Selection, Appraisal, and Retention of Digital Scientific Data: Highlights of an ERPANET/CODATA Workshop

J. Esanu, [1] **** J. Davidson and S. Ross,*** [2] ***and W. Anderson***[3]

[1]*U.S. National Committee for CODATA, The National Academies, 500 Fifth Street, NW, Washington, DC 20001*
Email jesanu@nas.edu
[2]*ERPANET, Humanities Advanced Technology and Information Institute (HATII), George Service House, 11 University Gardens, University of Glasgow G12 8QQ*
Email britisheditor@erpanet.org and s.ross@hatii.arts.gla.ac.uk
[3]*Praxis101, 12 Hayward Place, Rye, NY 10580*
Email band@acm.org

ABSTRACT

CODATA and ERPANET collaborated to convene an international archiving workshop on the selection, appraisal, and retention of digital scientific data, which was held on 15-17 December 2003 at the Biblioteca Nacional in Lisbon, Portugal. The workshop brought together more than 65 researchers, data and information managers, archivists, and librarians from 13 countries to discuss the issues involved in making critical decisions regarding the long-term preservation of the scientific record. One of the major aims for this workshop was to provide an international forum to exchange information about data archiving policies and practices across different scientific, institutional, and national contexts. Highlights from the workshop discussions are presented.

Keywords: Scientific data; Digital archiving; Long-term preservation; Selection and retention guidelines; Appraisal process; Data management

1 INTRODUCTION

The current state of archiving digital scientific data varies widely across disciplines with respect to the volume of data that have already been archived, the degree of standardization and interoperability among the data sets, and the frequency with which they contribute to ongoing scientific research. For example, researchers, data managers, and information specialists in the Human Genome Project in the biomedical sciences and at the U.S. Land Remote Sensing Archive in the Earth and environmental sciences began consultations from a fairly early date to identify selection and appraisal guidelines to determine which data would be preserved, along with relevant retention schedules. They established standards for metadata and software/hardware that would help ensure that the enormous quantities of data being collected would be maintained in compatible and accessible media. They also developed common policies for collaborating organizations for the deposit and dissemination of the data, and formal institutional and financial plans to make their data archiving activities sustainable. As a result, very large, centralized databases exist in these disciplines that are funded and maintained on a full-time basis and are available for remote and collaborative research.

In other disciplines, the volume, availability, and accessibility of archived data are more variable, with patchwork endeavors ranging from established policies at the local and national levels at various research institutions and universities to joint disciplinary efforts to larger international programs. Some of these discipline-based activities have become formally networked across institutions and nations, while others have not.

In light of the diverse disciplinary practices related to the archiving of digital scientific data, ERPANET—the Electronic Resource Preservation and Access Network—and CODATA—the Committee on Data for Science and Technology, an interdisciplinary committee of the International Council for Science—collaborated to convene an international workshop that focused on the selection, appraisal, and retention of such data. The workshop was held on 15-17 December 2003 at the Biblioteca Nacional in Lisbon, Portugal, and brought together more than 65

researchers, data managers, information specialists, archivists, and librarians from 13 countries to discuss the issues involved in making critical decisions regarding the long-term preservation of the scientific record.

One of the major aims for this workshop was to provide an international forum to exchange information about data archiving policies and practices across different scientific, institutional, and national contexts. The objectives were to:

- Use disciplinary and interdisciplinary case studies to assess the commonalities and differences among disciplines and determine the extent to which common selection, appraisal, and retention principles and policies can be identified and applied, regardless of discipline, and those that are unique to a discipline or a category of data.
- Identify and discuss some of the key scientific, technical, management, and policy considerations for the successful implementation of appraisal, selection and retention principles and policies, with particular attention to issues of efficiency, effectiveness, and the broad range of potential benefits – economic and other – that can be achieved.
- Provide a networking opportunity for workshop participants to meet with other researchers, data managers, information specialists, archivists, and science policy experts across discipline and national boundaries.

A final report was published by ERPANET that summarizes the workshop (ERPANET, 2004). As such the scope of this article focuses on the highlights of the workshop's discussions as described below.

2 MANAGEMENT AND POLICY CONTEXT FOR ARCHIVING DIGITAL SCIENTIFIC DATA

The selection and appraisal of scientific data for long-term preservation is embedded in a larger data management and policy context (Hodge and Frangakis, 2004; National Research Council, 1995a; Anderson, 2004). Some important policy issues related to archiving scientific data include accommodating the needs and practices of different scientific disciplines, as well as encouraging the development of interdisciplinary research values and methods. Differences in nomenclature and taxonomy of data exist within, as well as across, scientific communities. Names and concepts may change over time, and the preservation of data requires preserving historical contexts. Within a specific region there likely will be differences among the mandates and objectives of individual archives. Archives for different types of data (e.g., observational vs. experimental, physical science vs. biological science, human subjects or not) may have disparate procedures and metrics for data quality, and differing criteria for appraising value and selecting data for preservation. In addition, building digital repositories of scientific data requires resources and expertise to transfer paper-based data to electronic forms.

Relevant data management issues include the implementation of practical, day-to-day operational procedures and practices needed to preserve and maintain access to the data resources (Anderson, 2004). First and foremost is the ongoing need to secure and maintain the requisite funding. A general problem common to many scientific disciplines is the low priority attached to data management and preservation. Experience indicates that new research projects tend to get much more attention than projects that attempt to preserve and reuse existing data, even though the payoff from optimal utilization of existing data may be greater. In addition to managing the issues of data selection and appraisal, archives must also contend with issues of ownership and control. Operational issues for scientific data archives include properly managing data volume (the number of bits is enormous and growing); dealing with the diversity of sources, formats, and documentation; and maintaining a sufficiently long time horizon for access in the face of continual change in data definitions, digital data media and formats, and hardware and software obsolescence. Planning and developing requirements for data archives must accommodate the continual change and evolution in the practice of science; the local variability in focus, practice, and available technology and other physical and human infrastructure; the differing mandates and objectives of different data producers, as well as a diversity of potential users, including scientists, educational institutions, businesses, policymakers, and ordinary citizens.

Effective management of scientific data archives also depends on an investment in training and education. The dependency of science on technology makes archiving considerations part of experimental and observational data

collection and reporting. Archiving requirements no longer apply solely to post-research publication activities. Properly documenting data for long-term preservation and access must become part of the daily practice of scientists. Promoting these changes to established practice requires collaboration among scientists, data managers, educators, and archive users.

The U.S. National Research Council (1995a, 1995b) addressed the issues of retention criteria and the appraisal process when examining long-term retention of scientific and technical records for the U.S. federal government at the request of the U.S. National Archives and Records Administration. The National Research Council found that scientific data create special problems for appraising their long-term value, particularly beyond the primary user community. They are voluminous, constantly increasing, difficult to label, and often difficult to access and use by all but the original project scientists. In addition, they are often very expensive to collect, but provide baselines for future collections and enhance understanding of other data. They are frequently of immense importance for advancing scientific endeavors, for educating new scientists, and for many other social and economic applications. At the same time, it may be difficult to ascertain the full value of data to researchers and other users in the long term.

The National Research Council also noted that the appraisal process must take into account the whether the data are raw (primary) or processed. Retention criteria must be established to guide the creators and managers of scientific data. Creators must also be motivated to consider the needs of secondary users so that data are retained in a manner that provides maximum benefit to researchers and users in the long term. As such, scientists need to become active agents in the appraisal process.

The workshop examined these aforementioned issues through the identification of disciplinary and interdisciplinary case studies to identify common themes regarding the selection, appraisal, and retention of digital scientific data for long-term preservation, regardless of discipline.

3 WORKSHOP ON THE SELECTION, APPRAISAL, AND RETENTION OF DIGITAL SCIENTIFIC DATA

The workshop began with an introduction to CODATA activities related to archiving and preserving scientific data by William Anderson of Praxis101. In 2002, a CODATA Task Group on Archiving and Preserving Scientific and Technical Data in Developing Countries was established; Dr. Anderson serves as a Task Group co-chair. This Task Group has been working to identify the scientific, management, technical, and policy issues related to archiving scientific data. They have compiled a comprehensive, annotated bibliography of archiving resources and initiated a series of workshop focused on preserving and accessing digital scientific resources. This presentation was followed by an overview of the selection, appraisal, and retention of digital records from an archivist's perspective by Terry Eastwood of the University of British Columbia (Eastwood, 2004). An example of deriving the maximum economic benefit from the long-term retention of scientific data was then presented by Peter Weiss of the U.S. National Weather Service. These three talks helped to set the context of current archival activity and to identify potential benefits that the long-term preservation of digital scientific data may offer.

The workshop moved on to examine the selection and retention practices and appraisal guidelines, and archiving policies in general, in a variety of scientific disciplines through five disciplinary and interdisciplinary case studies. Three disciplinary case studies were presented. The first, by Jürgen Knobloch, focused on experimental laboratory data in the physical sciences with the example of the Large Electron-Positron (LEP) Collider at CERN. The second case study focused on biological sciences with an example of the Global Biodiversity Information Facility (GBIF). Meredith Lane of GBIF provided the archival experiences at GBIF, and Weber Amaral of the International Plant Genetics Research Institute provided the high-value user perspective for this case study. The International Virtual Observatory in Astronomy was the final disciplinary case study in the space sciences; Francoise Genova from the Strasbourg Astronomical Data Centre provided the data center perspective and Alex Szalay from the Johns Hopkins University gave the high-value user perspective. The first interdisciplinary case study presented examples of archiving practices in the social sciences in both Europe and the United States with a description of the activities at the U.K. Data Archive and Inter-university Consortium of Political and Social Research in the United States by Kevin Schürer and Myron Gutmann, respectively (Gutmann, et al., 2004). The final interdisciplinary case study focused on the archiving of large-volume data sets from remotely sensed observations in the Earth and

environmental sciences. John Faundeen described the appraisal process at the U.S. Land Remote Sensing Archive at the U.S. Geological Survey's (USGS) Earth Resources Observing System Data Center, and Luigi Fusco of the European Space Agency discussed the use of Earth observation archives in virtual digital libraries and GRID infrastructures (Fusco and van Bemmelen, 2004). The case studies illustrated the commonalities and differences among various scientific disciplines by offering a range of archiving, end-user, and national perspectives. In many cases, more than one of these perspectives was explored.

The case studies illustrated that there is a great range of archival activity being undertaken in the various scientific disciplines. While different disciplines focus their activity in different areas of the archival life cycle, all have valuable experience to share. Through improved communication and collaboration, a great deal can be learned from each other. The space sciences showed that metadata and interoperability can have a major impact on the accessibility of data. They also demonstrated the value of disciplinary cooperation in the creation and adoption of standards and strategies. The social sciences case study suggested that metadata can be used as a means of appraising the long-term value of digital data. In addition, the case studies provided valuable insights into the concepts of rejecting or disposing of digital data. The physical sciences case study demonstrated the importance of retaining contextual information with digital data for their long-term value and reuse. The case study in the biological sciences also outlined the value of maintaining context and demonstrated the potential social benefits of making digital scientific data more widely accessible, especially among developing countries. The presentations on Earth observation sciences showed that interdisciplinary collaboration can be of great benefit as was seen with the development of the USGS Appraisal Tool. The archival community demonstrated that appraisal is effectively a judgment on the value of digital data. As such, appraisal can have a huge impact on justifying future preservation activity being carried out on the digital data. The potential benefits of the widespread adoption of the Open Archival Information System (OAIS) Reference Model (Consultative Committee, 2002) were also illustrated by the archival discipline.

One of the main goals of this workshop was to determine any commonalities and differences that exist with regards to the selection, appraisal, and retention of digital scientific data among various scientific disciplines. Several commonalities, focusing primarily on the importance of maximizing the value of digital resources, were highlighted throughout the workshop. It was commonly agreed that maximizing value was best achieved through the reuse of digital data sets. It was universally agreed that context is of crucial importance in enabling reuse of digital data and that this could only be guaranteed through the application of quality metadata or documentation. As this will involve greater financial investment in the initial creation or retrospective documentation of digital data sets, the workshop participants believed that it would be necessary to increase awareness of the value of the reuse of data among funding bodies. Many participants thought that the lobbying of funding institutions to include data curation and preservation costs in scientific grants should be actively pursued as a result of the workshop.

As mentioned, the case studies also illustrated differences between the disciplines. For example, the amount of data collected and the method by which they are generated differed greatly by discipline from vast amounts of astronomical data collected by large telescopes to individual biological specimens collected by field researchers. The level of data (i.e., raw or processed) archived also varied by discipline. In addition, practices for sharing data, including publishing models and dissemination strategies, varied not only by discipline, but also by institutions and national policies. Finally, the commercial value of the data also differed by discipline.

A moderated plenary discussion followed the case studies, which allowed participants and speakers to engage in discussions on the issues raised during the presentations. The dialogue was lively and produced very interesting discussions on the issues of publication models, the roles and responsibilities of researchers and archivists in the preservation process, and the costs associated with archiving data. However, as the workshop progressed, it became clear that the concept of appraisal was neither universally understood nor agreed on by all participants. In fact, issues of semantics were at the root of some of the challenges encountered during the workshop. Definitions and use of key terms such as archiving often differed based on discipline and data perspectives. For example, the archivists used the term "electronic record" when describing the object for preservation. This term is usually defined as a digital document (Eastwood, 2004). However, scientists used the term "data," which in the scientific and technical community are usually defined as "measurements, values calculated therefrom, and observation or facts that can be represented by numbers, tables, graphs, models, text, or symbols that are used as a basis for reasoning or further calculation" (NRC, 1996). These two terms should not be used interchangeably.

Given these difficulties, an impromptu panel was convened with Jürgen Knobloch, Kevin Schürer, John Faundeen, and Terry Eastwood respectively providing the perspectives from the physical, social, and Earth and environmental sciences and from the archival community. This panel provided the speakers an opportunity to process the information from the case studies and subsequent discussions and recast the issues in a different light. The panel recommended development of an acquisition policy in all digital archives as a way of ensuring that data accepted into an archive comply with the organizations' overall mission. Acquisition policies enhance the transparency of the appraisal process and provide solid reasons for the rejection of data from the archive as well as the eventual disposal of data. The panel suggested that an acquisition committee be established for each digital archive to provide a suitable level of objectivity and accountability. The general consensus of the panel was that the appraisal process should be collaborative, involving scientists, archivists, and data managers and that appraisal should take place as early in the life cycle of the digital data as possible.

4 WORKSHOP CONCLUSIONS

The workshop proved to be extremely successful in enabling discussions between the scientific and archival communities. The workshop also highlighted some conceptual challenges to effective collaboration between the diverse communities. These challenges include the economic sustainability of data preservation activities; varying archival practices among researchers, institutions, and disciplines; and the issue of semantics. However, the workshop succeeded in its main goal of bringing together different communities of practice and different disciplines to openly discuss these issues across disciplinary and national boundaries.

There was strong agreement that data are the basis of the scientific endeavor, and that scientific data often have more than one life as scientific ideas advance and new concepts emerge. The curation and reuse of existing digital data also maximizes initial investments. Thus it is essential to properly preserve these data. Workshop participants noted that proper preservation strategies include ensuring that the data are regularly appraised; the related metadata are preserved with the data; and awareness is raised among funding agencies of the importance of digital preservation. In addition, when considering long-term retention of digital scientific resources, it is critical to consider long-term access to these resources.

There are varying levels of archival practice in use among the scientific disciplines based on the various needs and practices of the different communities. The space sciences, for example, have a long history of using selection and appraisal guidelines and have been active in establishing metadata and software standards within their discipline. But in other disciplines, the efforts have been less coordinated and sustained. More effective collaboration between the various scientific communities and archivists is needed to harmonize efforts.

The costs associated with collecting and storing data can vary depending on discipline. For example, the space and Earth observation sciences often use sophisticated and expensive equipment to capture huge amounts of data daily (often measuring in the terabytes) whereas digital data are generated on a smaller scale in the social sciences and many laboratory sciences through different techniques. The costs related to data migration or the storage media, for example, vary as well. The selection, appraisal, and retention needs of the various disciplines will reflect this diversity.

The need to establish a basic appraisal framework for digital scientific data was universally agreed. The specific criteria and appropriate timing, however, are less easily defined.

As previously stated, the terminology and semantics of archiving posed some confusion among many of the workshop participants. Many standard archiving terms—such as archive, appraisal, or record, for example—were used by different participants with different meanings depending on the discipline or area of practice. Progress in tackling the issues of selection, appraisal, and retention depends on addressing practical issues, such as semantics. This will be an important step in developing channels of interdisciplinary communication and collaboration in this area.

This workshop was an important step in the dialogue between scientists, archivists, data managers, and other information specialists in the area of data curation and preservation. The workshop illustrated areas where each can

learn from the others in establishing common frameworks and guidelines that will enable the effective selection, appraisal, and long-term retention of digital scientific data.

5 RELATED ACTIVITIES AND NEXT STEPS

In order to capitalize on the discussions at the workshop, CODATA, through its Task Group on Archiving and Preserving Scientific and Technical Data in Developing Countries, focused on issues related to archiving at the 19th International CODATA Conference—The Information Society: New Horizons for Science, which was held on 7-10 November in Berlin.. Plenary presentations on digital libraries and data archiving and on the European Space Agency's long-term archiving and access solutions were followed by a session focused on the InterPARES project. The conference also included a general session with contributed papers on data archiving.

A series of CODATA workshops have addressed these issues in more depth, including a Workshop on Archiving Scientific and Technical Data convened at the National Research Foundation in Pretoria, South Africa in May 2002 (CODATA, 2002); the Inter-American Workshop on Access to Environmental Data, which was held in Campinas, Brazil (Canhos, et al., 2004); and the international Workshop on Strategies for the Preservation of and Open Access to Digital Scientific Data, which was held in Beijing, China in June 2004. Another workshop focusing on permanent access to scientific information in southern Africa with a focus on health and environmental information for sustainable development is being planned by CODATA and its U.S. and South African National Committees in September 2005. This series of workshops continues to highlight the importance of permanent access to scientific information resources, and to examine the policy and legal, management and technical, and institutional and economic issues that impact the archiving of digital scientific data. CODATA is also working with the International Council for Scientific and Technical Information (ICSTI) to develop an Internet portal for online resources related to permanent access to scientific data and information. The primary task of this portal is to greatly raise the visibility of the issues related to archiving and preserving access to scientific and technical data and information internationally; the need for such an online resource was evident from the discussions at the workshop. Finally, ERPANET continues to address and raise awareness of the issues related to preservation of various resources, not only scientific data, through its continuing series of workshops and seminars.

It is clear through this workshop and the related activities undertaken by CODATA, ERPANET, the U.S. National Academies, and other organizations (e.g., ICSTI and the U.S. National Science Board) that permanent access to digital resources should be at the forefront of the global scientific agenda. The ever-increasing volumes of data collected are the lifeblood of the scientific endeavor and provide the foundation for future research in areas not yet identified. However, not all of the data collected and generated can be preserved indefinitely. Effective selection and retention guidelines along with defined appraisal policies are required to ensure that digital scientific assets are widely available for future generations.

The importance of these issues was also highlighted in the Plan of Action developed at the first phase of the World Summit on the Information Society (WSIS) in Geneva in December 2003; the idea of promoting "the long-term systematic and efficient collection, dissemination, and preservation of essential scientific digital data" was raised explicitly (WSIS 2004). It is recognized that the long-term availability of and access to scientific information resources will help accelerate progress of the issues related to bridging the divide—economic, technical, and cultural—between developed and developing countries. These issues will be highlighted at the second phase of WSIS, which will take place in Tunis in November 2005. WSIS II provides an opportunity to emphasize the importance of permanent access to digital scientific resources and to highlight examples of successful archiving practices in many different disciplines, such as those presented during the EPRANET/CODATA workshop.

6 ACKNOWLEDGEMENTS

The authors would like to thank Paul Uhlir, director of the U.S. National Committee for CODATA and of the Office of International Scientific and Technical Information Programs of the U.S. National Academies, and Gail Hodge of Information International Associates for their assistance in planning the workshop and identifying potential speakers, as well as their insights into the issues that impact permanent access to digital scientific resources. We

would also like to thank Fernando Campos of the Biblioteca Nacional in Lisbon for providing not only a great venue for the workshop, but also the library perspective on these issues. Finally, we would like to CODATA and the U.S. National Institute for Standards and Technology for providing additional funding for this workshop.

7 REFERENCES

Anderson, W.L. (2004), Some Challenges and Issues in Managing, and Preserving Access to, Long-lived Collections of Scientific and Technical Data. *Data Science Journal 3*(December 30), 202 -213. Retrieved December 30, 2004 from the World Wide Web: http://www.datasciencejournal.org.

Canhos, D.L, Uhlir, P.F., & Esanu, J.M., eds. (2004), *Access to Environmental Data: Summary of an Inter-American Workshop*, Paris: CODATA. Retrieved December 13, 2004 from the *Reference Center on Environmental Information* Web site: http://www.cria.org.br/eventos/iaed/WorkshopFinalReport.pdf.

Consultative Committee for Space Data Systems (2002), Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-B-1, Blue Book, January. Retrieved December 13, 2004 from the NASA/Science Office of Standards and Technology Home Page:
http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf.

Eastwood, T. (2004), Appraising Digital Records for Long-term Preservation. *Data Science Journal 3*(December 30), 214 - 220. Retrieved December 30, 2004 from the World Wide Web: http://www.datasciencejournal.org.

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

Fusco, L. & van Bemmelen, J. (2004), Earth Observations Archives in Digital Library and GRID Infrastructures. *Data Science Journal* (under review).

Gutmann, M. Schürer, K., Donakowski, D. & Beedham, H. (2004), The Selection, Appraisal, and Retention of Social Science Data. *Data Science Journal 3*(December 30), 221- 233. Retrieved December 30, 2004 from the World Wide Web: http://www.datasciencejournal.org.

Hodge, G. & Frangakis, E. (2004), *Digital Preservation and Permanent Access to Scientific Information: The State of the Art and the State of the Practice*, a report sponsored by the International Council of Scientific and Technical Information and CENDI, CENDI—2004-3: Rev 05/04, Washington, DC: CENDI.

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

NRC (1995b), *Study on the Long-Term Retention of Selected Scientific and Technical Records of the Federal Government: Working Papers*, Washington, DC: National Academy Press.

NRC (1996), *Bits of Power: Issues in the Global Access to Scientific Data*, Washington, DC: National Academy Press.

World Summit on the Information Society (2004), *Plan of Action Plan* (Item 22b), WSIS-03/GENEVA/DOC/5-E, Geneva: International Telecommunications Union. Retrieved from the World Summit on the Information Society Web site: http://www.itu.int/dms_pub/itu-s/md/03/wsis/doc/S03-WSIS-DOC-0005!!MSW-E.doc.