SOME CHALLENGES AND ISSUES IN MANAGING, AND PRESERVING ACCESS TO, LONG-LIVED COLLECTIONS OF DIGITAL SCIENTIFIC AND TECHNICAL DATA

W. L. Anderson
Praxis101, LLC, 12 Hayward Place, Rye, NY, 10580, and U. S. National Committee for CODATA
Email: band@acm.org

Abstract

One goal of the Committee on Data for Science and Technology[*] *is to solicit information about, promote discussion of, and support action on the many issues related to scientific and technical data preservation, archiving, and access. This brief paper describes four broad categories of issues that help to organize discussion, learning, and action regarding the work needed to support the long-term preservation of, and access to, scientific and technical data. In each category, some specific issues and areas of concern are described.*

Keywords: Scientific and technical data, Digital archiving, Data management, Long-term preservation, Data access

1 INTRODUCTION

Hardly a day goes by in which the discovery, correction, or use of scientific, technical, or business data is not featured in the news. It is well known that scientific and engineering research and development is completely dependent on observational and experimental data. Collecting data, generating data, reviewing and correcting data, analyzing data, and saving, sharing, and re-using data are activities that sustain a crucial information infrastructure for science, engineering, medicine, education, and governance. Recent reports show the contested discussion of climate changes and environmental pollution that grows out of analyses of current and historical data collected about the sea, sky, and land (Osumi-Sutherland, 2004; New View of Data Supports Human Link to Global Warming, 2003, November 18; Lower Atmosphere Temperature May Be Rising, 2003, September 12). Advancements in biology and treatments in medicine are likewise grounded in empirical data; individual patient experience, short- and long-term experiments, and clinical studies contribute to a growing database of information that informs clinical practice, treatment development, and government regulation.

There are three aspects of modern science that highlight the importance of maintaining long-term access to scientific and technical data. One is the growing interdependence among the traditional scientific disciplines. Much current science uses and depends on data captured in diverse fields. Data collected in one scientific discipline are likely to be used in other disciplines. This has long been true for areas such as biochemistry, chemical physics, and for large systems disciplines like ecology. Second, the requirements for scientific data management are changing in response to new and continuing developments of technology. Computation and networked computers are an integral part of the operational infrastructure of modern research and engineering. The proliferation of, and dependence on, computer-based technology is changing both the practices and products of science (Szalay & Gray, 2001). Third, these changes in practices and products are changing the role and nature of data preservation and access (Ginsparg, 2004, February 4). Archiving and preservation of S&T data can no longer be thought of as a post project activity. Data preservation and access practices must be part of the everyday practices of science and engineering.

[*] The Committee on Data for Science and Technology (CODATA) is an interdisciplinary Scientific Committee of the International Council for Science (ICSU). Information about CODATA and its activities can be found at www.codata.org.

CODATA works generally to foster worldwide cooperation to improve the quality, reliability, management, and accessibility of data in all fields of science and technology. In 2000, a CODATA Working Group on Archiving Scientific Data was established to focus on archiving, preservation, and access. The Working Group, along with the U.S. National Committee (USNC) for CODATA and the South African National Committee, organized a Workshop on Archiving Scientific & Technical Data, which was held on 20-21 May 2002 in Pretoria, South Africa (CODATA, 2002). In 2002, the Task Group on Preservation and Archiving of Scientific and Technical Data in Developing Countries was formed to continue the archiving efforts and to focus on developing countries. The Task Group, in collaboration with the U.S. and Chinese National CODATA Committees, organized a workshop on Strategies for Preservation of and Open Access to Digital Scientific Data in China, which was held on 22-24 June 2004, in Beijing, China (USNC, 2004). The Task Group and its activities have been presented and highlighted at other international workshops (ERPANET, 2004; CRIA, 2004). In addition, CODATA is collaborating with the International Council for Scientific and Technical Information (ICSTI, n.d.) to develop an Internet portal to online resources related to archiving and preserving access to digital scientific data and information.

2 FOUR CATEGORIES OF ISSUES

In exploring the many issues and requirements attendant to collecting, archiving, and preserving access to S&T data four categories emerged that are useful for capturing the breadth of discussion: science, management, policy, and technical. This section defines of each of these categories and subsequent sections describe key issues in each one.

Science-based issues focus on the discipline specific, as well as interdisciplinary and pan-disciplinary needs, values, methods, and practices of collecting and using data. Management issues are associated with the practices and procedures of individuals, archival institutions, and communities. Policy issues focus on the rules, regulations, and laws external to the archive, and include the ways in which they inform, assist, and constrain management and administration. Technical issues encompass standards, hardware, and software that support data preservation, archiving, and access functions. Technical issues are mostly discipline independent.

Applying these distinctions separates issues associated with collecting scientific data from those of managing the stored information, and both of these from the policies that require open access and wide distribution. However, the issues associated with managing and maintaining long-term availability of S&T data are often inextricably linked and co-defining. Even though issues often fit into more than one category, this classification has proved helpful in organizing CODATA archiving-related workshops, reports, and action plans.

3 SCIENCE-BASED ISSUES

3.1 What are scientific data?

This is a primary question that arises when talking about the science-based needs and issues of archiving scientific data. Most science is based on observations that include direct observations of the natural world and observations of experimental output and results. Increasingly, data are being generated from computer models. Sometimes “data” are equated with “observations”. In other situations, the word “data” is used as a descriptive term for the starting point of the analysis or research of interest. For example, to a seismologist interested in seismic waves, seismometer outputs are data. However, to a civil engineer concerned with evaluating the seismic risk that will influence the design of a new building, an earthquake catalog provides the basic data. What constitutes data is determined by the activities and objectives of the users.

From the perspectives of archiving, observational data are distinguished from experimental and computer-generated data by their uniqueness. Observational data are collected once at specific times and places, and are either remotely sensed from satellites or ground-based instruments or collected *in situ*. On the other hand, laboratory experiments are usually reproducible, and often the resulting experimental data need not be stored. In some cases, experimental data also include compilations of evaluated data. However, some experiments, such as the Large Hadron Collider at CERN, are very expensive to perform, and require access to limited resources that are in demand by many people and projects. The data from these experiments are, for all intents and purposes, unique, and therefore need to be saved and to remain accessible.

In contrast to physical science data, human-related data have their own characteristics. Human-related experimental data may be observational in nature (observational refers to the mode of data collection). The meaning and utility of the data are inextricably linked to how something is measured (by thermometer, questionnaire, ruler, human judge, etc.), and the context under which it is measured (highly constrained laboratory environment, in vivo, etc.). In addition, the tools and context of measurement are crucial pieces of information (metadata) that need to be associated with, and stored with, the raw data points. For all of these examples, the data generally consist of numeric values. However, in many disciplines, such as climate research, biology, or psychology, the primary collections include materials such as ice core samples, tissues samples, language corpora, interactions captured on video, as well as observational measurements.

3.2 Data and metadata

Discussions regarding the archiving of scientific data must be clear as to the kind of data being collected and preserved, and include all descriptive data and information that needs to be associated with those kind of data. Likewise, collections of S&T data must include data values and descriptive information. Metadata descriptions are as important as the data values in providing meaning to the data, and thereby enabling sharing and potential future useful access.

In addition, different data collections have different mandates and objectives that yield different metadata requirements. Collections that are available for the active phase of a research project often only produce simple, and possibly idiosyncratic, metadata. Data collections that support many different research projects and communities require more standardized descriptive documentation. Data collections that must remain available for many years, and for diverse user populations, have even more demanding documentation requirements.

3.3 Developing and developed country differences

Scientific research is a universal activity that is considered to be important for worldwide social and economic development. However, what qualifies as useful science and scientific data varies by country. Scientific work in developing countries often needs to be focused on basic health, agriculture, and environmental concerns. More developed countries often pursue capital intensive, observational and theoretical science and technology projects. In addition, data collections, even within a particular discipline, are often created and made available only in one spoken language. Spoken language differences are clear barriers to data access and sharing.

3.4 Nomenclature and taxonomy

Persistent and complex data issues that need to be resolved arise from the differences that exist around naming data. For example, in microbiology a recent study illustrates that reproducible, correct identification of microorganism species is dictated by “consensus and usage.” (Krichevsky, De Vos, Dejsirilert, Henry, Lalucat, Moore et al, 2004) In addition to the subjective assignment of species names, these names, and the concepts that inform them, change over time. Accommodating these natural developments in any scientific discipline requires that the historical context of names and concepts also needs to be saved.

3.5 Barriers to preservation and access

S&T data collections vary enormously in terms of written language, level of data aggregation, and standards for description, use, and storage, to name a few collection attributes. Sharing and preserving future access to these data collections raise issues of interoperability and compatibility. These issues include practical concerns about connectivity and electronic network infrastructure, as well as practices, tools, and standards to support language translation and consistent data description and use.

Different scientific communities have different approaches and attitudes to preservation. Astronomers have long appreciated their dependence on data collected by others as well as data collected in the past. They have a well-developed culture of sharing. Psychology, on the other hand, due in part to the personal and private nature of data collected from human subjects, is just beginning to develop broadly shared data resources.

Another barrier is that substantial amounts of original data are in a form that makes them less accessible (e.g., on paper, in videotape). In addition, substantial data are also buried in other products such as spreadsheets, databases, documents, etc. How many old floppy disks, that remain the primary data backup medium, are simply packed away in drawers, boxes, and closets? These data are kept locally and often are not identified, organized, or documented in standard, usable ways. Data are not always tabular. And finally, as mentioned earlier, not all data in a given discipline are published or available in the same language.

Furthermore, science and discipline related issues around collection management and preservation are complicated by the growth of interdisciplinary work that reflects the new research at the boundaries of established disciplines. These efforts require access to sometimes-incompatible data from different disciplines. Additionally, interdisciplinary science needs to support a diversity of users (Fox, Garcia & Kellogg, 2004).

4 DATA MANAGEMENT ISSUES

4.1 Archiving is changing

Archiving is often thought of as a post-research project activity and is divorced from the day-to-day work of research. Archiving scientific and technical data is associated with operational aspects of collection management. It is not considered part of research activity; it is seen as necessary overhead. However, as the amount and diversity of digital data collected and used in scientific research has grown, the needs of effectively managing these data collections have become part of daily research practice. In addition to providing the information infrastructures that enable and support data collection and analysis, data managers are also required to provide robust and reliable, short- and long-term, access to data collections of ever increasing volumes.

The growth of the Internet, and support for access to diverse and distributed digital collections of S&T data, has also raised expectations on the availability and quality of the data and its documentation. Expectations of the quality of metadata, and the usability of data are higher for digital materials than for paper materials. Since much S&T data is born digital, expectations of access and usability are present from the beginning of the project.

The convergence of data management responsibilities and scientific work practices requires changes in the work and attitudes of both data managers and researchers. First, the entire notion of archiving as being separated from research needs to be replaced by a notion of shared S&T data collection management. Second, reward structures and mechanisms need to be established to support and acknowledge the publication of data collections and repositories. Third, publication of datasets must be rewarded professionally in a manner similar to publication of scientific papers and technical reports. Finally, standards will be needed to insure consistent data management practices. The Open Archival Information System (OAIS) Reference Model standard provides a framework for developing such processes and practices (NASA, n.d.).

Prototypes are being developed using different technologies and policies that might manage collections for individual researchers and small projects. These attempt to accommodate diversity by being flexible about metadata and documentation. The MIT DSpace project is one example of potentially useful technology (DSpace, n.d.). Another example is the prototype work on establishing a public commons for geographic data (Onsrud, et al., 2004). In addition, an informative survey of operational digital preservation systems in science that identifies trends, issues, and activities of a wide range of organizations interested in preservation and permanent access was updated and released this year by CENDI and ICSTI (Hodge & Frangakis, 2004).

4.2 Requirements for effective data management

One thing that has been learned from existing scientific data collection management is that these archives must be initiated by respected scientists working with people trained and skilled in data management (CODATA, 2002). Domain knowledge is needed in order for collections to be managed, documented, and kept usable by scientists and researchers.

Maintaining effective access to S&T data collections requires the participation of all data collection users, researchers, data managers, laypeople and interested citizens. All stakeholders need to be involved in providing requirements for data quality assessment and assurance, as well as long-term use and usability. Documentation and metadata generation and maintenance are critical to enable effective and efficient preservation and access. The digital library community has done an enormous amount of work in exploring and developing metadata standards for access and for preservation (Woodyard, 2002). S&T data collection users and managers need to work with and build on these existing digital library efforts.

Standards efforts supporting a diversity of digital resources have been bolstered by the broad acceptance of an OAIS reference model in the digital library and numerous scientific communities. One of the great strengths of the OAIS reference model is that it provides a framework for system architecture and management and policy practices. Of interest in the area of collection management is the OAIS separation of concerns between (1) archive administration and (2) external management and community issues (NASA, n.d.).

4.3 Funding

One primary and unrelenting issue facing managers of S&T data is that of funding. First, maintaining collections of digitized data, is in many ways more expensive than storing paper records. Maintaining usable and accessible digital resources requires maintenance and upkeep of hardware and software. Changes in hardware and software often require transformation of data sets from one format, or configuration, to another. Data saved on old media or captured in obsolete and proprietary formats need special attention to be kept available. The primary ways available today of maintaining availability of old digital resources include (1) refresh (periodic copy of data from one storage media to another and possible data transformation between formats), (2) archive and maintenance of out of date hardware transformation, and (3) emulation (construction and maintenance of software that supports old and obsolete data formats). All these activities cost money.

Second, data management expenses, particularly those associated with long-term preservation and access, are often considered as a perhaps necessary cost, but one that takes money away from research activities. This is especially true when archiving and long-term data management are considered post-project activities. It is imperative that researchers and the agencies that fund research come to grips with the actual costs associated with long-term maintenance of, and preserving access to, S&T data.

4.4 Business and organizational models

Closely associated with the funding issues are the challenges of developing practicable business and organizational models for long-term, accessible data collections. Explorations of different kinds of partnerships — commercial and public — are needed to discover the economics and costs of different models. These partnerships will require negotiations of ownership and control of collected and managed data. The U.S. National Weather Service has been working on partnerships with the private sector for some time (NRC, 2001). Participation and input is required from all stakeholders to create working business and organizational models (NRC, 2003). Research into existing, and proposed, business arrangements is needed to make visible, the requirements, benefits, and costs of data archiving, and to secure the required resources.

4.5 Selection and appraisal

A key issue common to many, if not all, archives is the selection and appraisal activity. How are data chosen for retention? A recent CODATA/ERPANET workshop brought archivists and records managers and scientists and data managers together to discuss the objectives and practices of data selection, appraisal, and retention (Ross, et al., 2004; Esanu, Davidson, Ross & Anderson,2004). Although the two groups share many of the same values, the practices and procedures for selecting and appraising data differ across the archival and scientific communities.

Differences exist in data quality and in selection and appraisal criteria used to review and assess data value. The interchange at the CODATA/ERPANET was fruitful, and the difference in perspectives about records and data was striking. (ERPANET, 2004; Esanu, et al., 2004)

The retention of selected and appraised data raises other issues. These include maintenance of data quality, enforcement of data security, and, migration of data sets to current, available and maintainable, hardware and software systems.

4.6 Planning and requirements issues

Specifying requirements for managing digital collections of S&T data is problematic because these requirements change in response to changes in the practices of science and information management. In addition, practice changes are specific by discipline, geography, community, and time. Planning for these changes is difficult without relative consensus on activities and objectives. And as more and more collections are made available to the general public, supporting access for diverse customers and groups of users produces new requirements.

4.7 Training and education

To many researchers archiving and preservation is provided by others and ensures secure and long-term data backup. This view is understandable because scientific research results have most often been published in journals and books that have been collected and made available at public and private libraries. This remains as important practice of scientific research, and publications and citations remain important incentives and attributes of a successful research career. Earlier scientific research papers often contained actual experimental and observational data as well as the interpreted conclusions. This enabled others to duplicate experiments, and check calculations and conclusions.

Modern scientific research, however, is based, in many cases, on computer-generated and interpreted data. These data sets are large, stored, and preserved on diverse collections of computer hardware and software. It is not practicable to publish the data as part of research conclusions, and as a result the primary data are not accessible to other researchers. However, scientific progress, and increasingly governmental policy, depends on access to and reuse of primary data. Several recent news items regarding new analyses of climate data show how important preservation of, and access to, S&T data is becoming (Osumi-Sutherland, 2004; New York Times, 2003, November 18; Lower Atmosphere Temperature May Be Rising (2003, September 12).

Making a transition to practices that are grounded in computer and information systems will require training and education of researchers and S&T data managers. Some understanding of information systems operations and management needs to be part of a basic education for scientists.

4.8 Some operational considerations

Archiving and preservation of data collections also includes management of operational issues. These include size, diversity, and longevity. The sheer amount of data being collected and requiring preservation, is enormous (NRC, 1995). In addition, scientific disciplines such as astronomy rely almost entirely on computer-collected data. The volume of the collected data increases as computing power increases, although some studies show that increases in computing power and storage density can support projected data volumes (Szalay, Gray & Vandenberg, 2002). On the other hand, the availability of inexpensive mass storage reduces the pressure on research projects to select and document the data they save; it is easier to simply purchase more disks space. One result is growing collections of poorly documented data.

Scientific and technical data also vary widely in their sources, formats, and, perhaps most importantly, documentation. The lack of standard and usable metadata and descriptive documentation can render much S&T data useless. The recognition of the need for standard and usable metadata is generating much useful activity; a key challenge is developing and implementing reliable and reproducible data documentation practices.

Issues associated with longevity include the time horizon for future access. As the time horizon requirements increase so does the uncertainty of the preserved information. Data definitions, data formats, and metadata content and standards all change over time. There is no way to avoid the fact that future generations may need to address their own versions of these issues.

Another important time horizon issue is managing hardware and software obsolescence. As mentioned earlier, the major strategies available to deal with the continuing obsolescence of hardware and software are maintenance, migration, and emulation. None of these solutions is ideal.

5 POLICY ISSUES

Policy issues arise from the external rules, regulations, and laws that apply to managing data collections. These issues are complex because there are several differing rationales for archiving and preserving access to S&T data. One major rationale is to support research and scientific exploration and discovery. However, S&T data preservation serves cultural, economic, and political needs as well.

Several symposia and workshops have been held in the past ten years to explore and define the many policy and practical issues associated with maintaining open and fair access to scientific and technical data and information (See Esanu & Uhlir, 2003, for a summary).

5.1 National, regional, and global perspectives

Data policies encompass and represent national, regional, and global perspectives, and the issues include (1) challenges of determining and insuring data privacy and confidentiality, particularly those of human subjects; (2) questions of cultural ownership and use of data; (3) national security; (4) determinations of intellectual property protection, limits, and exceptions; and (5) general ethical issues, particularly when dealing with personal and public health data.

The United States has a long history of providing open access to publicly funded research data. Exceptions to this policy are associated with national security and privacy concerns. However, as the U.S. government partners with private industry to provide more cost effective and efficient delivery of, for example, weather data, new questions arise about what data are public, what are private, and how to make collaborations work (NRC, 2003). National policies relating to data access differ greatly, from the open policy embraced by the United States, to the more closed approach preferred by the European Union. Other countries are beginning to look at the positive societal and economic benefits of an open data policy. China's establishment of a new program focused on sharing of scientific data is one such initiative.

5.2 Open and fair access

Open and fair access to scientific information has been a controversial issue for some time. Traditional paper publication of scientific research has become increasingly expensive for libraries and individuals. The growth of the Internet and World Wide Web has spawned online archives and initiatives that provide free and open access to scientific publications in different disciplines (arXiv.org, n.d., Public Library of Science, n.d.)

The expectation of "fair use" in education and research for both scientific data and the resulting publications often conflicts with the development of technologies that seek to enforce digital rights management. Resolution of these questions requires clarifying data rights in the entire data stream, including collection, redistribution, transformation, and use in derivative products.

The policy issues involved in finding practicable solutions include (1) crafting enabling legislation, and defining controlling authorities; (2) supporting policies, regulations, and practices for freedom of information and access authorization; (3) financing and cost recovery policies that support economies of scale, accommodate unfunded mandates, and provide incentives that support S&T data archives; and (4) developing and implementing policy enforcement mechanisms (Reichman & Uhlir, 2003).

6 TECHNICAL ISSUES

6.1 Scientific data and databases are different than literature

While it is true that at the level of bits there is no difference between digital information (published papers, reports, proceedings) and digital scientific and technical data it is nonetheless important to notice the differences that do matter now, and will continue to matter in the future. First and foremost there are size and volume differences; the volume of digital data far exceeds that of published information. Many observation-based scientific research areas collect, analyze, and store enormous amounts of information daily. For example, the United States Geological Survey (USGS) land remote sensing archive holds "2 petabytes representing 14,000,000 individual satellite observational records." An additional "one to two terabytes of satellite observations … arrive daily." (Faundeen, 2003) By contrast, "it is estimated that the text of the print holdings of the Library of Congress would, if digitized, constitute 17 to 20 terabytes of information." (Library of Congress, n.d.)

Second, the metadata required to describe data can be more complex than that required to describe written texts. Scientific and technical data can be raw or processed, intermediate or final. In addition, the information about the data sensing equipment is sometimes crucial; for example, is it calibrated? If yes, then what specific procedures were used? What are the other environmental conditions? In these cases, details about the transformations of the raw data into recorded values also need to be captured. A good analogy in the digital library world is to the kind of metadata required for preserving scanned images. In addition to scanning resolutions (absolute or interpolated), information indicating color or grayscale half-tones needs to be carried along with the image to allow preservation of rendering the image on a display, printing it on paper or other media, and migrating the data from old, and perhaps obsolete, technology to new technology infrastructures.

A common factor for both digital information and data is that in most cases accessibility to the original material is through applications that are continually evolving. For textual material stored in a human readable format such as ASCII, computer accessibility is almost guaranteed. However, ASCII formatted data is not necessarily the norm. One current trend is to choose a readily available application format such as PDF. PDF is a convenient document format today, due to the work and success of Adobe Systems, Inc., the company that invented and distributes PDF applications. Continued accessibility to PDF formatted documents depends on the stability of the format and of the Adobe Corporation. Scientific and technical data, on the other hand, are almost always stored in application-specific formats such as spreadsheets and databases. The accessibility of these data is contingent on the availability of the appropriate applications.

6.2 Diversity of data types, formats, media types, formats, and standards

Technology continues to develop and evolve rapidly. As previously mentioned, data and information are buried in proprietary formats, commercial databases, and idiosyncratic software that are hard to maintain and access. With what kinds of systems, and by what methods and practices, will data quality and security be maintained, data storage provided, and data migration supported, in the face of these changes and differences?

Supporting usable data access requires technology for search, analysis, and visualization. Large-scale data mining and analysis projects require interoperability among distributed collections. Current trends in network protocols and web services will provide flexible distributed systems architectures (W3C, 2004). But who needs what capabilities and to what ends? What are the appropriate and usable collection content discovery, and user authentication and authorization mechanisms? These questions are particularly relevant to efforts to build data collection and sharing capacity in the developing world.

Finally, continuing work is needed on standards, and initiatives are underway. OAIS provides a reference framework and architecture for S&T data collections and is now an ISO standard. The Open GIS Consortium has developed a reference model and technical specifications to support the sharing of geospatial data. From the digital library community the Online Computer Library Center (OCLC, n.d.) is developing preservation metadata standards. The Open Archives Initiative provides standards that foster access. These, and many related efforts, need to be supported; financial incentives are required to develop useful and usable standards, and to encourage their application to S&T data collections. Standards development and maintenance entail a great deal of work, as is made very clear in a recent Science essay on the history of meteorological standards (Edwards, 2004, May 7).

7 SUMMARY

The effort required for management and preservation of access to collections of scientific and technical data to be a routine aspect of scientific and engineering research and development is substantial. More and more S&T data are born digital and become digitized. At the same time, studies show that, in some cases, as many as 20% of Internet addresses become inactive within a year (Dellavalle, Hester, Heilig, Drake, Kuntzman, Graber et al, 2003). This lack of stability in Internet document identifiers will impact the long-term access to S&T data. In some ways the digital data sets and information are more fragile than paper-based or physical specimen collections and archives. As Bernard Smith pointed out, “digital resources will not survive or remain accessible by accident.”(Smith, 2002)

Several areas of opportunity exist to ensure long-term archiving and preservation of access to scientific and technical data and information. First, it is possible and practical to leverage the common digital properties of scientific data and information. The publication of scientific research results that reference and contain digital collections of primary data is putting pressure on researchers and publishers alike to enable access to both primary data and research reports. The requirements for systems supporting research and publication activities are converging. Second, past and ongoing experiences with managing large and growing collections of digital data are contributing to learning and capacity building in scientific disciplines and in the developing world. CODATA and other scientific and non-governmental organizations support information sharing and the development of interdisciplinary and international social networks of researchers and data managers committed to preserving the data and the record of science. Third, scientific and technical data managers are collaborating with librarians, archivists, and existing cultural heritage digital preservation activities to use the systems that are already in place and to incorporate unique S&T data requirements into existing and emerging standards.

CODATA is committed to collaborating with other scientific unions and organizations to resolve the issues outlined here and to insure that the data and the record of science are preserved and remain accessible to scientists, engineers, educators, and interested individuals and organizations.

8 ACKNOWLEDGEMENTS

This paper is a summary of a compilation of major archiving, preservation, and access issues identified by an ad-hoc working group of the U.S. National Committee for CODATA. The author would like to thank the members of the working group for their participation and contributions in identifying these issues. Specific thanks are due M. Bullock, J. Esanu, J. Faundeen, J. Rumble, and the reviewers for their suggestions and improvements. The author accepts full responsibility for any errors or weakness in the overall presentation. In addition, the author wishes to thank Paul Uhlir, Director of the U.S. National Committee for CODATA, and Julie Esanu, Program Officer at the U.S. National Research Council, for their support and active participation in the working group.

9 REFERENCES

arXiv.org (n.d.) Homepage of arXiv.org ePrint archive. Available from: http://arXiv.org/

CODATA (2002) CODATA Workshop on Archiving Scientific & Technical (S&T) Data Report (20-21 May 2002, Pretoria, South Africa, Section 3.2.1). Retrieved December 7, 2004 from the *CODATA* website: http://www.codata.org/

CRIA (2004), Inter-American Workshop on Environmental Data Access. Retrieved December 7, 2004 from the *Reference Center on Environmental Information* website: http://www.cria.org.br/eventos/iaed/

Dellavalle, R.P., Hester, A.J., Heilig, L.F., Drake, A.L., Kuntzman, J.W., Graber, M., & Schilling, L.M., (2003) Going, Going, Gone: Lost Internet References. *Science 302*, 787-788.

DSpace (n.d.) Homepage of Dspace. Available from http://www.dspace.org/

Edwards, P. N. (2004, May 7) A Vast Machine: Standards as Social Technology, *Science 304*, 827-828.

ERPANET (2004) The Selection, Appraisal, and Retention of Digital Scientific Data (ERPANET/CODATA Workshop). Retrieved December 7, 2004 from the *ERPANET* website: http://www.erpanet.org/www/products/lisbon/LisbonReportFinal.pdf

Esanu, J., Davidson, J., Ross, S. & Anderson, W. (2004) Selection, Appraisal, and Retention of Digital Scientific Data: Highlights of a CODATA/ERPANET Workshop. *Data Science Journal* (under review).

Esanu, J.M. & Uhlir, P.F., (Eds.) (2003) *The Role of Scientific and Technical Data and Information in the Public Domain*. Washington, DC: National Academies Press.

Faundeen, J. (2003) Interdisciplinary Case Study 2: Earth & Environmental Sciences U.S. Geological Survey/EROS Data Center Archiving Perspective, Presentation to the *ERPANET/CODATA Workshop on the Selection, Appraisal and Retention of Digital Scientific Data*, Lisbon, Portugal.

Fox, P., Garcia, J. & Kellogg P. (2004) The High Altitude Observatory Data Service: experience in interdisciplinary data delivery. Retrieved August 1, 2004 from the *High Altitude Observatory* website: http://dods.hao.ucar.edu/papers/codata2000_paper.pdf

Ginsparg, P, Scholarly Information Architecture, 1989 – 2015, *Data Science Journal, 3*, 29-37.

Hodge, G. & Frangakis, E. (2004) Digital Preservation and Permanent Access to Scientific Information: The State of the Practice. Retrieved December 7, 2004 from the *CENDI* website: http://cendi.dtic.mil/publications/04-3dig_preserv.html

ICSTI (n.d.) Homepage of the International Council for Scientific and Technical Information. Available from http://www.icsti.org

Krichevsky, M.I., De Vos, P., Dejsirilert, G, P., Henry, D., Lalucat, J., Moore, E., Sega, M., Tang, J., Whitehead, S., Zhou, Y. & Yu, H. (2004) Consistency of Identification of Pseudomonads and Related Organisms. Presentation to the *American Society of Microbiology Meeting*, New Orleans, USA.

Library of Congress (n.d.), Retrieved December 7, 2004 from the *Wikipedia* website: http://en.wikipedia.org/wiki/Library_of_Congress

Lower Atmosphere Temperature May Be Rising (2003, September 12) *Nature*, Retrieved December 6, 2003 from: http://www.nature.com/news/2003/030908/full/030908-17.html.

NASA (n.d.) ISO Archiving Standards – Overview. Retrieved December 7, 2004 from the *NASA* website: http://ssdoo.gsfc.nasa.gov/nost/isoas/overview.html

National Research Council (NRC) (1995) *Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources* (pp 19-30). Washington, DC: National Academy Press.

New View of Data Supports Human Link to Global Warming (2003, November 18) *The New York Times*, Late Edition, Final, Section F, Page 2.

NRC (2001) *Resolving Conflicts Arising from the Privatization of Environmental Data*. Washington, DC: National Academies Press.

NRC (2003) *Fair Weather: Effective Partnerships in Weather and Climate Services*. Washington, DC: National Academies Press.

OCLC (n.d.) Homepage of the Online Computer Library Center. Available from http://www.oclc.org

Onsrud, H., Camara, G., Campbell, J. & Chakravarthy, N. S., (2004) Public Commons of Geographic Data: Research and Development Challenges. In Egenhofer, M.J., Freska C. & Miller, H.J., (Eds.), *Geographic Information Science,* Berlin; Springer-Verlag. See the prototype at http://www.spatial.maine.edu/geodatacommons

Osumi-Sutherland, D. (2004, August 11) Climate predictions gain surer footing. *Nature*, Retrieved December 7, 2004 from: http://rolos.nature.com/news/2004/040809//full/040809-9.html [doi:10.1038/news040809-9]

Public Library of Science (n.d.) Homepage of the Public Library of Science. Available from http://www.plos.org

Reichman J.H. & Uhlir, P.F. (2003) A Contractually Reconstructed Research Commons for Scientific Data in a Highly Protectionist Intellectual Property Environment. *66 Law & Contemporary Problems* (Winter/Spring), 315-462. Retrieved December 7, 2004 from: http://www.law.duke.edu/journals/66LCPReichman/

Szalay, A.S. & Gray, J. (2001) The World-Wide Telescope. *Science*, *293*, 2037-2040.

Szalay, A.S., Gray, J. & Vandenberg, J. (2002) Petabyte Scale Data Mining: Dream or Reality? *SPIE Astronomy Telescopes and Instruments*,Waikoloa, Hawaii.

Smith, B. (2002) Preserving tomorrow's memory: Preserving digital content for future generations. *Information Services and Use*, *22*(2, 3*)* 133-139.

USNC (2004) Strategies for Preservation of and Open Access to Digital Scientific Data in China. Retrieved December 7, 2004 from *US National Committee for CODATA* website: http://www7.nationalacademies.org/usnc-codata/chinese_workshop.html

W3C (2004) Web Services Architecture: W3C Working Group Note 11 February 2004. Retrieved December 7, 2004 from the *W3C* website: http://www.w3.org/TR/2004/NOTE-ws-arch-20040211/

Woodyard, D. (2002), Metadata and Preservation. *Information Services and Use*, *22*(2, 3) 121-125.