Preserving Scientific Information on the Physical Universe

by Kenneth Thibodeau [1]

Background

Over the past fifteen years, there have been several collaborative studies of the archival value of scientific records in the United States. Between 1978 and 1983, representatives of the History of Science Society, the Society of American Archivists, the Society for the History of Technology, and the Association of Records Managers and Administrators worked together on the Joint Committee for the Archives of Science and Technology (JCAST), assessing the state of documentation on research and development, the dissemination of ideas, technology transfer, and professional education in science and technology[2]. A self-acknowledged sequel to the JCAST project, was the collaboration of Joan Haas, Helen Samuels and Barbara Simmons at the Massachusetts Institute of Technology (MIT) which resulted in the publication of *Appraising the records of modern science and technology: a guide* in 1985[3]. Beginning in 1989, the Center for the History of Physics of the American Institute of Physics (AIP) inaugurated a long term study of fields of physics and related sciences where multi-institutional collaborations are prominent[4].

While the three projects were undertaken in different organizational contexts, with varying focus and goals, they share a primary concern with the records of research and development activities as resources for historical research. In these endeavors, the potential long-term value of the records of science and technology for further research in these fields themselves has been recognized, but not explored in depth. Generally, consideration of enduring value for science has focused on the data records generated in research and development activities. In fact, JCAST declared, "The first consideration regarding retention of data must be the needs scientists themselves have for these records." Both the JCAST and MIT publications recognized that the actual retention of data for science is usually in the hands of the scientists themselves or of specialized scientific data centers, although archivists may occasionally face the necessity of deciding on the retention of scientific data for scientific purposes[5]. The AIP project took into account the future needs of physicists. It identified categories of records which should be retained by scientific laboratories and science libraries, but did not articulate criteria for identifying records with continuing value for science.

Both the JCAST report and the MIT *Guide* drew attention to the distinction between observational and experimental data, and suggested that long-term value is more often found in observational data than in experimental results. The argument which supports this generalization is that experiments are repeatable, while observational data of ten relate to unique or rare events or sequences of events. With respect to scientific data, the AIP has concluded that, in high energy physics at least, very little data should be preserved for long periods, and then for purposes of exhibit, rather than scientific research[6].

Scientific Records in the National Archives of the United States

The National Archives and Records Administration (NARA) of the United States has been involved in appraising and preserving the records of science and technology since its inception. The types of scientific records in the National Archives include project case files, technical reports, laboratory notebooks, drawings and specifications, maps, charts, graphs, aerial photography, motion pictures, sound recordings, and digital data files. The subjects reflect the broad range of scientific and technical activities in which the Government of the United States has engaged, including astronomical, geological and meteorological observations, land and stream classifications, patents, weights and measures, nuclear energy, mineral deposits, weapons systems, aircraft and spacecraft, epidemiology and biometry, entomology, and many other subjects.

It is probably impossible to categorize in general terms the reasons why such scientific and technical records have been accessioned into the National Archives. However, among other factors, NARA has been concerned with the continuing value of these records for science and technology themselves[7]. Concern with the long term scientific value of records derives from a crucial provision of U.S. law, the definition of a federal records. This definition, articulated in title 44 of the United States Code, states that federal records are preserved or appropriate for preservation either as evidence or "because of the informational value of the data in them."

As JCAST recognized, the primary informational value of scientific data is for scientists. In appraising the records of science, then, NARA has an obligation to consider their continuing value to scientists. The most obvious means of exploring this value is to consult with scientists, as was the case in all three f the earlier projects I mentioned. In fact, the reports of all three projects recommended this practice as standard. In the case of research funded by the U.S. Govern-

ment, there are especially important reasons to include the perspective of specialists: (1) the functions of Federal agencies in sponsoring research, (2) the role of researchers outside of the Government in the life-cycle of the data, and (3) the knowledge of the provenance and life-cycle of the data that these scientists have.

(1) Agency Functions: Commonly, Federal records are documents used by agencies in the exercise of mission functions. The Internal Revenue Service, for example, collects tax returns as instruments critical to its function of collecting taxes. Often, however, the science agencies do not use the data that result from the research they sponsor; rather, their function is to sponsor research. In many cases where the funding agency requires the researchers to deliver the resultant data to the Government, the primary, if not sole, purpose for this is to make the data available to yet other researchers outside of the Government, in many cases for long periods of time.

(2) Role of outside researchers: A large proportion of the scientific data generated by the Federal Government is created as a result of the initiatives of investigators outside of the Government. Outside scientists of ten originate the research proposals and are responsible for the organization and conduct of the research. Through peer review of proposals, other members of the research community have a decisive role in determining what data are collected and how they are collected and organized. Even in cases where research is conducted by scientists employed by an agency, it is not uncommon to have government laboratories reviewed by peer groups composed of outside scientists.

(3) The Life-Cycle of Scientific Data: In many cases, scientific data are in the custody of researchers outside of the government for a major part of their life-cycle. These researchers collect the data, calibrate and refine it, and analyze it, definitively shaping the record. In many cases, the records are not transferred to government custody until the records cease to be active. In other cases, the records - although owned by the government - remain in the custody of outside researchers throughout their life-cycle.

The relevant framework of appraising scientific data sets, thus, is not defined by the business activities or the need for corporate memory of the sponsoring agency, but by the research community. Seeking the input of scientists in the appraisal of the data recognizes that the roles and the actions of academic researchers are at least as important as the functions of the agency that funded the research or launched the satellite.

Since 1990, NARA has sponsored two important efforts to obtain the advice of subject matter experts on the retention of data. The first study, undertaken by the National Academy of Public Administration (NAPA), focused on major federal databases used in support of mission activities. This study had a twofold purpose: first, to identify these databases and, second, to recommend what data should be preserved in the National Archives[8].

The NAPA project included the review of some scientific and technical data, notably in the areas of natural resources, the environment, and health. However, large collections of scientific data were intentionally excluded from consideration in the NAPA project because NARA felt that such a large and complex area as "big science" merited separate attention. Scientific data are the focus of the second recent study sponsored by NARA. This project, inaugurated in 1992, was undertaken by the National Academy of Sciences' National Research council (NRC).

The NRC study was divided into five subject areas: (1) space sciences; (2) physical, chemical and materials sciences; (3) earth sciences, (4) atmospheric sciences, and (5) ocean sciences. Panels of experts were organized to develop recommendations for the preservation of records in each of these areas of research. A steering committee oversaw the work of the panels and formulated generalized recommendations and criteria for the retention of scientific records based on the work of the panels.

The NRC project gave the National Archives the opportunity to interact with the records creators and to engage them in a dialogue on the long term value of the data for secondary use. It was hoped that the NRC project would serve to raise the addressing the complete potential life-cycle of the records during the development and performance of research projects. The final report:

Preserving scientific data on our physical universe. A new strategy for archiving the Nation's scientific information resources<end underline>, moves towards fulfilling that hope[9]. The report, which was completed in March of this year, makes several sweeping recommendations which can be grouped under the twin headings of retention and responsibilities.

Retention

Recommendation: "As a general rule, all observational data that are nonredundant, useful, and documented well enough for most primary uses should be permanently maintained. Laboratory data sets are candidates for long-term preservation if there is no realistic chance of repeating the experiment, or if the cost and intellectual effort required to collect and validate the data were so great that long-term retention is clearly justified[10]."

The report makes two procedural suggestions related to appraisal. The first is that each program or project should have a data management plan established at the origin and governing the entire life-cycle of the data: "Planning activities at the point of data origin must include long-term data management and archiving.[11]" The second is that

appraisal itself is a multifaceted, continuing process: "Formal appraisals should be kept to minimum, appraisals should be performed according to the data management plan established for each project1[2]." The first step in this process would be an interdisciplinary consensus regarding broad classes of data:

"All stakeholders... should be represented in the broad, overarching decisions regarding each class of data[13]."

"Scientists, information technology professionals, data managers, librarians, and archivists must unify their expertise in the establishment of a coherent strategy for end-to-end data and information management[14]."

Principal investigators and program managers would then appraise the long term value of individual data sets:

"The appraisal of individual data sets ... should be seen as an ongoing, informal process associated with the active research use of the data, and therefore should be performed by the most knowledgeable about the particular data.... In some cases, they may need to involve an archivist or information resources manager to help with issues of long-term retention.[15]"

Finally, the judgements of the primary users would be supplemented with some sort of peer review. The purpose of this review would not be to appraise the data, but to determine if they are as purported and if they are adequately documented. Several options for peer review are identified, ranging from "a formal peer review to certify integrity and completeness," to "documented evidence of the use of the data set in publications in peer-reviewed journals," or evidence from expert users that the data set "is as described in the documentation."

Responsibilities

Recommendation: "As a general principle, data collected by an agency should remain with that agency indefinitely.[16]" "Collection" in this context refer to data collected under agency sponsorship, through contracts and grants, as well as data created within the agency. The proposal is for scientific data to be held for the long term in distributed archives, typically in discipline oriented data centers, such as the National Space Science Data Center in NASA and the National Geophysical and Solar-Terrestrial Data Center in NOAA. The novelty in the approach advocated by the report is a proposal for coordination of the activities of such distributed archives "The federal government should create a National Scientific Information Resource Federation — an evolutionary and collaborative network of scientific and technical data centers and archives....[17]"

This recommendation to establish the federation suggests action by the Clinton Administration's Information Infrastructure Task Force, the National Science and Technology Council, and/or the Office of Management and Budget to initiate the federation. The report recommends that either an independent commission or an agency with an established mandate in both the physical sciences and information technology, such as the National Science Foundation, provide executive support for the federation. However, the organization is to be true federation; i.e., a collaboration among equals, not a top-down activity directed by the government.

Hypothetical Profile of the Life Cycle of a Data Set

{In an effort to understand the recommendations of the NRC report, I have constructed the following hypothetical profile of the life-cycle of a data set in accordance with these recommendations. This profile has been reviewed by both the project director and the chair of the steering committee. They both agreed that it accurately reflects the intent of the recommendations.}

When a research project is proposed, the principle investigators, with appropriate collaboration by the program manger in the funding agency, draft a data management plan. In evaluating the long term value of the data, the principals consult with NARA to learn of any recommendations that may have been made by diverse groups of stakeholders about the enduring value of data in the relevant class. The plan is completed no later than project initiation. Following generally accepted criteria, the plan assumes that most observational data produced by the project will be subject to long term retention. Laboratory or engineering data sets, however, are candidates for long-term preservation only if there is no realistic chance of repeating the work, of if the cost and intellectual effort required to collect and validate the data would be so great that long term retention is clearly justified. Other experimental laboratory data or engineering data generally will not need to be retained after completion of the project. The provisions of the plan conform to well established standards for information technology and documentation, as endorsed by the NSIR Federation.

NARA maintains liaison with the sponsoring science agency and, periodically or when necessary, consults with the agency and the investigators to remind them of their responsibilities for long term retention, management, and access.

During the conduct of the research, the investigators and managers will occasionally and informally consider whether the data actually collected merits retention and also whether the history of the project gives rise to any special requirements for documentation.

At the end of the project, the data is deposited in a data center or field archives, as specified in the data management plan. This repository is designated, or operated, by the lead agency in the subject area, where staff have appropriate expertise and close ties with the relevant researcher community. The repository has mechanisms for access to the data

by individuals beyond the primary users.

For the data set to be accepted into the data center of distributed archives, it must undergo some form of peer review to ensure that it adequately meets the standards of uniqueness and accessibility. Along with the data, metadata essential for others to use the data is transferred to the repository. When any required services related to retention or access to the data are available elsewhere in the Federation at lower costs than would be incurred by the organization with primary responsibility, those alternative services are used when feasible.

From the time the data set becomes available for use outside of the project, it is identified in a hierarchical information locator system. The data should be available for remote access and/or file transfer, ideally as an extension of the locator system.

NARA is informed of the existence and location of the data. NARA monitors the preservation and accessibility of the data over time and, acting in an advisory capacity, helps the custodians with any problems in these areas. If the custodians can no longer meet the needs of the user community, or if the data is no longer in regular use, the data should be considered for transfer to some other federal science agency or, as a last resort, to the National Archives.

The NRC report has been received too recently for NARA to have been able to take a position concerning its recommendations. However, as an individual archivist, I would like to offer some observations. They are entirely my own; they do not represent the position of the National Archives not, as far as I know, of any other individual in the National Archives.

Preserving scientific data on our physical universe purports to offer "a new strategy for archiving the Nation's scientific information resources." There are certainly new elements in this strategy. One is the articulation of the general principle that data collected in the observational sciences should be preserved permanently. Another is the creation of the NSIR Federation, conceived as a collaboration facilitated, but not directed, by the Federal Government. A major innovation entailed by the recommendations is that the scientific community would have to recognize data management, data retention and data access as valuable activities by scientists, and would have to adjust the culture of science to include rewards for these activities, on a par with the publication of research results.

These things would be significant changes, but they would be changes in the scientific community. From an archival perspective, the impact would be different. As the report recognizes, very little scientific data has been deposited in the National Archives, and there is no grounds for expecting this to change. The assertion that entities within the scientific community should be responsible for the long-term retention of scientific data sets and for access to them is not only consonant with the status quo, but also consistent with archival concepts, as articulated by the three projects I described at the start of this paper. Furthermore, the National Archives does not play a forceful role in the management, retention, or access to scientific data. The report, in fact, argues that such a role would be counter-productive. It suggests that the appropriate role for NARA would be that of consultant and collaborator on archival preservation and access issues. On empirical grounds, one might say that this is role that NARA has been playing in the domain of scientific data. Thus, one might argue that, from an archival perspective, the recommendations of the NRC report reduce, by and large, to a confirmation of the status quo.

References

1 The views stated in this paper are those of the author and do not represent the position of the National Archives and Redords administration. (Paper presented at IASSIST95 May 1995 Quebec City, Quebec, Canada.)

2 Clark A. Elliott, editor. Understanding progress as process. Documentation of the history of post-war science and technology in the United States. Final Report of the Joint Committee for the Archives of Science and Technology. Chicago. Society of American Archivists, 1983.

3 Joan K. Haas, Helen Willa Samuels and Barbara Trippel Simmons. Appraising the records of modern science and technology: a guide. Cambridge, MA. Massachusetts Institute of Technology, 1985.

4 Joan Warnow-Blewett and Spencer Weart. AIP Study of Multi-Institutional Collaborations, Phase I: High-Energy Physics. Report No. 1: Summary of Project Activities and Findings. Project Recommendations. New York: American Institute of Physics, 1992.

5 Elliott, p. 33-34. Haas et al., pp. 60-61.

6 Joan Warnow-Blewett, Lynn Maloney, and Roxanne Nilan. AIP Study of Multi-Institutional Collaborations, Phase I: High-Energy Physics. Report No. 2: Documenting Collaborations in High-Energy Physics. New York: American Institute of Physics, 1992. Pp. 75-76, 89-91.

7 Trudy Huskamp Peterson. Presentation on the National Archives and the Records of Science for the National Academy of Science/National Research Council STudy on the Long Term Retention of Scientific and Technical Records. Plenary Session, July 7, 1993.

8 National Academy of Public Administration. The Archives of the Future: Archival Strategies for the Treatment of Electronic Databases. A report for the National Archives and Records Administration. Washington, D.C. National Academy of Public Administration. 1991.

9 National Research Council. Preserving scientific data on our physical universe. A new strategy for archiving the Nation's scientific information resources. Commission on Physical Sciences, Mathematics, and Applications. Washington D.C. National Academy Press. 1995.

10 Ibid. p. 4. The report argues that technological developments make it possible to save everything and to provide access to it. However, the report recognizes that data management activities in general, not just preservation, are chronically underfunded and that they are at best a secondary concern in scientific culture. it does not adequately address how these difficulties can be overcome.

11 Ibid. p. 50

12 Ibid. p. 40

13 Ibid.

14 Ibid. p. 50

15 Ibid. p. 4

16 Ibid. p. 56

17 Ibid. p. 51