Proceedings of the 6th Annual
Federal Depository Library Conference

April 14-17, 1997

Key Considerations in the Long-Term Retention of Digital Information

Paul F. Uhlir, National Academy of Sciences/National Research Council
Washington, DC

Introduction

My presentation will provide an overview of the key elements that need to be considered in developing and implementing a plan for the long-term retention of digital information. Some of these elements, of course, are the same as for paper documents. Others are similar or analogous, while others are unique.

I have organized the discussion according to four topical areas. The first is the development of retention and purging criteria and processes. The second involves technical document management issues. The third focuses on considerations of long-term user access and retrieval of public Government information. And the final area encompasses institutional roles and responsibilities, including financial aspects.

Because the preservation plan that is being developed by the U.S. Department of Agriculture USDA) is still very much in a preliminary draft form, I will not discuss the specific actions and recommendations in that project. Rather, I will provide a summary of the major elements that need to be considered in the development of any new program for the preservation of public digital documents, and will draw on examples from the USDA and other Government agency activities to illustrate my points.

Development of Retention and Purging Criteria and Processes

We begin with the questions what? and why? Certainly the threshold question in developing a preservation plan for digital documents is defining the universe of information that will be preserved, and providing a well-understood rationale for its preservation. The broad reasons for saving digital information are the same as for paper documents: it is their historical significance, the intellectual achievement they represent, and their potential social, cultural, or economic value. Of course, there are more specific reasons that are dependent on institutional context, the nature of the information product, and the anticipated end users. You all are familiar with the criteria for published literature--books, reports, articles, and the

like--so let me focus briefly instead on retention criteria for electronic scientific data, which may be less familiar to you.

For example, there are significant differences in the need for long-term retention of experimental and observational data. Data from laboratory experiments in such areas as chemistry or materials science typically are reproducible, since the validity of any experiment depends on whether its results can be reproduced independently by other researchers. Thus, except perhaps in very large and expensive experiments, there is little need to keep the original, primary data once the results have been published and the experiment independently verified. Instead, the data that are of greatest long-term value are compilations of highly evaluated data that can be used repeatedly by other researchers as reliable reference points.

In the observational sciences, however, such as astronomy or environmental sciences, the research itself is dependent on the data themselves, which can be processed and interpreted at different levels of complexity.[1] Typically, each level of processing adds value to the original, raw data by summarizing the original product, synthesizing a new data product, or providing some interpretation of the original data. The processing of data leads to an inherent paradox that might not be readily apparent. The original unprocessed, or minimally processed, data are usually the most difficult to understand or use by anyone other than the expert primary user. With every successive level of processing or interpretation, the data tend to become more understandable and better documented for the nonexpert, general user. One might therefore assume that it is the most highly processed or evaluated observational data that have the greatest value for long-term preservation, as in the case of the experimental laboratory sciences, because such data are more easily understood by a broad spectrum of potential end users. In fact, just the opposite is usually the case for observational data, because it is only with the original, unprocessed data that it will be possible to recreate all other levels of processed data and data products. Thus, while laboratory scientists value most highly the evaluated data compilations, researchers in the observational sciences typically want all reliable original observations to be saved, because most observations are unique and non-reproducible, and the original data can be used repeatedly and in different ways in future research.

This one example highlights not only some of the differences in developing preservation criteria for different types of information products, but also between paper and digital products. Digitally generated observational data pose a significant challenge in volume and in proper documentation and preparation, that are not inherent in data recorded on paper or even in most other digital information products. The situation is becoming increasingly complicated by the generation of hybrid, multimedia information that may include text, numerical data, animation, sound and video all in one product, and that furthermore may include self-executing programs that will automatically update or revise that product over time.

The development of retention criteria for digital information is thus more complex and less straightforward than for paper publications, although some of the basic considerations will remain the same in both types of media.

Issues that might be considered in the long-term retention and life-cycle management of digital information products include:

- Legal restrictions
- Cost
- Documentation/metadata
- Quality control/quality assurance
- Provenance/authority/authentication, and
- Other context-specific issues

Legal restrictions include national security, privacy, and various intellectual property rights, similar to the paper paradigm. A potential significant difference may arise with regard to adequately sorting out intellectual property rights in hybrid digital information products which might integrate dozens or even hundreds of sources.

The costs arise from the labor required to evaluate and subsequently manage the digital information, as well from the technological infrastructure, as discussed later.

Documentation, also referred to as metadata, is especially important for scientific data and other esoteric information products that require some ancillary explanation to facilitate their use. Digital data that are so lacking in documentation that even an expert in the same discipline is unable to understand them are obvious candidates for the trash bin, unless their originator can be found and persuaded to make them intelligible. The physical separation of explanatory documentation from the data themselves should be avoided.

Quality control and assurance is another retention criterion that needs to be considered in whether to preserve an information product. One method appropriate for both paper and digital information is peer review. In contrast to paper products, however, electronic information may become corrupted due to technical deterioration or anomaly, or through the intentional or accidental introduction of errors as a result of use. What makes the quality control even more difficult for electronic information is that sometimes the problems, such as viruses, are not readily apparent and may lie dormant until some future point.

Provenance and authentication have parallel importance for both paper and digital forms, but pose more problems in the electronic context. As in the case of quality control, the original and authentic version may be difficult to ascertain, and fraudulent or illegal modifications can be made that are difficult or impossible to detect.

Issues that might be considered in purging or deeper archiving of documents include:

- Age of document
- Physical condition

- Cost
- Use history, and again
- Other context-specific issues

The implementation procedures for retaining and purging documents are also likely to differ from the paper model. Digital information products are more voluminous, varied, and complex than their paper counterparts, and therefore require a broader range of expertise for their proper evaluation and become more labor intensive and costly to screen.

Technical Document Management Issues

A detailed discussion of the hardware and software requirements for long-term retention of digital publications is beyond the scope of this presentation, and of course in any event is largely determined by the technological infrastructure that is already in place. Certainly one bit of good advice is to spend the time to do thorough background research to find out what are the technical "best practices" for long-term retention that can be derived from the experiences of other similar programs. Choosing the right technologies is a decision that should not be made lightly and there are many well-known horror stories. The acquisition or upgrading of the necessary information technologies is likely to be the single largest cost associated with the preservation of digital information, although many of those costs can be shared and integrated with the institution's overall information technology requirements. Indeed, it is essential that the preservation function--or, more accurately, the information life-cycle management considerations--be expressly included in the planning and procurement of information technologies for the entire institution.

There are several technical requirements or functions that are especially important to long-term preservation that should be mentioned here. Acceptable document formats and media for long-term retention need to be chosen in conjunction with the institution's information creators and information technologists. Costs can be reduced if the formats for both creating and preserving the information are the same, and interoperable technologies are used.

The transfer of all digital information products from old media to new media on a regularly scheduled basis is essential. There have been many instances of old tapes deteriorating and becoming unreadable, or of lacking equipment that can read the information stored on obsolete media. This is a non-trivial problem, as I'm sure you are all aware. How many different word processing programs have you used in the past 10 years just in the course of your daily office work, and how much information do you still have on 5 1/4 inch diskettes that you have not migrated onto 3 1/4 inch diskettes? Institutions that have several decades of large-scale experience in this, such as the NOAA National Data Centers, currently transfer the information on 10-year cycles. Related to this requirement is the need for providing physically separate back-up facilities and environmentally controlled storage conditions for both the primary and the back-up locations.

Finally, system security protocols must be established that effectively balance the need for open systems that allow for easy user access against the need for security against accidental or intentional destruction of either the technology or the information itself.

Long-term User Access to and Retrieval of Public Digital Information

The next important issue area involves the planning and implementation of long-term user access to and retrieval of public digital information. As in the other topical areas there is a lot of overlap between paper and digital information, particularly with regard to legal and policy requirements that you all know better than I. I will focus instead on some of the key differences.

Undoubtedly, the most significant difference is the vastly expanded universe of users who are now able to access and retrieve information remotely. This is a true shift in paradigm from the paper model. Although it is true that only a small percent of the population has ready desk-top access to on-line information, that number will grow inexorably, and in fact, almost everyone can now go to a library or other Internet source and establish remote access. Thus the focus of planning for providing access to the information and related user services needs to shift from perhaps dozens of on-site clients using the stacks to thousands of remote clients on a daily basis.

The following guidelines are useful to adopt, consistent with the need to maintain a customer orientation:

- Provide equitable access and retrieval services to all potential users;
- Minimize technical, regulatory, and cost barriers to access and retrieval;
- Make the information as easy to find and use as possible, while protecting confidential or proprietary information, and
- Establish a means for users to provide input and for your organization to respond to that input.

Starting with this last guideline first, one of the most difficult tasks is to be responsive to the vastly enlarged body of end users in the networked environment, particularly when you first provide on-line access. One mistake is to assume that the distribution of categories of end users will remain the same as with those who physically visited your facility. While you can be reasonably certain that your on-site visitors have a very specific objective and information need in coming to your facility, your on-line visitors are much more likely to be more diffuse and less focused than individuals who have to make a substantial commitment in time and perhaps expense in making their trip to you. Also, the demographics will change, with an obvious emphasis on those user groups who have ready access to the Internet. Although it may be difficult to anticipate at the outset what the primary on-line user requirements and interests may be, the good news is that you can easily track the types of users electronically and develop a customer distribution profile quite quickly.

One absolute necessity, whether the information products are all available on-line or not, is a comprehensive on-line directory or catalog, preferably in some multi-level format that will bring the user from the general to the specific. This service, although time consuming and expensive to develop, is invaluable to fully realizing the information transfer potential.

Another important piece of advice is to use a proven professional Web designer, rather than an in-house technologist. A seasoned expert will be sure to cover all the essential features--and many you may not even think of by yourself--in working with you to optimize your Web site for your needs.

One feature that ought to be included is a means for customers to provide feedback and useful suggestions. In addition, it may be important to appoint an advisory body of knowledgeable representatives from major end user groups. Such a formal advisory mechanism can be helpful not only with successfully maintaining a customer orientation, but in providing advice on major decision points such as the development of retention and purging criteria.

Institutional Responsibilities and Relationships

Finally, there are the various organizational roles, responsibilities, and relationships that need to be worked out. Again, many of these will be similar, or at least build upon, the organizational models used in the paper paradigm. Within a large Federal department or agency, there are many internal institutional links that need to be established and responsibilities agreed upon. Under current Federal law, the principal information policy and planning function resides in the Office of the Chief Information Officer. However, the lead entity for developing and implementing a preservation plan within each Federal organization will likely vary. In the Department of Agriculture, for example, the logical focal point is the National Agricultural Library. In addition, the successful implementation of a preservation plan is dependent on the active participation of the information creators throughout the entire institution and even outside it, to the extent that the institution preserves information products that are created by contractors or grantees. All of these parties need to be involved in the planning process and claim some ownership to its results in order to make it work.

Of course, there are some essential external responsibilities and relationships that need to be considered. Governmental organizations outside a Federal department such as the USDA that have an important role in the preservation and dissemination of public digital information include the National Archives and Records Administration, the Government Printing Office and its Federal Depository Library Program, the National Technical Information Service, and various other Federal and State Government institutions. Among the nongovernmental entities that have an important function vis-a-vis the Department of Agriculture are the land-grant university libraries, and the aforementioned user groups and contractors and grantees.

A key difference between the paper and digital organizational considerations is that the electronic networked environment allows for a more highly distributed system with specialized functions, without having to physically locate all documents that need to be preserved in a centralized repository. Indeed, the physical location of digital information can

be completely transparent to the end user, allowing for more flexible and responsive organizational structures that are optimized for function and cost. The challenge for the Department of Agriculture now is to create a management structure that will take advantage of these distributed attributes both internally and externally, while maintaining just enough authority and control to realize all its important objectives and requirements.

This brings me to the last issue, the unavoidable financial aspects. The good news is that a carefully designed and implemented preservation plan that takes advantage of broadly distributed functions can minimize the need for additional funding and spread the costs across a large number of organizations. The bad news is that it will not come without a price, and that new funds will have to be found in an era of reduced Federal funding. Because of the public good nature of this activity, the preferred option would be to seek an augmentation to the annual appropriations. In the event that direct appropriations or reprogramming of funds cannot cover the full costs, it may be necessary to charge user fees for certain products or services. In that case, some level of basic access should be kept free if at all possible.

---

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