles and practices of information-repository and information-system design and administration, including methods for data discovery, access, preservation, and visualization. Information-science personnel active in scientific data management may or may not have some familiarity with the underlying science of the data with which they work, depending upon the part of the data lifecycle in which they are active.[21]

The line between the IT and information-science communities is sometimes unclear. For example, some types of computer scientists straddle both communities, and organizationally may be placed in either library/archive or IT units. Generally speaking and in keeping with the definitions given above, this report uses the term "IT" primarily to refer to the more technological side of information systems and "information science" to the more conceptual side. However, some ambiguity is unavoidable.

In external usage, scientific **cyberinfrastructure** denotes the distributed national or supra-national complex of technologies and technology systems that supports the scientific research enterprise. In this report, the focus is on the elements of cyberinfrastructure specifically relevant to data management and data sharing. Terms like "organizational cyberinfrastructure" or "organization-level cyberinfrastructure" are used to refer to the technological assets of individual organizations that tie into, and collectively comprise, a nation's cyberinfrastructure. According to Gold (2007a), the term **e-science** has the same meaning as "scientific cyberinfrastructure," but is more widely used outside of the United States, particularly in the United Kingdom and Europe.

Organization of the Report

The findings section of this report is divided into three major parts.

[21] One interviewee particularly emphasized that it is imperative that information-science (and IT) personnel concerned with data *storage* understand how the data are used, so as to meet researchers' needs.

- The first, "A Changing Digital Data-sharing Environment," looks at the approach to data management and sharing that has historically predominated in biology, explores the forces undermining this traditional approach, and discusses how organizations, collaborative efforts, and professions are responding to the changed environment.

- The second, "Functional Areas," looks at four major aspects of the data-sharing process—discovery, access, use, and preservation—and the issues that surround them. It also explores several important factors that influence all four aspects: economics, cyberinfrastructure, and human resources.

- The third, "The Smithsonian," looks specifically at how the Smithsonian is responding in biology research to the new imperatives of data sharing and data management at the levels of individual units, cross-unit initiatives, and involvement with external partners and collaborative efforts.

The "Conclusions" section draws out some of the major points suggested by the findings for both data sharing in general and the Smithsonian specifically.

It is followed by "Recommendations," the study team's suggestions for how the Smithsonian might increase easy access to and use of its digital scientific research data.

At the end of the report, four appendices provide, respectively, a list of the literature reviewed for the study (Appendix A); a list of the organizations contacted for the study (Appendix B); descriptions of a number of inter-organizational collaborative efforts that address aspects of the data-sharing challenge (Appendix C); and a glimpse of the international dimension of the issues covered here, with a focus on efforts undertaken in a few nations and by international organizations such as the United Nations (Appendix D). An addendum provides the results of research the OP&A study team conducted into the extent to which social media play a role in digital data sharing in biology.

Findings: A Changing Digital Data-sharing Environment

Small Science and Data Management

The following description of scientific data management applies primarily to the principal investigator (PI)-based, curiosity-driven research done at universities and in similar settings—the sort of research that usually falls under the heading of "small science" discussed above. The description is broadly drawn, and there have always been exceptions to the generalizations. That said, the description captures some of the salient points that emerged from the literature and interviews about historical attitudes toward scientific data management in these settings.

An important 2005 report from the National Science Board (NSB) of the National Science Foundation (NSF), *Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century*, provides a succinct summary of the characteristics of small-science data sets, which it describes as:

> *… the products of one or more focused research projects [which] 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. [These data sets] are intended to serve a specific group, often limited to immediate participants. There may be no intention to preserve the [data] beyond the end of a project. One reason for this is funding. These [data sets] are supported by relatively small budgets, often through research grants funding a specific project.* (National Science Foundation. National Science Board 2005, p. 20)[22]

An important characteristic of small-science data management is that the researcher in essence has full control over the data. Historically, most small-science researchers have tended to see their data primarily as a means to an end: answering specific research questions and producing associated publications. From this perspective, time devoted to data management is time taken away from the more important tasks of data collection and analysis. Thus, researchers have typically managed their data only to the extent needed

[22] The terminology used in the NSB report is somewhat different from that used in this report. What this report calls "small-science data sets" the NSB report refers to as "research data collections," which it contrasts with two other categories of data that are marked by higher levels of data management and durability and which it calls "resource collections" and "reference collections." ("Durability" in this context refers to the length of time that a data set is preserved in usable form.)

to make them usable for their own analytical purposes, although if informal norms existed in their field, they might follow those. Once the PIs published their results, which might involve only a subset of the data collected, they moved onto other research projects. At this point, they had few incentives for further curation of the data, which they often left stranded in a variety of formats, with limited or no descriptive metadata, and on a variety of storage media, some of which were susceptible to deterioration or obsolescence. From the individual researcher's perspective, the costs of maintaining existing data sets over time (both direct monetary costs, and opportunity costs in terms of the alternative of collecting new data) vastly outweighed any potential benefits in terms of future use by others.

If an external party wanted to use another researcher's data, he or she would contact the PI, who would grant permission on a case-by-case basis. Many PIs were reluctant to make their data generally accessible, lest others misinterpret it, use it for invalid purposes, identify errors in the collectors' own analysis, or beat them to the publication of important results. In any case, potential users were often limited to a relatively small professional community. Because the existence and characteristics of a given data set were not widely known, researchers from different disciplines tended not to aggregate different data sets (what today the scientific community sometimes calls "mashups"). Funders and parent organizations accepted this approach to data sharing.

Several factors reinforced—and continue to reinforce—this approach to data management and sharing. The first, and perhaps most fundamental, factor is the absence of professional incentives for systematic data management or sharing. Even now, scientists are rarely given credit in performance evaluations for producing well-curated data sets of potential future value. Nor are there professional rewards for producing data that other researchers widely use or cite—at least nothing comparable to the rewards received for peer-reviewed publications that were cited frequently.

A second factor is the absence of support from the information-science and IT communities, which stood outside the domain-science research community and provided only minimal support in a few well-defined areas. For example, libraries provided access to the publications that resulted from data analysis, but had little to do with managing the data underlying the publications. Institutional archives might have preserved some data of potential historical interest, but not necessarily the descriptive metadata that would facilitate scientific re-use.[23] The role of IT departments was to provide the basic hardware, software, and technical support to facilitate research. That role did not extend

[23] Archives have archival metadata specific to their functions; in general, their metadata are intended to facilitate discovery and access rather than scientific interpretation.

to facilitating data management or sharing—or even, in most cases, to long-term network storage of small research data sets, because of the expense.

A third factor is that small-science projects traditionally did not require sharing expensive research equipment across organizations, integrating data from multiple sources across a range of geographic areas, or dealing with massive volumes of data—three characteristics of big-science research that pushed these fields toward more systematic and standardized data management. The scale of small-science projects was usually relatively modest in terms of equipment requirements, geographic scope, and data volume. By definition, the latter was no more than what could be managed by an independent research team armed with the computing technology of the day; translated into text and numbers, this might mean kilobytes or at most a few megabytes of data (Table 1).[24]

The end result has been that small-science biology has generated a large number of disconnected data sets located at many different locations in a great variety of forms, whose long-term preservation and use have often been ignored.

Forces of Change

The development of the internet, the increased availability of high-speed computing, and the reduced costs of data storage promise to revolutionize the small-science approach to data management. Many practitioners in small-science research have realized that new opportunities are being created that will allow the hard work they put into collecting data to be leveraged in exciting and novel ways—even if others still see the data as "theirs" and resist opening them up to broad scrutiny and use.

On the whole, however, it is also seen as increasingly anachronistic to continue the small-science approach to data *management*, because it inhibits the re-use and integration of data. Small-science research teams can expect to come under increasing pressure to break the habit of recording, documenting, and storing their findings in ways that do not facilitate external discovery, access, and use. This section briefly reviews some of the major forces contributing to this pressure.

[24] The qualification "translated into text and numbers" is required because digital representations of audio, images, and video must be thought of on a different scale. For example, a single high-resolution photograph or an hour of audio recording could be many megabytes in digital format, while an hour of film or video could be several gigabytes or more.

Table 1: Digital Data Magnitudes

Unit	Size	How to think of it
Bit (b)	One character (1 or 0)	Short for "binary digit," after the binary code computers used to store and process data.
Byte (B)	8 bits	Enough information to create an English letter or number in binary code.
Kilobyte (KB)	1,000 (2^{10}) bytes	One page of typed text is about 2 KB.
Megabyte (MB)	1,000 KB = (2^{20}) bytes	The complete works of Shakespeare total about 5 MB. A pop song audio file is about 4 MB.
Gigabyte (GB)	1,000 MB = (2^{30}) bytes	All the text from an entire floor of library books is about 50 GB. A two-hour video can be compressed into about 1-2 GB.
Terabyte (TB)	1,000 GB = (2^{40}) bytes	The text of all the catalogued books in the Library of Congress totals about 15 TB.
Petabyte (PB)	1,000 TB = (2^{50}) bytes	All letters delivered by the U.S. Post Office in 2010 amounted to about 5 PB. Google processes about 1 PB of data per hour.
Exabyte (EB)	1,000 PB = (2^{60}) bytes	All the words ever spoken in human history total about 15 EB.
Zettabyte (ZB)	1,000 EB = (2^{70}) bytes	The total amount of information in existence in 2010 is predicted to be around 1.2 ZB.
Yottabyte (YB)	1,000 ZB = (2^{80}) bytes	Currently too big to imagine.

Source: Adapted from Cukier (2010) and Van Garderen (2006).

The "Data Deluge"

Recent years have witnessed exponential growth in the volume of scientific data collected, even in historically small-science fields. For example, original research in systematic biology once focused mainly on the collection and study of biological specimens, which generated field and lab notes and measurements of specimen characteristics. The field now uses far more data-intensive methodologies, such as new techniques of genomic sequencing, to understand the relationships among species.[25]

[25] According to a recent *New York Times* article (Markoff 2010): "Genetic sequencing systems are capable of generating as much as a terabyte, 1,000 gigabytes, of information a minute."

Likewise, ecologists whose data once consisted of observation and measurement of environmental elements by researchers in the field increasingly make use of huge sets of monitoring data beamed down from satellites or collected by terrestrial sensors capable of measuring everything from atmospheric gas levels to subterranean microbial characteristics, perhaps on a continuous basis.

The data deluge results not only from the increasing data intensity of individual research projects, but it is also a cumulative phenomenon. For example, understanding today's environmental and conservation challenges often requires the aggregation of data from individual projects at many locations over the course of many years into larger data sets covering a wider geographic footprint and longer timeframes. A 2003 NSF report on cyberinfrastructure for biological science describes this shift:

> *The biological sciences are at a critical junction in their history, having absorbed over several decades the tremendous successes of "reductionist" experimentation, that is, of carefully focused investigations on simpler systems, model organisms and biological abstractions/models. Today, as the direct consequence of such extraordinary and even unanticipated successes, a new era of synthesis pervades thinking about the future of biological research, from macromolecules to ecosystems. To inform the process and deliver this synthesis, biological scientists must collect, organize, analyze and comprehend unprecedented volumes of highly heterogeneous, hierarchical information obtained by different means or modalities, with different standards, widely varying kinds (types) of data, over vast scales of time, space and organizational complexity.* (National Science Foundation. Directorate for Biological Sciences Advisory Committee 2003, p. 4)

Methods that may be adequate for managing data sets consisting of a few thousand observational data points quickly break down when confronted with megabytes, gigabytes, or even terabytes of data. Such data sets require a more systematic approach to managing incoming data streams from monitoring equipment and processing them for use; documenting the resulting raw data with descriptive metadata to facilitate re-use, recalibration, and interpretation; and preserving over the long term data of future value.

Interdisciplinary Global Challenges

Scientific data often have value beyond the context in which they were collected, and new, unforeseen uses for them can emerge as the future unfolds. This factor has been a major force behind the recent drive for better data management. Several interviewees discussed instances in which data collected for a particular purpose later proved to have

unexpected significance in an entirely different field or for a totally different research question. For example, one interviewee described how data gathered by seismologists in ocean environments have turned out to be of great value to oceanographers. Similarly, the website of the Biodiversity Information Standards (TDWG) organization offers this anecdote:

> *Some scientists assume that they are the only people able fully to understand the data that they have collected; however, data can often be used in the most unexpected ways. For example, before remote sensing from satellites started in 1973, nobody knew the extent to which Antarctic sea-ice had changed. Bill de la Mare had a hypothesis: he wondered if the southernmost catch limit of whales in the last centuries could be a surrogate for the northernmost extent of sea-ice. Whales (and the whalers) tended to hug the sea-ice edges. Bill studied 40,000 records of whale kills in the Southern Ocean and concluded that the sea-ice had retreated 2 degrees latitude (~320 kilometers) between the mid 50s and the early 70s.* (http://www.tdwg.org/)

Data re-use has become particularly relevant in the face of global environmental challenges whose solutions call for deeper understanding of complex systems across multiple disciplines, places, and time periods. Data gathered independently over the years to answer specific research questions in biology can, when aggregated, help address present challenges such as climate change, loss of biodiversity, fisheries and wildlife management, zoonotic disease transmission, and wetland destruction. A prime example is the application of paleobiology data from the distant past to understand climate change in the present.

In-depth understanding of such global challenges typically requires data not only from multiple fields within biology, but also from other scientific and social-scientific fields. To this end, data covering a wide range of times, locations, and disciplines need to be easily discoverable, accessible, and usable by specialists in a variety of fields.

Some information-science theoreticians have even suggested that the scientific research enterprise is moving toward a new paradigm of "data-intensive scientific discovery." In this paradigm, the traditional focus on gathering and analyzing new data to answer relatively narrow research questions will give way to an emphasis on analyzing accumulated historical data to address broader scientific challenges (Hey, Tansley, and Tolle 2009). Such a shift would move data management to a central place in the research enterprise, and fundamentally change the small-science approach of autonomous management of project data.

Technology

Technology has become a powerful force for change in the small-science data-management approach in two ways. The first is its role in the data deluge discussed above. The rapid evolution of remote-sensing technologies increasingly enables biology researchers to collect and work with enormous data sets. The rapid evolution of information technologies makes it easier to pull together large amounts of dispersed data (although, as discussed below, there is still a great deal that needs to be done to enable integration of data across different fields). In short, the data deluge issue is deeply intertwined with the progress of research and information technologies.[26]

Second is the way in which technological change has created the need for new skills and infrastructure to support data management. To take an obvious example, for most of the history of science, the predominant data storage medium was paper, which has proven very stable and robust when subjected to proper environmental controls. Paper data sets that are well-cared for can survive centuries—although in practice, they rarely did, since few researchers went to the trouble to systematically preserve their data once results were published. By contrast, digital data, for all their advantages in manipulability, are comparatively fragile and require frequent interventions to remain usable. For instance, frequent integrity checks are required to prevent bit decay, and migration from obsolete hardware and software platforms is necessary as new generations of technology come online.[27] Researchers can no longer adopt a "file-it-and-forget-it" approach to non-active data if they want any possibility of future re-use. Rather, maintaining the usability of digital data requires reliance on a costly complex of infrastructure and skilled IT personnel.

Legacy Data

Any scientific organization that has been around for any length of time has significant amounts of unmanaged legacy data—broadly speaking, data from past research that are at risk of loss because they have not been systematically managed. Legacy data date back years, decades, and even longer. They reside on all manner of storage media (hard-copy, analogue, and digital) in many formats. Often, they are scattered among the personal effects and equipment of individual scientists. Some have become virtually inaccessible due to outdated software or hardware, or a lack of adequate metadata documentation.

---

[26] Something of a chicken-and-egg dynamic exists here. A rapid increase in collected data can spur technological advances, but technological advances can also lead to rapid growth in the data collected. Similarly, new technologies are sometimes created in response to recognized research needs, while in other cases technological progress in areas unrelated to scientific research is appropriated for that purpose.

[27] For example, if a WordPerfect document stored on a 5½-inch data diskette 20 years ago had not been migrated, it would likely be very difficult and costly to retrieve and use today.

Most organizations have no idea what legacy data they hold and what condition these data are in, let alone whether they might be of value to future researchers. Few organizations have systematic plans to catalogue their legacy data, assess their future value, or preserve them.

The challenge presented by legacy data likely will become more acute over the next decade or so, with the imminent retirement of the first generation of "digital researchers"—those who lived through the widespread introduction and diffusion of digital technologies into small-science research, with all of the trial-and-error that this entailed. Without the knowledge of the original data collectors, their data, especially those from the earlier days of the digital era, will be particularly difficult or impossible for others to reconstruct. In some cases, preservation efforts will have to be undertaken soon to prevent the loss of legacy data.

Awareness of the vulnerabilities of digital data and the importance of preserving them for future use has grown substantially among newer generations of researchers. As a result, many parties are working to define standards for data collection, management, and documentation that will prevent the accumulation of future unmanaged legacy data. In the meantime, however, funding limitations and a lack of systematic support structures encourage researchers operating within a small-science paradigm to continue to create data that are vulnerable.

Ethical Imperatives

A number of ethical factors are creating pressure for widespread, easy data sharing. Three stand out. First is the public value of research data. In the United States, as in most nations, taxpayers pay for much basic research in biology. In light of the strains on government budgets at all levels, there is growing pressure on scientific research organizations to demonstrate societal benefits that justify this public investment. One possible response is to make the data more accessible to a range of users—scientists in applied fields, engineers, resource managers, economists, and others—who can apply them to new scientific explorations and to solving practical problems.

Second, and also linked to the public value of data, is the increasing concern over the bifurcation of the world's scientific communities into "haves" and "have-nots."[28] The former, which include the scientific communities of wealthier nations such as the United

[28] Concern about a "digital divide" has also arisen within the U.S. scientific community, with major research universities and other high-profile organizations on one side and less well-heeled peers—including minority-serving institutions such as Historically Black Colleges and Universities and American Indian Tribal Colleges—on the other. See, for example, National Science Foundation, Blue-Ribbon Advisory Panel on Cyberinfrastructure 2003, pp. 28-29.

States, have access to a wide range of scientific data and instruments. The latter, based for the most part in developing nations, often lack the infrastructure, expertise, and resources to access and use the work of the global scientific community or to share the valuable data that they have collected themselves. In response to this troubling gap, the scientific communities in more advanced nations are undertaking efforts to reduce this digital divide and to develop programs and institutions for the sharing of data, information, knowledge, and expertise. Improving the ease and decreasing the costs of international data sharing are obvious strategies for achieving this end. (Other means include collaborative research projects and capacity-building efforts such as training courses, scholarships, consultation, and development of infrastructure.)

A third issue is an increasing interest in understanding different ways of knowing, and integrating non-Western cosmologies and ontologies into the larger knowledge-base infrastructure. This interest is driven not only by the ethical imperative to preserve traditional cultures, but also by a pragmatic interest in gleaning insights from non-Western knowledge systems for further testing and study.

The Costs of Systematic Data Management

In light of these forces, pressure has grown in recent years for systematic data management efforts that display some degree of uniformity across projects, organizations, and even nations. But the costs of developing and implementing the required data management infrastructure at the level of individual organizations and nations, let alone at a global scale, are huge and continue to grow. They most certainly exceed the capacity of individual organizations and even nations to fund.

The resulting funding limitations have led to significant efforts to pursue collaborative approaches that achieve economies of scale in the management and preservation of data.[29] Both interviewees and the literature have pointed to a fundamental need to leverage resources across organizations and governments. In this collaborative world, autonomous or idiosyncratic data management by individual research teams is inefficient, even if it sometimes may be unavoidable in new fields.

Responses to the Changing Environment

The forces pushing for more systematic approaches to data management have resulted in a variety of responses both within individual research organizations, among collaborative

[29] In the jargon of economics, "economies of scale" refer to cases where the per-unit costs of production decrease as the number of units produced increases—in other words, production on a larger scale is more efficient.

groups of organizations, and at higher levels such as among national scientific communities and across different domain science fields. This section provides a brief overview of some of these efforts.

Individual Organizations

There is considerable activity by individual organizations that carry out biology research to come to grips with the growing challenges of data management. Efforts range from small-scale changes in workflows and incentive structures that impact only an organization's own researchers to massive projects by key players that affect whole fields. Examples of the latter include the National Institute of Health's (NIH) GenBank repository of DNA sequences,[30] the National Aeronautics and Space Administration's (NASA) online metadata Global Change Master Directory (GCMD), and Columbia University's Center for International Earth Science Information Networks (CIESIN), which hosts several important databases on human interactions with the environment. The more ambitious initiatives inevitably draw in other organizations as collaborators, partners, and sponsors, thus blurring the boundary between organizational and inter-organizational efforts.[31]

Interviewees and the literature agreed that much can be learned from activities at individual organizations. However, if such efforts are to be valuable to the wider world, they need to be undertaken with an awareness of other relevant efforts, so they do not waste resources reinventing the wheel. Further, the results and lessons learned need to be communicated to other organizations—for example, through channels such as participation in inter-organizational initiatives, formal publications, and presentations at professional conferences.

In most organizations, existing data-management arrangements are the result of an accumulated series of opportunistic decisions, path-dependent iterative processes, improvised solutions, and historical accidents. Efforts to change these arrangements frequently require dealing with obstacles arising from longstanding and strongly defended organizational workflows, divisions of labor, and turf. The costs of change (retooling, re-education, recalibrating existing data, and so on) also raise barriers to organizational reform. A further issue is that the locus of responsibility for scientific data

[30] The development of GenBank is interesting, in that it began as a repository for data that users employed at their own risk; there was no quality control. It has evolved into a system that offers levels of quality control and assurance, including a "gold standard" based on voucher-specimen availability—a standard developed by the Smithsonian-based Consortium for the Barcode of Life (CBOL).

[31] This study defined the dividing line between the two categories on the basis of whether a given initiative—no matter how many partners it eventually includes—started with a single, parent research (as opposed to funding) organization. Even using this definition, the status of some efforts is ambiguous.

management is often highly distributed within an organization.[32] Individual scientists, their departments, and various other units such as libraries, archives, bioinformatics offices, and IT departments all have roles and responsibilities that are not always aligned. Across the whole organization, both overlap and gaps frequently exist; various parts may duplicate efforts or work at cross-purposes (for example, using incompatible software for similar tasks); and these parts may regard one another not as partners for leveraging the resources of the whole, but as competitors for "their" slice of those resources.

There are, of course, examples of organizations with more systematic approaches to data management, some involving quite complex organizational structures. One is NASA, which has made the sharing and management of its voluminous satellite earth observation data an organizational priority. As part of this effort, it has established clear lines of responsibility for different aspects of data management.[33] Yet despite the progress NASA has made with data management, interviewees acknowledged that much hard work remains to be done to extend discovery, access, and usability of its data. While the basic data streams are generally well-archived, NASA continues to have problems managing derivative products, recalibration, and assimilation of data from research done by PIs outside the organization, whether funded by NASA or by other agencies. Other large government agencies such as NSF, NOAA, and EPA face similar challenges.

As discussed further in the subsection on "Professional Roles" below, libraries at some universities have moved to take on more responsibility for data management, working with their organizations' research communities. NSF and library/archival professional associations have provided some impetus for this development, in the belief that the traditional library role of facilitating discovery, access, and preservation of paper documents extends logically to electronic documents and to digital data.

Collaborations

In addition to working to get their own data-management houses in order, organizations have been engaging in a number of significant cooperative projects. Of particular note is the growth of collaborations among organizations/governments in the developed world

[32] In some cases, centralization may have been considered and rejected for good reasons. Nonetheless, the basic issue remains: distributed structures are typically more difficult to reform in response to changes in the external environment.

[33] NASA's efforts include some good examples of initiatives where the line between organizational and inter-organizational efforts is blurred. A case in point is that, while responsibility for NASA's GCMD resides in-house at NASA's Goddard Space Flight Center, some of its Data Centers are administered through other organizations, such as the Department of Energy (Environmental Dynamics Data Center at Oak Ridge National Laboratory), National Oceanic and Atmospheric Administration (National Snow and Ice Data Center), and Columbia University (Socioeconomic Data and Applications Data Center at CIESIN).

and their counterparts in the developing world. Such collaborations aim both at addressing the equity concerns discussed above, and at ensuring that data sets are globally available, regardless of their nation of origin. These efforts are often mediated by established international organizations such as the International Council for Science (ICSU) and agencies of the United Nations (UN).

The number and variety of inter-organizational partnerships, groups, and initiatives are a testament to the increasing attention that governments and the scientific community are paying to data management and sharing. However, there was widespread agreement among interviewees that

- The growing array of inter-organizational actors and efforts can be bewildering;

- Their work usually is not coordinated and in some cases appears redundant—they are not necessarily aware of each other's efforts, let alone systematically coordinating related work;[34] and

- Gathering accurate information on their work is time-consuming and difficult. While websites may suffice for a general overview of goals and functions, further investigation often reveals that some of the valuable resources and activities described on websites are still at a developmental stage, or have a very narrow focus.

The study team initially hoped to construct a summative typology to help the uninitiated reader make sense of the welter of interlocking, overlapping, and largely uncoordinated collective efforts, but found the challenges insurmountable. For example, a typology based on the functional areas of data discovery, access, usability, and preservation was unsatisfactory because most collective efforts are active in multiple areas. No other typology considered by the study team worked better, whether based on disciplines (sub-disciplinary, single disciplinary, interdisciplinary, etc.); geographic or geopolitical scope (sub-national, national, international, etc.); or type of participating organizations (governments, universities, libraries, etc.). However, to give a general sense of some of the efforts underway, Appendix C provides an unstructured, and certainly far from exhaustive, glimpse at some major collaborative efforts relevant to this study.

[34] There are exceptions. For example, TDWG has formally established links with the Global Biodiversity Information Facility (GBIF) and Open Geospatial Consortium (OGC). The Encyclopedia of Life (EOL) collaborates with GBIF, the Consortium for the Barcode of Life (CBOL), the Atlas of Living Australia, and a number of other collaborative initiatives that focus on species information. See Appendix C for more information.

One important reason for the confusing, patchwork-quilt of collective efforts is that "biology" has grown to encompass a complex and multifaceted array of disciplines and approaches. Some efforts are in the early stages of developing standards, infrastructure, and resources for data management and sharing, while others are more advanced. Across the whole, a lot of experimentation is now taking place in data management and sharing.

Another reason is that successful data sharing involves a complex set of tasks extending throughout the data life cycle, beginning before data are collected (with decisions about data-collection protocols and methodologies) and continuing through either long-term preservation of data or a conscious decision not to retain them. Tackling the entire process is an enormous venture, and obstacles need to be addressed at every stage: competing data-collection protocols and methodologies; differing metadata standards; hardware and software incompatibilities; financial obstacles to sustainable preservation and access; cultural, policy, or legal barriers to sharing data; and so on. Thus, different inter-organizational entities tend to address different parts of the puzzle. They place differing weights on the functional areas of data discovery, access, usability, and preservation—with some, for example, concentrating more on the challenges of immediate discovery and access, and others on long-term preservation. Examples of other dimensions along which the activities of collaborative efforts diverge include:

- An emphasis on technology (for example, cyberinfrastructure) versus non-technological issues (for example, data-sharing policy);
- A strategy of building data-management capacity from the ground up to address specific needs (as with the National Ecological Observatory Network, NEON) versus providing forums to facilitate the coordination of existing efforts (as with ESIP);
- A focus on a single field (for example, the Paleobiology Database, PBDB) versus a wider concern with interdisciplinary data sharing (for example, The Data Conservancy);
- "Top-down" origins (such as the DataNet projects initiated and funded by NSF) versus "bottom-up" origins (grassroots initiatives).

The bottom line is there is a wide range of highly varied collective efforts. Even when considered individually, these can be difficult to keep track of, at least for people other than those directly involved in them. Collectively, the challenge is exponentially greater. Thus, the study team found it exceedingly challenging to piece together who is doing what, which organizations are working on which initiatives, where efforts overlap, where gaps exist, and, from the perspective of the Smithsonian, which efforts merit careful

attention and perhaps participation. Although many interviewees thought that some degree of consolidation of current efforts is inevitable, few were willing to hazard predictions about where such consolidation was likely.

Professional Roles

Both within and across organizations, one of the most important responses to the forces for change discussed above has involved shifts in the roles, relationships, and responsibilities of three distinct professional groups—domain scientists, library and archives personnel, and IT personnel—as well as the new class of information-science personnel whose roles, expertise, and organizational placement overlap to some extent with those of library-and-archive and IT professionals.

The Scientific Community. Interviewees and the literature universally agreed that scientists need to become more supportive of and involved in data management from the inception of a project—even at the stage of writing grant proposals. It is almost always more efficient to manage data to a high standard from the moment of creation than to hammer poorly-managed data into shape after the fact, so funding for data management needs to be included in budgets for proposed projects. In the new model envisioned by many observers, and now being introduced in some organizations and collaborative efforts, scientists work systematically from the beginning of a research project with information-science personnel to manage data in a way that simultaneously serves researchers' analytical needs and supports external discovery, access, and use of the data down the road. In this approach, parent organizations provide scientists with IT and information-science staff support, user-friendly tools and templates, and training so they can effectively manage their data with a minimal investment of time and effort.

However, even with such support, interviewees agreed that the factors for change discussed above will require most small-science researchers to devote more time than has historically been the case to data-management tasks, such as formatting data in prescribed ways and writing descriptive metadata. They will also have to give up some of the control they have in the past exercised over access to their data.

The literature and interviewees widely agreed that professional incentives to manage and share data are needed at both the organizational and professional levels. For example, many funding organizations are now requiring data-management and data-sharing plans as a condition of grant awards. Interviewees also mentioned the need to incorporate meaningful requirements for data management and sharing as performance criteria in researchers' performance reviews. Finally, there appears to be movement in some fields toward according greater professional recognition to data publishing and citation.

Data publishing is a relatively new concept that is analogous to traditional scientific scholarly publishing, but that involves peer review of raw data sets rather than analytical articles. In its fully-realized form, data publishing requires that organizations professionally reward public dissemination of data in much the same way they reward the publication of articles. A data-publishing system would include:

- Peer-review mechanisms for vetting and validating data sets, similar to the current mechanisms for peer review of journal submissions;
- The provision of trusted channels through which data sets judged by professional peers to be of sufficient quality and interest can be easily discovered, accessed, and used by other researchers;
- Formal requirements that users of these data sets credit the originator of the data, similar to the requirements surrounding citation of analytical literature; and
- Organizational policies that grant the creators of data sets credit in their performance reviews for data publishing and data citations comparable to what they receive for peer-reviewed journal articles.

To some extent, the challenge of getting scientists more involved in data management may resolve itself over time, with new technologies that make the task easier and rise of new generations of researchers nurtured in an age of greater attention to data sharing and interoperability. In general, young scientists tend to be more cognizant than their older counterparts of the opportunities presented by data sharing, the value of systematic data management, and the need to incorporate some level of data management into their work.[35] In part, this is an offshoot of the growing interest in collaboration and interdisciplinary research that prevails among younger generations of researchers.

However, generational change by itself will probably not be enough. As long as success in publishing peer-reviewed analytical articles remains the near-exclusive criterion for professional reputation, researchers will have incentives to spend as little time as possible on data management and to control access to their data. Indeed, these incentives tend to be even stronger for young researchers, who need to establish a publication track record. Also, as more than one interviewee pointed out, even if young researchers are more *willing* to think about data management, they may not be *able* to do so without considerable institutional support, given the massively increasing volume and complexity of the data with which they must work.

[35] One interviewee, however, also noted that older scientists tend to become more attentive to these things as they approach retirement and confront the reality that, without proper management, a large part of their life's work—the non-reproducible observational data that their research has yielded—could be lost.

Libraries and Archives. The role of libraries and archives with respect to digital data sharing, management, and preservation continues to evolve. Gold (2007b) describes some possible areas of activity for libraries as follows:

> *Within the "downstream" side of the research cycle, librarians can play roles in the selection, acquisition, and licensing of data and data sets; in creating metadata (or metadata standards) for discovery and description of data sets; in creating or organizing documentation related to data; and in offering preservation services for digital data …[.] A role associated more with archives than with libraries, but common to both, is advising in the appraisal and selection of what data to keep for the long term. Another role libraries are well positioned to play is assisting users with finding data relevant to their research, using third-party high level directories and data discovery sources such as the Global Change Master Directory or the National Space Science Data Center.*
>
> *… Key to libraries or librarians playing more "upstream" roles in data science is their ability to position themselves as partners in research. By collaborating closely, and early, in the research process, librarians may become involved in creating data curation prototypes, or otherwise supporting the use of documentation, practices, or standards that will assure the longevity of the data downstream. Such close collaborations are far from common, but examples do exist, including the work of the Johns Hopkins University Libraries with the National Virtual Observatory; and the establishment at Purdue University in 2006 of a Digital Data Curation Center as an incubator of data collaborations between librarians and faculty. Another potential library role, building on libraries' experience with institutional repositories, might be to create more dynamic repositories that support pre-publication workflows, including collaboration environments supporting data integration, analysis, and visualization* [no page number].

Several years ago NSF awarded the Association of Research Libraries (ARL) a grant for a series of conferences on the possible roles of libraries in scientific data management. There have been other significant initiatives to integrate libraries into digital data management, including, but not limited to, scientific data management. Two key examples are the National Digital Information Infrastructure Program (NDIIP)[36] at the Library of Congress (LOC) and the Coalition for Networked Information (CNI) (see Appendix C).

[36] Technically, the inter-organizational collaborative element of NDIIP is the National Digital Stewardship Alliance; NDIIP itself is a program office in the LOC.

Many research university libraries are moving beyond their traditional focus on published end products and are establishing organizational repositories for scientific data and other unpublished materials.[37] Because of cost and staffing issues, some are beginning with the modest goal of short-term preservation to prevent data loss, and are simply storing data in whatever form they happen to arrive. However, others are taking a much more active role. In addition to the two universities mentioned in the quoted passage above, some leaders in this area include the Massachusetts Institute of Technology and Columbia University. Library personnel at these institutions engage in activities such as educating researchers on the importance of long-term data preservation and the benefits of proper data management,[38] promulgating best practices in these areas, and providing direct support for data management efforts. This support can be either project-specific or generic—for example, providing tools such as metadata templates.

In addition to working with researchers on the curation of scientific data from the moment of creation, some research libraries are taking the initiative in digitizing legacy literature and making data discoverable and accessible via online channels, as well usable by today's scientists. A major initiative in which the Smithsonian is playing a leadership role is the Biodiversity Heritage Library (BHL) (discussed in more detail in the section on Smithsonian collaborative efforts below).

Archives also seem to be moving toward greater involvement in the stewardship of digital data, although this evolution may be slowed somewhat because organizational archives tend to be smaller and less well-resourced than libraries, the nature of their holdings is often less-structured, and the concept of treating scientific data as institutional records is fairly new and not yet widely adopted. That said, archival organizations and professionals have skills and experience that are highly relevant to the emerging challenges of scientific data management and sharing, particularly in the area of long-term preservation. For example, archival workflow and systems-architecture models frequently appear in the literature on scientific data management, and archival criteria governing metadata for discovery purposes and the appropriate extent of preservation efforts for various kinds of institutional records could inform similar decisions with regard to scientific data.

At the national level, while the National Archives and Records Administration (NARA) has participated in the general research and development of systems for stewardship of

[37] For idiosyncratic historical reasons, libraries are already widely involved in the management of data in a few specific areas; one prominent example is Geographic Information Systems (GIS).
[38] These include not only the general benefits to the world of preserving data for possible future use, but also more tangible benefits to researchers themselves, such as satisfying funders that data management requirements are being met and safekeeping the data in case the researcher wishes to revisit them.

electronic records, the study team saw little evidence that it has been a major force to date in scientific data management per se. An interviewee at NOAA noted, however, that NARA has recognized three NOAA environmental data centers—the National Oceanographic Data Center, National Geophysical Data Center, and National Climatic Data Center—as national archives for certain types of big-science data. However, the study team does not know whether this is part of a larger strategy to systematically delegate the task of preserving federal research data to agencies with relevant experience. Overall, the case of NARA seems to reflect the general judgment that archives, for now, are relatively marginal players in the area of scientific data management.

One interviewee, musing on long-term developments, suggested that as data and other unpublished digital records increasingly become part of the scientific dialogue previously conducted through formal publications and conferences, the distinction between libraries and archives will erode. The result he envisioned would be unified "organizational memory" units that preserve and provide access to a wide range of published and unpublished materials, from raw data through formal peer-reviewed publications.

The IT and Computer-Science Communities. As information technology itself has grown more complex, so has the relationship between scientific researchers and IT personnel. Interestingly, many first-generation IT and computer-science innovators came from domain-science disciplines, having been forced to develop expertise in these areas to deal with questions that arose in their scientific work. However, subsequent generations of IT and computer-science personnel have tended toward more specialization and to have little domain-science knowledge. The divide has continued to grow, and greater turf consciousness has contributed to a more conflictual relationship that sometimes has repercussions for the ability of the IT and domain-science communities to work together productively.

The conduct of cutting-edge scientific research, the management of scientific data, and the design of cyberinfrastructure to support science are increasingly intertwined processes. Progress in these areas requires domain scientists to work with IT and computer-science professionals who can mediate between research goals and the technological realities that constrain or enable these goals. In big-science fields, research teams already include IT and computer-science professionals as core personnel who work hand-in-hand with scientists to design data collection and management systems, as well as to process data for analysis.

A cornerstone document on the challenges of cyberinfrastructure—*Revolutionizing Science and Engineering Through Cyberinfrastructure,* a 2003 NSF advisory study widely known as the Atkins Report—describes the delicate balance between domain

scientists and computer-science personnel in designing and implementing systems to support computationally intensive science and data management:

> *If the organization is weighted too heavily toward the domain scientists, the focus overemphasizes procurement of existing technologies, and computer scientists become viewed as "merely" consultants and implementers. If the weight shifts too heavily toward computer science, the needs of end users may not be sufficiently addressed, or effort shifts too heavily toward creating new technologies with insufficient attention to stability and user support.* (National Science Foundation. Blue-Ribbon Advisory Panel on Cyberinfrastructure 2003, p. 51)

The forces for change described above are pushing biology along the same path traveled by big-science fields in years past. For example, biodiversity informatics has become a recognized field of expertise. This hybrid specialization, which brings computer science and biological science together, applies computationally intensive techniques to unlock hidden patterns in enormous specimen, species, and ecosystem data sets. The modeling of environmental systems is another area where IT and computer-science expertise have become inextricably intertwined with scientific practice. Also, as discussed, automated data-collection equipment capable of producing enormous amounts of data is now being used in some areas of biology. Without appropriate processing and user tools to isolate relevant information and put it in a form amenable to analysis, the torrents of data collected by such automated processes are incomprehensible jumbles of bytes to most domain scientists.

Top-Down and Bottom-Up

Most sources agree that scientific communities tend to resist top-down solutions to their data-management challenges. Solutions generally need to be worked out through an often messy and protracted process of dialogue, discussion, and consensus building among parties that may have very different interests and preferences. In some cases, this process takes place at least partly through organized forums, like the Inter-governmental Working Group on Digital Data (IGWDD) for federal data-management policy and GEO/GEOSS for earth-observation data. In others, it is nudged along by powerful players, such as NASA or NSF's Cyberinfrastructure Program, with funding a common incentive. In still others, the process is essentially uncoordinated, involving the identification of best practices through trial-and-error and subsequent sharing of these practices informally or through professional channels.

At least one interviewee argued that a bottom-up approach is both necessary and desirable. For example, with regard to the process by which data standards emerge, he argued forcefully against attempts by the federal government to push developments in specific directions:

> *The best ideas … continuously emerge from the bottom. Science changes, hopefully. Government is not really good at [guiding] those changes, so why should it try? It should be in the business of supporting the winners in a process that lets [solutions] emerge from a lower level, where there is always some chaos. As you have new fields interacting in ways they haven't interacted before, the old standards on both sides aren't going to work. So there will be a period of chaos when you haven't got a new standard and nobody wants to let go of their old standard. You have to let that process work itself out while trying to manage it as best you can during that time of churn. It's kind of like capitalism—you need to have creative destruction of standards to get new, better versions.*

According to this view, the current, unstructured proliferation of efforts and initiatives will self-sort into a preferred set of tools that is both stronger and more widely accepted for having survived the sorting process.

By contrast, some interviewees expressed concern with what they saw as a process perilously lacking in overall direction and coordination, and in need of consolidation and leadership. At least one person pointed out a serious practical problem of taking a "wait and see" approach: while the process sorts itself out, additional legacy data are being accumulated at a prodigious rate:

> *It would be fine if we weren't continuing to generate data in such huge volumes. And with the advent of [new] sequencing technology, the volume of biological data we are generating is becoming increasingly large and will continue to grow exponentially. The problem is that we are generating giant amounts of data right now that don't have any place to go. If they all get put into different places and formats and architectures, there's a good chance that if those places and formats and architectures fade, the data will fade with them. At some point in the process, very early on, [a winnowing] process is probably necessary. But I think we are past that point now, and we need to make a set of decisions.*

Policy

The policy environment for data sharing is in flux at all levels—organizational, national, and international, as well as at the level of professions and fields. Nevertheless, the trend

is, according to Francine Berman, director of the San Diego Supercomputer Center, toward "[m]ore and more policies and regulations [that] require the access, stewardship, and/or preservation of digital data" (Berman 2008, p. 52).

Open Access

A central principle in any discussion of data-sharing policy is the concept of "open access," which stands as a widely-recognized global ideal for scientific data. In a world of pure open access, scientific data, to the extent technically and economically feasible, would be available to all external users, without condition or cost, immediately after initial data cleaning and checking are complete. This concept has found its way into international guidelines, agreements, and statements of principle from influential organizations such GEO, ICSU, the United Nations Educational, Scientific, and Cultural Organization (UNESCO), and the Organization for Economic Cooperation and Development (OECD). It shows up in governmental policies on data access, including those of the European Union (EU), and has increasingly been stressed by the U.S. federal government with regard to research supported by taxpayer funds.

However, apart from purely technical and economic obstacles, issues surrounding intellectual property rights, proprietary holds, security, intended data use, data citation, misinterpretation, and privacy often stand in the way of the open access ideal.[39]

- **Intellectual property rights.** In some cases, the question of who owns and has the legal authority to grant access to a data set is unclear, particularly when research involves collaboration among multiple research organizations or is supported by multiple funders. In other cases, intellectual property rights are controlled by private-sector funders or research organizations that have a commercial interest in restricting access. Likewise, publishers of scientific literature may impose restrictions on access to the data that underlie (and are submitted with) articles they publish.

- **Proprietary holds.** Even those with a deep commitment to open access acknowledge the ethical claim of researchers to some period of exclusive use of the data they collect. While the appropriate length of such periods may vary among fields or on the basis of other considerations, the principle that scientists should have first claim on analyzing and publishing results from their own data is

[39] Another factor worth noting is that until a few years ago publishers would not accept data sets for publication and frequently allowed taxonomists to cite only selected or representative specimens, which is considered inadequate documentation for a species concept. Publishers are beginning to accept additional data, but the issue of their long-term availability and curation persists.

widely accepted. Data-sharing policies often address this by building in some period of exclusive use of data, which may be expressed in terms of calendar time (for example, one year) or milestone events (for example, the initial publication of major findings).

- **Security.** For some types of data, open access raises security concerns. An obvious example is high-resolution satellite imagery of military installations or deployments. While data in biology tend to be less obviously sensitive, interviewees did raise examples where open access could have undesirable consequences from a security perspective, broadly defined. For example, conservation biologists usually want to block access to information that identifies the location of endangered species of interest to poachers.

- **Intended data use.** A related set of issues that may argue for limits on open access revolves around the purposes for which the data are to be used, and by whom. For example, research organizations that might be happy to provide unrestricted access to the scientific and educational communities for non-commercial use might balk at making their data available to for-profit firms that want to use them for commercial gain. Where such restrictions on data use are relevant, the trend is toward permitting access (or not) based on the type of organization making the request and the intended use. In an era when access is increasingly through the internet, this calls for technical measures that can provide positive identification of the organization or individual seeking access.[40]

- **Data citation.** Some observers are concerned that open access might result in uncited or improperly cited data, thus denying data collectors the credit that is their due. For this reason, the success of open access depends on the establishment of professional norms governing the acknowledgement of original data sources—for example, through formal citation of a data set or the inclusion of data collectors as co-authors of analytical work that used their data.

[40] The UN Convention on Biological Diversity will most probably impose restrictions on access to genetic resources, although what those restrictions will look like is unclear.

- **Misinterpretation.** Regardless of the willingness and ability of data owners to release it, open access may not be advisable if the data are liable to misinterpretation or misuse.[41] As one interviewee noted:

> *Open access won't happen unless we have adequately described the data in terms of quality, collection, intended purpose, and other internal factors that tell you what it's good for. If you throw it out there [without such documentation], it leads to bad conclusions and inferences, apparent contradictions, or missed connections through misanalysis or misunderstanding of the data.*

- **Privacy.** With research involving human subjects, the imperative to protect subjects' privacy places ethical and legal limits on data access.[42]

U.S. Federal Agencies

As noted, U.S. federal government agencies are increasingly expected to make their scientific data accessible for general use, and to justify research budgets in terms of societal benefits. This was true even before the Obama administration's strong emphasis on science as a tool for addressing societal problems.

- The E-Government Act of 2002 created the position of Federal Chief Information Officer within the Office of Management and Budget (OMB) and initiated a framework for improving online public access to federal-government information and services. One part of this framework is the requirement to create "a repository that fully integrates, to the maximum extent feasible, information about research and development funded by the Federal Government," as well as "[one] or more websites upon which all or part of the repository of Federal research and development shall be made available to and searchable by Federal agencies and non-Federal entities, including the general public." (U.S. Congress. E-Government Act 2002, p. 22)

[41] Data cleaning, recalibration, reinterpretation, and assimilation are necessary steps in the development of any integrated data set. In the case of some fields, such as taxonomy and weather/climate data, there has been considerable examination of the process to understand the implications for misinterpretation over time.

[42] This is a huge issue in fields such as medical research and psychology. While it is less of a concern in biology research, it is increasingly relevant in some other fields in which Smithsonian researchers are active, such as anthropology and ethnic studies.

- Among its many provisions for strengthening science, technology, and scientific education, the America COMPETES Act,[43] signed by President George W. Bush in August 2007, requires civilian federal agencies to create guidelines, policies, and procedures to promote the open exchange of data and research among agencies, the public, and policy makers.

- More recently, OMB Open Government Directive of December 8, 2009 took up data sharing among other issues related to transparency in government. It included this specific injunction:

> *Within 45 days, each agency shall identify and publish online in an open format at least three high-value data sets … and register those data sets via Data.gov. These must be data sets not previously available online or in a downloadable format.* (Office of Management and Budget 2009, p. 2)

There appears to be widespread recognition of the need for better policies on data sharing at the level of individual agencies, and legislation such as the E-Government Act implicitly commits the federal government to addressing this issue. However, action toward framing such policies for small-science research, let alone implementing them, is very much in the early testing stages in most agencies.[44] There appears to be little appetite for the top-down imposition of uniform policy across the federal community; rather, the expectation is that agencies will work out their own policies, with some degree of consultation through forums such as IWGDD (discussed below).

An important foundational document for digital data management within the U.S. federal government, *Harnessing the Power of Digital Data for Science and Society* (Interagency Working Group on Digital Data 2009) recommended that "… appropriate departments and agencies [should] lay the foundations for agency digital scientific data policy and make the policy publicly available." The precise content of such policy it left in the hands of the agencies. Perhaps the only specific requirement imposed across the board was that agencies were expected to have effective data management plans for all projects that yield data of potential long-term interest, but even here the report allowed great flexibility. As one interviewee described it:

> *Every project or proposal that is going to generate digital data for preservation should have a [data management] plan. … But [the IWGDD] doesn't say what should be in that plan. It just creates a framework [to address the question,]*

[43] America Creating Opportunities to Meaningfully Promote Excellence in Technology, Education, and Science Act.
[44] In this area as well, big-science research is well-ahead.

> *"How are you going to make data available now and in the future?" [Agencies can] answer that question in any way that [relevant scientific peers] find credible. Any answer [the scientific peer community] finds credible is okay; you just need a plan going in. This business of doing your project, getting to the end, and saying, "I have all this data on discs; what do I do now?"—that is yesterday. … You might not think certain data merits preservation, so there is no data management plan for it. If peer review says that's fine, then it's fine.*

IWGDD, which recently was in effect transformed into a permanent sub-committee of the National Science and Technology Council's (NSTC) Science Committee, remains a key venue where federal agencies come together to discuss data management policy. Other important inter-agency venues relevant to the subject of this study include the U.S. Group on Earth Observations (USGEO, the U.S. partner of the international GEO/GEOSS initiative) and the U.S. Global Change Research Program (USGCRP). More generally, the White House Office of Science and Technology Policy (OSTP), of which NSTC is a part, provides a forum for exchange and coordination across agencies.

Federal research organizations are typically active in wider collaborative entities that deal with data-management policy issues and that draw in private sector organizations such as universities, non-profits, commercial firms, and non-governmental organizations (NGOs). For example, many federal agencies with science portfolios are active in ESIP and the National Biological Information Infrastructure (NBII). Typically, these agencies are also involved in other, more mission-specific venues for policy coordination.

Despite all the legislation, directives, and activity, interviewees suggested that creating and implementing a coordinated, systematic, pan-governmental set of policies on data management will be a major challenge that will take time. It will also be influenced in unpredictable ways by the emergence of best practices, advances in technology and cyberinfrastructure, and shifts in the economics of data management and preservation. One interviewee likened the introduction of policies that impact traditional research norms and practices to steering an oil tanker. The existence of administrative and programmatic units within each agency—centers, projects, programs, departments, labs, and so on—will inevitably slow the formulation of overall agency policies. Moreover, even when policies exist, interviewees indicated that it takes time for scientists to accept them, become aware of their implications, and build them into their research.

Funding Organizations

In the United States and elsewhere, major funding organizations are creating enhanced norms for scientific data management and sharing. This process is developing rapidly as advances in technology and reductions in costs expand the boundaries of what is feasible.

NSF is the largest single funding source for basic-science research in the United States. Increasingly, it fosters improved data-management practices through a combination of carrots and sticks. The carrots include grant support for meetings, working groups, collaborative forums, and pilot programs that address key data-management and -sharing issues. For example, NSF's Cyberinfrastructure Program provides substantial resources for improving the national infrastructure for digital data management; NSF's domain-science directorates are also active in digital data initiatives specific to their fields.

The main stick wielded by NSF is a new policy, introduced in October 2010, requiring grant recipients to submit data management plans with their proposals. Most NSF directorates already had some type of data-management requirement for grantees, but interviewees agreed that the requirements had generally been vague and were rarely enforced. The new agency-wide policy signals a stronger commitment within NSF to setting and enforcing data-management and data-sharing policy.

Other U.S. federal agencies that fund research at universities and other external organizations also often have explicit requirements for data management and sharing. For example, some agencies require that grantees agree to make their data accessible to external users—possibly by posting the data on an agency website with few restrictions on access and use. NASA, for instance, requires that grantees deposit data in designated repositories, with standard embargo periods that vary by data type. Interviewees indicated that grant disbursements may be withheld until recipients fulfill the stated data-management requirements, such as writing metadata in a specific format or depositing data in a repository. As another example, NIH's detailed written data-sharing policy is premised on the principle that "data should be made as widely and freely available as possible, while safeguarding the privacy of participants and protecting confidential and proprietary data." It addresses, with some precision, key issues such as the timeliness of data release; expectations for data documentation; preparation of and reporting on compliance with data-sharing plans; treatment of confidential and proprietary data; acceptable methods for data sharing; and funding to support data-sharing plans.

Private-foundation policies for data sharing vary greatly, but the trend is clearly toward greater emphasis on making data widely available. For illustrative purposes, a few

excerpts from the Gordon and Betty Moore Foundation data-sharing policy, which runs nine single-spaced pages, are provided in the text box below.

Gordon and Betty Moore Foundation Data-Sharing Policy: Excerpts

It is the policy of the Gordon and Betty Moore Foundation [GBMF] that data produced with Foundation grants and support will be freely shared and made widely available for charitable purposes, thereby enabling the frictionless flow of data within and between fields. Data will be shared consistent with applicable laws and with … attribution to the data provider.

… [Grantee recipients] will be expected to include a description of their data-sharing strategy and implementation plan … or state why data sharing is not possible.

… *[D]ata includes not only summary statistics or tables, but also all the data on which those statistics and tables are based.* For most studies, GBMF expects final research data will be a computerized dataset. For some, but not all, scientific areas, the final dataset might include both raw and derived data. In all cases data must be accompanied by documentation (metadata)[.]

Given the variety of projects that GBMF supports, neither the precise content for the data documentation, nor the formatting, presentation, or transport mode for data is stipulated. … *However, grantees must plan for data sharing, and be aware of the current state of data sharing activities and data-management best practices within their disciplines and fields.*

- Relevant online data repositories (e.g., GenBank) and data federations
- Procedures for data documentation
- Data formatting and data exchange standards
- Software (or online data services) that conform to data format and exchange standards
- Procedures for quantifying the demand (use) for the data (i.e., number and rate of users and records per data repository and/or data federation per year)
- Procedures for the data owner, or provider of the datasets, to receive feedback about the quality of the data

… [D]ata sharing should occur in a timely fashion. Timeliness is influenced by the nature of the data collected, but GBMF expects data to be released and shared no later than the acceptance for publication of the main findings from the final dataset. Data from small projects can be analyzed and submitted for publication relatively quickly. If data from larger projects are collected over several discrete time periods or phases, it is reasonable to expect that the data be released in phases as they become available or as main findings from a research phase are published.

… Issues related to proprietary data can also arise when co-funding is provided by other donors or the for-profit sector … with corresponding constraints on public disclosure. GBMF recognizes the need to protect patentable and other proprietary data. Any restrictions on data sharing due to co-funding arrangements should be discussed in the data sharing plan [.]

Findings: Functional Areas
(Discovery and Access, Use, Preservation)

Discovery and Access

Discovery—knowing what data exist and where—is an important first step toward use. Even if data of interest are available in usable form, they are of little value if scientists cannot find them. Next, scientists need to be able to access, or retrieve, the data. While these two operations are distinct, with somewhat different enabling requirements, they are so interwoven in practice that they are discussed together here.[45]

In the pre-internet age, scientists relied on discovery channels such as

- Professional literature, networks, conferences, information on ongoing research projects, and word-of-mouth;
- Consultation with specialists in data discovery in libraries, archives, and specialized repositories, whether at their own institutions or elsewhere.

As discussed, the historical norm for post-discovery data access in small-science fields involved direct queries to data collectors, who handled them case by case. This afforded collectors a great deal of control over the use of their data, but also created additional burdens for them. Because requested data may not have been initially recorded or managed in a form usable by external parties, sharing the data often required collectors to engage in some degree of data processing beyond what they needed for their own purposes. Such processing could be onerous, and usually offered little professional reward.[46] (Some fields stand out as exceptions to this general rule; for example, practitioners of revisionary taxonomy receive tangible benefits—publications, grants, better-managed collections—for their data-management skills, as do some climate researchers.) Interviewees indicated that while specific requests from scientific colleagues for data were rarely denied, they could be delayed or derailed because of the time required for processing. In many cases, even after processing was complete, direct

[45] Many relevant points pertaining specifically to access were also discussed in the policy section on "Open Access" above.

[46] Data collectors usually want to be cited as the source in any analytical articles resulting from data they share. In some cases, they may require data users to list them as co-authors of such articles. One interviewee commented that the continual clean-up and reinterpretation of data contribute greatly to advancement in science, and needed to be recognized and encouraged.

consultation between the user and the collector about the limitations and idiosyncracies of the data was required.[47]

Of course, the preceding description presumes that the data collector was still alive, or that the relevant data had been archived. When these conditions did not pertain, the only access to data was usually through the historical literature, most of which was not readily accessible, at least by modern standards.[48] Moreover, even if the data were preserved in archives, older data tended to be hard to find because they rarely were indexed or had finding aids. (This is precisely the issue that the BHL and similar initiatives that digitize legacy publications are intended to address.)

Today, many mechanisms for effective online data discovery exist, although discovery can still be problematic in some ways, as discussed below.[49] In terms of access, while individual queries to data originators are still common, the potential is growing for direct access to data via online links. Online discovery and access services are typically provided via a mediating organization such as a data portal/center that either stores data of interest on its own servers, or directs interested parties to the servers of federated organizations. Such organizations are described in greater detail below.

While the internet offers new opportunities for easy access to scientific data, it is unclear how far these opportunities have been exploited. One reason is that some researchers are wary of direct online links to their data, because these require them to cede some degree of control over their data to the mediating organizations—although exactly how much depends on factors such as the policies of these organizations and their ability to verify the identities of external parties requesting data. Certainly, data collectors who consent to making their data accessible through mediating organizations can (and do) stipulate restrictions on when, to whom, or for what purposes these data may be released, and metadata schemas often include a field detailing such restrictions. Nevertheless, some collectors are reluctant to share their data in this way.

[47] When the original data collector was not available, the person requesting the data might have to resort to research notes to uncover metadata or understand the particularities of the data.

[48] One interviewee noted that as researchers born in the internet age take over, they are not as likely as older researchers to be familiar with how to use traditional libraries for discovery of and access to older data.

[49] For example, some potentially valuable data sets do not have the metadata necessary either to show up in a web search or to allow potential users to judge quality and relevance. Note that it is possible for non-digital data to be discoverable through the web, as long as digital *metadata* exist (but the actual data will not be web-accessible). For example, digital metadata may exist for handwritten research data that have not been digitized.

Mediating Entities

In an increasing number of fields, data portals (or data registries) and data centers exist that aid in the discovery of and access to data from multiple sources. Other mediating entities for discovery and access are research libraries and archives, as well as online search engines.

Operating Principles for Data Access Regimes

- Openness
- Transparency and active data dissemination
- Assignment and assumption of formal responsibilities
- Technical and semantic interoperability of databases
- Quality control, data validation, authentication, and authorization
- Operational efficiency and flexibility
- Respect for intellectual property and other ethical and legal requirements
- Management accountability, including funding approaches.

Source: Arzberger, et al. 2004a.

Where access is directly available, it usually means the mediating entity has secured permission from collectors (or other controlling parties) to release the data, albeit typically with stated restrictions on use where applicable. In many cases, mediating entities undertake some level of data curation, possibly in collaboration with collectors, to make the data more discoverable, accessible, and usable by external parties.[50]

Data Portals. Broadly, data portals are gateways to a range of information on, and sometimes direct access to, scientific data in a particular field or fields. In terms of specifics such as the type and scope of databases covered and the research organizations from which these are drawn, they vary widely.[51] The central elements of a good portal are a catalog of summary metadata to facilitate search-and-query discovery, and more detailed metadata for each data set aimed at making the data usable. Many portals offer direct links to the data themselves, either through a repository database maintained by the organization that administers the portal, or through links to databases at other organizations. Alternatively, data portals might provide users with relevant contacts in organizations that host data. For some interviewees, the question of whether a portal provides links to the data themselves, or at least to contact information for access to

[50] Some organizational repositories, however, are primarily focused on ensuring that data are not lost entirely, and will accept and store data in whatever form they arrive, providing only a minimum of processing and leaving it up to interested external parties to figure out how to use it.

[51] Individual organizations often refer to their own online catalogues of internal data as "portals."In this report, "data portal" primarily refers to online catalogues that bring together data from multiple sources.

unlinked data, is critical. One summed up the situation by saying that data portals today often "promise a lot, but deliver less."

Also important to the success of a portal is that it be widely known and trusted among practitioners in the relevant field(s), and reasonably comprehensive in its coverage. To determine which portals, if any, clear this bar in the various fields of biology is beyond the scope of this report—although candidates abound, including portals associated with major national and international entities such as NASA, NBII, the Long-Term Ecological Research (LTER) network, ICSU, and the Ecological Society of America (ESA), as well as numerous more field- and regionally-specific organizations and initiatives. Based largely on interviews, the study team gathers that present web portal efforts in biology have generally been fragmented, with relevant data scattered across multiple portals, and that there is relatively little agreement in most fields on which portals are definitive. Further, even when a portal is widely known to practitioners in its primary target field, it may be little known outside of that field. At the same time, it should be noted that many of the efforts to establish portals are in their early stages, and improvements are likely.

Some organizational websites that are not data portals per se can also provide useful search-and-query functions for data discovery and access—assuming a user has gotten to the point where he/she knows that useful data may reside with a particular organization.

Data Centers. The concept of a data center is somewhat more general than that of a data portal, and refers to an organization that offers some combination of data discovery, access, processing, and preservation services. Indeed, many data portals are administered by data centers, which are typically collaborative initiatives hosted by a parent organization with some claim to prominence in relevant fields.

The great-grandfather of data centers, which still serves as a model for similar initiatives in other fields, is the Inter-University Consortium for Political and Social Research (ICPSR), founded in 1962 and hosted by the University of Michigan, Ann Arbor. ICPSR has established itself as a central node of interest to social scientists looking for historical data collected by others, or as the place to preserve their own data. For data seekers, it offers an enormous database, convenient online access to research data sets, user-friendly search and analytical tools, technical support, and other resources. For data depositors, it offers curation services that ensure their data will stay safe, accessible, and usable.

Data centers are playing an increasing role in data-management and -sharing efforts in biology, and the situation continues to evolve. Some examples are:

- **NASA.** Formerly known as Distributed Active Archive Centers (DAACs), NASA's 12 Data Centers, many administered in collaboration with federal and university partners, exist to process, archive, document, and distribute data from NASA's earth-observing satellites and field measurement programs. Each serves a specific earth-system science discipline, providing users with raw data, data products, services, and user tools unique to the Center's specialty. Collectively, the Centers offer an extensive selection of earth-observation and related data.

- **World Data System (WDS).** The WDS, an initiative of ICSU, is a globally distributed network of World Data Centers. As with the NASA system, each facility specializes in data from a specific field (atmospheric, solar, geophysical, ecological, and human-environment interaction). The World Data Center for Biodiversity and Ecology is located at the Center for Biological Informatics in Denver, Colorado.

- **NOAA.** As noted, NARA has officially recognized three NOAA environmental data centers as national archives for specific types of data: the National Oceanographic Data Center, National Geophysical Data Center, and National Climatic Data Center.

Web Search Tools. The use of web search tools is another option for data discovery and access. The most well-known of these is, of course, Google, with its Google Scholar service for academic searches. However, many other search tools are now available, some of which are specialized to the needs of scientific researchers.

The web search tools of today are, for the most part, fairly blunt instruments for discovery that can err by returning a high proportion of irrelevant results, by failing to return relevant results, or both (lyrics from *South Pacific* returned in response to an ornithologist's web search for data on tropical birds is an oft-cited example). At their best, web search tools can narrow a user's search to a manageable short list that does not leave out anything important. But they do not offer means of judging the relevance of resources to a user's specific needs or the quality of the data.

A major current development that could radically transform future possibilities for web searching is the so-called "semantic web." While the term encompasses a broad range of concepts, protocols, and technologies—it refers to a vision of future web functionality, rather than to a specific technology or tool—the basic idea is straightforward. Currently, machine-readable metadata descriptions are largely limited to describing whole files or web pages. In the semantic web of the future, web documents would be written in formats that allow extensive machine-readable descriptions of specific elements within files or web pages, such as text, numbers, and images.[52] Thus, the meanings of and relationships among these elements, which currently must be determined by human users through direct inspection, could be "understood" by computers. This would allow much more sophisticated and discriminating search and query capabilities—in effect, the automation of a large part of the painstaking process of sorting through web pages of potential interest to find and combine relevant information.

Metadata Catalog Services

Metadata catalog services ("metcats") are mechanisms for storing and accessing descriptive metadata that allow users to query data based on desired attributes. They can be thought of as "phonebooks" that list data sources and their content and locations.

Metadata catalogs are often used in federated systems, where the actual data sit on remote computers at a variety of locations and a central server with the catalog connects user queries to the dispersed data. The catalogs usually produce very accurate returns on queries for data listed in them. Their limitation is the typically restricted number of databases they list.

The Centrality of Metadata

Web-based discovery mechanisms are currently only as good as the metadata attached to the target data.[53] Clear, comprehensive metadata are the indispensible key to discovery—and subsequent use—of data. At a minimum, metadata need to provide basic descriptive information (such as what the data are about; when, where, and how they were collected; and what format they are in) and retrieval information (such as whether they are open-access or restricted-use; who to contact for access; and so on), as well as labels in data tables that define the precise nature of each field (e.g., type, units, and range). Without

[52] The format most closely associated with efforts to build a semantic web is RDF (Resource Descriptive Format), a data model that according to Gold (2007a) "encode[s] data in a standard that is highly flexible and yields significant economies of scale."

[53] One interviewee noted that the next generation of data miners questions this assertion. They believe that they can and will overcome the metadata need through next-generation semantic ontology creation. This theoretical concept is not yet well-developed in practice, but many luminaries in the field believe it is the future.

such information, a potential user has no way of knowing whether a given data set might be of interest, how it can be accessed, or how to interpret it.

Higher-level metadata are also key to users' subsequent determination of the *quality* and *relevance* of discovered data.

- "Quality" refers to the basic validity, accuracy, and reliability of the data. Data identified by web search tools, data portals, and other discovery mechanisms may fall short on this score, but that may be difficult for searchers to determine.[54] Some data centers and repositories today vouch for the quality of data in their care with respect to certain criteria, and a recent trend has been to include quality-control information as part of raw data tables and metadata—that is, information that tells potential users what quality-control checks have been performed, thus allowing them to better judge soundness. For the future, there has been talk of developing systems of peer review for data quality, comparable to the existing peer review system for publications, as a means of quality control. At present, however, users often must judge quality on the basis of metadata detailing the methods, protocols, equipment, and circumstances of data collection and processing, or on the basis of informal considerations such as the reputation of the collectors and their parent institutions.

- "Relevance" can only be judged in the context of the intended use. Determining the relevance of data to users' specific purposes can be even more difficult than discerning quality. For example, a search may point the user to data that, while of impeccable quality, are nonetheless inappropriate for the intended use. While these discrepancies may be obvious in some cases, in many cases determinations of relevance must turn on careful inspection of contextual metadata.

This leads to the issue of the consistency and mutual intelligibility of metadata schemas in biology. Technical formats are not typically a problem. Rather, the different metadata schemas used with similar types of data can result in quirks such as descriptive fields that go by different names but contain equivalent types of information, or descriptive fields that are not parallel.

Scientists and information-science professionals who work on data-sharing issues at a high level are acutely aware of the problems raised by inconsistent and competing metadata standards, and establishing consistent standards is a high priority in data-

[54] "Bad quality" data do not necessarily imply a lack of competence or care on the part of data collectors. For example, data may have been collected using instruments or methodologies that were the best available at the time of collection, but would not be considered adequate for the needs of current research.

management forums.[55] For example, the development and promulgation of common standards are an important focus in collaborative biology initiatives such as NBII, TDWG, the GEO Biodiversity Observation Network, and the Global Biodiversity Information Facility (GBIF), as well as more general data-sharing forums such as GEO/GEOSS, the Open Geospatial Consortium (OGC), and the Federal Geographic Data Committee (FGDC).

However, individual researchers tend to be less cognizant of and concerned about the issue, and may retain idiosyncratic or unique schemas for any number of reasons, including substantively important ones. Thus, progress has been slow and often tentative; for example, new standards are developed that, when tested and implemented, reveal unsuspected needs and anomalies that limit their general appeal, especially for cutting-edge small science. Interviewees suggested that established and widely applied metadata standards remain the exception rather than the rule in many fields of biology. For example, standards such as Ecological Metadata Language (EML) have been developed for making data gathered from disparate ecological research projects widely discoverable and mutually intelligible, but their adoption by research organizations, let alone their consistent use by individual researchers, is proceeding slowly.

To be sure, in some areas of biology there are well-established metadata standards, such as the Darwin Core schema for systematic biology. And in fields where a limited number of different standards are in wide use, users often have tools or guidelines for cross-walking them. Even in these cases, however, it is impossible to avoid the problem of historical data sets that were documented before standards evolved or were not documented at all.

Use

As noted in the "Definitions and Scope" section above, data are considered "usable" in the context of this study if they are both *technically* and *scientifically* intelligible.

- "Technically" refers to the software, technology, and technology-tool requirements for reading and manipulating the data. (Thus, for example, data in unreadable formats, software, or storage media would not be considered usable.) It also refers to the data-compatibility issues associated with the creation of larger, functional data sets by combining data collected by different researchers in different times and places.

[55] The problem was recognized nearly 15 years ago in a seminal article by William Michener and his co-authors (Michener, et al. 1997), which proposed a metadata framework to address it.

- "Scientifically" refers to whether data are sufficiently defined and documented to allow scientific users to understand them and their limitations (and in turn, to modify or process them so that others might understand them as well). This turns on nuts-and-bolts documentation issues such as whether spreadsheet cells are clearly labeled, whether standard nomenclature and units of measurement are employed, and whether thorough metadata documentation that may affect interpretation has been provided—for example, information on the methods, protocols, and context of data collection.

Gradations of usability exist. At one end of the spectrum are data that function on standard technology platforms, use widely-accepted data and metadata standards, and provide adequate metadata for potential users to determine relevance and avoid misinterpretation. At the other end are hard cases like data stranded on obsolete media and data that are so poorly documented their meaning and limitations are impossible to decipher. In between is a vast gray zone.

In big-science fields where technical, data, and metadata standards are well-established, most data from the present and recent past are readily usable by other specialists in the field, although potential users from outside the field may not find it so easy. On the other hand, in fields where standards are not well-established, there may be obstacles to using others' data, even among specialists within the field. Although possibly dated in terms of some specifics, the following excerpt from a 1997 National Research Council (NRC) report on access to scientific data gives a sense of the issues that arise with respect to usability, all of which are still broadly relevant today:

> *Perennial problems affecting access to data in the observational sciences, for example, include gaps in quality control, incompatibility of data streams, inadequate documentation of data sets, and difficulty in meeting the requirements for long-term retention of data. In the biological sciences, the variety of attributes and qualifiers included with each observation and differences in terminology and usage put a heavy burden on any supplier of data to identify and specify the character of the data precisely enough to prevent misinterpretation.* (National Research Council 1997)

As noted, usability requires some degree of processing to bring data into conformity with applicable standards and to provide the metadata required for correct interpretation. When data are processed well, there is less need (optimally, no need) for external users to consult with the original data collectors. Ideally, data collectors and support personnel undertake systematic data processing as a standard part of research workflows, as is done in big-science fields. It is also possible for others to undertake systematic data-

processing efforts retrospectively—although even here, some minimum amount of descriptive information usually needs to come from the data originator. As discussed, however, the traditional small-science approach to data processing is not systematic; to the extent processing is undertaken, it tends to be case-by-case in response to specific requests.

Beyond the question of the usability of data considered in isolation, there is the question of whether data from multiple sources can be integrated for use together. Aggregating separately usable data sets may require additional processing, and in some cases may not be practically possible. For example, integrating data sets that employ competing metadata schemas requires the translation of nomenclature and data elements from one schema to the other. This may or may not be a task that can be undertaken by automated user tools and, if not, may pose enormous practical and validity challenges. Likewise, integrated data sets collected using different collection methodologies and measurements can pose difficult conceptual or practical questions.[56]

Another important question involves the usability of data across disciplines. Even data that are easily usable by specialists within a particular field may be inscrutable to, or misunderstood by, specialists outside that field. Indeed, interviewees suggested this is the normal state of affairs today. While some fields have made considerable progress toward the adoption of standards that allow general data usability within that field, mutual incomprehensibility remains the norm across fields.

Interest in the use of data from different disciplines is likely to grow along with the growth of scientific and policy interest in the global environmental challenges described in the section on "Forces of Change." With respect to interdisciplinary use, the Holy Grail would be an overarching schema that would make data from natural sciences, social sciences, engineering, and other relevant fields comprehensible to specialists from any of those fields, and capable of integration into a single, workable data set. This would enable, for example, an economist doing cost-benefit analysis of a proposed carbon offset program to integrate contingent-valuation studies of how people value environmental services with domain-science data from diverse fields on how the natural world responds to climate change. Unfortunately, such a schema is a long way from realization. All data are to some extent contextual, and require continual reassessment based on changes in understanding, needs, and models.

[56] For instance, what is the "right" way to integrate a data series of temperature readings taken at the same time every day over several years with a data series of temperatures that represent daily averages calculated from multiple daily readings?

Still, somewhere between the current Tower of Babel and the unattainable Holy Grail is a balance that can increase cross-disciplinary usability without sacrificing the scientific context that gives data their meaning. The study team is aware of at least one effort currently underway to move in the direction of such a schema for interdisciplinary data use: according to an interviewee, the Data Conservancy is trying to develop a high-level "data framework for observation." A major 2007 NSF report on cyberinfrastructure also noted:

> *NSF will continue to promote the coalescence of appropriate [data] with overlapping interests, approaches, and services. This reduces data-driven fragmentation of science and engineering domains. Progress is already being made in some areas. For example, NSF has been working with the environmental science and engineering community to promote collaboration across disciplines ranging from ecology and hydrology to environmental engineering*. (National Science Foundation 2007, p. 28)

At present, however, interviewees suggested that efforts to improve interdisciplinary usability are patchy. Moreover, schemas that would allow for broad integration of diverse data would have to function at a fairly gross level of abstraction; field-specific data idiosyncracies will always remain a part of the picture for those who wish to dig deeper. Interviewees suggested that one way to address this issue involves data and metadata schemas that provide several layers of information, with the topmost layers focused on the general information most likely to be of interest for interdisciplinary purposes, and the deeper layers progressively focusing on field-specific details primarily of interest to niche specialists.

Long-term Preservation

For data to remain accessible over time, active long-term preservation efforts are required, because digital data degrade and become unusable in the absence of such efforts. Even more so than for discovery, access, and use, the small-science approach to data management is incompatible with the requirements for successful long-term preservation.

Facilities for holding digital data are generically referred to as "repositories," but not all repositories are in the long-term preservation game. For example, some organization-level repositories provide little curation and serve mainly as a place to park data so that they are not lost, pending further decisions on their future. Successful long-term preservation requires a specialized and usually very expensive complex of technological

infrastructure, skilled personnel, and systematic workflows capable of reliably ingesting, documenting, storing, quality-checking, and migrating possibly enormous amounts of data over an indefinite period of time.

As a practical matter, there is a great deal of overlap between data-access organizations and data-preservation organizations, with most of the latter seeing the provision of ongoing access as an integral part of their mandate.[57] However, the two roles are conceptually distinct, and at least a few sources discussed the possible benefits of greater specialization among organizations with respect to these two areas. Some major issues and questions associated specifically with long-term preservation are addressed here.

Trusted Digital Repositories

Trusted digital repositories (TDRs) are long-term data-preservation facilities that meet certain criteria, beginning with a commitment to stewardship for an indefinite duration of the data for which they assume responsibility. At this time, there is broad international agreement on the requirements for a TDR, and efforts are underway at the International Organization for Standardization (ISO) and the Consultative Committee for Space Data Systems (CCSDS) to ratify *Audit and Certification of Trustworthy Digital Repositories* (ISO 16363 and CCSDS 652.0-R-1, respectively), which codifies these requirements.

A TDR must comply with the Reference Model for an Open Archival Information System (OAIS), a worldwide archival standard. The OAIS Reference Model outlines a set of guidelines for digital preservation that provides both a functional model for implementing key tasks and an information model for metadata to support long-term discovery, access, and use. It imposes upon compliant repositories such responsibilities as securing legal control of stored data; determining "designated communities" of target users; ensuring that preserved data remain usable by designated communities; and following documented policies and procedures to ensure data are preserved against reasonable contingencies—for example, maintaining backup systems and copies.

In addition to OAIS compliance, TDRs are expected to meet requirements in six broad areas: administrative responsibility, organizational viability, financial sustainability, technical and procedural suitability, system security, and procedural accountability. These requirements collectively aim to ensure that organizations that take on the responsibility for long-term data stewardship do so with a clear understanding of the weight of this commitment, and possess the financial wherewithal, technological and

[57] Some data are preserved without providing external access; they are sometimes referred to as "dark storage."

organizational infrastructure, administrative processes, institutional stability, and other necessary resources. As information-science scholar Clifford Lynch notes, "Stewardship is easy and inexpensive to claim; it is expensive and difficult to honor, and perhaps will prove to be all too easy to later abdicate" (as quoted in Caplan 2005).

The Economics of Preservation

Economic considerations of sustainability loom large in any discussion of stewardship and long-term preservation. A high-profile commission, the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, recently completed a final report that discusses in great depth the economic challenges of long-term preservation (Blue Ribbon Task Force on Sustainable Digital Preservation and Access 2010).

For a repository to qualify as a TDR, it must have a reliable, stable source of funding "forever." Some interviewees suggested that *only* national governments are in a position to credibly commit to funding in perpetuity, although others thought that consortiums of universities, research organizations, and intergovernmental organizations are also able to do so. Some well-endowed individual universities, such as major private and state research universities, may also have the potential to take on this role, as they are seen as stable and permanent institutions. Indeed, in a number of cases, university-based facilities already are, in effect, functioning as TDRs for certain kinds of data, albeit in many cases with support from federal agencies.[58]

However, even when hosted by well-financed parent institutions, there is concern that repositories cannot necessarily rely on discretionary organizational budget allocations alone to support ongoing operating expenses, infrastructure maintenance, and technology upgrades. At the moment, preservation organizations are experimenting with a variety of models to produce a reliable stream of additional, dedicated revenues. These include consortium-membership models, subscription models (much like scholarly journal subscriptions), fee-for-access models, fee-for-service models (charging data originators for services related to the preservation of their data), and various combinations of these.

Preservation Decision Making

Another major issue surrounding long-term data preservation concerns criteria for selecting which data to preserve, and for how long. Preserving everything forever is impossible; the quantity of digital data produced now outstrips available storage space, and the imbalance is growing (Gantz, et al. 2008). But who is to make the decision about what is preserved and for how long, and based on what criteria?

[58] For example, Columbia University hosts one of NASA's Data Centers.

Some guidelines exist for such decisions. First, it is generally agreed that replicability is a key criterion. Data that can be recovered, reconstructed, or regenerated at a reasonable cost are less in need of long-term preservation than data that cannot be. As a practical matter, this means that observational data are more likely candidates for long-term preservation than data derived from replicable experiments or modeling.

Second, the value of data to future generations is a prime consideration, although this can be very difficult to judge prospectively. It also raises the obvious question of "value to whom?" Scientific researchers in a specific field? The scientific enterprise more generally? Commercial firms? Society as a whole?

Third, there is widespread agreement in principle that the scientific community itself must decide what is worthy of preservation in various areas of investigation. In practice, however, this approach raises questions.

- It is not always clear who speaks for a given scientific field. To the study team's knowledge, widely-accepted, formal mechanisms for deciding what data to preserve and for how long do not currently exist in any field.

- Some interviewees suggested that, as a matter of professional culture, scientists are reluctant to weigh the costs of preserving data against the benefits. In other words, their default position would literally be to keep "everything" of interest in their field, which is not a viable strategy.

- If decisions about preserving data in a particular field are made by practitioners in that field, important considerations of interdisciplinary or societal value may be lost. For example, data that may be of limited interest from the perspective of a particular field might be of immense interest to researchers in other fields who can use those data to address global challenges.

In light of the unsettled state of mechanisms and criteria for selecting data for long-term preservation, the Blue Ribbon Task Force offers a recommendation that it calls an "option strategy." It suggests that for data that do not currently have an obvious long-term preservation home (like a GenBank), the default policy should be safe storage at a minimal level of curation sufficient to ensure they are not lost, coupled with periodic reassessment—say, every five years—of the case for continued preservation:

> *When future conditions are particularly uncertain—as they are for prospective future demand for many categories of digital materials—it is often economically justified to make a small current investment that in effect purchases the option to make a choice sometime in the future. … Preservation decision makers facing*

> *uncertain future demand for at-risk digital content should consider purchasing an option to postpone the final preservation decision until future uncertainties are at least partially resolved. … The purpose of the option strategy is simple—to buy time and wait until better information is available about the value of the [data] in question or preservation techniques have become more efficient.* (Blue Ribbon Task Force on Sustainable Digital Preservation and Access 2010, pp. 37-38)

Pressure to build infrastructure for long-term preservation of data is certain to grow (Table 2). Despite the increasing attention to data management, scientists in some fields and at some institutions still find they simply have no safe place to park the data from a research project that is winding down. When grant funding for a research project expires, so does the funding for managing those data, and researchers must turn to repositories at the organizational, field, or higher levels. If such repositories do not exist, the fate of such data is uncertain; they may become tomorrow's at-risk legacy data. While the Blue Ribbon Task Force's "option strategy" offers a possible solution to this dilemma, it cannot work in the absence of "option repositories"—whether at the individual research organization level or the inter-organizational level.

Cross-Cutting Factors

The pressures to improve all phases of scientific data management and sharing are raising a number of issues that cut across the functional areas of discovery, access, use, and preservation. Prominent among these are issues pertaining to human resources, cyberinfrastructure, and economics.

Human Resources

As noted, IT and information-science professionals will increasingly be expected to work closely with scientists in an integrated process of data management that begins with the project design and initial collection of data. For this process to achieve its full potential, changes in the human-resources status quo will be required, particularly with respect to training, hiring decisions, and criteria for performance evaluation.

Hybrid Professionals. Effective management of research data, particularly in the initial stages, often requires personnel with both scientific expertise and IT or information-science skills. The need for such personnel is likely to grow as data management

Table 2: Long-Term Threats to Digital Data Access and Preservation

Threat	Solution
Users may be unable to understand/use data (semantics, format, processes, or algorithms)	Create and maintain adequate representation information[a]
Non-maintenance of essential hardware, software, or support environment may make information inaccessible	Migrate data to replacement hardware, software, and support environments as necessary, using methods that ensure reliability and authenticity of data. Share information about the availability of hardware and software and their replacements and substitutes.
Chain of evidence may be lost; lack of certainty about provenance or authenticity	Bring together evidence from diverse sources about provenance and authenticity of a digital object
Access and use restrictions may make it difficult to re-use data, or alternatively, may not be respected	Maintain ability to deal with digital rights correctly in a changing and evolving environment
Loss of ability to identify location of data	Persistent identifier system that has adequate organizational, financial, and professional backing to survive for the very long term
Current custodian of data may cease to exist	Maintain ability to package information needed to transfer data across organizations and a system to facilitate such transfers
Those trusted to look after digital holdings may fail to do so	Certification process for long-term repositories

[a] "Representation information" is the OAIS term for everything needed to understand a digital object; it includes metadata documentation of a digital object's format, semantics, software, algorithms, processes, and anything else required to ensure usability.

Source: Adapted from PARSE.Insight 2009.

becomes a more central aspect of the research enterprise.[59] Depending upon the field, the project, and the point in the data lifecycle, the scientific competence required might be fairly superficial, or it might be the equivalent of an undergraduate major, a master's degree, or even doctoral-level expertise. Some interviewees and literature sources suggested that in some circumstances, it may be more practical to retrain content specialists in the appropriate IT and information-science skills, rather than the other way around.[60]

At present, it is difficult to find individuals with this mix of skills, and efforts to increase capacity in this area are in their infancy. If organizations are not able to hire such personnel "off-the-shelf," an alternative is to set up internal processes for cross-training staff. This could entail both training research scientists in data management, and training IT and information-science personnel in the basics of the relevant domain sciences.

Professional Balance. Even with scientists taking greater responsibility for the initial stages of data management and an increase in the ranks of hybrid professionals, at some point in their life cycle most data sets still must be handed off to personnel whose primary expertise lies in data management (IT or information science)—especially at the stage of long-term preservation. Information-science professionals are ultimately responsible for ensuring that the data collected by scientists are globally discoverable, accessible, and usable for as long as they are deemed of value.

Some interviewees suggested that the ratio of data-management personnel to research personnel needs to rise in many organizations and fields. The balance between data producers and data managers in most areas of biology is currently skewed toward the former, to the point that more data tend to be produced than can be properly managed.

A professional rebalancing that simply adds data-management capacity to existing research capacity would be the ideal. However, in an era of constrained resources, such rebalancing might at many organizations require trading off scientific research positions for data-management positions, which is a contentious proposition. For deep-seated

[59] Some observers (see, for example, National Science Foundation 2005) distinguish between two types of personnel whose services will increasingly be called upon as the pressure for more systematic data management grows. "Data scientists" are primarily involved in the initial stages of data curation, working with domain scientists; they need to have a relatively strong understanding of the science in the relevant field(s). "Data managers" (or "data curators"), by contrast, are active in the downstream phases, after collectors hand off their data to data-access and -preservation facilities; they do not need to be as conversant with the science.

[60] According to some interviewees, an informal variant of such cross-training commonly occurs at research laboratories throughout the nation, whenever a graduate fellow or post-doc (who may have little or no formal background in IT or information science) is assigned to be the project data manager, and must pick up the relevant IT and information-science skills on the fly.

cultural reasons, this is not a trade-off that most research organizations are inclined to make, as their institutional reputations depend on how much first-class research they produce, not on how well the data underlying that research are curated.[61]

Incentive Structures. As noted, research scientists have few formal incentives to devote time to data management. Both as scholars and employees, they receive credit primarily for analytical publications. Changes in the professional modus operandi to give scholars meaningful credit for data management and sharing are beyond the power of any individual organization, but organizations can choose what they reward in employee performance. If improved data sharing is seen as an institutional priority, organizations will have to consider introducing performance incentives for scientists to devote more time and attention to data management.

Cyberinfrastructure

The technologies and systems that support and enable digital data management and sharing comprise cyberinfrastructure.[62] A complete catalog of the elements of cyberinfrastructure is impossible, not only because of the number and complexity of the relevant assets, but also because of unavoidable ambiguity about the scope of assets covered by the term. However, cyberinfrastructure is usually taken to include a set of overlapping elements that includes, but is not necessarily limited to:

- Core technologies for data processing and storage;[63]
- Data repositories;
- Databases;
- Communications technologies and servers that link digital assets to users;
- Software and user tools for data discovery, access, and analysis;
- Virtual structures, such as platforms that provide shared workspace across organizations, clouds, grid computing, and social networking media;

[61] At least in small-science fields, research scientists have typically been regarded as the personnel who add value, while data-management personnel have been seen as "hired help."
[62] Data management and sharing are by no means the only parts of the scientific enterprise supported by cyberinfrastructure; for example, elements of cyberinfrastructure are geared toward supporting computationally intensive data analysis. However, the focus in this study is on those elements of cyberinfrastructure that pertain to data management and sharing.
[63] Data collection technologies may also be considered part of cyberinfrastructure, to the extent that these are automated and tied directly into the cyberinfrastructure network.

- Protocols and middleware that enable elements of the system to communicate;
- System architectures that define relationships and workflows across components of cyberinfrastructure; and
- Organizational relationships and workflows that allow the system to function reliably.

Cyberinfrastructure has been a major focus of discussion over the past decade, particularly since the landmark Atkins Report of 2003. As envisioned by the Atkins Report, cyberinfrastructure is more than the sum total of individual organizational assets. Rather, it would comprise an interoperable network that functions seamlessly across organizations and, eventually, across fields. The report holds up the internet as a model for cyberinfrastructure: a network of distributed assets that functions so cohesively that end users are not aware of the enormous feats of engineering and technological prowess that tie them together.

At the highest level, cyberinfrastructure consists of distributed generic assets and systems that are not inherently tailored to the needs of a specific field, and that in principle can provide the backbone for digital interaction across fields, organizations, and researchers of all descriptions. At the next level are distributed assets and systems that are custom-designed to serve a specific field or user community.[64] At the lowest level are the assets and systems of individual organizations, which are typically a mix of generic and customized elements. At all levels, systems run on a mix of open-source and commercial/proprietary software.[65]

Without doubt, the scientific research enterprise would be served poorly by cyberinfrastructure that consisted of a myriad of isolated organization-level assets that cannot "talk" to each other. Thus, the Atkins Report and subsequent discussions of cyberinfrastructure have stressed the overriding need for designing interoperability into

[64] Note that elements of cyberinfrastructure initially designed to serve one field can be, and often are, adapted for use in other areas.

[65] There seemed to be agreement that open-source software is superior from the perspective of cross-organizational interoperability, and collaborative library and archive infrastructure development projects tend to build core systems around open-source models such as DSpace, Fedora, and iRODS. Nonetheless, organization-specific considerations such as existing enterprise architectures, specific user needs, and whether in-house personnel are available to customize and support open-source software might sometimes favor proprietary solutions. It is possible, however, to create federated networks based on open-source software from individual organization systems based on proprietary products; this is the case, for example, with NARA's Transcontinental Persistent Archive Prototype. Some interviewees noted that the Smithsonian is currently short on the personnel needed to customize and manage open-source software (which typically requires more in-house support than commercial software, because the firms that supply the latter also supply user support services).

the system—that is, designing individual components at all levels to work together as a seamless, reliable network. At the same time, organizations clearly must consider their own unique circumstances, as well as the greater good of interoperability, in making decisions about organizational cyberinfrastructure. While the two are not necessarily incompatible, neither do they always dovetail.

This raises the issue of the need for some degree of coordination among organizations in planning and designing their organizational cyberinfrastructure assets. The Atkins Report envisioned NSF, with its central position in the U.S. scientific community, as a focal point in guiding the development of U.S. cyberinfrastructure. One major result of the report was the establishment of a new Office for Cyberinfrastructure within NSF to orchestrate the Foundation's activities in support of such infrastructure, which had been scattered across a number of NSF offices, including both domain-science funding units and units dedicated specifically to the application of cutting-edge computing to scientific work.[66] NSF has subsequently issued a number of high-profile reports, funded two DataNet pilot programs focused on cyberinfrastructure (DataONE and The Data Conservancy—see text box below), and provided support for numerous other cyberinfrastructure initiatives and assets.

However, while NSF has been active in providing start-up funds and other support for cyberinfrastructure initiatives, some interviewees suggested it is extremely reluctant to commit to ongoing support for permanent cyberinfrastructure upon completion of such initiatives. This arises from an institutional philosophy that money directed to infrastructure maintenance is, in effect, money diverted from research. In the words of one well-known information-science thinker interviewed by the study team:

> *NSF has always prided itself on **not** funding infrastructure, and dumping all of its money into cutting-edge science. They like to pay for research, then move on. There are a few pieces of infrastructure they might grudgingly pay for—the polar facilities, a few big scientific instruments, and a couple of databases. Much as it pains them to admit it, they do help to pay for the Protein Data Bank. But they really hate to do this; it's against everything in their culture. The universities are getting uncomfortable that NSF is getting more and more prescriptive about the need for systematic data sharing and management plans, but on the other hand says, "Don't talk to us about funding data after the grant runs out." … So NSF is in effect saying that if you are going to get any permanent funding for this function, it is going to come out of your overhead rates, in the same way that you*

[66] Cyberinfrastructure efforts continue to take place in NSF offices other than the Office of Cyberinfrastructure, but according to interviewees at NSF, these now are subject to a greater degree of internal coordination.

NSF DataNet Program

NSF created its Office of Cyberinfrastructure to serve as a focal point for its efforts to build IT infrastructure to support the scientific enterprise. One of the Office's major initiatives is the DataNet grant program, which aims to foster the development of a set of exemplar research data infrastructure organizations. The grant solicitation (NSF-07-601) for the program notes:

> *The new types of organizations envisioned in this solicitation will integrate library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise to:*
>
> - *Provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline;*
> - *Continuously anticipate and adapt to changes in technologies and in user needs and expectations;*
> - *Engage at the frontiers of computer and information science and cyberinfrastructure with research and development to drive the leading edge forward; and*
> - *Serve as component elements of an interoperable data preservation and access network.*
>
> *By demonstrating feasibility, identifying best practices, establishing viable models for long-term technical and economic sustainability, and incorporating frontier research, these exemplar organizations can serve as the basis for rational investment in digital preservation and access…, paving the way for a robust and resilient national and global digital data framework.*

NSF plans to make five DataNet grants, two of which it has already awarded: DataONE (based at the University of New Mexico) and The Data Conservancy (based at Johns Hopkins University). The five-year grants provide up to $20 million annually, with the possibility of a five-year renewal.

DataNet grants are specifically *not* intended to provide permanent funding for cyberinfrastructure. Rather, they are oriented toward capacity building by providing start-up funding for pilot projects. Although a five-year renewal of the initial grant is possible, annual funding levels in the second five-year period are to be progressively reduced, with no funding to be provided beyond the tenth year. Thus, DataNet grant recipients are expected to have plans for maintaining sustainable funding after the expiration of the grant.

> *have overhead rates to support, say, libraries.*[67] *But there are all kinds of problems with that, starting with how you correlate the amount of data your institution is trying to care for with your overhead rate. Just because you are unfortunate enough to have a researcher who is really into creating massive data sets, should that be reflected in an unusual overhead rate going forward?*

For example, NSF now wants to curtail funding for that organization because the concept has been proven and GBIF has in effect become an infrastructural program. How to transition from grant-supported to ongoing operational data management following the conclusion of grants is a hotly debated issue.

The Atkins Report estimates that about 65 percent of the total budget for cyberinfrastructure goes for "recurring costs of professional staff and researchers, as opposed to the acquisition of hardware and software," and that "a substantial portion of these recurring costs is devoted to developing, maintaining, distributing, upgrading, and supporting software" (National Science Foundation. Blue-Ribbon Advisory Panel on Cyberinfrastructure 2003, p. 81).

If this Atkins Report estimate of the breakdown of cyberinfrastructure costs is even approximately correct, it suggests that the funding model for cyberinfrastructure must be very different from the model that applies to physical infrastructure. The latter involves large up-front construction costs, coupled with relatively modest and predictable recurring maintenance costs over a long asset lifetime. By contrast, while the costs of initially installing elements of cyberinfrastructure can also be considerable, the funding model for cyberinfrastructure must place much greater emphasis on relatively large and often less-predictable flows of funding over time. This is not only because cyberinfrastructure typically requires considerable ongoing maintenance and user support, but also because major parts of it are in effect mutable works-in-progress, subject to regular reassessment, testing, redesign, and patching. Further, the useful lifetime of cyberinfrastructure elements tends to be both shorter than the lifetime of traditional physical infrastructure and less predictable, being subject to the inexorable but uneven process of technological progress. As the Atkins Report puts it:

> *The continuing exponential improvement of the hardware underlying cyberinfrastructure provides accelerating opportunities for exercising creativity, but can be daunting in terms of managing the attendant rapid obsolescence of facilities. Maintaining leading-edge cyberinfrastructure requires continuing*

[67] Note that at the Smithsonian, the Office of the Chief Information Officer (OCIO), which, as discussed below, has recently been playing a major coordinating role in addressing scientific data-management issues across units, does not receive funds from the overhead in grants.

> *investment, not one-time purchase. Cyberinfrastructure ("bit-based") investments differ from most other [physical] kinds. Delaying the start of construction of an accelerator or telescope or research vessel normally increases the cost of the acquisition. Frequently, the opposite is true for computing equipment, which becomes cheaper by waiting a year but becomes obsolete soon thereafter. One way to quantify this is through replacement schedules. Major research equipment may have a realistic lifetime of 10-25 years. The appropriate replacement interval for information technology at the frontiers of performance is closer to 3-5 years.* (National Science Foundation. Blue-Ribbon Advisory Panel on Cyberinfrastructure 2003, pp. 40-41)

While everyone agrees that constructing the sort of cyberinfrastructure envisioned by the Atkins Report is an enormous challenge, there are some differences of opinion about the precise nature of the challenge. Some interviewees tended to downplay the technical challenges, noting that while technology might be expensive and complicated, the more daunting obstacles arise from culture, politics, and attitudes—for example, getting researchers, organizations, and nations to agree on technical standards and software when doing so might mean sacrificing some of their own preferences to the greater good of interoperability. One interviewee at a major collaborative data center project, however, offered a different perspective on the relative importance of technical challenges:

> *There are major technical obstacles here. I stay up at night when I hear people say, "Oh, how hard is this? It can't be that bad." No, it* ***is*** *that bad. We don't even know if the sort of systems we have today will actually work properly for [their intended purposes]. We have vendors who normally are quite happy to tell us, "I have a solution, and I'll sell it to you and your life will be good," who now come to us and say, "Oh, boy. We don't really [have a solution]. Can we work with you and figure this out together?"*

Economics

As already discussed at several points in this report, it is costly to curate, preserve, and provide access to scientific data, and many key questions about who pays have yet to be answered. For example, will data access be explicitly subsidized by governments as a public good? If so, will this be done in the United States through existing organizations such as NSF, or through new organizations dedicated specifically to data access and preservation? Can access providers identify revenue streams that balance the moral demands of open access with the practical imperatives of financial sustainability? Moreover, while to some extent research organizations are increasingly accepting the

need to treat data management as a routine overhead cost, discussion continues about how much money can and should be made available for it.

Some distinctive economic issues in the management of scientific data that affect the question of who pays were discussed in the sections on "Forces of Change," "Long-Term Preservation," and "Cyberinfrastructure." They include:

- Economies of scale—in general, it is exceedingly inefficient to fragment the processes for curating, providing access to, and preserving data across hundreds of organizations and thousands of research teams.[68]
- The need for reliable long-term funding streams for preservation; and
- The differences in funding flows for traditional physical infrastructure versus cyberinfrastructure.

In addition, a number of other relevant economic factors apply to some or all of the stages of data curation, access, and preservation. These are briefly discussed here.

Public Goods. In the parlance of economics, a "public good" is one that has certain features that lead private producers to undersupply it relative to the socially optimal level of production.[69] Institutions that provide access to basic scientific data have a strong public goods character, in the sense that the potential benefits of such institutions (accelerating scientific progress within a field; enabling synthetic research across fields; providing information that improves management of natural resources; enabling commercial spin-offs in some cases; and so on) are distributed widely across the research community and the general population, and are by no means confined to those who actually access and work with the data. Thus, while it may be technically possible to limit access to paying users—this is the fee-for-access data center model—the expected result would be a socially sub-optimal level of use of these data.

This public goods character provides a case for government subsidization, or even outright government provision, of scientific data-access services. Indeed, the public

[68] However, IT history is replete with examples of the failure or overselling of centralized solutions that removed data management responsibilities from collectors and others close to the source who understand, at a deep level, how the data are used. In addition, considerations of security, such as firewalls, can raise technical barriers that hinder collaboration in data management and data sharing among scientists from different organizations.

[69] These features are (1) non-excludability (the inability of producers to control access to their product by those who do not pay for it) and (2) non-rivalry (the ability of a product to be shared by any number of consumers simultaneously, without reduction in the benefits that any individual consumer enjoys from that product). The "socially optimal" level of production is defined as the level at which the marginal cost of producing an additional unit equals the marginal benefit (willingness to pay) to consumers.

goods argument for supporting *access* to basic scientific data is identical to the public goods argument for supporting *production* of these data, and the latter has long been the widely-accepted rationale behind the decision of the federal government to generously fund basic scientific research through NSF and other channels.

Network Effects. Network effects arise when the value of a good or service is directly correlated with the number of users. For example, Alexander Graham Bell's first working telephone prototype in 1876, while clearly a marvelous invention, had little value to potential users until telephone lines were laid to form a network through which a large number of widely dispersed individuals could communicate. Moreover, a telephone network that connects 1,000 homes and businesses is less valuable to users than a network that connects 100,000, which is in turn less valuable than a network that connects 100,000,000, and so on.

Where network effects are operative, there is a tendency for dominant suppliers to emerge—for example, Microsoft Windows and Active Directory for institutional networking; internet giants such as Facebook for social networking, YouTube for media sharing, and eBay for online auctions; AT&T for voice telephony before it was broken up by court order; Google for web searches; and the dominant technical formats for consumer products such as DVDs. While in some cases this market dominance may be buttressed by anti-competitive practices, the more fundamental reason for it is that users *want* a dominant supplier because it increases the value of the network to them.

At the same time, dominant suppliers and standards tend to emerge in such markets only after a process of shaking out among competing candidates. Prematurely forcing the use of standards may therefore stymie innovation and lead to sub-optimal outcomes in the long run. The adoption of a dominant standard also has costs in the short run, such as those associated with writing off investments in technologies that are not interoperable with chosen standards.

Network effects are relevant to the "market" for scientific data centers. Ideally, researchers would like to know there is a single place they can go to find a comprehensive selection of a particular type of data, processed for usability and coupled with analytical user tools. In a few scientific fields, a dominant supplier has emerged or appears to be emerging, examples being the National Virtual Observatory (astronomy), Protein Data Bank (protein structures), GBIF (species/specimen data), and GenBank (genomics). However, in the case of most biology research, data centers and data portals provide only fragmentary coverage, and no dominant organization has emerged, a situation that is less than optimal from the perspective of users. Network effects also apply to standards for metadata and data-management practices. Off-the-shelf standards

that are widely understood throughout the relevant professional community reduce the need for researchers to expend resources formulating their own approaches to data management and reprocessing data in the face of external requests for their use.

At the same time, the costs of imposing standards prematurely must not be overlooked, especially with regard to cutting-edge areas of scientific research where technologies and practices remain in flux. As with the related issues of centralization versus autonomy, the economics of networks suggest the need to strike a balance between the benefits and costs of standardization, rather than looking for definitive, once-and-for-all solutions.

Cost Accounting. One of the most frustrating issues that confront those who wish to analyze or plan data-management services and infrastructure is that it is exceedingly difficult at this time to get reliable cost figures. Despite some promising work, estimates of how much it costs an organization to provide a given level of data curation, access, or preservation remain largely guesswork. This issue is covered in some detail in the 2008 Interim Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, which summed up the situation as follows:

> *For the most part, when seeking to develop detailed cost assessments, organizations have only had their own data to fall back on, and this is reflected in the literature by cost models and assessments that are largely atomistic. Studies also structure themselves differently; they define costs differently and assign different units of measurement; different formats are captured; and decisions regarding which costs and cost adjustments to include and exclude vary from project to project. When publishing project updates authors typically do not create economic "crosswalks" between their and others' frameworks. Even for those projects that explicitly build on earlier work, it is clear that within any given project the costs captured are generally focused upon only a small subset of activities within the digital preservation lifecycle (for example, storage costs). In short, the structure for previous studies rarely supports direct comparisons.*
>
> *Nonetheless, over time, the discussion has become more sophisticated. … In particular, two recent projects highlight the increasing economic sophistication we are now seeing: (1) the cost model developed by the LIFE (Life Cycle Information for E-Literature) project, and (2) the recently published model developed by Beagrie, Chruszcz, and Lavoie (2008).* (Blue Ribbon Task Force on Sustainable Digital Preservation and Access 2008, pp. 36-37)[70]

[70] The works by Ayris, et al. 2008 and Beagrie, Chruszcz, and Lavoie 2008 referenced in the quote are listed in Appendix A. Bibliography.

Findings: The Smithsonian

Background

The general consensus of interviewees is that the Smithsonian encompasses some important oases of innovation, initiative, and leadership in data management and sharing, for example in systematics, genomics, and Geographic Information Systems (GIS).[71] As an Institution, it is a major player in key international and national initiatives, such as GEO and USGEO, the Encyclopedia of Life (EOL), the Consortium for the Barcode of Life (CBOL), IWGDD, and GBIF. It has taken significant steps in recent years to strengthen data sharing and management. Some units, notably OUSS and the Office of the Chief Information Officer (OCIO), have begun to work collaboratively to develop a more systematic and comprehensive approach.[72] A number of individual researchers have, on their own initiative, been very active in important efforts related to their own specific fields.

That said, there was a sense among many—although not all—internal and external interviewees that much of the activity relating to data management and sharing at the Institution is fragmented and opportunistic, and depends too much on the efforts of committed individuals rather than organizational strategy. Moreover, the OP&A study team found little evidence that the Smithsonian was systematically keeping itself informed about, or taking advantage of, developments in scientific data management and sharing in the outside world. On the positive side, the comments of external interviewees indicated significant interest in teaming up with the Smithsonian to further data management and sharing.

Collaborative Engagement

There was a sense among many interviewees that a research organization of the Smithsonian's stature could and should be playing a greater, more strategic, and more systematic role in engaging with other organizations to address the pressing data issues of

[71] The 2010-15 Smithsonian Information Technology Plan explains GIS as follows:
A GIS organizes geographically referenced information into a visual form. It can combine map, satellite, and sensor information sources with spatial databases for spatial and temporal analyses that otherwise would be difficult. It also automates most of the archiving and display operations typically required to interpret data obtained in a geographic context. GIS databases constitute baseline data, the worth of which increases with reuse. (Smithsonian Institution. Office of the Chief Information Officer 2010, 132)

[72] OCIO has a vision for a more coordinated and systematic approach to scientific data management, but it does not currently have the resources (people and funding) to implement it as a whole. Instead, it pursues elements of the vision as opportunities and funding become available.

the day and to strengthen its own data management. One important symptom of the Smithsonian's limited engagement in this regard is that the corpus of staff who participate as official Smithsonian representatives in relevant collaborative forums is very small, which undermines the Institution's ability to influence and engage with these organizations. For example, only one official Smithsonian representative attended the 2009 USGEO/GEOSS meeting in Washington, DC, despite the numerous plenary sessions, working groups, committee meetings, and other meetings on aspects of data management and sharing of direct relevance to the Smithsonian. Likewise, the Smithsonian funds only a single representative to official GBIF meetings, even though a number of unit-level staff are actively working on their own initiative to support GBIF's work. As of 2010, the Smithsonian did not have any representation at all at the annual ESIP conference, which is a major national nexus of diverse organizations engaged in collaborative solutions to digital data-related issues.

Legacy Data

As noted, when it comes to data management, most of the biology research done at the Smithsonian has closely followed the small-science approach. For the most part, individual researchers and research teams have formatted and managed their data based on the needs of their projects, with little attention to future re-use or preservation. As a result, the study team heard stories about Smithsonian scientists retiring or dying and leaving behind masses of impenetrable data that lacked any of the prerequisites for discovery, access, and usability. While departments have sought to manage some scientific data after the departure of data collectors, the approach has generally been to warehouse them in the state in which they were received, rather than to systematically curate them to maximize the potential for sharing.

The study team has heard repeatedly (not only for this study, but also on other projects dealing with Smithsonian science) that one of the Institution's comparative advantages vis-à-vis most other scientific research organizations is its ability to undertake long-term, sustained monitoring and experimental biology research. Organizations whose researchers are more tied to the time horizons of temporary research grants are less able to collect data continuously over lengthy or indefinite timeframes. In some cases, the Smithsonian has also provided a home for historical data sets from government agencies that no longer have any direct use for them. As a result, the Smithsonian has, over the years, acquired many data sets covering long periods of time that are now of potentially great interest to a scientific community increasingly concerned with long-term environmental change. Making the data available in usable form will, however, require much greater efforts in data management and sharing than have hitherto been in evidence.

Neither the Smithsonian nor any of its individual units has developed formal repositories for scientific research data, although OCIO is in the early stages of planning a TDR and the Smithsonian Institution Archives (SIA) is examining whether it might expand its role in the preservation of digital scientific data.[73] SIL has historically not regarded the curation or preservation of digital scientific data as a part of its mandate. Because these data are not considered part of the Smithsonian National Collections, they are not subject to the policies and standards of SD 600 (Collections Management) that govern such collections, and no comparable policy exists regarding the curation and preservation of digital scientific data.

The unsystematic state of data management at the Smithsonian has resulted in very large amounts of legacy data. Exactly how much is impossible to say, as there is no Smithsonian-level, or even unit-level, inventory. Many interviewees suggested the problem is considerable, and is poised to become much worse, given the rapidly increasing rate of data collection and the large cohort of Smithsonian scientists likely to retire in the next decade or so. There was consensus among interviewees that the Smithsonian needs to begin systematic documentation and at least temporary network storage (if not full-service preservation) of these data sooner rather than later, and to take measures to prevent further accumulation. One interviewee's comments were typical:

> *The Smithsonian continues to pile up more legacy data; it's not getting ahead. … To get ahead, the main challenge is more staffing. You also need to involve [researchers] up front so they budget for data storage, figure out how to create the assets, what standards they're going to follow in generating them, and how much data will be generated.*

Discovery, Access, and Usability

Discoverability of and access to Smithsonian scientific data are highly variable. At present and for the foreseeable future, there is no central listing or catalog of what data the Smithsonian has, what state they are in, which are accessible or could be made

[73] SIA collects, preserves, and provides access to historical data that meet the criteria set forth in its appraisal methodology. See the discussion of SIA below.

accessible on demand, and which have the highest priority for access and preservation.[74] At present, many Smithsonian data are not discoverable—for example, forgotten legacy data or active data that Smithsonian scientists are still analyzing for their own publications. Although scientists may be willing to share their data upon request, the existence and availability of data are not typically advertised on unit or departmental websites, or listed on federal portals such as data.gov and Science.gov. Even when metadata descriptions of data sets are posted on a website or portal, accessing the data may still require contacting a researcher. And there is no guarantee the data set will be in standard formats or will have the additional metadata necessary to make them easily usable. Even in the small number of cases where the data themselves are directly accessible via a website or portal, they may not be usable by external parties without consulting the data collectors. For the most part, neither the units nor the central Smithsonian administration have yet issued clear policies or guidelines for researchers on data management or sharing, or paid much attention to these functions. There are, however, signs that this attitude is changing slowly.

OCIO and Pan-Institutional Initiatives

In the past several years, OCIO has become the focal point for pan-Institutional efforts to develop policies, priorities, and guidelines for digitization and the treatment of digital assets.[75] It has moved forward in a collaborative manner through committees and work groups with broad representation across Smithsonian disciplines and units (including science units, SIA, and SIL).

[74] Inventories have been attempted in the past, but have failed to meet expectations because they did not lead to additional resources for dealing with the issues identified through them—i.e., additional commitments of financial and personnel to support the management and organization of the data. One interviewee thought that undertaking another inventory would not be a good use of resources and would create ill-will within the research community, especially if once again resources for dealing with priority data were not made available. This interviewee also thought that the setting of priorities alone would also be a major, time-consuming process. An alternative would be for the Smithsonian to start with a concrete commitment of resources to the management of databases that most people would agree need curation in order to generate buy-in and develop capacity in data management, and then decide whether to proceed with a full-scale inventory. The interviewee added that if such a pilot endeavor showed the value of better data management, it might generate some of the needed self-correction in how data are handled on the part of individual researchers and teams.

[75] Although the Smithsonian uses the term "digitization" to encompass both the management of born-digital data (for example, the Smithsonian digitization strategic plan for 2010-2015 defines "digitization" as including "data collected by an electronic measuring device") and the creation of new digital assets from non-digital collections such as three-dimensional objects, photographs, and journals, there has been a tendency to associate the term primarily with activities in the latter area.

Biology units have been represented in the organized pan-Institutional conversation on digitization that began in 2006. For example, the pan-Institutional Digitization Strategic Plan Committee that issued the 2010 report *Creating a Digital Smithsonian* (see below) included two staff from NMNH and one from NZP, and the permanent Digitization Program Advisory Committee includes one representative from NMNH. (In fiscal year 2011, the latter group will create a formal charter that specifies the Committee's composition and how selections will be made.) Some interviewees thought, however, that among the science units, SAO was overrepresented and the biology units underrepresented in the relevant forums. One result, they believed, is that the unique data-management and IT needs of biology were not being adequately understood and addressed.

In addition, interviewees' comments revealed a concern that the pan-Institutional digitization program concentrated primarily on the National Collections held by the museums, libraries, and archives, rather than on digital scientific research data. One reason may be the higher public profile of the physical collections compared with research data. Another reason, according to one interviewee, is that the collections digitization process is supported by an extensive network of existing resources, policies, and systems, including collections information systems (CISs) and CIS managers, the SD 600 policy for collections management, and many personnel within individual units whose work focuses in whole or in part on the digitization of physical collections. There is no analog for these resources on the digital research data side.

The existing and rapidly growing data-storage challenge is a key factor driving OCIO's recent efforts. It is clear the accelerating data deluge will quickly overwhelm the Smithsonian's digital storage capacity without a combination of proactive investments in infrastructure and a more rational approach to data management. This remains a strong concern, even with the steps being taken to address it. As one interviewee put it:

> *Data stewardship roles need to be defined within the organization, and we are reaching the point where this is beginning to become clear to everyone. We … need to get more systematic about data management. We can't treat every individual data set as God's own junk drawer, where we can throw every random, partially-computed instance of data and keep it in perpetuity. Somebody needs to say, "We need to make a DVD of that and cut it off the hard disk." But that discussion is just emerging as we are coming to terms with the Smithsonian's voracious appetite for digital storage.*

The best overview of pan-Institutional digitization and IT developments is OCIO's Smithsonian Information Technology Plan (SITP), which covers a five-year time horizon

that is updated on a rolling annual basis (Smithsonian Institution. Office of the Chief Information Officer 2010).[76] The SITP at the time this report was researched, covering fiscal years 2010-15, laid out the current status of a variety of IT functions, systems, and initiatives across the Institution, as well as OCIO's plans for the coming years.

Intra-Organizational Coordination

External interviews and the literature both suggest that effective scientific data management within (as well as across) complex research organizations usually calls for coordinated efforts among research, IT, library, and possibly archive divisions, as well as support from the central administration. The actual mechanisms for such coordination, however, vary from organization to organization. Because widespread concern with data management is a relatively new phenomenon and systematic arrangements are just beginning to emerge, it is difficult to identify norms or best practices.

One interesting contrast between the Smithsonian and the external world that struck the study team is that libraries are spearheading collaborative internal efforts in this area at a number of major research universities. The study spoke with some of these, including the Massachusetts of Technology (MIT), Columbia, and Purdue. In these organizations, the information-science work is done primarily by the libraries, with IT departments mainly providing their traditional technical-support functions. Further, the libraries have taken the lead in engaging with researchers to explain the importance of data management and preservation, discussing the support that the libraries can provide, and developing tools to facilitate the data-management work that must still be done by researchers. At these organizations, involvement of organizational archives has been marginal.

By contrast, at the Smithsonian, OCIO has deliberately expanded its role beyond the technical-support function to play a key part in leading and coordinating efforts to address scientific data-management and -sharing issues. SIA is very interested in this area as well, as it considers which scientific data sets to treat as Institutional records. However, SIL has emphasized the "downstream" parts of the scientific enterprise, such as published results, and overall has paid less attention to the curation and preservation of digital data per se—although SIL interviewees did indicate a growing interest in linking online catalogs of Smithsonian research publications to the underlying research data sets. In addition, SIL has been very active in ongoing efforts to digitize the biology legacy literature in its collections and those of peer institutions; the BHL initiative, discussed later in this study, is a major instance of this.

[76] The SITP does have limitations as a source of Smithsonian-wide information. Its broad scope necessitates some trade-offs in terms of detailed information about particular initiatives.

The following sections discuss some of the major OCIO-coordinated pan-Institutional efforts relevant to the subject of this study.

Scientific Computing Needs Assessment

One of OCIO's early initiatives to engage more systematically in scientific research support was a 2004 assessment of the computing needs of the science research community at the Smithsonian. The survey included questions about the amount of extant data; how much was likely to be generated over the coming 10 years; how data are stored and backed up; what each unit's primary IT needs were; and what the main obstacles were to data management. While not all units comprehensively responded to the survey, the information collected nonetheless provided a valuable baseline for consideration of future priorities.

Digitization Strategic Plan

The Smithsonian digitization strategic plan for 2010-2015 lists three major goals. The first is to "Provide unparalleled access to Smithsonian collections, research, and programs by creating, managing, and promoting the Institution's digital assets" (Smithsonian Institution. Digitization Strategic Plan Committee 2010, p. 11). Objectives under this goal include:

- "Protect and enhance the value of all Smithsonian digital assets through coordinated digital asset management"—which entails measures such as identifying existing digital assets at the unit level; defining criteria for creating new digital assets; developing plans for life-cycle management of digital assets to ensure access and preservation; building partnerships where needed; constructing a "Smithsonian common data model"; clarifying any restrictions on access and use of digital assets; and developing tools to facilitate use.

- "Freely exchange Smithsonian digital assets, regardless of the systems on which these assets reside, by developing the necessary information technology infrastructure"—which would involve surveying existing and planned IT systems; implementing technical best practices and standards; assessing the feasibility of shared facilities; determining infrastructure and systems requirements for Smithsonian-wide interoperability; developing a methodology for projecting digital storage and back-up needs; and ensuring that TDRs are available for digital assets requiring long-term preservation.

- "Raise awareness and increase use of Smithsonian digital resources, both within and outside the Institution."

The second goal of the plan is to "integrate digitization into the core functions of the Smithsonian." The preamble to this goal notes that the Institution "will move digitization from an activity handled differently in each museum and center to an integrated Smithsonian digitization program that meets both internal needs and external expectations" (Smithsonian Institution. Digitization Strategic Plan Committee 2010, p. 12). Objectives of this goal include:

- "Create a pan-Institutional policy to guide the Institution in its digitization activities"—including both Institution-wide principles and criteria for the division of labor on digitization between the units and the central administration.

- "Cultivate an internal culture that embraces digitization and sharing of Smithsonian collections, research, and expertise"—including the development of a central mechanism for communication and exchange; the provision of resources for digitization needs; and the creation of rewards and incentives to promote collaboration on digitization projects.

The third goal pertains specifically to resources: "Through novel, innovative approaches, secure sufficient resources and build capacity to create and sustain a digital Smithsonian" (Smithsonian Institution. Digitization Strategic Plan Committee 2010, p. 13).

The 2010-2015 digitization strategic plan is a high-level framework document whose scope includes but extends far beyond scientific data management and sharing. OCIO, individual units, and various pan-Institutional committees are now beginning to undertake concrete steps toward achieving the plan's goals, including drafting two Smithsonian directives on, respectively, management of and access to digital assets (Smithsonian Directive [SD] 609, "Digital Asset Access and Use," and SD 610, "Digitization and Digital Asset Management").[77]

Although internal interviewees for this study generally applauded the new pan-Institutional, strategic approach to digitization, some thought, as noted, that the specific needs of biology researchers had not been well-represented on the pan-Institutional digitization working groups and committees. For example, while representatives from all the relevant biology research units with the exception of SERC had participated in the groups, their representatives came primarily from the IT, biodiversity informatics, and administrative areas, rather than the research side. Why this has been the case is

[77] As of this writing, the directives had not been finalized and adopted.

unclear—one interviewee suggested that it reflected the lack of incentives for researchers to take part in data-management efforts.[78]

While the digitization strategic plan and associated activities implicitly cover scientific research data, as noted, in practice the focus has been mainly on the National Collections. In the science realm, this means the physical collections of NMNH and NZP, both of which are recognized collecting units under SD 600 (SERC and STRI are not). Some research-unit interviewees suggested that the emphasis on collections digitization has been detrimental to their units' digital scientific data needs. As one put it:

> *My worry, ever since this digitization initiative began several years ago, has been that [our needs] are not being taken enough into consideration. Solutions are going to be designed that are really museum-centric, and don't take into account what [a research unit] is all about. … We have to be involved in whatever models are being created, so we don't have problems integrating ourselves into those models.*

Another noted:

> *[Managing digital research data] is a big, complicated job, and it may need to be done differently than the object folks. The Smithsonian takes pride in claiming millions of objects, but we also have millions of data.*

Proposed Smithsonian Institution DataNet

The Smithsonian Institution DataNet (SI DataNet) project,[79] which envisions a central repository for Smithsonian scientific data, is the central digitization initiative most directly relevant to the subject of this report. Indeed, the 2010-15 SITP description of the purpose of the proposed SI DataNet reads like a summary of this study's findings:

> *The world of Science, Research, Engineering and Education [is] increasingly digital and increasingly data-intensive. Currently there is no central repository for Smithsonian scientific data. Instead, much of it is kept on the scientist's hard drive or non-archival CDs and DVDs. In order for scientists to understand, interpret, and use data collected by another researcher it must be cataloged with [descriptive metadata], indexed, and stored in such a manner that it is easily recoverable. Among the worldwide scientific community there is a growing need to develop scientific curation policies, procedures, and systems to preserve data*

[78] By contrast, SAO regularly sent researchers to participate in these groups and committees.
[79] There is no connection between the Smithsonian DataNet project and the NSF DataNet initiative discussed above.

> *for the long term—centuries versus years. Any such polices and systems would need to address … [data interoperability] across not only data sets, but also across disciplines.*
>
> *The digitized data that is the product of current and future research activities can be used as the basis for new hypotheses and research. This extension of one research activity to another and the reuse of the digital products or data sets represents a challenge to the scientific community, to manage the provenance, provide interoperability, and ultimately maintain the structure and integrity of the data sets. … DataNet will ultimately make available data sets to the public in an easily accessible manner, whether it is another scientist, an elementary science class, or an amateur scientist. The data sets will be managed … as a collection of objects that … can be stored and retrieved in a variety of digital formats through time and technology changes.* (Smithsonian Institution. Office of the Chief Information Officer 2010, p. 135)

While the SITP lists the SI DataNet project as "partially funded," at the time of this writing no dedicated funds had been allocated to the program. Its "partially funded" status reflected the fact that OCIO had allocated some staff time to pilot efforts to test approaches to, and demonstrate the potential of, such a repository. Success at this stage would lead to further steps toward implementation, which would culminate in the unveiling of a secure web-based access portal and the application of TDR policies to the data preserved in the repository. While the SITP projects an initial introduction date of 2014, much depends on funding for the program.

In a related development, the Digitization Program Advisory Committee recently launched a Trusted Digital Repository Team with cross-unit and cross-discipline representation to systematically address issues of long-term data preservation. It included representatives from OCIO, SIA, SIL, NMNH, and STRI, as well as several other units. In addition, OCIO recently filled a new Director of Research and Scientific Data Management position that is specifically dedicated to developing scholarly and scientific research repositories at the Smithsonian. The candidate selected for this position had broad experience developing open-source repositories at many institutions, and planned to have in-depth conversations with Smithsonian scientists, librarians, and archivists on how the Institution might move forward in this area.

Digital Asset Management System (DAMS)

Like the proposed SI DataNet, the Smithsonian Digital Asset Management System (DAMS) provides a system that allows the Smithsonian to store, organize, and provide

access to digital assets. The digital assets covered by DAMS are images, audio, and video. Thus, DAMS is not explicitly concerned with research data sets, although it can deal with some kinds of digital assets that serve as research data—audio files of frog vocalizations, for example. In the past several years, the DAMS project has secured funding and is therefore much further along in its implementation than the SI DataNet project.[80]

Enterprise Digital Asset Network (EDAN)

The goal of the Smithsonian's Enterprise Digital Asset Network (EDAN) is to make digital collections (and a limited number of other digital assets) scattered across Smithsonian units web-searchable and accessible through a central, pan-Institutional portal. It addresses the considerable technical obstacles to pan-Institutional search-and-access that have resulted from the historically decentralized character of Smithsonian units and their IT systems, with digital assets stored on many different collections information systems (CISs) at different units that may employ different metadata schemas, as well as being stored informally on a wide range of media. As the 2010-15 SITP notes:

> *Until … the launch of EDAN, the Smithsonian lacked any way of providing the public, researchers, and staff a unified view into its collections and associated digital assets. Under the Smithsonian's collection data architecture, those searching for digital information and assets had to know which Smithsonian unit held the object and (often) in which collection system the object resided, regardless of whether the inquiry was being made by a scholar, the general public, a curator, or a scientist.* (Smithsonian Institution. Office of the Chief Information Officer 2010, p. 182)

The first phase of EDAN was successfully completed in fiscal year 2009 with the introduction of a Collections Search Center that can be accessed through the si.edu home page. The initial focus of the Collections Search Center has been overwhelmingly on museum, archival, and library collections. It includes the digital assets in DAMS, but offers links to only a small number of research databases at NMNH and STRI.

Despite being operational, EDAN remains a work-in-progress that will continue to expand its database coverage over time. For example, EDAN can potentially include assets in the SI DataNet, as well as those that reside on the CISs at individual units. Further work may also need to be done on the search-and-query functions of the system, which one interviewee complained returned too many "unanticipated results."

[80] Still under discussion is the application of TDR policies for the long-term preservation of the data.

Smithsonian Research Online

The Smithsonian Research Online (SRO) project combines two previously separate initiatives:

- The Smithsonian Research Bibliography, an initiative undertaken by SIL and OUSS to collect and provide access to descriptive metadata on research by Smithsonian scholars, both published and unpublished.[81] It has been a noteworthy first step in pursuing a coordinated approach across the units to making Smithsonian research more easily discoverable and accessible. In the case of the units coming under the OUSS, the bibliography is intended to be the universal database of record. At of the time of this writing, the Bibliography contained over 30,000 research citations covering all Smithsonian units. It is indexed on Google and Google Scholar. The SRO site, research.si.edu, offers a filter for searching the Bibliography by department, unit, and other criteria.

- The Smithsonian Digital Document Repository, which contains digital versions of publications referenced in the Research Bibliography, along with associated images and other supplementary materials. Examples of Repository scientific content available via the SRO site include articles, preprints, working papers, technical reports, conference papers, books, and theses.

SIL interviewees indicated that one goal of the SRO project is eventually to provide links to the research data underlying the referenced works. This would mimic the modus operandi of scholarly publishers that require authors to submit the data set underlying a published article, which is then archived and made accessible to interested external parties, often through an appropriate data center with which the publisher maintains a relationship. Interviewees pointed out, however, that the data associated with a published work are often only part of a larger data set, and it is unclear whether the Repository would offer access to the entire data set, or just the sub-set relevant to the referenced work. The SRO also provides direct access to a small number of research data sets that are independent of any publication.

Interviewees indicated that participation by scientific research personnel in the SRO (including submission of the required meta-documentation for the Bibliography and a digital copy of the work for the Repository) is widespread, but not universal (as intended, at least for the science units). At the time of this writing, the OP&A study team was told

[81] While the project began as an initiative of OUSS, non-science units have taken advantage of it as well.

that only about 20 percent of the works referenced in the Bibliography were linked to digital full-text copies in the Repository.[82]

Proposed Smithsonian Institution Geographic Information System (SIGIS)

All Smithsonian biology units make extensive use of GIS, because the precise locations of observations are key data for many research projects. For example, SCBI uses GIS to track migratory movements of endangered species and to delineate the extent of habitat loss; SERC uses it to track shoreline changes along the Chesapeake Bay and to monitor the spread of invasive aquatic species; and NMNH researchers use it for spatial analyses of genetic and human diversity.[83]

NMNH is currently leading efforts to combine existing unit-based GISs into a central SIGIS, having recently secured funds to purchase the required enterprise software package. In 2011, NMNH, working with the other science units, will develop milestone dates for the project. The significance of this critical piece of cyberinfrastructure is described in the 2010-15 SITP:

> *A Smithsonian-wide GIS can be used to integrate data gathered from individuals, provide an end-to-end system for analysis of multitudes of data fields, and facilitate the integration of Smithsonian data with other regional and global data sets through partnered data sharing agreements, in addition to facilitating the sharing of data between Smithsonian scientists and researchers.*
>
> *… Currently, there are many individual installations of GIS software throughout the Smithsonian. This non-integrated implementation of GISs hinders productivity by fostering an environment that has researchers and scientists working in isolation. It also endangers the preservation of this data due to personnel turnover, lack of training, and limited personnel, hardware, and software resources.*
>
> *… A Smithsonian-wide GIS [will support] the reuse and repurposing of data. Every time data are reused or shared, the value of the investment in obtaining [them] multiplies. Additionally, a robust scientific archiving and database*

[82] Digital copies come from a variety of sources, but primarily from authors (who have either scanned them or received them from their publishers). Many of the publications indexed in the Bibliography came out in the pre-digital era, so they would have to be individually scanned to be included in the Repository. Since the SRO project does not have the staff and other resources to scan legacy literature, it would fall to the scientists to do the scans, and time and resources are likely to be an issue there, too.

[83] Non-biology Smithsonian units also make extensive use of GIS. For example, NASM's Center for Earth and Planetary Studies and SAO both use it for multiple analyses of earth systems, and NMAI has hired a research geographer who is building on pan-Institutional GIS in new ways.

> *process will mean that more data are kept in a well-organized and well-protected fashion, and more readily accessed by a much larger community of interest.* (Smithsonian Institution. Office of the Chief Information Officer 2010, p. 131)

Individual Science Units

The state of data management and data sharing varies widely not only across, but also within, Smithsonian biology units. While most have a policy (or set of policies) concerning the terms and conditions for external use of their data and other scientific information, none currently has a policy detailing expectations for unit researchers when it comes to data management and sharing. To the extent that researchers systematically manage and share their data, it is usually at their own initiative or at the behest of publishers or funders, not in response to unit or Smithsonian requirements. Some units provide central resources to support researchers' own data-management and -sharing efforts, although in most cases these resources seem inadequate to the task.

The following sections provide greater detail on specific data-management and -sharing efforts underway at the four biology research units.

National Museum of Natural History

Of all the Smithsonian biology units, NMNH faces perhaps the most daunting digital data challenges, owing to its long history of collecting and research, the size of the physical collections to be digitized,[84] the amount of its accumulated research data, and the wide scope of science it supports. Although other biology units hold object collections, the scale of those at NMNH—over 126 million items—is unique.[85] In addition, NMNH has an intimidating backlog of legacy research data, both digital and analog, in need of systematic curation and preservation. It also has a very significant amount of handwritten field notes and research data dating back to the mid-1850s that require conversion into digital format.

NMNH's central Information Technology Office (ITO) provides central digital storage capacity for museum scientists, supports collections digitization, and houses several data-management projects such as the Integrated Taxonomic Information System (ITIS) and Integrated Open Taxonomic Access (INOTAXA) project (both are discussed in the

[84] In this context, "digitizing" refers to entering at least the basic textual information on items into a collections database.
[85] The current state of collections digitization is uneven across the Museum. Some departments are well advanced in digitizing their physical collections, but others have made much less progress, usually because of the size of the collections.

"Collaborations" section below). For online access, the museum maintains a Research and Collections Databases page (http://www.mnh.si.edu/rc/db/databases.html), which includes a link to collections records grouped by department (http://collections.nmnh.si.edu/). The page also offers online access to research bibliographies, reference databases, and two catalogs of scientific images and illustrations. It does not provide access to digital NMNH research data or metadata per se, except to the degree that these are incorporated in the available collections and reference databases. The ITO imposes some basic standards for collections data. For example, an interviewee noted:

> *You have standards, say, for an image and the resolution, [or for basic metadata] like the date and where it was taken. You can force those kinds of standards. Within the EMu system, you are following the Dublin Core, which gives you a data schema where you have this field and what it means. You make sure your data field is following what that core definition is, so that when people elsewhere use that same core data item, it means the same thing. … When you add new functionality to EMu, you're basing it on current industry standards of interoperability.*

Digital Storage Space Requirements. A sense of what NMNH is up against came through in the 2004 scientific research computing needs assessment. At that time, NMNH estimated that it needed about 93 TB of digital storage (compared to about 7.5 TB for STRI), and had only about 68 TB (compared to about 1.4 TB for SERC). At the time, these figures were thought to be low, given how hard it was to calculate them.

According to interviewees, the digital storage situation is made more complicated both by the rapidly rising tide of data and by an organizational culture in which data management is usually an afterthought:

> *[NMNH wants] to have processes in place for people to communicate with us [ITO] so we know what their needs are. … You might have a research scientist who has an ongoing project that creates huge amounts of data, and the project might last 10 years; this is a newer issue. … We have one scientist who generates terabytes of data, and had two terabytes on hard drives when he came to us. We love it when people come to us earlier.*

The ITO is not involved in decisions about what data are stored at the unit level and what are left to the discretion of individual researchers or departments. Rather, it provides each NMNH researcher with a certain amount of storage space to be used at his or her discretion, and responds to specific requests for additional space as needed. Data not saved at the unit level may be saved on researchers' hard drives, at the department level,

in the digital storage facilities of collaborators' institutions, or in a central repository such as GenBank. An interviewee indicated that it is not always clear how decisions about where to store data are made.

Collections Digitization. Digitization of collections—involving not only the creation of digital images, but also the writing of descriptive metadata to facilitate discovery and use—has been a priority for NMNH. For some collections of more manageable size, such as vertebrates and botany, digitization efforts are well advanced and data management has reached a relatively high standard. As would be expected, however, the very large collections such as entomology and invertebrates face exponentially greater challenges, and continue to lag in some areas. The main factor slowing progress, predictably, has been a lack of funding.

NMNH's Research and Collection Information System (RCIS), based on a commercial software product, KE Software's Electronic Museum (EMu), has been operational since August 2001. Currently, RCIS contains over 5.4 million records and about 650,000 digital images; current plans call for migrating more than 5.6 million additional records from legacy systems, as well as adding new records on an ongoing basis. An interviewee indicated, however, that at current rates of digitization, the backlog of undigitized collections is actually growing, noting that "the number of new records going into EMu is about 45,000-50,000 records [a year], which is well short of the number of items that come in per year." In short, while collections digitization has received a great deal of attention at NMNH, the gulf between current realities and sought-for ideals remains wide.

Minimally, digital collections records aim to collectively provide a good sense of what NMNH holds (so that interested researchers know what specimens and tissue samples might be available for loan or onsite study) and to individually provide the basic information about what an object is and where, when, and by whom it was collected. Progress has clearly been made toward these limited goals, although much remains to be done.

Ideally, digital records would offer enough rich information that physical access to the underlying objects would not be necessary for most scientific purposes—although interviewees indicated that the extent to which this is realistically achievable varies by field and by the scientific purpose for which the collections are being consulted. Such rich information might include links between biological specimens and associated genomics sequences; links to relevant literature and related collections; annotations and research notes; precise geospatial coordinates for where items were collected and where similar items have been observed; high-resolution two- and, where appropriate, three-dimensional images; and so on. This level of documentation, however, remains a distant

dream not only for NMNH, but for most organizations with natural history collections of any size.

Interviewees also indicated that, after many years in which collections digitization was pursued opportunistically, NMNH recently began to apply a more strategic approach. Among other goals, it will set priorities for collections digitization and use statistical techniques to determine how to adequately represent a collection without digitizing every specimen.[86]

One interviewee expressed concern that, while NMNH was one of the first organizations in the world to embark on a program of collections digitization, it has in recent years lagged in comparison with other organizations. That said, this interviewee agreed that most peer organizations are also struggling with collections digitization to some extent. While a few organizations with relatively small collections, such as the University of Kansas, have done a fine job of comprehensively digitizing their holdings, most major natural history collecting institutions are facing issues and obstacles similar to those encountered by NMNH.

Sharing of Research Data. In the context of digital research data, central NMNH IT systems are intended to serve more as platforms for internal storage and analysis than as platforms for external discovery and access. The ITO does not impose standards for data reporting or metadata schemas, leaving it to researchers to apply field-based standards and best practices where they exist, to develop their own standards and practices where they do not, or simply to do the bare minimum of documentation necessary for their own needs.

Even if standards to facilitate data discoverability, access, and use were to be introduced in the future at the Museum or field level, one interviewee at NMNH foresaw huge problems with applying any such standards retrospectively to the huge amounts of existing data that do not conform to them. In light of very real resource constraints, he suggested it might be necessary to simply draw a line under the past:

> *People follow best-practice standards in an informal fashion, but if we have terabytes of data that don't fit what a new standard [for data management] might be, we would have to get these new standards, and then they would trickle into the system. You have tons of historical data, and we need to decide whether we just go forward [or] try to play catch up—and we don't have a lot of resources. You have to make strategic decisions on what you want to do.*

[86] According to the 2010-15 SITP, about 50 million records will be needed to adequately represent all of the over 126 million objects and specimens in NMNH's collections.

Even prospectively, the ITO as currently configured is not staffed or resourced to provide systematic data curation services and other support to scientists who would like to process their research data to facilitate external discovery, access, and use. (That said, one interviewee noted that when researchers ask ITO for help in these areas, its staff do the best they can on a case-by-case basis, subject as always to the constraints of internal resources and of OCIO's infrastructure, firewalls, and servers.)

One external interviewee suggested that some of the issues with data sharing at NMNH arose from culture and institutional values, rather than resource constraints alone. He saw many NMNH researchers as still firmly rooted to small-science approaches to data sharing that have slowed progress toward better online access to data and collections information:

> *[At NMNH, there is a] collection-by-collection decision to hoard data. … Frankly, some of the curators are living in the Stone Age. They still think, "Oh my God, if I share my data and it isn't accurate, it might be misused, and we might get sued"—although, of course, this has never happened. Or, "It's my data, I still have a plan to do research on it, and I don't want to give it out." … Virtually every natural history museum in the country that is worth its salt is freely mobilizing and serving its data. … [Institutions that are not] serving the [collections] data are working against the very mission for which those collections were made. They were not made just so that curators could sit in their offices and [study them themselves]. That work is important, but the sharing of the collection information is critical. How can an institution like the Smithsonian claim to be the national institution of biodiversity when we have an extinction [crisis], and they are sitting on their hands hoarding some of the most important biodiversity collection information?*

National Zoological Park

NZP collects a wide variety of information of potential value to the wider scientific and zoological communities, including but by no means limited to the research data generated by the six centers that comprise SCBI. Other types of NZP scientific documentation of potential interest to external researchers and conservationists include animal records such as keeper reports[87] and veterinary records.[88] Based on interviews and the 2006-16 NZP strategic plan, the sharing of scientific data does not appear to be an explicit Zoo priority.

[87] NZP maintains daily records on the diet, behavior, health, and other characteristics of animals in its collection.

Animal Records. For many years, NZP, like many other zoos, has organized its animal records using a DOS-based suite of software from the non-profit International Species Information System (ISIS),[89] consisting of:

- The Animal Records Keeping System (ARKS) for basic registrarial information and keeper reports;

- The Medical Animal Record Keeping System (MedARKS) for veterinary records; and

- The Single Population Analysis Records Keeping System (SPARKS) for genetic sequencing information.

Collectively, these systems contain information on about 50,000 NZP animals, past and present.

The ARKS suite is mainly for internal records management, although data in the ARKS database are summarized by hand on a weekly basis, and this summary is shared with ISIS. ISIS maintains a current database of animal holdings at its member zoos, although the information contained in it is limited to species name and numbers of males, females, and recent births. Full information on a given animal is shared with any zoo to which that animal is transferred, but the full data records contained within the ARKS suite are not immediately accessible to or systematically shared with other ISIS members.

ISIS has been working on a web-based global Zoological Information Management System (ZIMS) to replace the obsolete ARKS suite, although the project has been plagued by delays. NZP, a founding member of ZIMS, has contributed substantially to its development through design consultation and testing support, and was originally to have acted as an alpha adopter in June 2010. However, concerns about security (related to Federal Information Security Management Act requirements) and systems architecture (NZP records would be stored on an ISIS server without local back-up) led the Zoo to decide against alpha adoption of the system.[90] The Zoo still expects to eventually adopt ZIMS, but the ongoing slippage in the project's timeline has led to a current focus on

[88] NZP maintains a database of all official records, such as clinical notes, diagnoses, parasitology, administration of anesthetics, administration of medicine, and pathology.
[89] ISIS provides global-standard zoological data collection and sharing software to more than 800 member zoos, aquariums, and related organizations in almost 80 countries. The ISIS global database for the zoological community contains information on 10,000 species and 2.4 million individual animals.
[90] As the federal zoo, the security concerns are unique to NZP. However, an interviewee raised the possibility that a substantial number of other zoos shared NZP's concerns about the systems architecture.

creating NZP-specific interim systems for safekeeping records currently stored in the increasingly anachronistic and fragile ARKS system.[91]

In principle, successful implementation of ZIMS would hugely increase the potential for external access to NZP animal records by interested parties such as other zoos, conservation organizations, and so on. The extent to which access would actually be permitted, however, is an open question. While plans exist for limited sharing of summary veterinary data collected through ZIMS, the Zoo has not adopted the principle of open access, and interviewees indicated that attitudes toward open access vary among NZP personnel. In general, geneticists were thought to have a relatively relaxed attitude toward access; the genetic information maintained in SPARKS is already freely shared upon request with other zoos for the purpose of captive breeding. By contrast, according to some interviewees, veterinarians may have reservations about sharing their records, based on concerns that external parties unaware of the specific circumstances in which animals were treated could draw incorrect conclusions about why NZP vets employed particular treatments, or could retrospectively criticize treatment decisions.[92]

Research Data. The issues surrounding the ARKS and ZIMS animal-record systems are not generally relevant to the work of SCBI researchers, except to the extent that researchers' work involves veterinary treatment or genetic sequencing of the Zoo's own animals. A review of the NZP website revealed no obvious mechanisms for searching or accessing SCBI research data—whether at the level of SCBI as a whole or of its six centers—although the site does offer a searchable bibliography of researchers' scientific publications.

Individual researchers and research programs at SCBI tend, as is the case with counterparts at other Smithsonian biology units, to share their data through professional channels within their own fields. For example, interviewees at the SCBI Center for Conservation and Evolutionary Genetics indicated that they deposit their results in GenBank as a matter of course. The study team also learned of active NZP participation in the Smithsonian cryo initiative, along with NMNH, STRI, and SERC. According to an unpublished NZP background document, the aim of this initiative is "to modernize and streamline the methods in which frozen collections are documented, processed, reported, and stored at the Smithsonian" in order to "provide staff with the means to efficiently preserve and make available valuable scientific collections." This project has a large

[91] To this end, the interim MedarX system for veterinary records replaced MedARKS in 2009.

[92] One interviewee suggested that the well-publicized controversies surrounding animal care at NZP in the early 2000s might have reinforced the veterinarians' general resistance to providing access to their records. This interviewee was also quick to note that NZP was not necessarily negligent in the cases that drew attention, because the personnel were often dealing with "unknown territory" where best practices did not exist and judgment calls were frequently required.

data-management component, as there is currently little consistency in how such samples are documented across—or in some cases, within—units. For example, the initiative is "implementing a global database, called Freezerpro, to properly manage and make accessible data about samples."

Digging into the past, the study team learned that the predecessor to the SCBI Center for Conservation Education and Sustainability was at one time active in efforts to establish standardized protocols for data management, analysis, and interpretation related to the biodiversity monitoring plots where it worked. To this end it developed the Biodiversity Monitoring Database (BioMon) software. BioMon is still being used, primarily for STRI's forest monitoring plots, but according to one interviewee, it needs to be upgraded.

Smithsonian Environmental Research Center

SERC efforts to improve the management and widespread sharing of its data go back to 1995. In that year, it developed a data management plan calling for a brief metadata listing of all data sets for external discovery purposes; more detailed descriptive metadata for each data set based on a NBII metadata format; enhanced use of database software; participation in national environmental data repositories like NBII and GCMD; and greater data sharing, including distribution of some data sets directly through the SERC website. In 1998, SERC embarked upon a Research Data Collection Initiative, which succeeded in documenting some key data sets and making them available both through SERC's website and through clearinghouses such as GCMD and NBII. It also developed a uniform metadata schema for SERC data and a prototype for a centralized relational database to facilitate functional integration of data from individual projects. Despite progress, funding for the effort ended in 2001.

SERC's current policy is to share its data with any responsible party who presents a reasonable request and gives appropriate credit to the Smithsonian. It has also sought to implement some of the data-management practices developed through LTER, an organization that shares SERC's emphasis on ecological and environmental research over time at dedicated research sites. SERC maintains a central "Data Table of Contents" on its website (http://www.serc.si.edu/research/longterm_data/dtocindx.aspx). More detailed metadata in a NBII standard format are available for a few data sets. Expanding both the metadata index and the detailed metadata descriptions to include new and historical data is an ongoing process that proceeds in fits and starts as funding becomes available.

The Data Table of Contents is one of several databases accessible through the SERC Research Databases web page (http://www.serc.si.edu/research/databases.aspx). Also included are links to a page that offers point-and-click access to

- A limited selection of long-term SERC data;
- The National Exotic Marine and Estuarine Species Information System (NEMESIS) database (see below);
- The National Ballast Information Clearinghouse database (see below); and
- The Aquatic Invasions Research Directory.[93]

With some exceptions, and as with other units, most data management is undertaken independently by individual research teams, whose approaches and outcomes vary widely, according to an interviewee:

> *How [well the data are] managed, proofread, standardized, formalized, all those kind of things depends a lot on variables [that differ] from investigator to investigator, even data set to data set. … The data themselves are very uneven—extremely variable in scope and quality and value. If you have limited resources, where should you put your limited data management resources? Those [decisions] are often being made on a case-by-case basis, varying from lab to lab or PI to PI.*

Similarly, access to SERC research data tends to take place on a case-by-case basis through requests to individual SERC scientists.

That said, individual SERC researchers and labs have undertaken some critical data-management and -sharing initiatives. For example, under the National Invasive Species Act of 1996, the U.S. Coast Guard and SERC created the National Ballast Water Information Clearinghouse (NBIC) to collect, analyze, and interpret data on the ballast-water management practices of commercial ships that operate in U.S. waters, with an eye to controlling ballast-related coastal marine species invasions. NBIC maintains an online database that offers free downloads of ballast water data from 1999 on. This database is accompanied by a formal data-use agreement and disclaimer that address some of the legal and re-use issues with open access discussed in the "Policy" section above.

[93] This last is a free database SERC created to promote information transfer, coordination, and collaborative research on aquatic invasions; it contains current information on relevant SERC and external people, research, technology, policy, and management issues.

The SERC Marine Invasion Research Program developed and maintains NEMESIS, a national database of marine and estuarine invasions in the continental United States and Alaska. It contains detailed information on approximately 500 non-native species of plants, fish, invertebrates, protists, and algae that have appeared in U.S. coastal waters.

In March 2010, SERC picked up one thread of the Research Data Collection Initiative. The data of a recently retired SERC researcher have become the initial focus of a collaborative project with OCIO aimed at creating a schema to standardize some basic observational data fields to support relational analysis across projects. If successful, there is hope this schema can be adopted more widely within SERC and across the Institution. An interviewee described the kind of problem the schema would address:

> *SERC has a huge backlog of information that needs to be organized, standardized, documented, and linked together through modern relational database management systems, with the expertise on how to structure that effectively. To use a simple example, if somebody is taking the string of biological data on variation and abundance of [say, blue crabs] running around the Maryland coastal plain, we are also taking data on rainfall and oxygen in the Chesapeake Bay and temperature and all those kinds of things, and those are all in separate data sets. How do you then set those up in a way that will allow you to test and put those together? Somebody calls up and says, "So, it's a drought year—are blue crabs up or down as a result?" That is not an easy question. … Setting up those kinds of things so you could do that in relational ways is important.*

Despite these noteworthy initiatives, a lack of resources allotted specifically for data management and sharing has hampered SERC's efforts in this area. In contrast to both STRI and NMNH, SERC does not have any unit staff dedicated to data management or bioinformatics.[94] SERC personnel repeatedly stressed that the largest obstacle to progress in the unit's data-management and data-sharing efforts is the lack of an onsite data manager, and overcoming this obstacle remains a top priority.[95] Interviewees also noted that some external requests for data cannot be met simply because they require substantial data processing that SERC cannot undertake with current resources.

[94] SERC personnel indicated that the ratio of funded support staff to scientists at NMNH is three to four times the ratio at SERC.

[95] SERC personnel elaborated on this position as follows:

> *Such a data manager would lead SERCs own data-management efforts, interface with central Smithsonian IT personnel and resources, and interact with peers among environmental research organizations. We view this as a higher-level job—GS-12 with promotion potential—requiring both experience in environmental research and strong IT skills.*

On the 2004 OCIO scientific research computing needs assessment survey and in other venues, SERC has indicated that its greatest IT challenge is "documenting, preserving, and sharing data," and that this task requires additional data-management personnel. It has been years since SERC has had a data manager or even a webmaster, and parts of its public website are out of date.

One interviewee at SERC expressed concern that central resources for digitization have tended to go to museum collections projects, which leaves units such as SERC, which is not a designated Smithsonian collecting unit, on the outside:

> *The focus has been on the objects and not on the kind of data that we have. Then the Castle turns around and says, "Why haven't you guys done it?" Because nobody's been willing to pay for it. It's not because we don't think it's important.*

A final point made by SERC personnel, which echoed a theme sometimes raised in the literature, is that distributed systems provide the most effective and sustainable way to manage and share research data. This means data should be accessible through the servers or repositories where they are most likely to be found by their target audiences of scientists or other users. For this reason, interviewees at SERC held that servers or repositories focused on particular fields are a more promising long-term home for the Center's data than a Smithsonian repository, which presumably would include data from many fields. Moreover, SERC personnel argued that living, evolving data sets (like those arising from long-term research) should be housed within the research units themselves to facilitate updating and maintenance.

Smithsonian Tropical Research Institute

In the past several years, STRI has been paying more attention to data sharing. Its 2008-13 strategic plan lists "put[ting] collections and data on the internet" as one of three major strategies under the goal of "increas[ing] access to STRI data and collections." Since 2006, STRI has had a bioinformatics office tasked with facilitating online access to STRI's data. According to the office's mission statement, it seeks to provide:

> *… the technical assistance and training necessary to migrate STRI's data to the web in a standardized data management system that provides global access to the data; simple-to-use, web-based data management analytical tools; and integration with other data sets, both STRI's and other organizations.*

The office's website (http://biogeodb.stri.si.edu/bioinformatics/en/) describes itself as a user-friendly, query-able portal to a wide range of internal data. However, the reality is that the office remains a work-in-progress.

Although STRI is not a designated Smithsonian collecting unit, it holds some of the largest collections of biological specimens, as well as archaeological and paleontological artifacts, in Panama. It is currently engaged in a systematic effort to digitize these collections and make the information available online. It has made significant progress over the last three to four years in digitizing textual information for several of these collections, but has made much less progress on creating digital images. STRI also possesses a unique collection of slides, photographs, and videos that collectively document scientific research in Panama over the past 100 years. Tens of thousands of these images need to be digitized, and almost all need metadata.

STRI recently decided that it would, as a matter of policy, make data from internally funded research projects accessible online. For example, data from the Smithsonian Institution Global Earth Observatory (SIGEO)[96] (and from the Center for Tropical Forest Science [CTFS] a parallel and sometimes overlapping program), and all collections digitization projects,[97] are now available through the STRI website. STRI has also made much of the long-term physical monitoring data from its Environmental Science Program available.[98]

Despite these stated policies and initiatives, for the most part STRI has not required data management and sharing from PIs doing research under its auspices:

> *STRI is, for the most part, a facilitator of research by independent scientists. Historically, STRI has been independent scientists doing whatever research they want, with very little mandated policy about how they use their data and where it should go. That has only changed in the very recent past. In the last five years, there has been a real push to get those data online, but it's going against a long cultural history. … There are special cases like SIGEO where the Director has*

[96] The former applies only to data generated by the STRI SIGEO plots. Although STRI manages the SIGEO portal with Harvard University, the SIGEO model is that all research stations own their own data and decide what, how, and when to put those data on the web.

[97] This mainly concerns the digitization of formal biological collections, as well as some informal collections such as researchers' slides.

[98] Unusually for the Smithsonian, this program has a detailed written data-sharing policy that discusses the program's own objectives and responsibilities for providing access. (Other units have formal policies governing use of accessible data, focused on the limitations on external use.) An excerpt reads:

> *Our ideal goal is to make all Meteorological and Hydrological data available within 1-2 months of collection. Other data sets should be available via the Internet 2-3 years after collection. Realistically, however, it is not possible to make all data available in that time frame. We therefore recognize two access categories for electronic data sets:*
> ***Type 1** data are freely available over the Internet once the terms of the Data Use Agreement are agreed to.*
> ***Type 2** data are not freely available, but may be distributed under specific restrictions. Metadata for Type 2 data will generally be made freely available via the Internet.*

mandated that those data get put onto the web. There are other situations where scientists, by their own volition, are putting some of their data online. … [But] unless there is a financial incentive or a specific mandate [from a funding agency], there is little interest in putting raw data online.

However, the same interviewee was optimistic that the future would look very different with regard to data access:

What percentage [of existing STRI data is accessible online]? Probably a very small one. STRI has 60 years of research, and probably much less than 1 percent of the data that has been collected has gotten to [the Bioinformatics Office]. But if we start at zero today and look into the future, I would say it will be a very large percentage. [I can't give you] a number, [because] I would just be inventing [it]. But an ever-growing percentage of the data will make it to the web.

With regard to data preservation, STRI does not currently have formal organizational arrangements, apart from those services (such as migration across changing technology platforms) that its commitment to maintaining access requires:

There hasn't been any institutional entity to do [long-term preservation]. The Office of Bioinformatics has begun to do it in an ad hoc way, because for the first time, we have a large network-attached storage device where people can put their data and make sure it's backed up. A number of projects that are producing large amounts of data have been given access to this system. Most researchers, however, are on their own when it comes to long-term data storage. Few researchers approach the Office specifically with the idea of archiving their data. More frequently, their primary interest is in the creation of a website to facilitate public access, and they leave a copy of their data for this purpose.

A related issue is that the STRI Bioinformatics Office does not carry out extensive curation, so data that are accessible in principle may not be easily discoverable or usable in practice. Most of the data that come through the Bioinformatics Office have little or no metadata documentation, let alone the kind of extensive documentation that would allow re-use without consulting the data originator:

Metadata is a hope and an aspiration. It is even worse off than long-term archiving. [Almost] the only things that have metadata are the collections and the physical monitoring data. Most other data sets and stored files have virtually nothing.

Like their counterparts at SERC, interviewees at STRI complained that, despite statements by the central Smithsonian administration about the importance of data management and sharing, longstanding requests for funds to bring in additional staff to support these functions have not yielded results. In the meantime, the 2010-15 SITP indicates that STRI continues to accumulate new data faster than its Bioinformatics Office can ingest it.

Other Units

Smithsonian Institution Archives

SIA's basic mandate is to preserve and provide access to permanent Institutional records in perpetuity. SIA has been acquiring digital information since 1994, and as of 2003 has had an official Electronic Records Program dedicated to the preservation, curation, management, and provision of access to the digital records in its collections. In support of this responsibility, SIA has collaborated with other organizations on research efforts to develop new digital preservation techniques and to define and disseminate preservation standards for data formats, TDRs, and digitization guidelines. In this capacity, it also undertakes some data-management education functions similar to those carried out, as described, by some university libraries. As one SIA interviewee put it:

> *SIA is responsible for lifecycle integrity from the point at which material comes to us. But with digital materials, the problem is that integrity begins at the point of creation. SIA educates people about this—for example, how to keep things, the various risks such as risks related to the media it is stored on. 8.5-inch floppies won't work because the technology is almost non-existent.*

In terms of scientific data specifically, SIA has organizational responsibility for the papers of Smithsonian scientists and the records of the science units. It has acquired, managed, preserved, and provided access to Smithsonian research data and other scientific documents since its establishment in the 1950s, in accordance with its

collecting mandate as the Institution's official records manager and its appraisal methodology for which documents merit archival preservation.[99]

However, not all data from research undertaken at the Smithsonian are transferred to SIA, and some that are may not be adequately processed or documented to ensure usability. There are a number of reasons:

- Data sets that are still being actively analyzed, used, or augmented at the research units remain in their custody as long as this is the case.
- Ownership issues can pose obstacles to transfer to SIA—for example, in the case of data generated by Smithsonian scientists from research projects funded by other organizations.
- Data held by a specific researcher usually do not transfer to SIA until that researcher retires or otherwise departs; if a departure is precipitous, data may be left to SIA in unusable condition.
- SIA collected digital records for almost nine years (1994-2003) before its Electronic Records Program was established. During this time, curation of digital records was less systematic than it has subsequently become, again raising issues about the usability of some of the archived data from this period.

The core question of whether all Smithsonian-generated digital scientific research data should be classified as Institutional records, and therefore should come under SIA's purview, has been raised, but neither the central administration nor SIA has given a formal answer. Absent that decision, there has been some discussion within the central

[99] The SIA appraisal methodology, available via the SIA website, notes that

> *Types of [research] records of enduring value, regardless of media, include:*
>
> - *Field/lab/research notes/photographs/illustrations/film footage*
> - *Correspondence with colleagues pertaining to research and collections*
> - *Research reports, including content based editorial comments/notes/correspondence*
> - *Unpublished manuscripts (final draft), including content based editorial comments/notes/correspondence*
> - *Nomenclature lists and notes*
> - *Professional activities files, including photos of colleagues, ephemera (e.g., program – not registration forms and logistics), professional conferences, symposia, and workshops attended in support of original research, papers presented including lectures, slides/photographs*
> - *Departmental records regarding departmental research planning*
> - *Raw data – data should reside in the department for use there (NOTE: If raw data is found among personal papers, determine if it is associated with SI collections. If so, consult with the department to gain a clear understanding of the significance of the data). Case by case appraisal.*

administration and among other units about what roles SIA might play in the Institution's digital scientific data-management and -sharing efforts, and how its roles would relate to those of OCIO, SIL, and the research units themselves.

More immediately, SIA has embarked on a project with the National Herbarium of the NMNH Botany Department to digitize and enhance access to botany field journals of Smithsonian scientists in the museum's and SIA's collections. Personnel involved with this initiative see it as a demonstration project that might serve as model for other NMNH departments and beyond. Parenthetically, an interviewee discussing this project provided a fine example of why domain scientists need to be intimately involved in data curation, at least until a certain point in the data lifecycle:

> *You have to include a subject matter specialist. Otherwise, you've got a field notebook from the 19th century, and they talk about a certain species—well, that species has been renamed; it's in a different genus now. If you're not a specialist, you're going to tag the old name, and people are going to be looking for the new name. … To really understand the type of data that's in there, what they're talking about, you really need somebody who knows the field.*

As with interviewees at SERC, STRI, and OCIO, SIA personnel emphasized the need to devote more resources to digital data management if the Institution is to keep up with growing demands, let alone make headway with backlogged materials. They noted that SIA currently manages to keep on top of incoming digital materials (including, but not limited to, some research data) only through a combination of triage and reliance on the volunteers and interns.

Smithsonian Institution Libraries

SIL's traditional focus has been on providing discovery and access to publications, physical and digital. Of course, as noted, the boundary between published results and underlying data is blurry at several points. For example, publications are a key means of identifying what data might exist where, and this role is likely to grow in the future as published results are increasingly linked to underlying data sets. Likewise, from the perspective of digital discovery, access, and use, legacy literature (such as SIL's extensive collections of hard-copy biodiversity literature) may have more in common with non-digital legacy data than with modern scientific publications. For example, legacy literature needs to be retrospectively captured in digital format, crosswalked so that old scientific names conform with modern usage, organized in databases that allow for online search-and-query, and so on.)

SIL is heavily involved in two critical initiatives that deal with data in this broad sense. The SRO, discussed above, is a pan-Institutional Smithsonian project. The Biodiversity Heritage Library (BHL), which is discussed in the following section, is a broader initiative that draws in a number of major external partners. In those initiatives and more generally, SIL is now paying more attention to the issue of providing links among collections, publications, and research data, particularly semantic links from published material to the related source data (such as specimen collections data).[100] As yet, however, SIL has no policy or plan to become systematically involved in scientific digital data management and sharing.

Collaborative Smithsonian Endeavors

In addition to these internal Smithsonian scientific data-management and -sharing activities, many collaborative efforts exist among the units and with external organizations. Most of these efforts are undertaken at the level of units, sub-units, or individual researchers. To the study team's knowledge, no centralized listing exists of past or present collaborative data-management and -sharing efforts across the Smithsonian. What follows are examples of collaborations brought to the study team's attention, meaning that the list is not exhaustive.

NMNH appears to be particularly active in collaborative projects, in many cases as a matter of unit policy and in others thanks to the initiative of individual staff. In terms of data-management standards, access portals, and search tools, interviewees indicated that systematic biology is generally one of the more advanced areas within biology, and NMNH has made substantial contributions to a number of collaborative efforts in this area.

- NMNH hosts the secretariat of EOL, a high-profile collaborative project with the Field Museum, Marine Biological Laboratory (Woods Hole), Missouri Botanical Garden, and Harvard University, funded in part by major grants from the MacArthur and Sloan Foundations. EOL has set itself the ambitious task of creating a website that provides, for every known species, a web page of consolidated, expert-reviewed information resources, including links to relevant publications, data sets, and images. EOL aspires to be a centralized resource for

[100] Related to the possible futures of libraries, Helly, Staudigel, and Koppers (2003) have called for a "new protocol [in which] the traditional publication of scientific results is accompanied by publication of its data and metadata making the 'complete' scientific product available in a consistent and coherent manner." This protocol envisions an "augmented library [that] will protect the future of science against any loss of valuable research data that normally resides in private files [of researchers]."

both scientists and lay people who need consolidated, reliable information on species.[101]

- Another NMNH-hosted entity, CBOL, has more than 170 members in 50 countries. CBOL promotes DNA barcoding as a practical, reliable method for identifying and distinguishing among species and tissues derived from them.[102] Data management and sharing are major priorities for CBOL. For example, it has developed standardized protocols for DNA barcoding and is compiling a public library of high-quality DNA barcode records. Further, it works with GenBank to establish data standards; the two organizations have promulgated a "barcode data standard" that represents a major step toward linking research data, specimen data, and publications.[103]

- NMNH is a lead player in ITIS, the consortium of primarily federal government organizations,[104] which seeks to improve the organization of, and access to, standardized species nomenclature. In collaboration with the Europe-based organization Species 2000, ITIS produces the Catalogue of Life (COL), widely recognized as the authoritative global standard for systematic scientific biological nomenclature. COL provides the "backbone" for more information-rich biological databases; EOL, for example, uses the COL database for accepted scientific names, recognized historical synonyms, and taxonomic placement of all species.

- SIL has, as noted, been a leader in the BHL consortium, a collaboration among 12 major natural-history museum libraries, botanical libraries, and research institutions. The goal of the project is to digitize, mark up for data retrieval, and disseminate the corpus of legacy biodiversity literature—primarily older books (including rare works) and journals—in the collections of consortium partners. According to one interviewee, BHL is now also working to link data from its legacy literature to other databases. (An example would be extracting specimen

[101] A number of interviewees mentioned that the project has been slowed by delays in the peer review of species pages, because scientists have few incentives to devote time to data review and management functions that are not traditionally rewarded at either the organizational or professional level. As of this writing, the vast majority of reviewers are volunteers, and the vast majority of those are not NMNH staff.
[102] Barcoding in this context refers to species identification on the basis of a small, distinctive segment of the overall genomic sequence for that species. While still somewhat controversial among traditional taxonomists, DNA barcoding holds enormous potential for a wide range of scientific research needs, as well as economic, forensic, public-health, and other practical applications.
[103] In a related effort, several scientists at NMNH are working under a NSF grant on ways to link field data to museum collections data and, in turn, to publications.
[104] Major partners also include the non-profit NatureServe, NBII, and government bureaus in Canada and Mexico.

information from the legacy literature text, and linking it to data on the same specimen or species from other sources—collections databases, images, DNA data in GenBank, and so on.) BHL, which also works with EOL (see "Addendum: Social Media and Science Research" at the end of this study) is widely considered to be a very important step in making published data from the legacy literature globally accessible in digital form.

- The INOTAXA project aims to create a model for global access to biodiversity data by providing a web-based "workspace" in which taxonomic descriptions, identification keys, catalogs, names, specimen data, images, and other resources can be seamlessly accessed from multiple servers globally, according to user-defined needs. INOTAXA lists SIL, STRI, GBIF, and a number of other prominent biodiversity organizations in the United States, United Kingdom, and Latin America as partners, but in practice it is largely driven by a small core of individuals working on their own initiative, including a NMNH scientist.

- SIGEO involves a network of forest observation plots that works in parallel with STRI's CTFS.[105] Plots involved in the SIGEO project are located in South, Central, and North America, the Caribbean, South and Southeast Asia, and Africa. Scientific partners hail not only from these nations, but also from universities and NGOs in North America, Europe, and Japan. SIGEO is at root a scientific research and capacity-building initiative, not a data-management or data-sharing initiative, but it addresses a variety of issues in the latter area in pursuit of its goals. Scientists from many SI units participate in SIGEO projects.

- Some important data-sharing initiatives in which STRI has played a major role include the Barcode of Life Data Systems (BOLD) project, which aids in the collection, management, analysis, and use of DNA barcodes;[106] the Global Plant Initiative (GPI), an international partnership of herbariums working to create a coordinated database with information on and images of plants worldwide; and the Inter-American Biodiversity Information Network (IABIN), a network dedicated to the adoption and promotion of ecoinformatics standards and

[105] According to the description on its website, SIGEO builds on and expands the CTFS global network of forest plots, transforming it into a platform for a broader range of scientific investigations. CTFS research on tropical forest dynamics continues, involving new initiatives to study carbon fluxes, temperate forests, and the impacts of climate change on biodiversity and forest function.
[106] BOLD is a close collaborator with the CBOL project discussed above.

protocols, and the sharing of biological information, in all the countries of the Americas.[107]

- The Smithsonian as a whole is represented on IWGDD. IWGDD's focus goes well beyond biology and includes digitization issues at a relatively high level of abstraction—comparable to the Smithsonian digitization strategic plan, which in fact is guided in part by IWGDD recommendations. The OUSS has coordinated with the designated Smithsonian representative to IWGDD (who hails from SAO) to ensure that Smithsonian biology is represented, as indicated in the keystone report of IWGDD.

- At the level of OUSS, the Smithsonian also participates in forums such as GBIF, GEO/GEOSS, and USGEO.[108] The Smithsonian OUSS representative to USGEO was recently appointed co-chair of that group, and interviewees described him as one of the driving forces behind Smithsonian participation in such forums.

- NEON has designated the Front Royal, Virginia campus of SCBI as a core site for its mid-Atlantic regional domain. NEON has a strong data-management component, in that all research at its 20 core sites must use a standard data-collection protocol and jointly calibrate all equipment to ensure consistent data across the sites. Similarly, it processes and manages all raw data in the same way, with easy access by other users a prime consideration. One interviewee indicated that NEON is seen as a more centrally-managed successor to the existing LTER network, which is a looser confederation of research sites.

- For the most part, individual Smithsonian scientists who are active in fields where established data repositories or platforms for data sharing exist take advantage of these resources. Although this is most obvious in the case of SAO's astronomers and astrophysicists, it also applies in areas of biology. For example, Smithsonian genetics researchers upload their data to GenBank, as noted, and Smithsonian paleontologists make use of PBDB.[109]

On the other hand, the Smithsonian is not participating, as far as the study team could tell, in several high-profile initiatives and organizations that deal with data management

[107] The central Smithsonian was one of the creators of IABIN, which along with GBIF and the North American Biodiversity Information Network (a prototype of interoperability funded by the Commission on Environmental Cooperation of the North American Free Trade Agreement), provides institutional, political, and capacity-building support for biodiversity informatics.
[108] The U.S. contact organization for the global GEO/GEOSS project.
[109] At one point, NMNH scientists made a bid for NMNH to host the administration of PBDB; this did not happen.

and sharing. Some interviewees cautioned that the Smithsonian was likely to suffer a variety of intangible losses by not participating—for example, not being aware of developments and lessons learned that could benefit its own efforts; not making contacts that might be helpful in its work; lower visibility in the field; a reputation for not being a collaborative organization; and an inability to capitalize on opportunities for partnering and leveraging resources. The study team has not attempted to select the key initiatives and organizations in which the Smithsonian has no presence, but simply passes on those that interviewees cited:

- The National Academies' Board on Research Data and Information (BRDI), a national forum for the discussion of data management and data application issues.
- ESIP, the forum for exchange whose membership includes the major federal science agencies and a wide assortment of university, non-profit, and industry organizations.
- DataONE and The Data Conservancy, the first two NSF DataNet projects, which are attempting to build cutting-edge models and pilot infrastructure for data management. The former is of particular interest because of its focus on environmental sciences.
- The federal government data portals Data.gov and science.gov, which are part of the ongoing efforts discussed above to make taxpayer-supported federal research more accessible.

Similarly, interviewees at two influential organizations, ESA and the LOC, expressed an interest in exploring possibilities for collaborating with the Smithsonian in certain areas of scientific data curation and sharing. Interviewees at the National Center for Ecological Analysis and Synthesis (NCEAS), a highly-respected center hosted by the University of California at Santa Barbara and dedicated to promoting cross-disciplinary research in ecology and allied disciplines using existing data, also indicated that the Smithsonian might consider getting involved with its efforts. (NCEAS was instrumental in the development of EML.)

Overall, many interviewees agreed that, in principle, the Smithsonian is the sort of high-profile, nationally-esteemed organization that could play a leadership role in formulating collective responses to the forces for change described at the beginning of this report. However, to do so, they suggested, would require identifying a few appropriate niches where the Smithsonian is uniquely positioned to contribute, coordinating its efforts much more closely with peer organizations to avoid redundancy and working at cross-purposes,

and, perhaps most importantly, committing resources to these areas. One interviewee offered this suggestion:

> *At the Smithsonian, we are in the interesting situation of being in the perpetuity business. Subject to budget constraints, we stand ready to play a part as a database of record. It seems that consortiums of organizations are emerging for distributed long-term preservation—for example, through the NSF DataNet projects—and the Smithsonian could play into that. We might be in a position to shepherd certain data types, so that if the consortium were to break down, we could still get to the stuff that we value. But we definitely want to avoid duplication and make clear where the repository of record is for specific items. In many cases, our role would be to stand shoulder-to-shoulder with our peers, and be ready to pick up the slack if some of them slip away. Where we play, we might have to be prepared to be the guy holding the bag at the end.*

Barriers to Data Management and Sharing

When asked about factors that slow the Smithsonian's progress toward addressing the pressing issues of data management and sharing in biology, interviewees repeatedly mentioned three general themes: resources, culture, and organization.

Resources

The study team often heard that a lack of resources at the level of both individual research units and central offices such as OUSS and OCIO was a major obstacle. There was a sense that the central administration's rhetorical support for digitization had not yet been translated into a major commitment of resources for data management and data sharing beyond CISs. Even here, the Smithsonian is really only funding the systems, and not the digitization of legacy data or initiatives for increasing access to new data. The emergence of insufficient resources—or, more accurately, a mismatch between current resources and growing needs—as an obstacle is not surprising; it comes up in most OP&A organizational and management studies.

Funding. At the most basic level, the shortage of resources is about money. As noted, sustained funding for data management is unlikely to come from individual project grants, which usually do not provide explicit support for long-term digital data

management. Even if they cover such costs during the lifetime of the grant, that leaves the issue of how to fund continued data management when the grant ends.[110]

Under current funding norms, research organizations must for the most part think of the cost of digital data management and sharing as overhead, to be funded out of the same general pool of resources that pays for administration, facilities, and operations. However, as expectations for better data management and wider data sharing inevitably grow, these functions will constitute a major new demand on scarce overhead funds at both the unit and Institutional levels. Where this money will come from is not clear. Possibilities that came up in various places over the course of the study included raising overhead rates on research grants and/or allocating a larger part of them to data management support infrastructure; diverting funds from lower priorities; and raising new funds explicitly for this purpose from sources such as individual donors, foundations, and the Congress. However, several interviewees thought that raising external money for digital infrastructure and data management/sharing is likely to be difficult, because they are a hard sell in terms of service to specific audiences (in contrast to education and outreach programs), and do not carry the glamour and publicity associated with physical facilities and exhibitions.

Personnel. There was widespread agreement among interviewees that the number of staff with specialized skills needed for scientific data management and associated IT support functions (such as programmers and developers familiar with relevant open-source software) is inadequate, both at the unit and the Institutional levels. However, some interviewees thought the issue was less a shortage of funds for personnel per se, so much as it was the strong preference of Smithsonian research units to use available funds to hire scientists rather than scientific support personnel.[111]

Whatever the reasons, at all research units, responsibility for the bulk of data management rests with the researchers, who, as noted, generally do what they need for their research and have few incentives to invest time and resources to make the data durable for multiple purposes. On research teams, data-management work is often delegated to junior team members such as post-docs and fellows. Central offices at NMNH and STRI provide some limited support for specific data-management projects (for example, collections digitization at both units) and functions (for example, online

[110] Unless, of course, the data can be handed off to an existing preservation organization such as a digital repository.

[111] One interviewee noted that some time back, the Congress mandated that NMNH develop a database for its collections, which led to the hiring of a large number of temporary staff with digitization skills. However, these numbers shrank over time as people retired or left—a process that was accelerated by organizational changes that assigned some of these personnel to tasks for which they were not always best suited.

access to selected data sets at STRI). However, interviewees were quick to point out that neither of those offices was staffed to provide ongoing data-curation or data-management support to scientific personnel. Nor are they generally geared toward supporting scientific staff participation in data-curation and data-management initiatives.

In the case of OCIO, until recently it had only one high-level staff member with a background in biology, and his portfolio was not limited to, or even focused on, scientific-computing and data-management issues. In November 2010, OCIO did, as noted, hire a Director of Research and Scientific Data Management—its first management position dedicated to scientific computing and data management. Initially, this person will work with the pan-Institutional TDR action team to define requirements in that area. OCIO also succeeded in getting a data curator position included in the Smithsonian's fiscal year 2012 federal budget request—a hybrid IT/information-science/domain-science professional of the type discussed under "Human Resources." But even with these two positions, the Smithsonian falls far short of the level of central personnel needed to support the creation and maintenance of a TDR for the Institution's scientific research data sets.

Another personnel issue is that the Smithsonian devotes few staff resources to liaising with external organizations on data management and sharing. OUSS, for example, has just one staff member responsible for working with external organizations to leverage Smithsonian scientific assets and develop partnerships, and data management and sharing are only a small part of his overall portfolio. In any case, most external collaboration in this area takes place at the unit level and below, and is not formally channeled through OUSS. Smithsonian liaisons to external data-management initiatives often take on this role as an additional task on top of their "day jobs" as researchers, IT and library-science personnel, collections managers, and so on—and may receive little professional credit and practical support or training. In contrast, according to participants at the January 2010 ESIP conference, other federal agencies often include external relationship building and participation with collaborative initiatives in their staffs' position descriptions, and provide travel budgets for such work.

IT Infrastructure and Technical Support. An extensive comparison of the Smithsonian's IT budget and organization with those of peer institutions was not undertaken as part of this study. Moreover, such a comparison would be very difficult and potentially misleading, as agencies such as NASA and NOAA are able to devote considerably more resources to IT as a result of their much larger total budgets. (As one interviewee put it, the Smithsonian would be a "rounding error in NASA's IT budget.") At the same time, of course, larger agencies often face commensurately larger IT challenges.

On the plus side, interviewees at the Smithsonian research units suggested that OCIO's responsiveness to their needs for basic support services has improved markedly over the past few years, although some stressed that it was starting from a low baseline. Within the limits imposed by available resources and competing needs, some volunteered the opinion that OCIO was making reasonable efforts to support digital data management and scientific computing at the Smithsonian. Many attributed positive changes in their relationship with OCIO to the current Chief Information Officer (CIO).[112]

This opinion was not universal, however. Some interviewees believed OCIO was too concerned with pan-Institutional standardization of processes and policies, and too slow, bureaucratic, or inflexible to keep up with the rapid changes and innovations in scientific computing needs.[113] Such criticisms were often framed in terms of the essentially federal agency character of the Smithsonian's IT processes and policies, which was contrasted unfavorably with the more flexible and responsive approach of universities.[114] As discussed, some interviewees at the science units also gave the impression they believed that the biology community was not adequately represented within OCIO (or that OCIO did not engage with it adequately); that collecting units received more attention than scientific research units;[115] and that SAO has been favored over other science units.[116] While OCIO interviewees conceded that biology research may not have been sufficiently represented in some of the digitization efforts it has coordinated, they attributed this

[112] The CIO has stressed on numerous occasions that she was fortunate to inherit strong basic infrastructure and is now in a position to strengthen mission-related support functions. Her predecessor successfully upgraded the Institution's network, implemented VoIP telephony, standardized email and network services, and developed a state-of-the-art data center—all prerequisites for developing a data-management and scientific-computing capacity at the Smithsonian.

[113] There were specific references to implementation of open-source software as an area where OCIO comes up short. OCIO personnel, by contrast, mentioned some important open-source software-based initiatives at the Smithsonian, including EDAN and the Ocean Portal and Human Origins websites at NMNH. (CIO granted NMNH a waiver to develop these sites using the open-source DRUPAL, and is working toward getting in-house resources to support the LAMP architecture on which DRUPAL works. The LAMP architecture is also needed for the scientific repositories currently being considered by OCIO.) The scientists who cited issues with open-source software, however, appeared to be thinking mainly about the difficulties that scientists encounter when working with field- or application-specific open-source software on Smithsonian computers, not about major Institutional IT initiatives.

[114] OCIO interviewees concurred but stated that the Smithsonian has little discretion in the matter, as it receives approximately 90 percent of its IT funding from federal appropriations, which come with an expectation it will enforce OMB standards, especially in the area of security.

[115] OCIO has dedicated funds for implementing and maintaining CISs. For example, it has funding for staff to support the SIRIS CIS used by SIL and SIA and The Museum System CIS used by the Smithsonian art museums, National Air Space Museum, and National Museum of African American History and Culture. It does not have similar dedicated funding to support research-data management systems or processes.

[116] SAO works in the big-science paradigm, and as a leading force in the national astronomy and astrophysics fields, it receives extensive external funding from NASA to support the high data-management and computing expectations of those fields. This could lead some observers to conclude that central internal resources are being disproportionately channeled to SAO.

mainly to the unwillingness of scientists to devote time and effort to these issues, in part because there were no incentives to do so.

Some interviews noted that many problems raised in the 2004 computing needs assessment persist. For example, OCIO security firewalls continue to make it difficult for collaborators at other institutions, or even Smithsonian researchers working from non-Smithsonian computers, to access and work with data stored on Smithsonian systems. Several interviewees cautioned that this was driving some Smithsonian researchers to store their data on personal computers or on the less-restrictive systems of external collaborators' organizations.

Summing up the issues relating to inadequate IT support staff, infrastructure, and services, one interviewee suggested that the generic IT elements available to Smithsonian researchers were quite good, but the specific needs of scientific research were not adequately met:

> *The Smithsonian has two good legs of a tripod for IT. One is at the enterprise level, dealing with our email, GovTrip, Prism, the Herndon data center, and all that. We have phenomenal resources [at the level of the] very big-picture stuff to maintain us as an Institution. … The other leg we have [is day-to-day support]: I need a new computer, I need Windows installed, I need a virus taken off my computer, I need a new account for one of my people, my printer needs a cartridge, my Ethernet port is jiggy today—just individual- and departmental-level maintenance of IT needs. … The missing leg of that tripod is IT support for science, and that's completely missing. I can't get somebody to build me a database or make two databases talk to each other. There isn't the knowledge, skills, and abilities. If they're there, they are being applied on the administrative or operations side of the house, not the research and collections side of the house. Nobody in-house either has the time or ability to make that happen for us. … We need to give the scientists IT people [to whom they can say], "This is what I want to happen—you make it happen." That is what has been missing ever since I've been at the Smithsonian. It applies much more to access to the data than to preservation. Preservation and long-term curation are more of an OCIO/ Herndon problem.*

Central Versus Unit-based Resources. Interviewees spoke of a seeming disconnect between the role of OCIO as the Institutional focal point for collaborative digitization efforts, and the continuing perception among research units that OCIO is primarily a

technical support organization.[117] Several interviewees at the units thought that if further data-management resources were to become available, they should be placed with the research units rather than, say, being pooled centrally and made available to the units as needs arise. Their rationale was that the science done at individual units is sufficiently complex and idiosyncratic to require IT services tailored to their needs. Interviewees from OCIO did not disagree with this assessment. While they certainly thought additional resources were needed centrally for generic, Institution-wide IT functions such as developing and maintaining repositories, they also saw the need for enhanced data-management capabilities at the unit level, where the specific needs of particular types of data must be addressed.

However, there was not universal agreement among interviewees with this conclusion. A few discussed other arrangements that would allow IT specialists and equipment to be deployed more flexibly across units on the basis of shifting needs and priorities. For example, one proposed a flexible pool of central IT management personnel who would be assigned, on a rotating, as-needed basis, to the units:

> *Much the way that NMNH can make an informatics department and populate it with compatible and complementary people, so can the Institution. They can place some at STRI, at NMNH, at SERC, at the Zoo, up at SAO—that can easily be done. The important thing is to make sure that those people have the resources they need and the ability to communicate so that each bureau isn't operating in isolation, but is mandated to work with the other bureaus so they can globally solve problems. … They as a group will be at NMNH for a three-day workshop in January. They will be at STRI for a three-day workshop in March. They communicate what they are doing and how they are doing it, and scientists can interact with the ones placed in their units as they see fit. I don't think anyone would have a problem with that. You are bringing small bottles of water to absolutely thirsting people in the desert. We are not going to complain that we have to open too many bottles to drink a gallon.*

Culture

The Smithsonian suffers from all the cultural barriers to data management already discussed with respect to the academic culture in general—although, as in the wider world, a generational change is underway. These barriers include researchers'

[117] All of the biology units except SERC have their own IT support staff, and staff typically do not bring their IT issues directly to OCIO.

- Sense of "ownership" of their data and lack of interest in their further use once they are done with them;
- Fear that by releasing a data set before they have finished working through it to their satisfaction, others will get publications from it that might have been theirs;
- Concern that data will be misinterpreted or misused;
- Concern that errors in their analyses or data collection methods will be uncovered;
- Perception of data management (beyond that needed for their own analyses) as an additional burden that has not historically been their responsibility, and that detracts from scarce research time; and
- Lack of incentives to devote time to better data management and sharing.

Other cultural obstacles are more specific to the Smithsonian. Chief among these is the relative insularity of the Smithsonian as an organization (as distinct from individual scientists) vis-à-vis external entities, and of the individual units with respect to one another. Despite an active leadership role in USGEO and regular participation in a number of other prominent national and international initiatives, the Smithsonian was portrayed by some outside interviewees as relatively aloof from potential allies in the federal science, biology research, and natural-history museum communities, and thought that it was missing opportunities to partner and leverage resources. Some internal interviewees suggested that the Smithsonian's biology units sometimes tended to see one another as competitors for grants and central Smithsonian resources, leading to a zero-sum mentality toward cross-unit or pan-Institutional efforts. This insularity also has a tendency, according to some interviewees, to manifest itself as relative indifference to audience and stakeholder needs.

Some interviewees also expressed concern that pan-Institutional digitization plans, policies, initiatives, and efforts too often were undertaken in a vacuum, without sufficient consideration for similar work underway externally. The Smithsonian, according to some, was not making sufficient use of products, approaches, tools, lessons, and resources developed elsewhere, and not taking advantage of opportunities to piggyback on or participate in work done externally. For example, OCIO representatives do not typically engage with outside organizations such as ESIP, even though IT issues can be a very prominent part of such organizations' agendas. Again, this may in part reflect the scarcity of Institutional funding to support such engagement; the study team heard of cases in which participation in relevant external groups had to be funded by staff members themselves.

The study team certainly saw some evidence, at both the unit and Institutional levels, of a tendency to think in terms of internal solutions, rather than looking for ways to leverage or learn from external efforts, lessons, and assets. However, it also saw evidence of a growing appreciation that the challenges of data management and sharing will require a greater willingness to connect with others' efforts and collaborate on solutions. Moreover, as one interviewee noted, the historical tendency to think in terms of internal solutions is by no means unique to the Smithsonian:

> *It's much more difficult usually to look around and try to understand existing standards than to start building your own from scratch. Using existing standards often forces people to think differently about their data. It might feel like you are shoehorning it into something where it doesn't really fit. But in many cases, that's because there are aspects of the data collection or data management processes that were not really considered by the investigator, but were considered in the development of those standards. So while it might seem harder to understand existing standards than to just come up with your own model, in the long run it almost always works out better to spend that initial effort on doing a literature review on existing models and frameworks that you might use.*

Organization

A number of interviewees mentioned that the federated nature of the Smithsonian scientific research enterprise is an obstacle to faster progress with data sharing and management. Not only are biology research efforts distributed across four major research units, but within each unit, there are numerous internal silos (laboratories, research centers, departments, even individual researchers) that function with a high degree of autonomy. Digitization and data-management efforts are fragmented across these silos and not always coordinated. In many cases, personnel operate with little or no knowledge of potentially complementary models, activities, and resources elsewhere within the Smithsonian.

For the most part, internal interviewees thought that a top-down approach with greater Institution-wide coordination is not feasible because of the Institution's organizational structure. Even within some units, it may not be feasible. Rather, the consensus seemed to be that leveraging complementary efforts will require getting buy-in from individual players, rather than knocking heads together.

The recent creation of a Digitization Program Office at OCIO has created a focal point for digital data issues at the Smithsonian. However, none of the major central players in the scientific data area—OUSS, OCIO, SIA, and SIL—currently has the authority or

staffing to effectively coordinate efforts across units or to set and enforce policy. Moreover, there is currently no forum or mechanism to bring these players together at the Smithsonian to develop a big-picture plan for scientific data sharing that accommodates the needs of individual units while leveraging collective strengths.[118]

The question of who, if anyone, speaks for the Smithsonian as a whole with regard to scientific data management and data sharing also arose. Key strategic and policy issues—such as digitization priorities and data access policies—for the most part remain unsettled at both the unit and central Smithsonian levels, although the 2010-2015 digitization strategic plan calls for addressing the issues in the next few years. Even in cases where unit leaders have attempted to introduce priorities or policies for their units, interviewees suggested that compliance is spotty. Across units, and even within, there is little consistency in terms of preferred metadata, ontologies, formats, and so on, even in similar areas of research.

The federated nature of the Smithsonian can also confuse external parties. They may not, for example, appreciate the difference between working with a representative of OUSS and working with a representative of an individual unit, and may not understand when it is appropriate to approach one versus the other. In interviews and conversations with individuals at external forums, the OP&A study team repeatedly heard some variant of the observation that "we just don't know who to call" to initiate a high-level dialogue or collaboration with the Smithsonian.

Taking Care of the Present, Looking to the Future

This report focuses heavily on current conditions at the Smithsonian, and on the need to address basic unmet data-management needs to facilitate discovery, access, use, and preservation. To a large extent, it addresses issues of how to move Smithsonian biology as a whole as rapidly as possible to the frontiers of what is already available and known in these areas.

[118] At the end of 2010, the Smithsonian established the new position of Deputy Under Secretary for Collections and Interdisciplinary Support, effective Dec. 19. Responsibilities include central planning and development of collections and oversight of SIA and SIL. In addition, the Deputy Under Secretary will serve as a liaison between the Smithsonian and various cultural and scientific organizations in the United States and around the world on collections management and cooperative programs. It was unclear at the time of this writing whether and to what extent digital data management and sharing might fall under the purview of the Deputy Under Secretary, but one interviewee noted that the position was created in part to address some of the issues that this report describes.

However, new concepts, tools, technologies, and techniques in data management and sharing are emerging at an increasingly rapid rate. Some that may seem very cutting-edge today—cloud computing, the semantic web, auto-generating ontologies, new techniques for data mining, and so on—could quickly become central to scientific data sharing, just as social networking went from a curiosity to an established tool for outreach and education in the span of a few short years.

Thus, a major and important question raised by one interviewee concerns how the Smithsonian can, at the same time it struggles to address its admittedly daunting current needs, prepare itself for emerging and future concepts, tools, technologies, and techniques, and perhaps even position itself as a leader in exploring and applying some of them. This returns to a wider question that has frequently come up in OP&A management and organizational studies of the Institution, and that will take time to effectively address: How can the Smithsonian transform itself into a 21st century learning organization that accepts rapid change as a norm, and systematically looks to and engages with the outside world to keep on top of change?

Conclusions

Conclusion: The small-science approach to the management and sharing of digital biology research data is anachronistic. It is at variance with the growing emphasis both in U.S. policy and around the world on open access to scientific data, and it may put the results of important research investments at risk and impede long-term access to valuable scientific resources.

Mitigating the severe global challenges society faces, such as climate change, species loss, and detrimental human effects on ecosystems, is a top priority nationally and internationally. Because of the geographic scope of these challenges, the variety of disciplines involved in understanding them, the large volume of data that needs to be collected on them, and the infrastructure requirements and costs of research, it is widely agreed that collaborative and often interdisciplinary research approaches have become imperative. Such approaches require that scientists be able to discover, access, and use interoperable data from different times, places, fields, and disciplines. Because of the potential long-term value of many data, they must be preserved for an indefinite future.

To meet these requirements, the Smithsonian needs to carry out systematic and thorough management of its biology data throughout their lifecycle[119]—that is, from the planning stage of data collection through the point at which the data are deemed no longer to be of value. Currently, most decisions about data management are made at the level of individual units, departments, projects, and often researchers. These decisions are usually based on the immediate needs of the data collectors themselves, and rarely place much weight on external use and long-term preservation. Some key decisions are not made until after data collection is well underway, or even after it has been completed.

The study team encountered a number of very noteworthy and important initiatives at the Smithsonian to further systematic management and data sharing, as well as active participation in external data-sharing efforts. It was concerned, however, about the absence of an Institutional strategy for participation in external efforts. Elements of such a strategy would include identification of priority external initiatives with which the Smithsonian or individual units should engage, and a well thought-out framework of policies, goals, processes, and resources to support participation. It was struck by the small number of

[119] The OP&A study team refers here only to biology data because they were the subject of its study. It recognizes that any steps toward open access and related data management likely would be relevant to other fields and should therefore address the full panoply of Smithsonian research data.

individuals and low level of support allocated to organizational networking with external entities engaged in furthering data management and open access. Internally, communication and coordination of like efforts across Smithsonian units (and sometimes within units) occurred infrequently. In short, to a significant extent, biology at the Smithsonian remains bound by the small-science approach to data management and sharing.

A particularly serious consequence of this situation is that, by all accounts, significant quantities of data are currently at risk of becoming lost or unusable. Some unknown quantity has undoubtedly already reached those states, as a result of being stranded on unstable and outdated storage media, the absence of descriptive metadata, or a lack of knowledge that they even exist. Moreover, without increased attention to data management in the present, the volume of unmanaged legacy data will continue to grow, and at an ever-greater rate, creating even more at-risk data.

Conclusion: To make its digital biology data easily discoverable, accessible, and usable by internal and external users, the Smithsonian needs to unequivocally articulate a policy of open access and systematically establish the capacity and tools to implement that policy.

The study team believes that the Smithsonian, a largely taxpayer-funded entity, has an obligation to provide open access to its biology data, subject where appropriate to reasonable proprietary waiting periods and to exceptions for security, intellectual property rights, and other such considerations. The study team further believes that the biology data Smithsonian scholars produce are national assets—this is particularly evident in the case of unique data sets such as those relating to the voucher specimen collections and those derived from long-term monitoring projects such as SERC's 25-year CO_2-level experiments and STRI's CTFS and SIGEO plot inventories. The status of Smithsonian digital biology data as national assets needs to be stressed at all levels of the organization to address the tendency of scientists to treat their data as proprietary.

In addition to the principle that publicly-funded research should generally be publicly available, there are practical reasons for the Smithsonian to support open access. The rest of the world, including the U.S. government and funders, is moving quickly in that direction, and the Smithsonian's reputation and competitiveness in raising funds may suffer if it pursues an insular approach to sharing its data. The trend toward funders requiring data-sharing and data-

management plans as a condition of grant awards also argues strongly for open access.

Until recently, the Smithsonian has shown ambivalence about open access. The 2010-15 Smithsonian Strategic Plan and the Digitization Strategic Plan, however, specifically call for increased sharing of digital data, although neither uses the term "open access." A starting point to bring this goal to fruition would be the promulgation of a policy of open access to Smithsonian data. The policy could take the form of a standalone Smithsonian Directive (SD) comparable to the one governing its National Collections (SD 600, "Collections Management") and complementary to SD 501, "Archives and Records of the Smithsonian." Alternatively, the policy could be integrated into proposed SD 609, "Digital Asset Access and Use" and SD 610, "Digitization and Digital Asset Management."

However the policy is established, it would need to clearly establish open access as a fundamental operating principle of the Smithsonian's research enterprise and establish external usability as a primary consideration in decisions regarding data-management processes, standards, infrastructure, and technology. It would also need to define the circumstances under which reasonable restrictions can be imposed on data access, and for how long.

A related policy matter is how much effort the Smithsonian should and can devote to supporting wider efforts to advance sharing of digital biology research data. The Smithsonian is already playing a leadership role in some forums and initiatives on data sharing, such as USGEO and EOL. The study team believes it is well-positioned to take on that role in other areas. For example, the important issue of improving the usability of data across fields might be an area where the Smithsonian could take the lead, given how many disciplines are represented within the Institution and the strong push here for interdisciplinary research. Yet another possible area for Smithsonian leadership might be data publishing. While significant work has already been done in this area, widely-accepted systems for peer-reviewing data sets, citing them, and tracking their use after publication do not yet exist. Achieving the critical mass that could lead to wider professional acceptance of the concept may require an organization of the Smithsonian's stature, in conjunction with other high-profile research institutions, to demonstrate how data publishing might work in practice and to push for broad acceptance.

Also on the table is how much representation and participation in external forums and with outside organizations are optimal. In the case of organizations like GEO/GEOSS, for example, existing central representation might usefully be

augmented by representatives from units that are deeply involved in the relevant issues and areas. Similarly, in the case of large forums such as ESIP, with their specialized committees and working groups, representatives from different units and sciences at the Smithsonian is preferable to a single representative of the Institution. Greater participation in external forums would need to be bolstered by inclusion of this activity in position descriptions and by increased resources for travel.

Conclusion: Sharing of Smithsonian biology data requires fundamental changes in its current data-management and -dissemination practices.

Data sharing requires that the Smithsonian undertake proactive efforts to manage its biology data in a manner that makes them easily discoverable, accessible, and usable by internal and external users. The current small-science, seat-of-the-pants, fragmented, and mostly unit- or department-based approach will have to give way to a set of core Institution-wide principles and standards. These will have to be framed with careful attention to the distinctive needs of different sub-disciplines and research projects, and will have to allow some flexibility in application, so as to accommodate particularly innovative or unique research.

In moving forward, it is important that IT staff, information-science personnel, and domain scientists work closely as a team to develop a supportive environment for researchers that offers a variety of appropriate, continuously-upgraded tools, systems, and services to facilitate their role in data management and sharing. Internal Smithsonian efforts will need to draw on and coordinate with experts and organizations in the external environment engaged in advancing data management and sharing.

The following are basic elements critical to increased data sharing, as well as long-term preservation where the data merit such treatment. These could usefully be identified in a Smithsonian-wide policy:

- *Data-management and -sharing standards for the entire life cycle of digital data*. Optimally the lifecycle begins at the time the project is being designed and ends only when a decision is made that the data no longer require preservation. The standards will need to take into account the increasingly common requirement of external funders for sound data-management plans. To the extent possible, the requirements would also apply to the management of legacy data.

- *Criteria for deciding the appropriate level of management and preservation for specific data*, given that the volume of data past, present, and future exceeds the Smithsonian's capacity for long-term stewardship. Development of such criteria is best accomplished in close consultation with both the internal Smithsonian research community and external organizations, both because research today is a multi-organization endeavor and because far more can be accomplished by leveraging resources and avoiding duplication.

- *Compliance of Smithsonian researchers with relevant internal and external data-management standards and requirements.* Some interviewees argued against standards and requirements, particularly those that were not yet universally accepted or that might change. The study team, however, found many examples of benefits from the application of existing standards, metadata, and other requirements. In most cases, their use does not preclude transition to modified or new standards and requirements, as a capacity to adapt is typically built into systems and tools. By contrast, waiting until some unknown time when consensus best practices emerge means perpetuating the inaccessibility of Smithsonian data, putting valuable data at risk, and greatly increasing the cost of dealing with legacy data. Here, too, the Smithsonian will need to keep abreast of external efforts in a variety of fields to develop and refine data-management standards and to avail itself of external expertise and experience.

- *A central record (ideally including the information necessary for discovery, access, and use) of the Smithsonian's digital biology data holdings, and a system for regularly updating the information.* To support discovery of its data, the Smithsonian needs a robust record of what it holds. Ideally, this record would include information such as: subject matter; date(s) and location(s) of data sets; format and medium on which the data are stored; back-up storage; metadata; condition of the data; value and risk of loss; restrictions on use; point of contact for queries; and other information critical to discovery, access, use, and appropriate preservation. Such a listing is an essential underpinning of data management and sharing. Moreover, it could help guide the allocation of resources on the basis of the potential value and level of risk of data sets.

- *Smithsonian tools that facilitate discovery and access.* As noted, successful data sharing requires easy discovery and access. EDAN and SRO are good examples of discovery tools that afford a single point of entry to Smithsonian resources in the areas of collections and research publications, respectively, and that are continually being improved. Something similar for digital biology research data would greatly advance the goal of sharing them with internal and external users. Implementation of such tools will depend in part on creation of the record of data holdings referenced above and the development of descriptive metadata where they are lacking. The Smithsonian can also promote discovery and access through collaboration with external players. As noted in the findings, many organizations maintain portals and other tools that list, provide metadata on, link to, and even host data from external sources. The study team could not find any valid reason for the Smithsonian's reluctance to participate in the federal government's data.gov and Science.gov portals. The Smithsonian would do well to identify and make use of well-established external portals and data centers relevant to its data—as always, subject to reasonable limitations on open access.

- *A trusted digital depository.* At the time of this writing, OCIO was moving to establish a critical piece of an effective data management system: a TDR for long-term storage of and access to usable scientific data. It is to be hoped that funds will be allocated for this initiative in the near future. Decisions about the role of any future Smithsonian TDR are best undertaken with extensive input from representatives of relevant internal and external biology research units. Here, too, the Smithsonian would do well to look for opportunities to leverage external resources. For example, if an established, field-specific TDR exists, relying on it for long-term preservation of relevant Smithsonian digital biology research data could be a quick and cost-effective solution. At the same time, the proposed Smithsonian TDR could host external as well as Smithsonian data. (This might involve cost-sharing arrangements.)

Conclusion: Systematically and immediately addressing the risk of Smithsonian legacy data loss and preventing further growth in the backlog of legacy data are high priorities.

The already very large and rapidly growing volume of legacy data at the Smithsonian requires near-term attention. As noted, the starting point is

identification of existing legacy data at serious risk of loss. Equally important is to ensure that the data of researchers soon to retire, and projects soon to end (or recently ended) receive near-term attention. Once a scientist is gone or a project is completed and the researchers move on, it can be very difficult to get the information needed to develop descriptive metadata. To lessen further growth of legacy data, the Smithsonian also needs in the near term to put in place processes for scientists with ongoing projects to list their data in the record of Smithsonian holdings and to routinely back them up on a stable medium to insure against accidental loss. Optimally, the data would include adequate descriptive metadata and an estimate of the size of the final data set to support forecasts of organizational data-management and -storage needs.

Conclusion: It is important that the central administration be proactive in reaching out to Smithsonian biology researchers to raise awareness of the value of managing and sharing digital data, alert them to the support available to facilitate those tasks, and obtain buy-in among research staff for data management.

The study team is sympathetic to the desire of researchers not to spend scarce time on data management and sharing. Neither can be accomplished effectively, however, without their participation. Fortunately, new technologies and approaches are available that make these tasks easier and more manageable.

- To get buy-in for routine data management and sharing, and to facilitate the work in these areas that must be done by researchers themselves, the Smithsonian might follow the lead of some university libraries that have been proactively reaching out to scientists to explain the personal and societal value of data management, long-term data preservation, and data sharing. At the same time, these libraries have furthered buy-in by providing a menu of resources to facilitate the work and minimize the time scientists have to spend on data management and sharing. Examples of resources are metadata templates with information on how to apply them; links to online sites where scientists can list, link, or store their data; and direct assistance from IT and information-science personnel for tasks such as designing data management plans at the beginning of a research project (and even for inclusion in grant proposals), developing metadata, choosing formats and software, registering data with appropriate data centers, and transferring the data to a central repository when a project is over.

- The Smithsonian needs to offer incentives to scientists to manage and share their data, such as formal professional credit for data publishing,

data sharing, and fulfillment of the front-end data management efforts that must be carried out by the researchers themselves.

- As to the point that scientists are reluctant to share data because other users might misinterpret or misuse them, the study team agrees with those interviewees who maintained that this has not proven to be a problem in practice. Similarly, the identification and correction of errors, should they exist, is in the interest of the larger scientific enterprise, and can be thought of as a variant of the widely-accepted concept of peer review. Overall, a willingness to expose its data to outside scrutiny can only enhance the Smithsonian's reputation for high-quality research.

Conclusion: Meeting the growing challenges of digital data management and sharing at the Smithsonian will require additional resources.

Where the additional funds needed for better data management and sharing will come from is the ever-present conundrum. While the federal government is likely to subsidize some parts of data-management infrastructure and operational costs, the extent, timing, and focus of its commitment are unclear, as are the institutional arrangements through which support would be channeled (including which agencies will have what responsibilities). The poor fiscal climate created by the recent period of recession and weak recovery compounds these uncertainties.

For the Smithsonian, some combination of the following strategies to boost the resources available for data-management support will likely be necessary and merit exploration:

- Pursuing a budget line item for digital data biology management and sharing, given that such data are core national assets that can be critical to the formulation and implementation of federal science policy in numerous areas.

- Approaching data management and sharing more systematically—for example, planning across all science units instead of one by one—so as to use limited resources more efficiently.

- Raising contributions specifically for this purpose from foundations, donors, the Congress, and so on.

- Shifting funds from lower-priority functions and activities to data management and sharing.

- Increasing grant overhead rates and allocating part of them to cyberinfrastructure, data management, data discovery, and new initiatives such as data publishing.

- Leveraging resources through partnerships with other organizations and participation in collaborative initiatives.

- Offering fee-based services (for example, for preserving data from other organizations in the Smithsonian TDR).

With respect to support staff, at all Smithsonian biology units the balance between "data producers" (research scientists) and "data managers" (IT and information-science personnel) is heavily skewed toward the former. As the study team has found in other studies within the Smithsonian science community, the overwhelming emphasis has been on researchers, rather than support personnel. But the increased attention to, and requirements for, effective data management and sharing have already created a demand for more of the latter at both the central and unit levels, and that demand is likely to grow. A policy requiring open access will necessitate the development of a cadre of research support personnel well-versed in and dedicated to data management and sharing. Collectively, that cadre needs to combine IT, information-science, and domain-science expertise. An open question is the appropriate balance between (1) staff permanently assigned to particular units and (2) central support unit staff "owned" by units such as OUSS, OCIO, SIL, and SIA, with the flexibility to rotate among units as needed.

Conclusion: Significant collaboration with external organizations will need to be part of the Smithsonian's approach to managing and sharing its data.

Two related themes have surfaced repeatedly in this report:

- Extensive and important work on data management and sharing is taking place outside the Smithsonian, and this work offers a wide array of valuable resources.

- The Smithsonian needs to take advantage of this work and become an active member in the key collaborative communities undertaking it.

The study team sees several reasons why it would benefit the Smithsonian to participate more fully in external data management and sharing initiatives. First, it does not have the expertise or resources to do everything that needs doing;

collaboration with other organizations offers potential shortcuts to its goals, as well as leveraging of resources. Second, being part of the process raises the Smithsonian's visibility and reputation as a major player in the national and international research communities, and as a collaborative research organization. Equally important, active participation ensures that Smithsonian interests are represented in final outcomes and products.

The study team did not feel that the Smithsonian has systematically explored and taken advantage of opportunities to engage with other organizations working on similar data-management and -sharing challenges. While outreach has occurred in a number of areas and venues, it has tended to be unsystematic. Certainly, there is no central plan or central organizational focal point (analogous to the Digitization Program Office with respect to digitization efforts more broadly) for tapping into and taking advantage of the experiences and resources of the external community.

Relatedly, the study team believes that the Smithsonian needs to give at least as much weight to external standards (technological, archival, documentation, etc.) as to internal Smithsonian preferences when it makes decisions about data-management policy, standards, and practices. It would benefit from staying abreast of external standards where they exist, and working with peer organizations to establish standards where they do not, so that its interests are taken into account. While internal needs or the consequences of past decisions sometimes may require departures from the ideals of external usability and interoperability, such departures should be rare and undertaken for compelling reasons.

Conclusion: The Smithsonian will need to put in place an organizational structure, with clear roles and responsibilities at the levels of the central administration and research units, to ensure coordinated implementation of sound data-management and data-sharing standards, systems, and practices.

Absent an Institutional focus on data management and sharing, activities in this area have been fragmented and often opportunistic. There has been relatively little Institutional attention to or support for these efforts, and no overarching strategy and framework to guide and link them. A more systematic, structured approach that rationally distributes responsibilities and leverages resources across units is needed.

One office—OCIO—has endeavored in the last few years to coordinate a systematic approach to all things digital, including digital research data. It is to be

commended for leading the development of a pan-Institutional digitization strategy and moving forward with plans for a TDR. It has done so through a participatory process that has included representation from the science community and other relevant internal stakeholders. Nevertheless, it is not clear that the specific characteristics and needs of biology data management and sharing are being adequately addressed. Moreover, notwithstanding language that is inclusive of digital research, it seems that OCIO's focus to date has tended to be on issues and infrastructure related to the Smithsonian's administrative operations, object and archival collections digitization, and public programming.

Further, however much OCIO is willing to do, some aspects of data management and sharing are beyond its area of operations and expertise, nor does it have organizational authority to mandate data management and sharing. It is unlikely to ever have, nor should it be expected to have, the specific domain-science expertise to address the unique data-management and -sharing requirements of biology. It cannot, for example, drive data management plans for research projects, make decisions about descriptive metadata, or resolve issues of interdisciplinary usability. Rather, digitization issues specific to particular fields or types of data need to be addressed by domain specialists. From the perspective of interdisciplinary usability, representatives of all the natural and physical sciences need to engage with the IT and information-science sides of the Institution.

One open organizational question is the appropriate roles for SIA and SIL. SIA is already somewhat involved with the management and preservation of scientific data, and will continue to be involved under any plausible future scenario. SIL currently supports the research community by facilitating the discovery and retrieval of scientific publications and other types of information (including in digital formats), and appears in a limited way to be taking up the important issue of linking published results to the relevant underlying data. But in general SIL has not been active in managing or providing access to research data sets per se, and has not shown a strong interest in actively promoting and facilitating data management among researchers, as some university libraries have. Regardless of what role ultimately falls to SIA and SIL, it is likely that digital data discovery, access, and sharing will require some level of training for existing staff and perhaps recruitment of specialist staff.

Because data management and sharing are functions that extend across research units while simultaneously involving common infrastructure such as servers and other IT assets managed by OCIO, the study teams believes that a single

Institutional focal point is needed to mediate among the needs of the various research units and the central administration. OUSS seems well-situated to take on this role, although to do so effectively would most likely require an infusion of resources—for example, augmenting its capacity for external networking and bringing in staff who combine domain-science and information-science expertise.

Recommendations

The OP&A study team has three overarching recommendations:

1. **The Smithsonian should unequivocally commit to a policy of open access to its digital biology data,[120] subject to reasonable restrictions, including an initial embargo period to allow researchers to publish. Data sharing and the systematic underlying data management needed to support it should be fully integrated into the Institution's biology research enterprise, and external usability should be a primary consideration in decisions regarding data-management processes, standards, infrastructure, and technology.**

2. **The Smithsonian should establish the capacity and tools to make its digital biology data easily discoverable, accessible, and usable by present and future users, internal and external.**

3. **The Smithsonian should engage more fully and systematically with external organizations working to advance data sharing and management, taking on a leadership role in areas where it has particular expertise and resources or where it is in the Institution's strategic interests.**

To accomplish these three core recommendations:

4. **OUSS should convene a working group to (1) develop a plan of action for managing and sharing digital biology data and (2) draft a policy to govern biology data management and sharing.**

    - The **working group** should include representatives from the Smithsonian biology research units, OCIO, SIL, SIA, the Office of Human Resources, central management, and other relevant personnel.

    - The working group should draft a **plan of action** that addresses the issues enumerated below, with the option of splitting the work into two parts so that priority issues such as the potential loss of legacy data can be addressed in the near term.

[120] The OP&A study team limits its recommendations to digital biology data, which were the topic of the study. It presumes, however, that any policy directive would likely encompass digital data from all science (and possibly social science) research conducted at the Smithsonian.

- Based on the plan of action, the working group should draft a **policy for digital biology data management and sharing**. This policy could be integrated into existing Smithsonian Directives, included in the proposed SDs 609 and 610, or issued as a stand-alone Directive.

- The plan of action referenced above should address the following:

    - *Definition of open access*—including when data would be made available, guidelines for restricting access, and processes for decision making on access.

    - *Provision of a status for Smithsonian digital biology data comparable to that of the National Collections covered in SD 600*.

    - *Core requirements for digital biology data management over their lifecycle to facilitate discovery, access, and use.* Such requirements might include common Smithsonian-wide data management standards, specifications for metadata (defined in the context of particular fields or types of research), and acceptable formats.

    - *Infrastructure needed to support digital biology data sharing*—including a data portal and a TDR for long-term preservation.

    - *Near-term secure storage of the Smithsonian's digital biology legacy data*—including storage as-is in a secure repository until the data can be assessed for future value and subsequently curated or disposed of as appropriate.

    - *Development and maintenance of a record of the Smithsonian's digital biology data holdings*—including adequate metadata to support discovery, access, and use.

    - *Measures to minimize the continued growth of legacy data*—including a requirement that researchers prepare a data management plan as part of project design; list the data they are collecting in a central record and update their status over the lifetime of the project; and ensure that their data meet at least the basic criteria for discovery, access, and use before being transferred to a secure central location.

- *Criteria for determining the appropriate level of data management for specific data sets.*

- *A central biodiversity informatics/data-management support capability for the Smithsonian biology research community.*

- *Professional incentives for biology researchers to engage in data management and sharing, and provision of tools and support services to assist them in these efforts:*

    * Include in staff position descriptions expectations for data management and sharing and for participation in forums on data management, and provide formal credit in performance evaluations for data publishing and citations, and for participation in external and internal data-management and -sharing initiatives;
    * Require that Smithsonian scientists, prior to leaving the Institution, consult with a responsible party to determine what should happen to their data, and engage with support personnel to carry out whatever data management is necessary to prepare these data for transfer for long-term preservation if appropriate;
    * Provide services and tools to minimize the time researchers need to spend on data management and sharing;
    * Raise researchers' awareness of the importance of data management, long-term preservation, and sharing, and of the personal and societal benefits.

- *Systems and tools for easy discovery, access, and use of Smithsonian data by internal and external users*.

- *Workforce requirements and deployments at the central and unit levels*—including consideration of the appropriate numbers and types of support personnel for major data-management and -sharing tasks, and the appropriate balance between researchers and research support staff.

- *Increased Smithsonian participation in relevant external initiatives and forums*—including taking on a leadership role in appropriate areas of Smithsonian strength and strategic interests.

- *Strategies to increase or leverage funding for data management and sharing:*

    * Pursuing federal allocations specifically for data management and sharing, including a line item in the Smithsonian's federal budget, particularly in the context of assuming responsibility for appropriate parts of federal cyberinfrastructure for scientific data;
    * Increasing and reallocating overhead allowance rates on scientific grants to reflect growing requirements and costs for data management and sharing.
    * Leveraging resources through partnerships and participation in collaborative initiatives.
    * Making more efficient use of internal resources, and shifting funds from lower-priority functions to data management and sharing.
    * Providing services for fees.
    * Coordinating with the National Science Foundation (NSF), Library of Congress (LOC), National Archives and Records Administration (NARA), and other major national scientific and library/archival organizations to raise awareness in Congress, the Office of Management and Budget (OMB), and the public about the wider societal benefits of data sharing (particularly with respect to addressing global environmental challenges), and the importance of federal investment to defray the costs of a national data-sharing infrastructure.
    * Participating actively in forums that discuss federal investment in data management and sharing.

- *Design of an organizational structure to support data management and sharing at all levels:*

    * Definition of roles and responsibilities for Smithsonian central support offices (particularly OUSS, OCIO, SIL, and SIA) and research units—including which unit(s) have primarily responsibility for implementing specific parts of the plan of action;
    * Development of communication and coordination mechanisms to leverage relevant resources across units, ensure smooth internal collaboration, and disseminate lessons learned across the Institution.

- *Identification of the highest near-term priorities:*

* Prevention of further loss of unmanaged legacy data;
* Identification of national and global initiatives in which the Smithsonian should participate;
* Explicit inclusion of participation in such initiatives in staff job descriptions, and/or professional credit for participation;
* Provision of funds for travel and other support for such participation;
* Development of a system for keeping the Smithsonian abreast of relevant external developments on an ongoing basis, and identifying promising opportunities for leveraging resources through new collaborations.
* Engagement with data.gov and Science.gov.

5. **The Smithsonian should issue a digital biology data-management and -sharing policy, based on the draft policy of the working group.**

6. **The Smithsonian should begin implementation of the near-term priorities identified in the plan of action as soon as possible following receipt of the working group's recommendations.**

Appendix A: Selected Bibliography

Anderson, Martha. 2008. "Evolving a Network of Networks: The Experience of Partnerships in the National Digital Information Infrastructure and Preservation Program." *International Journal of Digital Curation* 1(3): 4-14.

Anderson, W.L. 2004. "Some Challenges and Issues in Managing and Preserving Access to Long-lived Collections of Digital Scientific and Technical Data." *Data Science Journal* 3: 191-201.

Arzberger, Peter, Peter Schroeder, Anne Beaulieu, Geof Bowker, Kathleen Casey, Leif Laaksonen, David Moorman, Paul Uhlir, and Paul Wouters. 2004a. "An International Framework to Promote Access to Data." *Science* 303(19): 1777-1778.

——. 2004b. "Promoting Access to Public Research Data for Scientific, Economic, and Social Development." *Data Science Journal* 3(29): 135-152.

Association of Research Libraries (ARL). 2009. *The Research Library's Role in Digital Repository Services.* Final Report of the ARL Digital Repository Issues Task Force.

——. 2006. *To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering.* Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships.

Ayris, Paul, Richard Davies, Rory McLeod, Rui Miao, Helen Shenton, and Paul Wheatley. 2008. *The LIFE^2 Final Project Report.* Accessed September 27, 2010 at http://eprints.ucl.ac.uk/11758/.

Baker, Karen S., and Lynn Yarmey. 2009. "Data Stewardship: Environmental Data Curation and a Web-of-Repositories." *International Journal of Digital Curation* 4(2): 1-16.

Barksdale, Jim, and Francine Berman. 2007. "Saving Our Digital Heritage." *The Washington Post*, May 16: A15.

Beagrie, Neil, Julia Chruszcz, and Brian Lavoie. 2008. *Keeping Research Data Safe: A Cost Model and Guidance for UK Universities*. Accessed May 3, 2010 at http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf.

Bell, Gordon, Tony Hey, and Alex Szalay. 2009. "Beyond the Data Deluge." *Science* 323: 1297-1298.

Berman, Francine. 2008. "Got Data? A Guide to Data Preservation in the Information Age." *Communications of the ACM* 51(12): 50-56.

Berman, Francine, and Reagan Moore. 2006. "Designing and Supporting Data Management and Preservation Infrastructure." *CT Watch Quarterly*, May. Accessed May 7, 2010 at http://www.ctwatch.org/quarterly/articles/2006/05/designing-and-supporting-data-management-and-preservation-infrastructure/index.html.

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. 2010. *Sustainable Economics for a Digital Planet: Ensuring Long-term Access to Digital Information*. Final report.

——. 2008. *Sustainable Economics for a Digital Planet: Ensuring Long-term Access to Digital Information*. Interim report.

Borgman, Christine L. 2008. "Data, Disciplines, and Scholarly Publishing." *Learned Publishing* 21(1): 29-38.

——. 2007. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. MIT Press.

——. 2006. "Little Science Confronts the Data Deluge: Habitat Ecology, Embedded Sensor Networks, and Digital Libraries." *International Journal of Digital Libraries* 7(1-2): 17-30. Preprint (final paper revisions).

Borgman, Christine L., Jillian C. Wallis, and Noel Enyedy. 2006. "Building Digital Libraries for Scientific Data: An Exploratory Study of Data Practices in Habitat Ecology." In J. Gonzalo, et al. (eds). *Lecture Notes in Computer Science 4172*: 170-183. Springer-Verlag.

Brandt, D. Scott. 2007. "Librarians as Partners in e-research: Purdue University Libraries Promote Collaboration." *College and Research Libraries News* 68(6): 365-367.

Brandt, D. Scott, Jacob Carlson, Melissa Cragin, Bryan Heidorn, Carole Palmer, Sarah Shreeves, and Michael Witt. 2008. "Investigating Data Curation Profiles Across Multiple Research Disciplines." Purdue Libraries Research Publications.

Caplan, Priscilla. 2005. "Digital Preservation and Trusted Digital Repositories." PowerPoint presentation at the American Library Association annual conference, Chicago, Illinois. Accessed September 24, 2010 at http://www.fcla.edu/digitalArchive/presents/DigitalPreservationandTrustedDigitalRepositories.ppt.

CENDI (Federal Science and Technology Information Managers Group). Digital Preservation Task Group. 2007. *Formats for Digital Preservation: A Review of Alternatives and Issues*.

Chi, Ed H., Sean Munson, Gerhard Fischer, Sarah Vieweg, and Cynthia Parr. 2010. Design of Social Participation Systems. NSF Technology Mediated Social Participation Workshop. Version 10. July 21.

Choudhury, Sayeed. 2009. "The Data Conservancy." PowerPoint presentation at the Coalition for Networked Information spring forum, April 7. Accessed September 27, 2010 at http://www.cni.org/tfms/2009a.spring/CNI_Choudhury.pdf.

Choudhury, Sayeed, Benjamin Hobbs, Mark Lorie, and Nicholas Flores. 2002. "A Framework for Evaluating Digital Library Services." *D-Lib Magazine* 8(7/8).

CODATA. 2009. "The Importance of Data Sharing Within GEOSS: CODATA Speaks to José Achache, Director, GEO Secretariat." *CODATA Newsletter* May (special issue).

Consultative Committee for Space Data Systems. 2002. *Reference Model for an Open Archival Information System.*

Costello, Mark J. 2009. "Motivating Online Publication of Data." *BioScience* 59: 418-427.

Cukier, Kenneth. 2010. "Data, Data Everywhere: A Special Report on Managing Information." *The Economist*, February 27.

"Data's Shameful Neglect." 2009. Editorial. *Nature* 461: 145.

Eakin, Lorrain, Amy Friedlander, and Roger Schonfeld, with Sayeed Choudhury. n.d. "A Selective Literature Review on Digital Preservation Sustainability." Accessed May 3, 2010 at http://brtf.sdsc.edu/biblio/Cost_Literature_Review.pdf.

Eastwood, Terry. 2004. "Appraising Digital Records for Long-term Preservation." *Data Science Journal* 3: 202-208.

Ecological Society of America. 2009. Incentives for Data Sharing in Ecology, Evolution, and Organismal Biology workshop report (February 19-20).

——. 2006. Data Centers for Ecology, Evolution, and Organismal Biology workshop report (December 8-9).

Foster, Ian, et al. 2005. "Service-Oriented Science." *Science* 308: 814-817.

Fromer, Margot. 2001. "Public Access to Research Data: A Right to Know or Off Limits?" *Oncology Times* 23(5): 52-54.

Gantz, John F., Christopher Cox, Alex Manfrediz, Stephen Minton, David Reinsel, Wolfgang Schlichting, and Anna Toncheva. 2008. *The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011*. IDC International Data Corporation. Accessed May 7, 2010 at www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf.

Geospatial Multistate Archive and Preservation Partnership. 2009. "FGDC and Dublin Core Metadata Comparison." Accessed September 27, 2010 at http://www.geomapp.net/docs/MetadataComparison_200903.pdf.

Global Research Library 2020. 2009. *Outcomes and Recommendations of GRL2020 Asia.* Report on the 3rd Global Research Library 2020 Workshop, Taipei, Taiwan.

Gold, Anna. 2007a. "Cyberinfrastructure, Data, and Libraries, Part 1: A Cyberinfrastructure Primer for Librarians." *D-Lib Magazine* 13 (9/10).

——. 2007b. "Cyberinfrastructure, Data, and Libraries, Part 2: Libraries and the Data Challenge: Roles and Actions for Libraries." *D-Lib Magazine* 13 (9/10).

Harley, Diane, Sophia Krzys Acord, Sarah Earl-Novell, Shannon Lawrence, and C. Judson King. 2010. "Chapter 4: Biology Case Study." *Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines.* Center for Studies in Higher Education, University of California, Berkeley. Accessed April 26, 2010 at http://escholarship.org/uc/cshe_fsc.

Harvard University. 2009. Report of the Task Force on University Libraries.

Helly, J., H. Staudigel, and A. Koppers. 2003. "Scalable Models of Data Sharing in Earth Sciences." *Geochemistry, Geophysics, Geosystems (G^3) Technical Brief* 4(1)(January 25).

Hendler, James. 2003. "Science and the Semantic Web." *Science* 299: 520-521.

Hey, Tony, and Anne Trefethen. n.d. "The Data Deluge: An e-Science Perspective." Preprint of chapter in Francine Berman, Geoffrey Fox, and Tony Hey, eds., 2003, *Grid Computing: Making the Global Infrastructure a Reality*.

Hey, Tony, Stewart Tansley, and Kristin Tolle. 2009. *The Fourth Paradigm: Data-intensive Scientific Discovery*. Microsoft External Research. Accessed March 17, 2010 at http://research.microsoft.com/en-us/collaboration/fourthparadigm/default.aspx).

Hodge, Gail, and Evelyn Frangakis. 2004. *Digital Preservation and Permanent Access to Scientific Information: The State of the Practice.* Report sponsored by the International Council for Scientific and Technical Information (ICSTI) and the U.S. Federal Information Managers Group (CENDI). Accessed September 24, 2010 at http://www.icsti.org/documents/04-3dig_preserv.html.

Interagency Working Group on Digital Data. 2009. *Harnessing the Power of Digital Data for Science and Society*. Report to the Committee on Science of the National Science and Technology Council.

International Council for Science (ICSU). 2008. Final Report of the Ad Hoc Strategic Committee on Information and Data to the ICSU Committee on Scientific Planning and Review.

——. 2004. Scientific Data and Information. Report of the Committee on Scientific Planning and Review Assessment Panel.

International Council for Scientific and Technical Information (ICSTI). 2009. *Managing Data for Science.* Report of the ICSTI conference, Ottawa, Ontario, June 9-10. Accessed September 24, 2010 at http://www.icsti.org/reports.php.

Jäger, Efrat, Ilkay Altintas, Jianting Zhang, Bertram Ludäscher, Deana Pennington, and William Michener. 2005. "A Scientific Workflow Approach to Distributed Geospatial Data Processing using Web Services." Accessed September 27, 2010 at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.75.1569&rep=rep1&type=pdf.

Johnston, Leslie. 2009. "Identifying and Implementing Modular Repository Services: Transfer and Inventory." In *Proceedings of DigCCurr2009: Digital Curation—Practice, Promise and Prospects.* School of Library and Information Science, University of North Carolina at Chapel Hill.

Key Perspectives, Ltd. *Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long-term Viability: A Comparative Review Based on Sixteen Case Studies*. SCARP Synthesis Study commissioned by the UK Digital Curation Centre. Accessed September 27, 2010 at http://www.dcc.ac.uk/scarp.

Kuipers, Tom, and Jeffrey van der Hoeven. 2009. *PARSE.Insight: Insight into Digital Preservation of Research Output in Europe—Survey Report.* Deliverable D3.4, December.

Lavoie, Brian F. 2008. "The Fifth Blackbird: Some Thoughts on Economically Sustainable Digital Preservation." *D-Lib Magazine* 14(3/4).

Lavoie, Brian F., and Lorcan Dempsey. 2004. "Thirteen Ways of Looking at ... Digital Preservation." *D-Lib Magazine* 10(7/8).

Leake, Jonathan. 2008. "Captains' Logs Yield Climate Clues: Records Kept by Nelson and Cook Are Shedding Light on Climate Change." *The Sunday Times* (London), August 3. Accessed May 4, 2010 at http://www.timesonline.co.uk/tol/news/environment/article4449527.ece.

Lord, Philip, Alison MacDonald, Liz Lyon, and David Giaretta. n.d. "From Data Deluge to Data Curation." Summary findings of the JISC e-Science Data Curation report. Accessed September 24, 2010 at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.111.7425&rep=rep1&type=pdf

Markoff, John. 2010. "U.S. Scientists Given Access to Cloud Computing." *The New York Times*, February 4.

Mayernik, Matthew, Jillian C. Wallis, Alberto Pepe, and Christine L. Borgman. 2008. "Whose Data Do You Trust? Integrity Issues in the Preservation of Scientific Data." iConference paper. Accessed September 24, 2010 at https://www.ideals.illinois.edu/bitstream/handle/2142/15119/PA10-1_iconf08.pdf?sequence=2.

Michener, William K., James W. Brunt, John J. Helly, Thomas B. Kirchner, and Susan G. Stafford. 1997. "Nongeospatial Metadata for the Ecological Sciences." *Ecological Applications* 7(1): 330-342.

Mullins, James L. 2007. "Enabling International Access to Scientific Data Sets: Creation of the Distributed Data Curation Center (D2C2)." Purdue Libraries Research Publications.

National Academy of Sciences. 2009. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age.* National Academies Press.

National Research Council. 2003. *Sharing Publication-related Data and Materials: Responsibilities of Authorship in the Life Sciences.* National Academies Press.

——. 1999. *A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases.* National Academies Press.

——. 1997. *Bits of Power: Issues in Global Access to Scientific Data.* National Academies Press.

——. 1995. *Finding the Forest in the Trees: The Challenge of Combining Diverse Environmental Data.* National Academies Press.

National Science Foundation. 2007. Cyberinfrastructure Council. *Cyberinfrastructure Vision for 21st Century Discovery.*

National Science Foundation. Blue-Ribbon Advisory Panel on Cyberinfrastructure. 2003. *Revolutionizing Science and Engineering Through Cyberinfrastructure.*

National Science Foundation. Directorate for Biological Sciences Advisory Committee. 2003. *Building a Cyberinfrastructure for the Biological Sciences—2005 and Beyond: A Roadmap for Consolidation and Exponentiation.*

National Science Foundation. National Science Board. 2005. *Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century.*

Nelson, Bryn. 2009. "Empty Archives." *Nature,* 461: 160-163.

Oak Ridge National Laboratory. Environmental Sciences Division. 2001/2007. *Best Practices for Preparing Environmental Data Sets to Share and Archive.* Written 2001 by Cook, et al. and initially entitled *Best Practices for Preparing Ecological and Ground-based Data Sets to Share and Archive.* Updated 2007.

Office of Management and Budget. 2009. Open Government Directive (December 8).

——. 2002a. Circular number A-16 (Coordination of Geographic Information and Related Spatial Data Activities), revised.

——. 2002b. Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies. *Federal Register* 67(36): 8452-8460.

——. 2000. Circular number A-130 (Management of Federal Information Resources), revised.

Onsrud, Harlan, and James Campbell. 2007. "Big Opportunities in Access to 'Small Science' Data." *Data Science Journal* 6 (Open Data Issue, June 17).

Organization for Economic Cooperation and Development. 2007. OECD Principles and Guidelines for Access to Research Data from Public Funding. Accessed August 13, 2010 at http://www.oecd.org/dataoecd/9/61/38500813.pdf.

PARSE.Insight. 2009. Deliverable D2.1: Draft Roap Map.

Parr, Cynthia. TMSP Position Paper. Encyclopedia of Life. National Museum of Natural History, Smithsonian Institution. Unpublished.

Pepe, Alberto, Christine L. Borgman, Jillian C. Wallis, and Matthew S. Mayernik. 2007. "Knitting a Fabric of Sensor Data Resources." Accessed September 24, 2010 at http://polaris.gseis.ucla.edu/cborgman/pubs/pepe_ipsn_dsi_8.pdf.

Pilat, Dirk, and Yukiko Fukasaku. 2007. "OECD Principles and Guidelines for Access to Research Data from Public Funding." *Data Science Journal* 6 (Open Data Issue, June 17).

Research Data Strategy Working Group. 2008. *Stewardship of Research Data in Canada: A Gap Analysis.* Accessed September 7, 2010 at http://data-donnees.gc.ca/docs/GapAnalysis.pdf.

Research Information Network and the British Library. 2009. *Patterns of Information Use and Exchange: Case Studies of Researchers in the Life Sciences.*

Romanello, Samantha, James Beach, Shawn Bowers, Matthew Jones, Bertram Ludäscher, William Michener, Deana Pennington, Arcot Rajasekar, and Mark Schildhauer. 2005. "Creating and Providing Data Management Services for the Biological and Ecological Sciences: Science Environment for Ecological Knowledge." Accessed September 24, 2010 at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.60.2421&rep=rep1&type=pdf.

Sabourin, Michel, and Bernard Dumouchel. 2007. "Canadian National Consultation on Access to Scientific Research Data." *Data Science Journal* 6 (Open Data Issue, June 17).

Savage, Caroline J., and Andrew J. Vickers. 2009. "Empirical Study of Data Sharing by Authors Publishing in PLoS Journals." *PLoS ONE* 4(9): e7078.

Schofield, Paul, et al. 2009. "Post-publication Sharing of Data and Tools." *Nature* 461: 171-173.

Schröder, Peter. 2007. "Possible Downsides to Data Sharing in the Research Commons: Assets and Liabilities, Opportunities and Risks." *Data Science Journal* 6 (Open Data Issue, June 17).

Shadbolt, Nigel, Wendy Hall, and Tim Berners-Lee. 2006. "The Semantic Web Revisited." *IEEE Intelligent Systems* May/June: 96-101.

Sharp, James M., and James L. Gumnick. 1981. "A Method for Peer Group Appraisal and Interpretation of Data Developed in Interdisciplinary Research Programmes." *Journal of the Society of Research Administrators* 13(2): 51-66.

Skinner, Katherine, and Matt Schultz (eds). 2010. *A Guide to Distributed Digital Preservation.* Atlanta, Georgia: Educopia Institute.

Smithsonian Institution. n.d. Smithsonian Directive 609, Digital Asset Access and Use. Draft.

Smithsonian Institution. Digitization Strategic Plan Committee. 2010. *Creating the Digital Smithsonian: The Smithsonian Institution Digitization Strategic Plan. Fiscal Years 2010-2015*.

Smithsonian Institution. Office of the Chief Information Officer. 2010. *Smithsonian Institution Information Technology Plan FY 2010 to FY 2015*.

Smithsonian Institution. Office of Policy and Analysis. 2010. *Collaborations in Conserving Time-Based Art.*

——. 2009a. *Addressing Complexity: Fostering Collaboration and Interdisciplinary Science Research at the Smithsonian.*

——. 2009b. *Lessons for Tomorrow: A Study of Education at the Smithsonian.*

Staudigel, Hubert, John Helly, Anthony A.P. Koppers, Henry F. Shaw, William F. McDonough, Albrecht W. Hofmann, Charles H. Langmuir, Kerstin Lehnert, Baerbel Sarbas, Louis A. Derry, and Alan Zindler. 2003. "Electronic Data Publication in Geochemistry." *Geochemistry, Geophysics, Geosystems (G^3) Technical Brief* 4(3).

Thomas, Claire. 2009. "Biodiversity Databases Spread, Prompting Unification Call." *Science* 324: 1632-1633.

Toronto International Data Release Workshop. 2009. "Pre-publication Data Sharing." *Nature* 461: 168-170.

Uhlir, Paul F. 2009. "Revolution and Evolution in Scientific Information." Presentation to the Conference on Managing Data for Science, International Council for Scientific and Technical Information, Library and Archives Canada, Ottawa, June 10.

——. 2007. "Open Data for Global Science: A Review of Recent Developments in National and International Scientific Data Policies and Related Proposals." *Data Science Journal* 6 (Open Data Issue, June 17).

——. 2005. "Creating a Global Information Commons for Public Science." Presentation at the Institute for Advanced Studies, University of Sao Paolo, September 23.

Uhlir, Paul F., and Peter Schröder. 2007. "Open Data for Global Science." *Data Science Journal* 6 (Open Data Issue, June 17).

U. S. Congress. 2007. America COMPETES Act. Public Law No. 110-69. Accessed June 7, 2010 at http://thomas.loc.gov/cgi-bin/query/D?c110:5:./temp/~c110OuLt44.

——. 2002. E-Government Act. 2002. Public Law 107–347. Accessed July 28, 2010 at http://www.reg-group.com/library/E-GovLaw.pdf.

U.S. Long Term Ecological Research Network (LTER). 2007. *The Decadal Plan for LTER: Integrative Science for Society and the Environment.* LTER Network Office Publication Series, number 24.

USGEO. 2009. "Progress Report of the U.S. Group on Earth Observations." Paper prepared for the GEO-VI Plenary, Washington D.C., November 17-18.

——. 2008. Architecture and Data Management Working Group. "Exchanging Data for Social Benefit: An Integrated Earth Observation System Web Services Architecture." Accessed September 24, 2010 at http://usgeo.gov/images/USGEOMain/ExchangingDataforSocietalBenefit.pdf.

Van Garderen, Peter. 2006. "Digital Preservation: An Overview." Presentation at the Managing Information Assets in the Public Sector Conference, Edmonton, Alberta, October 12-13, 2006.

Various authors. 2006. Association of Research Libraries / National Science Foundation Workshop on Long-Term Stewardship of Digital Data Collections position papers. Accessed September 23, 2010 at http://www.arl.org/pp/access/nsfworkshop.shtml.

Wallis, Jillian C., Christine L. Borgman, Matthew S. Mayernik, and Alberto Pepe. 2008. "Moving Archival Practices Upstream: An Exploration of the Life Cycle of Ecological Sensing Data in Collaborative Field Research." *International Journal of Digital Curation* 3(1): 114-126.

Walpole, Matt, et al. 2009. "Tracking Progress Toward the 2010 Biodiversity Target and Beyond." *Science* 325: 1503-1504.

Walters, Tyler O., and Robert H. McDonald. 2008. "Creating Trust Relationships for Distributed Digital Preservation." Accessed September 27, 2010 at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.169.4920&rep=rep1&type=pdf.

Whitlock, Michael C., Mark A. McPeek, Mark D. Rausher, Loren Rieseberg, and Allen J. Moore. 2010. "Data Archiving." *The American Naturalist* 175(2): 145-146.

Witt, Michael. 2009. "Institutional Repositories and Research Data Curation in a Distributed Environment." *Library Trends* 57(2): 191-201.

Whitfield, John. 2008. "An Indifference to Boundaries." *Nature* 451: 872-873.

World Data Center Modernization Task Team. 2003. "Towards a New World Data Center System: Meeting Global Needs." Accessed April 26, 2010 at http://www.ngdc.noaa.gov/wdc/reports/Moderinzation_ReportFinal_121203.pdf.

Xu, Guan-Hua. 2007. "Open Access to Scientific Data: Promoting Science and Innovation." *Data Science Journal* 6 (Open Data Issue, June 17).

Young, Peter. 2010. "The Role of Libraries and Archives in Data Management." PowerPoint presentation to UCLA Department of Information Studies course "Data, Data Practices, and Data Curation," March 10, 2010.

Zimmerman, Ann S. 2008. "New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data." *Science, Technology, and Human Values* 33(5): 631-652.

Zorich, Diane M., Günther Waibel, and Ricky Erway. 2008. *Beyond the Silos of the LAMs: Collaboration Among Archives, Libraries, and Museums.* Dublin, Ohio: OCLC Research. Accessed September 24, 2010 at http://www.oclc.org/research/publications/library/2008/2008-05.pdf.

Appendix B: Organizations Interviewed for the Study

Individuals from the following organizations were interviewed for this study.

External Organizations

Board on Research Data and Information (BRDI) (The National Academies)

Center for Air Pollution Impact and Trend Analysis (CAPITA), Washington University

Center for International Earth Science Information Networks (CIESIN) (Columbia University)

Coalition for Networked Information (CNI)

Committee on Data for Science and Technology (CODATA) (ICSU)

The Data Conservancy (Johns Hopkins University)

Data Observation Network for Earth (DataONE) (University of New Mexico, Oak Ridge National Laboratory, U.S. Geological Survey (USGS), National Center for Ecological Analysis and Synthesis, and Duke University)

Digital Earth Watch (DEW Project)

Ecological Society of America (ESA)

Federation of Earth Science Information Partners (ESIP)

FRAMES (University of Idaho)

Global Biodiversity Information Facility (GBIF) (University of Kansas)

Group on Earth Observations / Global Earth Observations Systems of Systems (GEO/GEOSS)

Incorporated Research Institutions for Seismology (IRIS)

Interagency Working Group on Digital Data (ISGDD)

Intergovernmental Panel on Climate Change (IPCC)

Library of Congress

Long Term Ecological Research Network (LTER)

Massachusetts Institute of Technology (MIT)

National Academy of Sciences (NAS)

National Aeronautics and Space Administration (NASA)

National Archives and Records Administration (NARA)

National Biological Information Infrastructure (NBII)

National Center for Ecological Analysis and Synthesis (NCEAS) (University of California at Santa Barbara)

National Digital Information Infrastructure Program (NDIIP)

National Ecological Observation Network (NEON)

National Oceanic and Atmospheric Administration (NOAA) (U.S. Department of Commerce)

National Science Foundation (NSF)

National Science and Technology Council (NSTC)

Oak Ridge National Laboratory (ORNL)

Office of Science and Technology Policy (OSTP)

Purdue University Libraries

The New Media Studio

Graduate School of Education and Information Studies (University of California at Los Angeles)

Educational Leadership Program, School of Education and Human Development (University of Southern Maine)

U.S. Environmental Protection Agency (EPA)

U.S. Geological Survey (USGS) (U.S. Department of the Interior)

U.S. Group on Earth Observations (USGEO)

Smithsonian Institution

National Museum of Natural History (NMNH)

Office of Advancement

Office of the Chief Information Officer (OCIO)

Office of Communications and External Affairs

Office of International Relations

Office of the Under Secretary for Science (OUSS)

Smithsonian Astrophysical Observatory (SAO)

Smithsonian Conservation Biology Institute, National Zoological Park (SCBI/NZP)

Smithsonian Environmental Research Center (SERC)

Smithsonian Institution Archives (SIA)

Smithsonian Institution Libraries (SIL)

Smithsonian Tropical Research Institute (STRI)

Appendix C: Inter-organizational Efforts

As discussed in the main text of this study, many collaborative efforts are underway to address various aspects of data management and data sharing. This appendix presents an overview of some of the collaborative initiatives that came to the study team's attention. This selection should be considered illustrative; it is most certainly not exhaustive. Note that:

- Some of the entities discussed here (for example, DIVERSITAS and Long Term Ecological Research network [LTER]) are scientific research initiatives that have become involved to some extent with data-management and -sharing issues;
- Some—the Coalition for Networked Information (CNI), Open Geospatial Consortium (OGC), and Digital Preservation Europe—are data initiatives that deal with biology data among other areas; and
- Some—Data Observation Network for Earth (DataONE), Biodiversity Information Standards (TDWG), and Global Biodiversity Information Facility (GBIF)—are specifically focused on data management and sharing in biology.

Atlas of Living Australia	
Mission	Making biodiversity information on the flora and fauna of Australia more accessible and usable online.
Strategies	Main strategy is creation of an online portal and associated infrastructure to integrate data on Australian species currently dispersed across museum and botanical collections and databases. The Atlas seeks to combine taxonomic and distribution data with images, scholarly literature, maps, and species identification information.
Funding[123]	Funded by the Australian Government under the National Collaborative Research Infrastructure Strategy (NCRIS) and other grants. The Atlas is developing close collaborations with other related NCRIS projects to leverage its resources.
Structure	Collaborating organizations include 14 major Australian natural history museums, research organizations, universities, and scholarly councils, plus a government agency. Maintains ties with international biodiversity data organizations; for example, it is a participant node of GBIF (see below) and a partner of Encyclopedia of Life (EOL) (see below), Biodiversity Heritage Library, European Union's Distributed Dynamic Diversity Databases for Life, and DataONE (see below).
History	Project officially launched in July 2010; site will be publicly accessible in October 2010.

[123] For all collaborative initiatives described here, an implicit source of support is in-kind contributions from member organizations (i.e., through participation in initiatives, projects, and activities).

Biodiversity Information Standards (TDWG)	
Mission	Promoting a wider and more effective circulation of information about the world's heritage of biological organisms.
Strategies	Developing and promoting standards/guidelines for data exchange; acting as a forum for discussion (through conferences and publications).
Funding	Annual individual and institutional membership dues; operational support from major collaborators (e.g., GBIF has provided financial support for TDWG's work on standards development); miscellaneous other sources such as voluntary donations.
Structure	International, non-profit organization with individual and institutional members. Latter include museums, botanical gardens, universities, and non-profit organizations. Formally collaborates with GBIF and OGC (see below). GBIF has adopted many TDWG standards and protocols, e.g., TAPIR and Life Sciences Identifiers (LSID).
History	Founded in 1985 to establish international collaboration among biological database projects. Initially the International Union for Biological Sciences' (IUBS) Taxonomic Database Working Group, with a focus on standards for plant taxonomic databases. In 1994, role expanded to encompass taxonomic databases in general.
Board on Research Data and Information (BRDI)	
Mission	Improving the management of, policies governing, and use of digital data and information for science and society.
Strategies	Monitoring data issues of interest to BRDI sponsors; proposing National Research Council (NRC) initiatives; engaging in planning, program development, and administrative oversight of projects launched under its auspices.
Funding	Supported by parent organization (the National Academies) and by sponsoring organizations (National Science Foundation [NSF], National Institutes of Health [NIH], Defense Technical Information Center [DTIC], and Library of Congress [LOC]).
Structure	Part of the National Academies. Individual members drawn primarily from universities, with a smaller number from private industry. Also includes federal government liaisons from NSF, National Institute of Standards and Technology [NIST], LOC, DTIC, NIH, Institute for Museum and Library Services [IMLS], and National Science and Technology Council [NSTC]. Hosts U.S. National Committee for CODATA. Small administrative staff located in Washington, D.C.
History	Formed in response to a recognized need to improve systems for managing, sharing, re-using, and preserving government and private digital resources. First meeting in January 2009.
Coalitional for Networked Information (CNI)	
Mission	Supporting the promise of networked IT for the advancement of scholarly communication and the enrichment of intellectual productivity.
Strategies	Disseminating knowledge about architectures and standards for digital information; working to improve scholarly communication; engaging in projects to study the economics

	of digital information, advance Internet technology and infrastructure, enhance education and training, and understand the changing digital environment.
Funding	Primarily membership dues.
Structure	Institutional members include universities, publishers, scholarly and professional associations, libraries, government agencies, and network, telecommunications, and IT firms. Small administrative staff located in Washington, D.C. In addition to members, key collaborators and partners include major funding agencies, National Research Council (NRC), LOC, World Wide Web Consortium (W3C), and major data/IT organizations in countries such as the United Kingdom, Germany, and the Netherlands.
History	Founded in 1990 to bring the library and IT communities together to enhance scholarship.
Committee on Data for Science and Technology of the International Council for Science (CODATA)	
Mission	Improving the quality, reliability, management, and accessibility of data of importance to science, technology, and society.
Strategies	Fostering international collaboration (with a particular focus on developing countries); promoting data quality-control measures; and promoting the development of national and international data policies that serve the needs of both the research community and society as a whole. Particular emphasis on (1) data-management problems common to different disciplines; and (2) use of data outside the field in which they were generated.
Funding	Supported by parent organization, the International Council for Science (ICSU), and membership dues.
Structure	An interdisciplinary committee of ICSU (see below). Membership includes 20 national members (scientific academies and similar organizations that represent individual nations) and 16 international scientific unions, plus a wide variety of other supporting organizations. Small administrative secretariat based in Paris, France.
History	Established in 1966 by ICSU.
Consortium for the Barcode of Life (CBOL)	
Mission	Promoting DNA barcoding as a global standard for the identification of biological species.
Strategies	Contributing to compiling a public reference library of DNA barcode sequences; sponsoring working groups, networks, workshops, conferences, outreach, and training to advance DNA barcoding practices and applications.
Funding	Secretariat supported by host organization, the Smithsonian National Museum of Natural History (NMNH). Additional support from Alfred P. Sloan Foundation grants.
Structure	Two hundred member organizations from 50 countries, including universities, museums, botanical gardens, private firms, and various non-governmental organizations (NGOs). In addition to members, key partners on the international Barcode of Life project are the International Barcode of Life (iBOL), Barcode of Life Data Systems (BOLD), and GenBank. A small secretariat resides at that NMNH in Washington, D.C.
History	Established in 2004 with initial support from the Alfred P. Sloan Foundation.

data.gov	
Mission	Increasing online public access to data generated by agencies of the executive branch of the U.S. government.
Main strategy	Development of a web portal with metadata descriptions of selected data sets held by federal agencies, information on how to access these data sets, and user tools.
Funding	Federal appropriations.
Structure	Interagency federal initiative hosted by the General Services Administration.
History	Launched in May 2009 by the Federal Chief Information Officer of the United States. U.S. Open Government Directive of December 8, 2009 required that all agencies post at least three "high-value" data sets online and register them on data.gov.
Data Observation Network for Earth (DataONE)	
Mission	Preserving and providing access to multi-scale, multi-discipline, and multi-national data in the biological and environmental sciences, from the genome to ecosystem levels.
Strategies	Data-management and data-sharing infrastructure development; development and dissemination of standards; outreach and education to the researcher community and beyond.
Funding	Initial funding from a NSF DataNet grant; working to develop a sustainable post-grant funding model.
Structure	Coordinated by: a principal investigator (PI) affiliated with University of New Mexico; and co-PIs affiliated with Oak Ridge National Laboratory (ORNL), U.S. Geological Survey (USGS), National Center for Ecological Analysis and Synthesis (NCEAS), and Duke University. Individual collaborating partners associated with a wide range of universities, federal agencies, and a small number of NGOs. Mainly U.S.-based; a small number of international participants. Administration based in Albuquerque, New Mexico.
History	Began operation in August 2009. With The Data Conservancy, one of two initial recipients of NSF DataNet grants for pilot projects in integrated scientific data management and sharing.
Digital Preservation Europe	
Mission	Facilitating the pooling of digital preservation expertise across the academic research, cultural, public administration, and industry sectors in Europe.
Strategies	Fostering collaboration and exchange among existing national and international initiatives across Europe; funding and undertaking pilot projects.
Funding	Initially funded by a European Union Framework Programme for Research and Technological Development grant. Supported by partner institutions.
Structure	Core partners consist of 11 universities, libraries, archives, and government initiatives in the United Kingdom, Austria, Denmark, the Netherlands, Germany, Czech Republic, Italy, and Lithuania. Eighteen associated institutional partners from these and other European

	nations, plus Israel. Administered out of University of Glasgow in Glasgow, Scotland.
History	Founded under the European Union's Sixth Framework Programme for supporting European research and technology, 2002-2006.
DIVERSITAS ("An International Programme of Biodiversity Science")	
Mission	Promoting an integrative biodiversity science that links biological, ecological, and social disciplines to produce socially relevant knowledge; and providing a scientific basis for the conservation of biodiversity.
Strategies	Synthesizing existing scientific knowledge; identifying gaps and emerging issues; promoting new research initiatives; building bridges across countries and disciplines; investigating policy implications of biodiversity science and communicating these through policy forums.
Funding	Most funds come from annual contributions from full national members of DIVERSITAS, including the United States, received through national committees that represent scientific organizations within each nation. (The U.S. national committee is based at the National Academy of Sciences [NAS]). Some funding directly from sponsoring organizations and from donations and grants.
Structure	Sixteen full national members (which make annual financial contributions) and 14 affiliate members (which do not), with 4 additional affiliates in the process of setting up national committees. Sponsoring organizations include the United Nations Educational, Scientific, and Cultural Organization (UNESCO), Scientific Committee on Problems of the Environment (SCOPE), IUBS, and ICSU. A Secretariat is based at the Muséum National d'Histoire Naturelle in Paris, France.
History	Established in 1991 by UNESCO, SCOPE, and IUBS as an international, non-governmental umbrella program to address changes in global biodiversity. In 1996, ICSU became a sponsoring organization.
Encyclopedia of Life (EOL)	
Mission	Organizing and improving access to species information.
Main strategy	Creating an EOL portal website that aims to provide an individual, "infinitely expandable" web page for each of the planet's approximately 1.8 million known species. Will incorporate the Biodiversity Heritage Library, which contains digitized versions of print collections from the world's major natural history libraries.
Funding	Major funding provided by the Alfred P. Sloan Foundation and John D. and Catherine T. MacArthur Foundation, with additional support from partner organizations.
Structure	Major partners include the Smithsonian Institution, Harvard University, Missouri Botanical Garden, Field Museum, and Marine Biological Laboratory, as well as the Biodiversity Heritage Library consortium. EOL maintains collaborative agreements with a variety of taxonomic and biological database organizations that serve as sources of information, including CBOL and GBIF. The EOL Secretariat is based at NMNH in Washington, D.C.
History	Initial $50 million foundation grant received in May 2007. EOL website went live in February 2008.

Federation of Earth Science Information Partners (ESIP)	
Mission	Making Earth observation data and information more available and understandable to both scientific researchers and others, including educators, policy makers, and the general public. Fostering the application of these data and information to practical policy and resource-management issues.
Strategies	Providing forums for Federation members to exchange information, data, and ideas; facilitating collaborative pilot projects among members.
Funding	Primary source of funds is sponsorship from National Aeronautics and Space Administration (NASA) and National Oceanic and Atmospheric Administration (NOAA).
Structure	A consortium of over 110 members, including universities, government research laboratories, supercomputing facilities, education resource providers, information technology (IT) and other commercial firms, non-profit organizations, and NASA, NOAA, and USGS data centers. The Foundation for Earth Science, a non-profit corporation based in Raleigh, North Carolina, serves as the Secretariat.
History	In response to a recommendation by NRC, NASA created ESIP in 1998 and has served as a major financial supporter. The ESIP Secretariat, the Foundation for Earth Science, was established in 2001. In 2006, NOAA joined NASA to become a financial supporter.
Global Biodiversity Information Facility (GBIF)	
Mission	Making biodiversity data available for research, conservation, and sustainable development.
Strategies	Direct provision of information infrastructure (including a data portal); facilitation of the creation and dissemination of community-developed tools, standards, and protocols; capacity building.
Funding	Each participant is responsible for funding its own participation. The Secretariat, Advisory Committee, and work program are funded through annual dues paid by GBIF's voting-participant national members (including the United States, whose node is the National Biological Information Infrastructure [NBII]). Dues are received through national delegations, which represent scientific organizations within each nation.
Structure	Multilateral initiative established by an intergovernmental memorandum of understanding (MOU). Full voting members now include 32 nations, with 22 associate national members and nearly 50 organizational members. The latter are mainly other biodiversity-focused consortia, including DIVERSITAS, EOL, CBOL, International Species Information System (ISIS), and TDWG. The Secretariat is based in Copenhagen, Denmark.
History	MOU creating GBIF, initially including 17 countries, signed in 2000.
Group on Earth Observations / Global Earth Observation System of Systems (GEO/GEOSS)	
Mission	Connecting globally dispersed Earth observation facilities (satellite-based and terrestrial) to make their data interoperable and widely accessible.
Main strategy	Implementation of a 10-year plan to knit together these facilities into an interoperable Global Earth Observation System of Systems, with a particular focus on using the resulting

	capabilities to achieve social benefits in areas such as public health, agriculture, disaster management, resource management, and biodiversity conservation.
Funding	Most resources provided through existing national and international mechanisms and through voluntary contributions to special projects. Costs arising from GEO participation are borne by the member government or participating organization. Members (and other entities) may make voluntary financial or in-kind contributions for GEO activities and administration through a trust fund administered by the Secretariat.
Structure	Members include 81 national governments and the European Commission, plus 58 intergovernmental, international, and regional organizations active in Earth observations or related matters. National membership is contingent upon formal endorsement of the GEOSS 10-year implementation plan.[124] All members belong to one of GEO's five regional caucuses, which nominate members of the Executive Committee. The Secretariat is based in Geneva, Switzerland.
History	The Third Earth Observation Summit, held in Brussels in February 2005, endorsed the GEOSS 10-year implementation plan and established GEO to carry it out.
Interagency Working Group on Digital Data (IWGDD)	
Mission	Coordinating the development strategic data management plans within and across the U.S. federal government, to ensure preservation of and access to federal data in science, technology, engineering, and related areas.
Main strategy	Providing an organized forum for exchange of information and ideas among agency representatives.
Funding	Federal appropriations.
Structure	Under the auspices of NSTC of the Office of Science and Technology Policy (OSTP), part of the Executive Office of the President. Composed of representatives from over 20 federal agencies, plus the Smithsonian Institution. Functions of the temporary IWGDD were recently transferred to a permanent sub-committee of the NSTC Science Committee.
History	Established in 2006 by the NSTC Committee on Science. Released an influential report (*Harnessing the Power of Digital Data for Science and Society*) in March 2009 that recommended the creation of intra-agency and inter-agency data management policies and data digitization.

[124] The 10-year GEOSS implementation plan states the following data-sharing principles:
(1) There will be full and open exchange of data, metadata, and products shared within GEOSS, recognizing relevant international instruments and national policies and legislation. (2) All shared data, metadata, and products will be made available with minimum time delay and at minimum cost. (3) All shared data, metadata, and products free of charge or no more than cost of reproduction will be encouraged for research and education."

Inter-American Biodiversity Information Network (IABIN)	
Mission	Fostering technical collaboration and coordination among countries of the Americas in collection, sharing, and use of biodiversity information relevant to decision making on natural resources management and conservation; and education to promote sustainable development in the region.
Strategies	Building an infrastructure for biodiversity information exchange; strengthening technical capacity to exchange biodiversity information and expertise across political, linguistic, and institutional boundaries; providing access to biodiversity information useful to decision makers to improve biodiversity conservation; enhancing capacity to store, use, and distribute scientifically sound and up-to-date biodiversity information; producing information products for decision makers.
Funding	Funds from 78 regional and national institutions.
Structure	34 countries in the Americas have official national representatives to IABIN. IABIN also has numerous organizational members and non-member collaborators, including NGOs, universities, museums, private firms, and foundations. Secretariat is based at the Ciudad del Saber in Panama City, Panama.
History	IABIN was created in 1996 as an initiative of the Santa Cruz Summit of the Americas meeting of heads of state.
International Species Information System (ISIS)	
Mission	Facilitating international collaboration in the collection and sharing of knowledge on animals and their environments for zoos, aquariums, and related organizations.
Strategies	Supporting the development of animal-information software systems and tools; promoting the development of data standards and best practices; providing mechanisms for online sharing of data; promoting scientific usage of data beyond captive animal management; and fostering the development of relevant resources (human, financial, and technological).
Funding	Major source of income is membership dues. Other support from grants, donations, and in-kind contributions from federal and private funding organizations, corporations, individual donors, and member organizations.
Structure	An international non-profit with over 800 organizational members in 80 countries, the majority in Europe (350) and North America (300). Most members are zoos, aquariums, and related organizations. Central offices are located in Eagan, Minnesota, with branches in Amsterdam, The Netherlands; Bogotá, Colombia; Tokyo, Japan; and Gurgaon, India.
History	Founded in 1973; initially included 51 zoos and aquariums in Europe and the United States. ISIS's Zoological Information Management System (ZIMS), currently being deployed, is the world's first web-based, real-time, integrated animal records database, including records for animal health.
International Council for Scientific and Technical Information (ICSTI)	
Mission	Fostering cooperation among stakeholders engaged in scientific communication with the aim of improving the effectiveness of scientific research.

Strategies	Acting as a non-political, non-commercial forum for exchange among members from the public, non-profit, and commercial spheres; encouraging collaboration among organizations and stakeholders; sponsoring technical projects to address key issues and concerns.
Funding	Operational budget essentially funded from membership dues.
Structure	International non-profit association with members from 16 countries in Europe, North America, and Asia (for-profit, governmental, and non-profit), plus 13 international organizations. Sponsored by ICSU. Administration based in Paris, France.
History	Created in 1997 as the successor to the ICSU Abstracting Board.
Long Term Ecological Research Network (LTER)	
Mission	Providing the scientific community, policy makers, and society with data and information to conserve, protect, and manage U.S. ecosystems and their biodiversity.
Strategies	Administering a national network of research sites covering a wide range of ecosystems. Each site develops research programs in five core areas (plant growth; species populations; organic matter accumulation; movements of inorganic nutrients; and site disturbances); supporting collaborative and synthetic research across sites and with other research organizations; maintaining a data portal with entries for over 6,000 data sets from LTER research sites, plus thousands of non-LTER datasets; and supporting a coordinated program of information management across sites that entails common metadata standards and a centralized information architecture.
Funding	The network administrative office is funded by NSF; network sites are supported by site-specific affiliated organizations.
Structure	LTER network currently consists of 26 research sites, typically administered by affiliated organizations (universities, federal agencies, and research centers). Sites host researchers the host organizations and a wide range of institutions. A Science Council, with a representative from each site, establishes the general scientific direction and vision of the network. The network office is located at the University of New Mexico in Albuquerque.
History	NSF established the LTER program in 1980 to support research on long-term ecological phenomena in the United States. In 1993, LTER became a founding member of International Long-Term Environmental Research (ILTER), a global "network of networks" focused on long-term ecological research.
National Biological Information Infrastructure (NBII)	
Mission	Improving access to data and information on the nation's biological resources.
Strategies	Linking biological databases, information products, and analytical tools maintained by NBII partners and other contributors for central access through the NBII web portal; developing new standards, tools, and technologies to find, integrate, and apply biological resources information.
Funding	Federal appropriations.
Structure	Among NBII's large network of partners are federal agencies (most science and land

	management agencies); state government natural resource agencies; county and municipal governments; international initiatives at the bilateral, regional, hemispheric, and global scales (NBII is the U.S. representative for many international initiatives); non-profits and NGOs; private firms; and universities. Managed by the USGS Biological Informatics Office, based in Reston, Virginia. NBII consists of "nodes" across the country that serve as focal points for biological and regional issues.
History	Established in 1993 following the recommendation of a NRC special panel on national biological resource issues. Funds to develop the NBII network of nodes were first appropriated in 2001.
National Ecological Observatory Network (NEON)[125]	
Mission	Enabling understanding and forecasting of the impacts of climate change, land-use change, and invasive species on continental-scale ecology.
Strategies	Main strategy is providing infrastructure and consistent methodologies to support a nationwide network of research sites that functions as a coherent observation platform. Includes an outreach component to translate research data into information that non-specialists can understand and use.
Funding	Large start-up grant from NSF ($433.7 million over five years to construct the network); additional federal appropriations; and initiation fees, membership dues, special assessments, and in-kind support from member organizations.
Structure	A 501(c)(3) corporation created to manage a continental-scale ecological observation network on behalf of the scientific community. The network is divided into 20 regional eco-climatic domains covering the continental United States, Hawaii, Alaska, and Puerto Rico, each with a core site administered with a collaborating partner (university or research organization). Over 50 institutions are members of the project, including universities, museums, scientific associations, and environmental NGOs. The large central administrative operation is based in Boulder, Colorado.
History	The network is not yet functional. Construction was expected to begin in late 2010 and to take approximately five years to complete. The network will eventually consist of 20 core sites; 40 relocatable sites; 36 aquatic sites (including 10 experimental ones); and three airborne remote sensing systems.
National Phenology Network (NPN)	
Mission	Promoting understanding of plant and animal phenology (patterns of perennial change) and the relationship with environmental change.
Strategies	Collecting, processing, and disseminating phenology data, including from the general public; developing a phenology data portal; educating citizen scientists on how to observe

[125] NEON has a policy of open access to network data and information products. According to the NEON website: "Data generated by NEON will be freely available to all. NEON will endeavor to archive and distribute data generated by individual investigators at NEON sites, provided [they are] in accordance with NEON formats. Data collected based on funding from public agencies … will follow agency/NEON policies for public release. We anticipate priority periods for investigators, depending upon agency policies." NEON estimates its network will produce about 178 terabytes of data per year. (http://www.neoninc.org)

	and report phenological phenomena; and sponsoring projects to rescue and digitize historical data sets.
Funding	Mostly financial and major in-kind contributions from a core of sponsoring organizations, including NSF, ORNL, University of Arizona, U.S. Environmental Protection Agency (EPA), U.S. Fish and Wildlife Service, and USGS.
Structure	In addition to funding sponsors, NPN partners include over 20 organizational collaborators, plus individual citizen scientists, resource managers, educators, and scientists from organizations including public agencies, Native American tribes, NGOs, and universities. National coordinating office is in Tucson, Arizona.
History	The first meeting of the NPN National Coordinating Office was held in 2007.
Open Geospatial Consortium (OGC)	
Mission	Serving as a global forum for developers and users of spatial data products and services, and advancing the development of international standards for geospatial interoperability.
Strategies	Working to achieve consensus on open standards for geographic information systems (GIS); sponsoring a series of hands-on engineering initiatives to accelerate the development and acceptance these standards; offering resources (technical documents, training materials, test suites, reference implementations, and other interoperability resources) to help technology developers and users take advantage of the standards; and supporting publications, workshops, seminars, and conferences to help technology developers, integrators, and procurement managers introduce OGC into their architectures.
Funding	Primarily membership fees.
Structure	An international industry consortium of nearly 400 companies, government agencies, and universities participating in a consensus process to develop open standards. In addition to member organizations, OGC collaborates with a number of prominent international standard-setting, research, geospatial infrastructure, and educational consortiums and organizations. Administrative offices are located in Wayland, Massachusetts; Herndon, Virginia; and Bloomington, Indiana.
History	Founded in 1994 by eight charter members.
Science Environment for Ecological Knowledge (SEEK)	
Mission	Creating IT and related resources to facilitate global access to biodiversity and ecology data and information, and creating new methods for capturing, reproducing, and analyzing data.
Strategies	Sponsoring collaborative projects to create tools and infrastructure (e.g., EcoGrid network and Keplar scientific workflow tool) to overcome technical obstacles to sharing biology data (e.g., locating and connecting to distributed computational services and providing uniform interface for accessing dispersed databases running on different software).
Funding	NSF grant.
Structure	A project under the aegis of the Partnership for Biodiversity Informatics, a collaboration among NCEAS, LTER, University of Kansas, and University of California at San Diego. In

	addition to these core partners, other collaborating organizations include the University of New Mexico; Genome Center at the University of California, Davis; Arizona State University; University of North Carolina; University of Vermont; and Napier University in Scotland.
History	SEEK projects tend to be carried out by dispersed virtual teams that collaborate remotely.
Science.gov	
Mission	Increasing public access to data generated by the U.S. government.
Main strategy	Central Web portal providing access to scientific data and results from a range of federal agencies.
Funding	Federal appropriations.
Structure	Participating entities include 18 scientific and technical organizations from 14 federal agencies that contribute data and research to science.gov. These include the National Agricultural Library (U.S. Department of Agriculture, USDA) ; U.S. Forest Service, USDA; NIST; National Technical Information Service; DTIC; National Library of Education; U.S. Department of Energy; OSTP; Department of Health and Human Services (HHS); Food and Drug Administration (FDA) (HHS); National Library of Medicine (NIH); NBII (USGS); National Transportation Library (U.S. Department of Transportation); EPA; LOC; NASA; National Archives and Records Agency (NARA); NSF; and Government Printing Office. The portal accesses 42 databases and 200 million pages of scientific information.
History	Launched in December 2002 as the first unified search engine for federal scientific data. Four updated versions of science.gov have subsequently been launched. Science.gov is the U.S. contribution to WorldWideScience.org, an international science web portal.
The Data Conservancy	
Mission	Researching, designing, implementing, deploying, and sustaining data curation infrastructure for cross-disciplinary discovery, with an emphasis on observational data.
Strategies	Data-management and data-sharing infrastructure development; development and dissemination of standards; sponsorship of pilot projects to address issues in data management and sharing.
Funding	Initial funding from a NSF DataNet grant; working to develop a sustainable post-grant funding model.
Structure	Participating organizations include four universities, two digital-preservation non-profits, two research centers, one commercial firm, the National Virtual Observatory project, and the National Snow and Ice Data Center (University of Colorado, Boulder). An additional 13 partners (which do not receive NSF project funds) include the U.S., British, and Australian governments, libraries, research organizations, and IT organizations. The administrative office at Johns Hopkins University in Baltimore, Maryland.
History	Began operation in 2009. With DataONE, one of two initial recipients of NSF DataNet grants for pilot projects in integrated scientific data management and sharing.
U.S. Global Change Research Program (USGCRP)	

Mission	Coordinating and integrating federal research on changes in the global environment and their implications for society.
Strategies	Assembling a National Climate Assessment report for Congress and the President (which functions as a status report on climate science and impacts); supporting research and observational activities in collaboration with other national and international science programs; and integrating cross-agency federal research on climate-related health issues.
Funding	Federal appropriations.
Structure	Membership consists of 12 federal departments and agencies (Commerce, Defense, Interior, Agriculture, Energy, State, Transportation, HHS, NSF, NASA, U.S. Agency for International Development [USAID] [U.S. Department of State), and EPA), plus the Smithsonian Institution. The Integration and Coordination Office is in Washington, D.C.
History	Began as a presidential initiative in 1989; later mandated by Congress in the Global Change Research Act of 1990 (P.L. 101-606), which called for *"a comprehensive and integrated United States research program which will assist the Nation and the world to understand, assess, predict, and respond to human-induced and natural processes of global change."* Known as the U. S. Climate Change Program from 2002 to 2008.

Appendix D: Additional International Initiatives

This report focuses mainly on data sharing and dissemination in the United States and a small number of multinational organizations. This appendix provides information on the work of selected non-U.S. governments (the European Union [EU] is treated here as a national actor) and multilateral organizations aimed at influencing, coordinating, and facilitating global scientific data sharing, including discovery, access, usability, interoperability, and management.

Other Nations and the European Union

Canada

In 2008, the National Research Council-Canada Institute for Scientific and Technical Information (NRC-CISTI), Canada's national science library and leading scientific publisher, led the formation of the Research Data Strategy (RDS) Working Group. The group, which aims to provide recommendations for improving the state of data management and sharing in Canada, includes representatives of universities, institutes, libraries, and granting agencies, as well as individual researchers.

One of RDS's first actions was to release *Stewardship of Research Data in Canada: A Gap Analysis* (Research Data Strategy Working Group 2008). This document outlines numerous national shortcomings in such areas as data policy, funding, roles and responsibilities, data repositories, standards, skills, training, accessibility, and preservation.

NRC-CISTI is leading the push for a National Science Library Trusted Digital Repository, which is to be the foundation of a broader strategy to build a networked national digital information infrastructure. NRC-COSTI also provides the Gateway to Scientific Data, a central portal for access to Canadian scientific, technical, and medical data sets, as well as to information on selected policies and best practices in data curation.[1]

NRC-CISTI is a founding member of DataCite, an international collaboration of libraries dedicated to improving access to research data. DataCite enables organizations to register data sets and applies digital object identifiers to help researchers locate and cite the data sets.

[1] The study team did not investigate the comprehensiveness of the Gateway's coverage.

China

The study team found it difficult to research data management and sharing in China because the government is less transparent than is the case with the other nations considered here. However, there is no question that the government and scientific community of China are aware of the changing digital data dynamic and of the need for proactive initiatives at various levels to respond to its challenges.

In 2006, China's Ministry of Science and Technology revealed a plan for the years 2006-2020. To promote data access, the plan envisioned the creation by 2010 of 40 national scientific data centers, 300 master databases, and a public portal administered by the Ministry—collectively called the National Scientific Data Master Network. (The study team was unable to determine the status of the plan's implementation at the time of this report's writing.)

The plan ties into China's Scientific Data Sharing Project (SDSP), which dates from 2002. To foster interoperability, SDSP calls for a set of 32 principal national standards for data processing and storage, based on existing national and international standards. Some of SDSP's other goals include forming a more user-friendly scientific data-management and -sharing complex and developing a set of supportive laws, policies, and standards. SDSP's goal is for 80 percent of government-funded scientific data to be available to the public. SDSP will also collect data from other countries and other national programs, agencies, and institutes, which it will publicly share with the Chinese scientific community through the network of data centers.

European Union

In addition to efforts of individual EU countries, there has been a great deal of effort in recent years to address data sharing and management across the EU as a whole.

Alliance for Permanent Access (APA). APA is a non-profit organization registered in The Netherlands, but involving much of Europe. Its mission is "to develop a shared vision and framework for a sustainable organizational infrastructure for permanent access to scientific information."[2] It addresses issues of infrastructure interoperability in Europe, and offers joint advocacy and representation on behalf of scientific communities within the states of the EU. Within specific scientific fields, APA facilitates efforts to identify which digital data to preserve and to establish appropriate metadata schemas. More generally, it supports the development of a sustainable European Digital Information Infrastructure for permanent access to documents. APA members, which

[2] http://www.alliancepermanentaccess.eu/index.php?id=2.

include both national and international organizations, are required to contribute staff time, expertise, and funding, and to participate in Alliance meetings. In return, each is voting member of the Alliance's board.

One important APA project is the Permanent Access to the Records of Science in Europe initiative, or PARSE. PARSE.Insight, a two-year project co-funded by the EU that ended in February 2010, sought to develop a roadmap for science data infrastructure in Europe, with an emphasis on the preservation of scientific research data. An interesting aspect of the project was an international survey that asked researchers about the current state of their own data management and their opinions on the purposes of, requirements for, and obstacles to expanded data management and sharing (Kuipers and van der Hoeven 2009).

Shared Environmental Information System (SEIS). SEIS is an example of a proposed system for data sharing in biology at the EU level. A collaborative initiative of the European Commission, EU European Environment Agency (EEA), and SEIS's 32 member countries, its purpose is to improve the availability and quality of the data that inform EU environmental policies. To this end, SEIS aims to create a system in which these data are managed close to their source, then made widely available to users in an open and transparent way. Data and derived information will be stored in decentralized (but interconnected and fully interoperable) databases throughout the EU, building on existing systems. SEIS will make environmental data from member states easily available not only to EEA, but also to other international organizations that collect and disseminate global environmental data, such as the Organization for Economic Cooperation and Development (OECD) and the United Nations Environment Program (UNEP).

United Kingdom

Interviewees suggested that the United Kingdom (U.K.) has been one of the more advanced nations in terms of crafting a coordinated national strategy for e-science and supporting it with the requisite institutions.

Joint Information Systems Committee (JISC). JISC, under the British Library organizationally, aims to "provide world-class leadership in the innovative use of information and communications technology to support education, research, and institutional effectiveness."[3] It conducts research and infrastructure development projects, offers consulting and capacity-building services to the higher education community, and administers the Joint Academic Network (JANET) system, the U.K.'s

[3] http://www.jisc.ac.uk/publications/strategy/2010/jisc2010strategysummary.aspx.

national research and education communications network. JISC is also promotes the development of digital libraries, digital repositories, and data publishing, and maintains a policy of open access on all projects that it funds.

One JISC program that specifically addresses scientific data sharing is the Managing Research Data Programme (JISCMRD). This initiative targets a number of key areas related to the management and sharing of scientific data, including: infrastructure; data-management planning and tools to support it; data publication and methods for citing, linking, and integrating research data; and development of new skills among academics and research support staff.

A number of U.K. public higher-education funding entities support JISC.

Digital Curation Centre (DCC). DCC was created in 2004 in response to a JISC study that argued for the establishment of a national center to tackle digital curation challenges beyond the capabilities of a single institution or discipline. DCC is jointly administered by the Universities of Edinburgh and Glasgow (which together host the [Scottish] National e-Science Centre), U.K. Office for Library and Information Networking (UKOLN) at the University of Bath, and the government-funded Science and Technology Facilities Council. DCC is a leading national center of practical expertise on data management. Among its activities are providing free access to curation tools and resources; delivering at-cost training and network-building events; establishing communities of data curators to share best practices; contributing to the development of standards, tools, and practices; providing a range of learning resources to update curation skills; and offering fee-based consultancy services to organizations seeking more specific guidance.

Multilateral Initiatives

International organizations and initiatives are increasingly addressing scientific data issues because of the global nature of the scientific enterprise. A well-functioning set of policies, processes, and cyberinfrastructure for data management and sharing is understood to require coordination of the efforts taking place across national scientific communities, as well as among individual organizations. This section looks at a few illustrative examples of important multilateral efforts.

International Council for Science (ICSU)

ICSU was founded in 1931 as a non-governmental organization (NGO) comprising representatives of national scientific entities and international scientific unions, and

dedicated to advancing global scientific activity and its use for the benefit of society. Data sharing is integral to ICSU's mission: the universality of science—a concept that includes equitable access to data—is the fifth statute of ICSU's *Statutes and Rules of Procedure*. ICSU has undertaken numerous initiatives to address data and information issues, and a 2004 report from ICSU's Committee on Scientific Planning and Review set a number of priorities and directions for activities in this area (International Council for Science 2004).

World Data System (WDS). WDS, which was until recently known as the World Data Center System, dates to 1957. The World Data Centers comprise a network of repositories for data in a number of fields (such as atmospheric, solar, geophysical, ecological, and human-environment interaction); these provide open access to the data in their care. In 2008, the ICSU General Assembly decided to replace the World Data Center System with the WDS, which will incorporate data from both the old World Data Centers and the ICSU Federation of Astronomical and Geophysical Data Analysis Services (FAGS). The WDS promises a more centralized system that incorporates more disciplines and has a larger geographic base than the World Data Centers.

In addition to providing mechanisms for equal and open access to data, the WDS will encourage compliance with standards and conventions, review the quality of the data to which it provides access, and provide long-term data stewardship. As of this writing, however, WDS was still in its formative stage and had not yet developed a constitution or membership criteria.

CODATA. Another important ICSU initiative is the Committee on Data for Science and Technology (CODATA), established in 1966. CODATA is an interdisciplinary committee whose membership includes 20 national members (scientific academies and similar organizations that represent national scientific communities) and 16 international scientific unions, plus a wide variety of other supporting organizations.

CODATA's main focus is on fostering international collaboration to improve the quality, management, and accessibility of scientific data, with a particular focus on developing countries. It emphasizes data management issues that cut across scientific fields and the use of data outside the field in which they were generated. CODATA has pursued its goals through a variety of means, for example, establishing a number of task forces to deal with a wide range of issues. It is particularly active in efforts to frame data-sharing policies at national and international levels. For example, CODATA contributed to the development of OECD guidelines on access to data (discussed below) produced by publicly funded research.

Science Commons

The Science Commons is an effort to expand the Creative Commons model originally devised for creative and artistic works into the scientific realm. That model, developed by a San Francisco-based non-profit organization, is intended to ease legal and technical barriers to sharing creative intellectual property. To this end, it provides free, user-friendly, easy-to-understand licensing tools that let intellectual property owners specify the terms under which others can legally access and use their material.

Analogously, the goal of the Science Commons project is to remove barriers to the accessibility of scientific data, both through similar licensing tools for scientific assets and through the development of open-source platforms for data and information sharing. For example, the Science Commons has undertaken the Scholar's Copyright Project, which seeks to lower barriers to open access by addressing both legal and technical barriers to data access and integration.

United Nations Environment Programme (UNEP)

Open access to environmental data is one important area of activity for UNEP, a UN agency established in 1972 to act as the "environmental conscience" of the UN system. For example, while better known for its work evaluating the current state of climate change science, the Intergovernmental Panel on Climate Change (IPCC), established by UNEP and the World Meteorological Organization in 1988, was also instrumental in the creation in 1998 of a Data Distribution Centre (DDC) to facilitate access to climate-related data. When research organizations provide the DDC with data sets, they agree to place these data in the public domain and make them available online to registered users free of charge. The DDC provides four types of data, with accompanying technical guidelines: climate observations; global climate model data; socio-economic data and scenarios; and data and scenarios for other environmental changes.

UNEP has also worked to increase access to environmental data through its administration of the Nairobi Convention Clearinghouse, a data facility associated with the 1985 Nairobi Convention compact, whose aim is to protect the marine and coastal environments of the western Indian Ocean through regional coordination. The Clearinghouse facilitates the dissemination and use of data relevant to the protection of these environments: through its web portal, users can freely access a broad range of data and metadata from diverse sources on climate change, environmental disasters, ecosystem management, environmental governance, harmful substances, and resource efficiency. The Clearinghouse's data-sharing policy promotes open access to data, although it does indicate that some data sets, particularly those with detailed geographic

information, have "restricted access"—in these cases, the original data are not accessible online (although metadata documentation is provided).

GEO/GEOSS

The Group on Earth Observations (GEO) focuses on earth observation systems and data. Founded in 2005 following the 2002 World Summit on Sustainable Development in response to calls to improve earth observation systems, it counts among its members both nations and international organizations. Its mission is to coordinate and facilitate global efforts to create a Global Earth Observation System of Systems (GEOSS) that will knit together earth observation platforms scattered across the world and make their data sets widely accessible and interoperable.

To accomplish this, GEOSS is creating a common infrastructure with an online portal that will enable users to discover and access GEOSS data and services. The portal will also feature a standards-and-interoperability registry, with information about relevant data management standards, so that data contributors can configure their systems appropriately. GEO is still finalizing its data sharing policy.

Organization for Economic Cooperation and Development (OECD)

OECD, created in 1961 as an international organization of democracies with market economies, serves as a forum in which nations can cooperate to address common problems and work to coordinate policies. While by no means focused on scientific or data issues, in 2007 OECD published a document with important implications for this subject, *OECD Principles and Guidelines for Access to Research Data from Public Funding*. This document provides broad policy recommendations for data access based on the principles of openness, transparency, and interoperability. However, implementation of the guidelines has been problematic. One reason is that many nations and institutions already have, or are in the process of developing, their own data-sharing policies. Another reason is that the nature of "public funding" for research varies considerably across nations. Nevertheless, the *OECD Principles and Guidelines* serves as an important example of an international body's effort to push the international community in the direction of open access.

Addendum: Social Media and the Dissemination of Digital Biology Data

Given the widespread and constantly increasing use of social media around the world, the OP&A study team decided to look at the role those media played in the discovery and sharing of scientific research data in general. Social media would seem to have promise for discovery and access to data, as they have become a major tool for linking people to widely dispersed and varied information and to one another.

The study team found that a large number of government agencies and organizations use social media quite extensively to connect to the general public—this is a main thrust of the Obama administration. The study team also came across instances where social media were being used in scientific research, such as crowd sourcing, use of wikis to get input from colleagues and reviewers, and data collection by citizen scientists and the general public. Fewer examples emerged, however, of social media being used to further discovery, access, and sharing of digital scientific research data, despite the availability of social media platforms. One reason seems to be that social media platforms are not fully adapted to these uses, as the scientific community does not yet view them as useful for these purposes.

Despite the limited use of social media to date for digital data discovery and access, the study team includes this addendum because role and use of social media tools in general is growing exponentially, and it seems likely that it is only a matter of time before the scientific community exploits it for data discovery and access. The addendum looks at some examples of how the professional scientific community outside and within the Smithsonian is already taking advantage of these tools.

Background

Social media refers to

> *Any form of online publication or presence that … allows the creation and exchange of user-generated content. A common thread running through all definitions of social media is a blending of technology and social interaction for the co-creation of value. … Social media tools are generally available to anyone at little or no cost, and most anyone can use them.* (www.onlinematters.com/glossary.htm).

Social media take many forms, but have an underlying commonality: interactivity. In 2010, the National Archives and Records Administration (NARA) published *A Report on*

Federal Web 2.0 Use and Record Value, which groups online social media tools and platforms into the categories listed below (depending on how they are used, specific tools and platforms can fit into more than one category) and gives examples of the media that fall under the categories.

- *Web publishing*—sites that allow users to post or publish content and receive feedback. Examples include:
    - Blogs (WordPress, Blogger)
    - Microblogging[129] (Twitter, Plurk)
    - Wikis (Wikispaces, PBWiki)
    - Mashups[130] (Google Maps, popurls)
- *Social networking*—sites that allow users to establish interactive connections and share information. A social network service usually hosts profiles of each user, social or professional links among them, and a range of services to facilitate exchange. Common social networking platforms include:
    - Social networking services (Facebook, LinkedIn)
    - Social bookmarks[131] (Delicious, Digg)
    - Virtual worlds (Second Life, OpenSim)
    - Crowdsourcing[132]/social voting (IdeaScale, Chaordix)
- *File sharing and storage*—file-hosting services and online file-storage providers that hold content that users can manipulate and comment on. Common file sharing/storage platforms include:
    - Photo libraries (Flickr, Picasa)

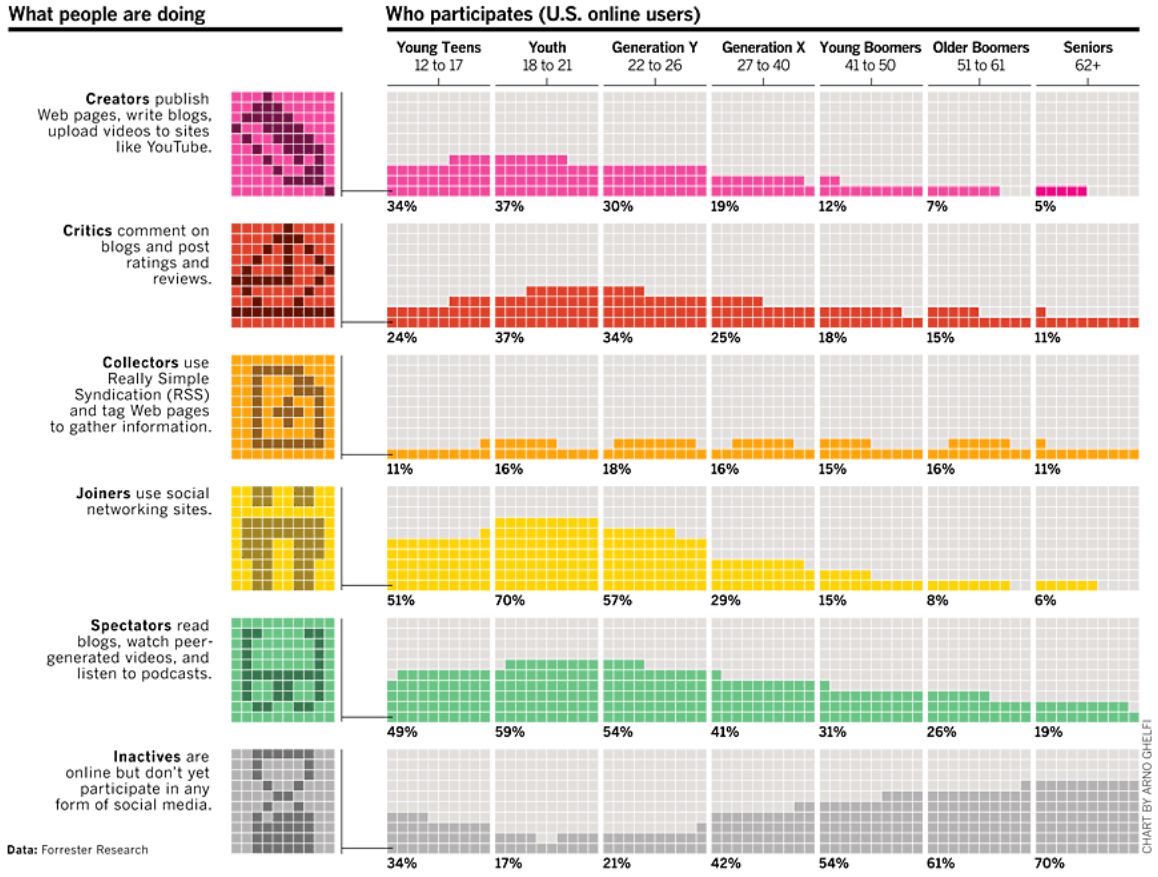
[129] The content on microblogs is typically much smaller than that on traditional blogs (http://en.wikipedia.org/wiki/Microblogging).
[130] A mashup is a web application hybrid that combines data and/or functionalities from more than one source (http://en.wikipedia.org/wiki/Mashups).
[131] Social bookmarks offer a way for Internet users to organize, store, manage, and search for bookmarks for, or references to, resources online (http://en.wikipedia.org/wiki/Social_bookmarking).
[132]Through crowdsourcing, individuals and organizations can outsource tasks they would normally handle themselves to some undefined group of people or communities (crowds) (http://en.wikipedia.org/wiki/Crowd_sourcing).

- Video libraries (YouTube, Vimeo)
- Storage (Google Docs, Drop.io)
- Content management (SharePoint, Drupal)

Social media have become increasingly popular as use of the internet has moved beyond the initial focus on transfer of information to facilitating interaction among individuals. Social media have become a cheap and effective way for organizations to communicate interactively with geographically dispersed stakeholders and for creating online networks of all descriptions and purposes.

Although social media tend to be associated in the public mind with younger people, they are in fact popular across all age groups, and becoming more so each year (Addendum Figure 1).

Addendum Figure 1. Uses of Social Media by Generations, 2007



Source: Inside Innovation—In Data, *Bloomberg BusinessWeek*, June 11, 2007 (http://www.businessweek.com/magazine/content/07_24/b4038405.htm).

Social Media for Scientific Data Sharing

External Organizations

Following are some examples of how federal agencies and external organizations are using social media to disseminate information on or share scientific data with professional audiences (as well as other audiences).

- **Digital Ocean.** Digital Ocean (DO), a collaborative project headed by the University of California, Santa Barbara, is a new initiative aimed at knitting together a wide range of specialist and non-specialist online communities with a common interest in the earth's ocean. The Digital Ocean website describes the target audience as "Communities of scientists, educators, students, policy makers, media specialists, ocean enthusiasts, people who make their living from the sea, and Google Ocean user groups," all of whom are expected to "add content and value to the DO system." Researchers can, for example, use the Digital Ocean site to share ideas for research projects and to make queries, for example, on whether other researchers have had issues with a particular methodology or have experience applying it in a different place or context. The nature of its target audience, in combination with its Web 2.0 infrastructure, social networking technology, and new media positions, allows Digital Ocean to facilitate "inter- and intra-disciplinary ocean science collaboration on local, regional, and global scales." Digital Ocean was started in part to address the difficulties that ocean science poses because it involves multiple disciplines, multiple types of data, and many researchers working in relative isolation. The idea was to provide these groups a single place for conversations across fields and geographic boundaries.

- **The U.S. Environmental Protection Agency (EPA).** EPA, a federal agency, is making effective use of social media within a clearly defined organizational structure. The Office of External Affairs and Environmental Education (OEAEE), located under the Office of the Administrator, manages EPA's social media in conjunction with the Office of Environmental Information and Office of Public Affairs. In January 2010 the latter two offices issued a set of guidelines for representing EPA on social media sites. The guidelines define social media, outline how employees should utilize them, and discuss how to present EPA in various situations. EPA also hosts an internal wiki that provides staff with guidance on the benefits and use of specific social media platforms such as Twitter and Facebook, one aim being to make staff feel comfortable using social

media. Staff also use the wiki to post work for review and comment by colleagues.

- **National Aeronautics and Space Administration (NASA).** NASA, another federal agency, is using social media quite extensively—they include over 30 Twitter accounts, 37 blogs, and sites for chatting with NASA scientists. The NASA home page has a prominent "Connect" button on the main menu that links to a social media portal that has further links to all NASA's social media platforms. The major NASA scientific projects have their own social media sites, administered by web managers. Typically staff with in-depth knowledge of a project or program run the associated Twitter feeds. The Office of Communications under the Office of the Administrator maintains the social media sites.

  NASA also runs an enhanced intranet called Spacebook that serves as an internal social networking site with user profiles, forums, groups, and social bookmarks. Each employee has a home page where he or she can publish information on his or her status, share files, "friend" others and follow their activity, and join communities of interest. Spacebook helps employees find expertise dispersed across the organization, form online groups, share files, write wikis, and even help locate or dispose of surplus gear.[133]

- **Federation of Earth Science Information Partners (ESIP).** ESIP's membership includes "data centers, research institutes, educational groups, commercial interests and other organizations that provide Earth science data and technology services." It uses an extensive wiki site to support collaboration among its many members and sub-entities and to keep people informed about its activities.[134]

- **Air Twitter.** Air Twitter, which ESIP initiated, is now run out of the Center for Air Pollution Impact and Trend Analysis at Washington University in St. Louis, Missouri. Project managers query social media sites such as Twitter to gather real-time air quality-related content and re-tweet this information through the project's own Twitter account. The data from the Twitter account are then

[133] See, for example, "Spacebook: Lessons Learned from NASA's Enterprise Social Network," by Emma Kolstad Antunes, an IT Specialist for NASA Goddard Space Flight Center's Office of the Chief Information Officer (http://www.gov2expo.com/gov2expo2009/public/schedule/detail/10307), and "Q&A: NASA's Stephanie Schierholz on navigating the frontiers of social media," in which Schierholz, NASA's social media manager, talks about the agency's digital outreach (http://econsultancy.com/us/blog/6302-q-a-nasa-s-stephanie-schierholz).

[134] ESIP website, http://www.esipfed.org/.

compiled on a wiki event space ("event" refers to something that is affecting air quality, such as a wildfire). Originally, the account was intended for storing the data that were collected, but over time its managers realized that people were following its Tweets to get information on air quality in their own regions. Nevertheless, the project's focus is firmly on the use of social media to support the work of scientists and researchers.

Smithsonian Institution

Social media projects are proliferating across the Smithsonian as people grasp their value for the "diffusion of knowledge." The science units that were the focus of this study—National Museum of Natural History (NMNH), National Zoological Park (NZP)/Smithsonian Conservation Biology Institute (SCBI), Smithsonian Environmental Research Center (SERC), and Smithsonian Tropical Research Institute (STRI)—all make use, some extensively, of Facebook, Twitter, YouTube, Flickr, and/or other social media.

Interviewees cited a number of reasons for the growing use of social media at the Institution. Most generally, the units employ them to foster interactive communication with audiences and other stakeholders—Smithsonian staff, the general public, citizen scientists and hobbyists, specialists and scientists, policy and decision makers, and any combination of these groups. Social media also provide a space for communities of interest to form around a topic, an example being the Smithsonian Migratory Bird Center's platforms that support an online community of bird watchers and enthusiasts. Going one step further, social media can be used to provide opportunities for the general public to get involved in scientific research. For example, citizen-science projects enlist the public in collecting and providing data, including photographs, to support the scientific enterprise. Social media are also used for crowd sourcing, an example being EOL (see below).

For the most part, the Smithsonian has used social media primarily for education and outreach, rather than for digital data discovery and/or sharing per se. Often what is being shared are publications or narrative compilations of information from different sources, and not the digital data underlying the science or links to the data. However, the following examples show how some units are using social media for discovery and data sharing among scientists. A number of these examples have a research component as their starting point, mainly data collection by citizen scientists.

- **SERC blog.** SERC maintains a blog aimed mainly at audiences with a specific interest in ecology, such as reporters, students, conservationists, scientists, and SERC's own staff and volunteers. Content includes interviews with experts,

accessible write-ups of scientific findings, highlights of events, and information on SERC programming. Currently, most blog posts receive few comments, except when stories have been picked up by the news media. SERC also disseminates information via Facebook and YouTube.

- **Encyclopedia of Life (EOL).** EOL, a highly collaborative initiative headquartered at NMNH that seeks to organize and make information available in one location on virtually all of Earth's life forms, makes extensive use of social media. The aim is, according to one interviewee, "to accelerate the pace of research and new species description by making freely available, searchable, and re-usable the information currently in libraries or in local databases inaccessible to most of the world's scientists." EOL's content includes "descriptions, photos, bibliographic links, distribution maps and other rare and specialized information that has traditionally been scattered around the world in libraries, museums, herbaria, colleges and universities, databases and other storehouses of expert knowledge." The heart of the project is a set of web pages—one for each of the approximately 1.9 million known species—that offer information on each species, compiled from different databases and serve as entry points to a great deal of additional knowledge. The goal is for the content of each webpage to be reviewed for accuracy and retention by authorized curators.[135]

  In addition to its own social-media tools, EOL now has a blog and leverages Twitter, Facebook, Wikipedia, Flickr, and Vimeo. Further, EOL recently entered into an agreement with Wikipedia whereby the latter's articles on species are included on the EOL species page by default. The background remains yellow until the information is reviewed and validated by an EOL curator. EOL makes all its review information available through an Atom feed, and its curators are encouraged to correct the same article on Wikipedia as well.

  EOL targets and is accessible to a wide range of users that include scientists; however, EOL does not currently track users by category and therefore cannot say how many scientists are using the site. According to one interviewee, at present "there is not enough detail to satisfy the needs of most scientists."

[135] EOL, as noted, has mechanisms for content review by specialists, and ways to display whether posted content has been verified for accuracy. However, interviewees acknowledged that the reviews are often delayed for prolonged periods, something EOL is trying to address (discussed later). An alternative to staff review is the wiki model, in which the user community as a whole reviews postings for accuracy and appropriateness. Generally speaking, Smithsonian specialists hold this model in low regard, often citing Wikipedia as an example of what they would *not* want official Smithsonian sites to become. In addition, the wiki model still requires "referees" in many cases, for example, to help resolve intractable differences of opinion.

- **CBOL.** CBOL is developing a mechanism for generating a unique genetic barcode for every species and making the entire set of barcodes available in what the project's executive secretary calls "a kind of telephone directory for all species." CBOL hosts blogs and forums where scientists can exchange information, pose questions, and solicit assistance.

- **Smithsonian Migratory Bird Center**. The Smithsonian Migratory Bird Center, part of SCBI, runs a number of social media platforms, including a blog, Facebook page, Twitter account, and YouTube site. The blog, which dates back to March 2005, features stories, pictures, and videos related to migratory birds and the research conducted at the Center. Content is geared toward bird enthusiasts, conservationists, and others with a specific interest in this area. The blog links back to the Center's website, where users can find links to the Center's Twitter, Facebook, and YouTube accounts.

- **SIGEO and CTFS.** The Smithsonian Institution Global Earth Observatory (SIGEO) and the Center for Tropical Forest Science (CTFS), part of STRI, maintain a blog featuring news and photographs of the Center's activities around the world. It also offers a Flickr photo stream, RSS feed, and links to recent journal articles and books by staff. The primary audience is staff and volunteers involved in SIGEO and CTFS's extensive global network of research plots in 21 countries, the goal being to help keep this audience—scattered across 21 countries—up-to-date on activities, publications, and developments, and to support communication.

- ***Smithsonian Wildlife*** **project.** An example of how one group at the Smithsonian is using social media to advance discovery and sharing of digital data is the *Smithsonian Wildlife* project. It came about because of the growing importance of camera traps as a research tool for wildlife ecologists, and the need to improve data sharing and workflow to enable research across large geographic areas where hundreds of cameras are deployed in the field by scientists and citizen scientists. The project provides a platform for aggregating, editing, and exporting standardized data across projects. One goal is to enable researchers (and the general public) to search and access the material from three different viewpoints:

    - Geography, for example, offering overlays of the trap locations on high-quality base maps and allowing queries by country, sub-region, ecoregion, etc.

- Taxonomy, with users able to query the collection by scientific/common name.
- Temporality, such as the year or season.

The first phase of *Smithsonian Wild* aggregated 201,770 images and associated metadata from nine research sites in Asia, Africa, North America, South America, and Central America, migrated the content to Flickr using the Amazon Cloud, built a web user interface to expose the research, images, and metadata to the public, and developed metadata standards with leading organizations using camera-trapping methods.

- **Global Volcanism Program.** The program, housed at NMNH's Department of Mineral Sciences, serves as an international clearinghouse for reports, data, and imagery, drawing on information from contributors who make up the Global Volcanism Network. For example, in the early stages of a volcanic eruption anywhere in the world, members of the network send real-time information to GVP staff, which organizes and makes the information available. Members of the network range from fishermen, farmers, and villagers who live near volcanoes and notice significant changes, to professionals such as scientists, technicians, professors, meteorologists, tour-operators, and devoted amateurs. Many have been contributing reports for many years.

- **Smithsonian Commons**. According to the Smithsonian Commons website, which is now being prototyped, a core premise underlying the Commons is that everyone should have free access to the tools of discovery and knowledge creation so as to maximize their social benefit. The Smithsonian Commons, which is characterized as "vast, findable, sharable, and free," is intended to provide anyone in the world virtual access to the entire Smithsonian, including "deep collections and the vitality, curiosity, and creativity of our staff, visitors, partners, and our extended global community."

The Barriers to Social Media as a Tool for Sharing

Interviewees brought up a number of barriers to the use of social media for data discovery and access at the Smithsonian. For the most part, the attitudes of scientists toward social media were characterized as similar to those toward data management—social media are not of interest because they are seen as "resource hogs" (in reality or potentially), as posing new costs in the form of learning, maintenance, and updating skills, and as taking scarce time away from research. Further, most scientists are used to

and satisfied with disseminating their work through peer-reviewed journals and professional conferences, and to conversing with colleagues via e-mail and in-person. At the same time, interviewees thought that younger scientists on the whole were more willing to participate in social media projects and share their work through such channels.

The difficulty experienced in getting Smithsonian scientists to participate in EOL is illustrative of the obstacles to data sharing, as one interviewee explained. At the time of this writing, over 600 people had signed up as EOL curators, and more than 200 were actively working, but nearly all were not NMNH staff. Interviewees cited a number of reasons. Although scientists are credited on EOL for posting their material and for reviewing and correcting content, their peers and supervisors may not recognize or reward their contributions unless these activities are explicitly included as outreach in performance plans. Scientists worry about other people scooping their research. They may have their own website and not want to participate on another one. They often don't see any benefit to participating in EOL, except in the case of some retiring scientists who realize that it is a way to save their work for posterity. Scientists are reluctant to share their work until, as one interviewee explained,

> *[it is] perfect. … Scientists are used to holding it back until is ready for the public. So we do have a lot of NMNH curators that have done some work, but they are not ready to share on EOL because is not the way they wanted it to look. … [Scientists] do not want to jump in too soon because [they are concerned] it will all disappear, and also the relationship between what EOL is doing and what they are already doing is not always clear.*

Finally, initially EOL did not offer training on working with EOL, although it plans to do so in the near future. EOL is also working on improved tools that should make reviewing material easier and help with recruiting more volunteers.

A further issue for many Smithsonian staff, both scientists and non-scientists, is a deep wariness about platforms that allow the public to post content or comments on Smithsonian sites. People might post inappropriate material that reflects badly on the Institution in the eyes of key stakeholders, or post scientifically inaccurate material that undermines the Smithsonian's reputation as a trusted source of authoritative content. While such problems can be addressed by maintaining, monitoring, and responding to posts on social media platforms, those tasks take up a lot of time, of which staff have little to spare. While lack of staff time is a perennial complaint, in the case of social media it has particular resonance. The demands of these technologies did not exist until recently and are growing very rapidly, without commensurate funding. Too often, responsibility for them has simply been added onto existing workloads.

A related issue is that the Smithsonian has limited capacity to communicate in languages other than English tend to be scarce across the Smithsonian (except at STRI), which stymies ambitions for creating a stronger international presence on social media platforms.

Some interviewees also noted technical obstacles. The Smithsonian's network infrastructure does not support chat ports because security concerns prohibit chatting during webcasts and video conferences. Audiences can engage only via email, which is less interactive.

In general, Smithsonian leadership is strongly behind use of social media as an important means of communicating with different audiences, in particular younger ones. This policy is reflected clearly in the Smithsonian's 2011-2015 strategic and digitization plans. The issue, not surprisingly, is the limited funding available to apply these media. Some funding is available through the SI Commons and Web 2.0 funds. However, there is concern that it is insufficient to allow the Smithsonian to keep up with emerging platforms and applications and with other organizations such as NASA. The risk is that the Smithsonian will be out-of-date and unresponsive, especially to new generations of scientists. A further issue, according to some interviewees, is that there is no clear leader/advocate for social media at the Smithsonian.