Archiving Electronic Publications – a Librarian's Point of View [1]

Uta Grothkopf, European Southern Observatory, esolib@eso.org

1 Introduction

Today's scientific literature is not limited to one single medium; various media have coexisted for some time. In recent years, an increasing number of publications have been published electronically. Electronic format alters the traditional way of publishing, retrieving, obtaining and preserving information, and it modifies the structure, composition and features of documents. Electronic publications are available in digital format and therefore require the use of a computer in order to access and use them. They can be delivered to users on physical media (diskette, CD-ROM, tape etc.), over networks (via electronic mail, ftp etc.) or may have to be accessed actively by users (typically on the World Wide Web). They can be made available in various formats with identical information and functionality in parallel. Various types of information (text, images, graphics, sound, software programs etc.) can be combined in one multimedia product. Embedded software may allow users to interactively modify underlying data or launch virtual experiments. Networked electronic publications can be accessed and used from anywhere at any time by several simultaneous users; interconnected documents provide links to other electronic resources. Depending on whether or not they can be modified after publication, electronic publications are dynamic or static documents.

In order to preserve the scientific and cultural knowledge of today, electronic publications have to be archived just as print publications have been. However, archiving electronic media is considerably different from archiving print publications. Up to now, archiving has been relatively easy and straight-forward. Libraries subscribe to journals and buy books, and after publications have arrived in our libraries, they are cataloged, classified, and technically prepared, and finally they are placed on the library shelves where they are available for users. References to print publications follow generally agreed-upon rules for citations that uniquely describe the source, and users can trust that a print document has not been changed after publication. Libraries own the publications they have bought, and unless it is decided to re-sell or remove the items, they are kept in the libraries indefinitely. Users can rely on a decentralized system of libraries where they can consult both current and historical publications whenever needed without additional costs.

In contrast, archiving electronic publications is much more complex. Up to now, no particular institution has been responsible for long-term preservation of electronic publications, hence future access to electronic documents is still uncertain. As of today, users don't know how they will be able to retrieve, obtain, and use electronic publications even a few years from now. Once archiving institutions have been selected, they must solve a large number of political, administrative and technological questions. What exactly has to be archived? Who provides and maintains the necessary tools and technological equipment? Who pays the personnel? Who will have

[1] An abridged version of this article will be published in the Proceedings of the 23rd IAU General Assembly, held in Kyoto, Japan, August 18-30, 1997; ©Kluwer Academic Publishers

access to the archives and at which costs? Is the technological infrastructure capable of meeting the demand for access to documents via networks?

Archives must guarantee that information resources can be identified, accessed, retrieved and used over time, and in the interest of the general public they must ensure access is available at reasonable costs. Archiving therefore is one of the most important and urgent problems arising from electronic publications and requires thoughtful solutions.

2 Archiving electronic publications

2.1 Who archives?

At present, originators (i.e., authors and publishers) are responsible for preserving the electronic documents they publish. There is a danger that publishers will not be willing or able to archive documents without commercial value. Therefore, in the future institutions whose interest in archiving goes beyond immediate economic consideration should be responsible. Electronic archiving is more expensive than print archiving, and small libraries may not be able to do it. Large institutions and national or state libraries, which can keep up with changing technology, remain the possible candidates. Legal deposit regulations, which currently exist for print media, could be extended so that also electronic publications are required to be deposited in national libraries. Legal deposit regulations in the United States have already been changed accordingly, so that a copy of all electronic publications must be deposited in the Library of Congress.

According to the Task Force on Archiving of Digital Information, institutions "must meet or exceed the standards for archival certification" (Garrett and Waters, 1996) if they want to become certified digital archives. Minimum requirements include the guarantee to provide access to electronic publications, even if the initial vendor has disappeared and the original system configuration is no longer available, as well as the provision of "a critical fail-safe system [...] in order to actively and aggressively rescue electronic publications which are in acute danger". Exact definitions to determine which institutions are eligible for becoming digital archives will vary in different countries. However, guidelines and criteria should be established in order to provide a general framework for international standards.

2.2 What shall be archived?

Documents always have been subject to appraisal in order to determine which material had to be preserved. Archiving institutions need to develop standards that define what exactly constitutes an "electronic publication" as opposed to electronic communication, databases and other ephemeral information, and which electronic publication fulfills the quality criteria for archiving. In addition to contents-related quality criteria, further factors may influence the decision to archive electronic media, for instance, functionality, usability, technological standard and adequacy of maintenance.

2.3 How many copies shall be archived?

For networked electronic publications, it may be attractive to archive only one copy of a document which can be accessed remotely, but this approach bears many risks with regard to loss or damage of that master copy. Redundant copies of electronic documents should be stored in various places in order to avoid loss or corruption of archived material. Providing multiple copies in a distributed network also accelerates access times as users can log into a server that is geographically close to them or to which they maintain particularly fast links.

2.4 Storing electronic publications

Currently, the majority of electronic publications are static documents, i.e., they are not changed once they are released. Users may be able to change underlying data interactively according to their needs, launch virtual experiments from embedded software or add comments to networked documents, thus creating a dialog with the authors and other readers. However, the original document remains unchanged. Archiving institutions will have to decide whether or not comments and other additions should be preserved together with the publications. A subset of networked electronic publications are dynamic documents which can be modified after publication. Archiving all versions of dynamic publications is difficult. Probably only "snapshots" of the available data can be taken at regular intervals. Alternatively, the entire record of changes will have to be logged in order to be able to reproduce all changes carried out after release of the original document.

Electronic publications can be stored on a variety of physical media. If high user demand can be expected, archiving institutions may decide to provide publications online on networked servers in order to reduce access times. Otherwise, offline storage might be more appropriate (and cheaper). Theoretically, electronic publications provided on networks can be accessed from anywhere at any time. Documents remain interconnected, i.e., they can contain active hypertext links to other digital documents and vice versa. In practise, however, global accessibility requires stable and fast telecommunication networks worldwide which are not yet in place.

2.5 Technology refreshing and migration

Electronic publications are not inseparably tied to a particular physical format, but can be copied to new storage media with little or no loss. However, experience about the long-time durability of the media is not yet available. An even greater danger than deterioration of the storage media lies in the technological obsolescence of the reading devices which are necessary to use electronic publications. Technology is changing rapidly; hence, in the digital environment "preservation means copying, not physical preservation" (Graham, 1994). This process usually is called technology refreshing; in an even broader concept, the Task Force on Archiving of Digital Information refers to it as "migration" which is defined as "a set of organized tasks to achieve the periodic transfer from one hardware/software configuration to another or from one generation of computer technology to a subsequent one" (Garrett and Waters, 1996). The purpose of migration is to retain the ability for clients to use digital objects in the face of constantly changing technology. Copying should take place at regular intervals in order to avoid unexpected and possibly unbridgeable gaps in the availability of necessary hardware and software. Information should be encoded in a system-independent format, for instance SGML, as the ability to migrate and refresh will be much greater for such a richly tagged format than it is for a page image format like PDF or Postscript.

2.6 Archival data

Just like printed documents, electronic publications must be retrievable in order to be used. Structured descriptive information about electronic documents, so-called metadata, must be mapped to retrieval systems (e.g., library catalogs). Metadata typically are contained in the documents themselves or accessible through them; they must include formal data (author, title, release date etc.) as well as contents-related data (keywords, time and geographic coverage). In addition, the reference system must provide a variety of further information, for instance about modes of access (WWW, telnet, ftp), tools necessary to use the electronic material (e.g., particular software required at the user's site), and computer file characteristics (size, format). The authenticity (identity) of electronic publications needs to be easily recognizable in order to determine which version of the document is being used. Electronic time stamping of documents (i.e., creating a permanent, indelible mark in the digital file that indicates the version), digital watermarking technologies (i.e., software that allows encoding an identification into a document that can be located by WWW search engines), digital fingerprints (i.e., creating hash codes that reflect every bit in a record) and other authentication methods are hoped to help assure the authenticity of electronic publications. Additional information should be available about embedded links to other digital objects, access rights and restrictions, as well as data to prove the integrity of documents.

2.7 Integrity of electronic documents

A critical distinction between printed and electronic publications is the ease with which their integrity can be determined. The Task Force on Archiving of Digital Information defined five components that constitute the integrity of digital documents: content (intellectual substance contained in information objects), fixity (content fixed in a discrete object as opposed to continuously updated documents), reference (reliable systems for locating and citing), provenance (a record of the document's origin and chain of custody), and context (a document's interaction with elements in the wider digital environment) (Garrett and Waters, 1996). Printed publications do not create doubt about their integrity; we take for granted that they have not been changed since they were published. Electronic publications, in contrast, can be changed in various ways: accidentally (for instance during copying to newer devices); on purpose, with well-meant intention (as is the case with dynamic documents); or on purpose, without well-meant intention, i.e., through fraud (Graham, 1994). Mechanisms to prove the integrity of data and information systems are not only necessary in order to trace fraud, but also to make sure the documents have been moved to a new storage medium without loss.

3 Libraries and Archiving

At first glance, electronic publications seem to be not only an alternative format for scientific documents, but also the solution to many problems libraries face today (Lehmann, 1996): they offer accelerated access to literature for several simultaneous users over networks regardless of their actual location; the growing need for expensive storage space on shelves is diminished; the problem of deterioration of books due to acid paper is solved. On the other hand, digital publications raise a large number of new questions and problems, many of which are not yet solved. The traditional librarian's role involved retrieving, obtaining, making accessible and archiving information needed by our users. As archiving is undergoing vast changes, this also has a major impact on the other classical library services.

3.1 Retrieving archived electronic publications

If electronic publications are received offline, i.e., on physical media, they will be treated in libraries similar to print media. Items will be cataloged and shelved and will be retrievable through the library catalog. Networked electronic media have to be cataloged, too, and library records must be enhanced so that all available archival information can be included in catalog entries. The Dublin Core Metadata Element set provides a resource description that will improve access to information on networks[2]. If national libraries become digital archives, electronic publications will appear in the respective national bibliographies which will make their retrieval easier than it is today (Deutsche Forschungsgemeinschaft, 1995). However, the ability to retrieve networked electronic publications in distributed systems depends not only on their appearance in library catalogs and bibliographies, but also on a reliable global naming and address system. Current network addresses still depend on the Uniform Resource Locator (URL). But URLs are easily mistyped and can lead to error messages when trying to access networked documents. Name resolvers which are already used by some electronic astronomical journals allow a unique name to be assigned to an online publication that will not be changed even if the document is moved to another server. The name resolver will keep track of the actual location of the requested file and translate the name into the current network address.

3.2 Obtaining archived electronic publications

In the digital environment, journal subscriptions and book purchases as we know them from the print environment are replaced by the concept of granting access to publications just for a given time. The library does not "own" the items, but "leases" access to the documents. The detailed access and usage conditions are described in contracts between publishers and libraries. These contracts, typically called license agreements, regulate, for instance, for how long the library has access to the publication, what users and librarians are allowed to do with it, how many simultaneous users are granted access, and how often the publication can be viewed, printed or otherwise used. License agreements must also determine what happens after the contract terminates.

[2]For more information, please visit the Dublin Core Metadata Element Set homepage at http://www.oclc.org:5046/research/dublin_core/

Since it is the nature of leased material that it remains physically with the provider, many librarians fear that they will be left without anything when a leasing contract ends.

As a result, publications requested by our users more and more often will not be available in our libraries, but will have to be obtained from external sources. No standard mode for access to back volumes of publications is yet being applied, and terms and conditions vary from publisher to publisher. It is possible, however, that libraries will have to pay twice (or more often) for the same product – once while a contract lasts, and again after cancellation or termination of a title if a user needs a particular article from the timeperiod when the library subscribed.

3.3 Accessing archived electronic publications

Once a requested electronic publication has been located, accessing and using it can be cumbersome. Electronic publications require that certain hardware and software be available at the user's site. At present, a large variety of physical storage media for electronic publications (CD-ROM, diskette, tape etc.) exist and will continue to exist, since newly evolved technologies do not always replace previous ones, but co-exist with them. A library that owns or leases electronic publications therefore will have to provide the necessary reading devices for each storage medium to be used. Electronic publications in HTML format on the World Wide Web not only require networked terminals at the user's site, but also often make it necessary that a particular version of the browser and other software be installed. Networked documents depend on a stable and reliable infrastructure. At present, bandwidth does not keep up with demand. Even in technologically developed countries, accessing networked documents can be painfully slow, and the situation is even worse in countries with less stable infrastructure. It is not yet clear who will be able to pay for new versions of software and hardware that are necessary at ever shorter intervals, and who will ensure reading devices for offline electronic publications are available in our libraries many years from now.

Further important changes in the electronic environment refer to copyright law and interlibrary loan. Copyright law is meant to grant the owners of creative and scientific works specific rights. For instance, those who want to use copies of a work must have the copyright holder's permission and usually have to pay a certain amount to the owner. In order to balance the rights of copyright-holders and those who want to use the information, there are some exceptions to this rule. Interlibrary loan (ILL) is one such exception. With regard to journal articles, this name is somewhat confusing, because libraries today do not usually lend bound volumes of journals, but rather send photocopies of individual articles. However, the concept of interlibrary loan has remained the same up to now – a library that needs a publication owned by another library may ask for a photocopy, provided the principle of fair use (or fair dealing) is adhered to, i.e., the copy is meant for personal, scientific or educational, but not commercial, use. Misuse of intellectual property seems to be all too easy in the electronic environment; therefore many publishers currently are trying to diminish some of the rights users had with print publications, in particular the right of interlibrary loan. But "there is a misconception that fair dealing does not apply to electronic databases. Fair dealing takes no regard of the medium" (Oppenheim, 1997). In the interest of the scientific and general public, interlibrary loan and the fair use concept must be maintained with regard to electronic publications.

Following a phase of free access for test purposes, we now witness a commercialization of electronic information resources on the Internet. Publishers are concerned about their earnings if they are not able to trace who has accessed, printed, or otherwise used the publications they market. Most probably, their concerns will vanish soon once reliable Electronic Copyright Management Systems (ECMS) become available that will be able to automatically tag copyrighted works and monitor their use (Oppenheim, 1997). There is another reason for publishers not to worry. Those libraries that can afford it will continue to order needed publications from commercial document delivery services (from where publishers can be sure to receive royalty revenues), as in many countries commercial services are quicker than interlibrary loan due to work overload at the lending library. However, those libraries whose budgets are too limited for purchasing articles from commercial services must be able to rely on ILL, or their users will be entirely excluded from access to electronic information resources.

3.4 Libraries in the electronic environment

One of the most important tasks of librarians always has been to provide access to information resources. This role will not change in the electronic environment. Users need not worry about the question from where a publication can be obtained; they can rely on a system of libraries to obtain needed information for them. In the print as well as in the electronic environment, librarians assist users in searching and retrieving information. They mediate between authors/readers, publishers, digital archives and other parties involved, provide appropriate tools capable of matching user needs with the available information, and deliver current as well as historic publications to the users.

4 Conclusion

Many of the complex, interrelated factors that influence archiving of electronic publications are not yet known. Decisions made now will have an enormous impact on the future availability of scientific literature. Archiving electronic publications requires thoughtfulness, vision, and longterm commitment from all parties involved. A stable and reliable infrastructure must exist. International collaboration is essential in order to develop standards that ensure preservation of the scientific and cultural record throughout time and regardless of national boundaries. The current increasing commercialization of information access must not turn information into a privilege of the rich. In the interest of the scientific and general public, archiving institutions must guarantee future access to and use of electronic publications at reasonable costs.

Acknowledgements

My sincere thanks go to Ellen Bouton (National Radio Astronomy Observatory, Charlottesville) and Sarah Stevens-Rayburn (Space Telescope Science Institute, Baltimore) who provided very helpful comments and suggestions. I also wish to thank Peter Boyce (American Astronomical Society, Washington, DC) for sharing his expertise and knowledge with me.

References

[1] Deutsche Forschungsgemeinschaft (DFG), Bibliotheksausschuß, 1995: Elektronische Publikationen im Literatur- und Informationsangebot wissenschaftlicher Bibliotheken. *Zeitschrift für Bibliothekswesen und Bibliographie (ZfBB)*, **Vol. 42, no. 5**, 445-463. Electronic version at http://www.dfg.de/foerder/biblio/web.urz.uni-heidelberg.de/epub.html

[2] Garrett, J. and D. Waters (co-chairs), 1996: Preserving digital information. Report of the Task Force on Archiving of Digital Information, commissioned by The Commission on Preservation and Access and The Research Libraries Group. Electronic version as of May 1, 1996 at http://www.rlg.org/ArchTF/

[3] Graham, Peter S., 1994: Intellectual preservation: electronic preservation of the third kind. Washington, DC: Commission on Preservation and Access. Electronic version at http://aultnis.rutgers.edu/texts/cpaintpres.html

[4] Lehmann, K.-D., 1996: Making the transitory permanent: the intellectual heritage in a digitized world of knowledge. *Daedalus*, **Vol. 125, no. 4 (Fall)**, 307-329

[5] Oppenheim, Charles, 1997: Copyright in the electronic age. In: *UNESCO World Information Report*. Unesco, Paris. Electronic version at http://www.unesco.org/cii/wirerpt/chap26.htm