Ensuring the Preservation and Use of Electronic Records

Shigeo Sugimoto, Professor

Research Center for Knowledge Communities
Graduate School of Library, Information and Media Studies,
University of Tsukuba

1. Introduction

At present, almost all document creation is done electronically, with word processor functions, etc. While many documents are printed on paper for the purpose of distribution and publication, many are also distributed electronically, via e-mail and the Internet. In many cases, e-mails and web documents are printed on paper only for the purpose of reading; these hard copies are discarded after reading and the documents are preserved on electronic media. I take this as an indication of the spread of electronic documents and paperless arrangements, which does not mean less consumption of paper but less reliance on paper as a media to record information. At the same time, however, we face the prospect of eventually not being able to read electronic documents created in the past.

There is a well-known awareness that the preservation of electronic documents, and particularly those that were originally created, distributed, and utilized electronically, is a major issue for contemporary society, with its expanding amount of network information. It is also well known that this issue will be a difficult one to resolve completely. The main reason for this difficulty is the gap between the length of time for which preservation is required and the service life of the hardware and software required for use of electronic documents. In addition, there are difficulties differing from those of paper documents in the aspect of collection of documents for preservation.

Over the last few years, I have had opportunities to attend committees and groups to discuss and study preservation of digital resources - a committee considering collection of network publications in the Legal Deposit System Council of the National Diet Library, a committee researching the ideal of the management, transfer, and preservation of governmental records using electronic media in the Cabinet Office, and other venues for discussion on preservation of electronic documents. I was also given opportunities for involvement with long-term preservation of electronic materials in the context of research regarding digital

libraries and metadata. In this paper, I would like to set forth mainly my own understanding of the issues obtained through these activities. I must preface my account with a disclaimer to the effect that my observations are based more on my own cogitation than on solidly established facts.

2. Digital archive

The term "digital archive" is often used these days. In some cases, it is used in reference to the digitization and accumulation of real-world articles, e.g., the digitization and provision of cultural legacy. In others, it refers to the collection and accumulation of various electronic documents. In the former case, the digitization activities are not confined to books and manuscripts, but extend to historical sites themselves and intangible cultural properties such as performing arts and crafts as well as the artifacts of both art and natural science museums.(1) As this suggests, digital archive has various objectives. In this paper, I shall deal mainly with the collection, accumulation, and long-term preservation of digital resources, including those created by the digitization of original materials and those originally created in digital form.

There are also several types of electronic document archives. This list includes web archives devoted to the collection and accumulation of documents disclosed on the Internet, archives for the long-term preservation of electronic documents collected by libraries and equivalent institutions, and archives for the preservation of documents created by certain organizations in accordance with their rules, e.g., governmental records and company records. Similarly, the term "electronic documents" encompasses everything from simple documents composed of plain texts and charts created with word-processing and spreadsheet software to complex documents with hyperlinks, motion pictures and/or three-dimensional images. Electronic documents are also divided into two general categories: those contained on compact disks or other packages and used in that form (packaged documents) and those accumulated on servers and used in electronic networks (network documents). In this paper, I have decided not to take up the subject of preserving packages per se, but to view issues from the standpoint of preserving contents, and therefore shall not make any particular distinction between the two.

Digital archives connected to networks enable us to use various documents and materials at any time and place. Unlike paper and concrete materials, digital ones require hardware and software for their use. To keep the accumulated contents in a

usable condition over the long term consequently requires preservation not only of the materials themselves but also of the hardware and software for their use. Amid the fast-paced advances in the related technology, preservation encompassing hardware and software is no easy task. These circumstances are making it hard to achieve long-term preservation of electronic materials, and particularly "born-digital" ones created digitally from the start and premised on use in an electronic environment.

3. Components for construction of digital archives

In this section I would like to view digital archives from the perspectives of collection, accumulation and preservation, and organization and use.

3.1 Collection

The basic construction methods for digital archives can be classified as follows.

- Collection of materials through electronic networks
  - Collection of materials publicly available on the Internet and other networks

    There are two methods: selective collection based on contents and comprehensive collection in a designated scope.
  - Collection of materials reflecting the policy of the materials provider (e.g., organization-specific collection)
- Collection of electronic files offline (collection of packages)
- Collection by digitization of (real-world) objects

Collection of materials on networks is generally done automatically, using a software robot collecting network resources. The Uniform Resource Identifier (URI) is ordinarily used to identify materials in such collection. Because many network materials are updated, robots repeatedly collect from the same URL in such cases. Generally speaking, the updating and collection are not synchronous. As a result, we must construct schemes adapted to the purpose at hand in the case of collection of all contents in each updating. Even some materials available for public access on the Internet may include materials that do not accept access by collection robots or stored in databases that cannot cope with automatic collection. On the other hand, if the material provider and the archives cooperate in material collection, it would be possible to collect materials in accordance with a policy determined by both in advance. In addition, to heighten the credibility of the

archives themselves, we must also consider issues such as deposition of collected materials at other archives and adaptation to changes in the organization owning the archives.

3.2 Accumulation and preservation

The objectives of electronic document preservation cannot be attained merely by proper accumulation and preservation of the digital data contained in the form of bit string. For this reason, archives must accumulate and preserve the information required for use of the data along with the documents. The Open Archival Information System (OAIS) describes an archival reference model. Besides the original data, the system contains information needed for their reproduction, and preserves them in the form of a package also containing information required for preservation.

Regardless of the type of digital material, we cannot expect preservation to be perfect. Even if archives precisely retain information required for reproduction and preservation, they will not be able to reproduce the materials if the hardware and software required for reproduction go out of existence. Furthermore, in the case of materials including hyperlinks to external materials, it would not be easy to keep the link in effect. Problems of this nature may readily be imagined from electronic documents in our midst today, such as text created by word processor. Therefore, preservation of materials must be preceded by studies to determine the permissible degree of loss of the functions possessed by the originals.

3.3 Organization and use

Although they do not necessarily enable long-term preservation, metadata are indispensable for it. Metadata description elements have been proposed for preservation based on the aforementioned OAIS. In addition, standards have been established for metadata directed to archives for electronic materials, as exemplified by Encoded Archival Description (EAD) and Metadata Encoding and Transmission Standard (METS). Preservation Metadata Implementation Strategy (PREMIS) is a new metadata standard for digital preservation. Thus, the library and archives community have made significant efforts to develop metadata standards to archive and preserve digital resources.

In general, preservation of electronic materials requires description of not only the

contents of the subject materials but also their form, the environment required for their preservation, their preservation history, and other matters. Metadata descriptions required for preservation are apt to be complicated. This creates a need to curtail costs entailed by them. While descriptions of file types and other such items can be automated in certain respects, it would be preferable to attach metadata at the stage of creation of the materials if possible, in order to hold down the cost of description of their contents.

Another key consideration is measures to heighten the utility of the materials preserved, by means of functions for search of materials and navigation within the archives. Such functions would depend on the nature and objective of the archives. The basic requisite ones would probably be a capability for searching based on the description of the content of materials digitized as image data and navigation resting on key words excerpted from the materials, for example. Archives would also have to enhance their utility for various types of users, including children and students on all levels, adult members of the general public, and experts. This would demand special displays on the Web profiling the archive materials with reference to some specific themes, and explanations, etc. telling prospective users about the archive contents and method of use. Another task that must not be overlooked will be assurance of archive accessibility to all regardless of disabilities.

As archives evolve, a capability for inter-archives access will presumably assume increasing importance. Because archives vary in respect of character, this points to a need for functions enabling access to and search of a wide range of archives while making the most of the features of each.

To endow archives with such additional value will require studies to determine methodology for archival organization that has arrangements for both sharing and distinguishing, with reflection of the unique personality of each one.

4. Electronic preservation of governmental records

In this section, I would like to outline what I have learned through the discussions in the aforementioned committee studying the ideal of the management, transfer, and preservation of governmental records by electronic media in the Cabinet Office and my research visits to the National Archives of Australia and others. Many of these observations may appear to be only natural on reflection, but I shall relate them here because they have stayed in my mind.

(1) Digital archives for libraries and digital archives for the National Archives

Libraries, museums, and the National Archives began to take approaches to digital archives at an early stage in the explosive spread of the Internet in the 1990s. The aim is to make precious information resources available to anyone at any time and place, through electronic networks.

The words "digital archives for a library" immediately conjure up the image of an accumulation of digitized historical records and precious materials. The Archives, too, are taking analogous approaches to digital archives of historical and precious materials. I do not see any substantial difference between the two in this respect. Both are also attempting to build archives for materials that were born digital. For this reason, they share various difficulties. The gaps between them instead surface as a difference of nature. For example, whereas libraries essentially collect published documents (materials in the public domain), Archives collect documents created in the process of work. This leads to differences in respect of the precision required in collection, scope of collection, and relationship between the document author and the archives. In spite of their mutually different personalities, however, the two have much in common, and it goes without saying that they must engage in the sharing of knowledge and technology supporting their digital archives and cooperate in provision of services.

(2) Electronic documents - definition, identification/discrimination, and other matters

The prospect of preserving materials that were born digital raises the question of how to define electronic documents as the subjects of preservation. Electronic documents have key attributes in their dynamic properties, i.e., their amenability to alteration of the form of display in correspondence with the user and use environment, production of a single document by dynamic combination of components contained in data bases, and possession of hyperlinks. This suggests a need for redefinition of the mass of documents that are the subjects of preservation. We must decide, for example, whether the forms displayed on specified browsers should be considered documents, or to confine usage of the term to those documents accumulated on servers. At present, it would presumably be appropriate to regard word-processed text, the output of spreadsheet software, and other works reflecting the conventional image of printed matter as documents. Further in the future, however, I suspect we will need to build a shared

understanding on the nature of the electronic documents to be preserved. This, in turn, will require a lot of experience and a good understanding of the rapidly evolving technology.

(3) Life cycle of governmental records and foundation of document distribution

Preservation of governmental records must be positioned in the context of the overall document life cycle beginning with creation and ending with discarding after retention. At the stage of creation, the attachment of a certain level of metadata for document distribution and retrieval (search) could be expected to lower the costs of metadata creation at the preservation stage. Determination of uniform metadata rules for retrieval of governmental records and establishment of standards for classification would make the preservation of electronic documents more efficient.

The desirable form of document at the active stage and that at the preservation stage are not necessarily consistent. It is sometimes necessary to remove some functions of active records in order to put them into a form adapted to preservation. It follows that we need a shared understanding on the document contents that must be preserved. To this end, it would be preferable to have a consistent, integrated management of electronic documents anticipating the environment of their distribution and use, and their entire life cycle. In other words, there must be cooperation between the side creating documents and that preserving them. Document management, however, must not present obstacles to the incorporation of new electronic document technology arising along with technical advances.

(4) From the survey visit to the National Archives of Australia - giving it a try

In October 2005, I visited the National Archives of Australia to conduct a survey on its electronic preservation of governmental records. I would like to outline some of what I learned from this visit.

In its services, documents are sent to the Archives in packages instead of through electronic networks, and are accumulated on a server after a certain period of time to check for viral contamination. The server is located in a site with full security measures, in a network environment. I found this consideration for the practical end quite interesting. In addition, for long-term preservation, they are developing

the tool XENA for document preservation based on XML. Although it cannot cope with all types of document at present, it is characterized by an ability to accommodate document formats in widespread use, assurance of extensibility, and integrated accumulation of metadata and documents. The Archives is also engaged in development through open sources, and is looking to cooperation with other organizations. Such efforts are the best that can be taken right now.

I was deeply impressed with the stance of the Archives, which asserted that it was only natural to preserve governmental records electronically as they were already being created and used electronically. Similarly, while it was easy to imagine the emergence of all sorts of problems with preservation of electronic documents given their nature, the Archives took the position that difficult problems were no reason for inaction, and was determined to work on them, one at a time.

At a relatively early date, Australia established the Australian Government Locator Service (AGLS), a classification vocabulary for organization of governmental records, and embarked on promotion of metadata attachment at the stage of document creation based on AGLS. It strikes me that these efforts are being linked to the efficient distribution, accumulation, and preservation of governmental records.

5. Conclusion

The preservation of electronic materials is one of the important and difficult tasks of digital libraries. The National Diet Library, for example, has been conducting investigative research on the long-term preservation of electronic materials.(2) The results of the discussions in the aforementioned study group on electronic management, transfer, and preservation of governmental records instituted in fiscal 2005 in the Cabinet Office are presented in the report submitted by the committee on proper management of governmental records including preservation and use.(3)

Archives receive documents from various government offices and agencies. While they basically have had to take measures for preservation of only paper materials thus far, preservation of electronic materials will presumably compel them to handle an extremely great variety of documents. Resolution of the related problems will demand efforts on their part, but the Archives will not be able to solve all of them on their own. There is a need for cooperation over a larger scale. I should also note that it will not do to abandon the idea of electronic preservation

just because of the difficulties entailed; all must be willing to take the initiative and start doing what they can now.

It is becoming the general practice to offer an interface for provision of online services not only to people but also to computers. Meanwhile, the user pool and the use environment are becoming more diverse. Circumstances are demanding an adequate provision of information to anyone at any time and place, and in a manner suited to the environment of the user in question. As I see it, collaboration with various institutions and organizations will be indispensable for achievement of services adapted to this new environment.

I do not have practical experiences in archives and the life cycle management of governmental records, and the opportunity to participate in the aforementioned Cabinet Office study committee was indeed a golden one for me in my examinations on electronic preservation of such records. In closing, I would like to express my deep appreciation to all concerned members of that committee, to whom I feel much indebted.

References

(1) "Cultural Heritage Online," http://bunka.nii.ac.jp/ (in Japanese, accessed September 2007).
(2) National Diet Library, "Long-term Preservation and Use of Digital Information," report and website at:
http://www.ndl.go.jp/jp/aboutus/preservation.html
(In Japanese, accessed September 2007).
(3) Committee on proper management, preservation, and use of governmental records (Cabinet Office), "Report on centralized management in the intermediate stage and on management, transfer, and preservation by means of electronic media," 2006, p32.
http://www8.cao.go.jp/chosei/koubun/kondankai14/houkoku.pdf
(In Japanese, accessed September 2007).