Archiving Scientific Literature in the Digital Age

M. R. V. Sahyun
Consultant and Editor, Journal of Imaging Science and Technology
Maplewood, Minnesota, USA

Abstract

Within the context of current strategies for preservation of digital archives, we consider the special issues unique to maintaining our heritage of scientific publication. Responsibilities of authors, new functions for the peer review process, and the custodial role of publishers are considered. A distributed model of archiving scientific literature has been developed, with the social science community in a leadership role.

Introduction

With the trajectory of scientific publication to electronic media now well established, the issue of maintaining the archival quality of the scientific literature needs to be put into the larger context of the preservation of electronic records.

Background

According to Tom Strang of the Canadian Conservation Institute,[1] three main principles govern the preservation of digital information:

1. We don't throw out the original form of the record, e.g., the print edition of a journal, keeping in mind that "preservation is not the same thing as accessibility".
2. Within the digital library, continuous migration is necessary to keep ahead of media deterioration, format and machine obsolescence. In traditional archiving we are familiar with the migration from print records to microform to digital libraries. The principle here is "move it or lose it".
3. For the sake of feasibility, prioritization is inevitably implemented at each migration step, i.e., we have to ask the question:, which records do we take the time and trouble to update to the newer format, and what is the best use of our limited storage capacity and custodial resources. We must understand that those records we don't update are inevitably lost. The guiding philosophy is "something is better than nothing".

Archiving of digital information is defined as "…the managed activities necessary to ensure continued access" to the information.[2] Therefore archiving and preservation are not the same thing, though the former presumes the latter. For instance, file formats which are appropriate for web search, i.e., access, may not be the most reliable and robust for preservation.

Saving information is not a passive activity, it is a pro-active one, requiring continuous attention and monitoring. Inherent in this whole concept of migration is the realization that it will be accompanied by widespread loss of information. Of course, such loss of information has been well-known in the pre-digital era, owing to such phenomena as material deterioration, catastrophic events, etc. These losses tend to be random and are often controllable, e.g., by duplication of collections, whereas the losses associated with migration are deliberate and involve prioritization, selection and appraisal at each step of the process.

Special Concerns

In the context of the scientific literature the natural trajectory is for the appraisal process to be integrated into peer review. This provides for a modicum of knowledge, expertise and, it is hoped, objectivity to be brought to the selection process. It has been suggested that peer review, as it has been practiced over the past century of scientific publication, is outmoded, given that the internet allows a broad spectrum of the scientific community to react to any new contribution to the literature at a very preliminary stage. I propose that the contrary is more likely to be the case; selection and appraisal are more overtly integral to archiving of information in digital form. Reviewers then must take on an added burden of responsibility insofar as their recommendations will become the basis on which a particular contribution is preserved for posterity or not, and reviews may (and should) be preserved as metadata accompanying the research report archive. I further envision that post-publication peer review may also come into play, insofar as selection decisions are made at each migration step in active preservation.

Responsibilities of the Author

The electronic author must consciously address archiving issues at the time of writing and submission. Creation of an electronic paper involves choice of digital format and designation of a custodian, i.e., publisher. In light of the role of peer review, above, it may be presumed that choice of a non-reviewed forum, e.g., conference proceedings, personal or institutional web page, etc., may facilitate short term

availability of the information but preclude long term preservation, thus consigning the information to oblivion. Internet-only publication risks widespread loss of information and of records of value.[3] These risks include media (and organizational) deterioration, technical and fiscal obsolescence, and custodial neglect,[4] and these consequences are virtually guaranteed in the case of do-it-yourself publication.

Among the responsibilities of the electronic author, selection of an internet publisher is arguably the most critical. In the paradigm of print publication, the custodial role, with archiving responsibilities, has rested largely with libraries and academic institutions. In digital publication this responsibility shifts to the publisher. For reasons including the ability to authenticate and identify sources, degradation associated with compression for file transmission and storage, and, perhaps most significantly, intellectual property considerations,[5] the electronic publisher must also take on the role of custodian of the published information.

It is doubtful whether *ad hoc* consortia of scientists who seek to avoid some of the costs and control involved in traditional publication, commercial sponsors of topical conferences and short courses, and even some smaller professional societies have the resources or longevity required to be good custodians of electronic scientific information.

The Distributed Model

Smaller societies may overcome this limitation by forming consortia or collaborative associations. In this respect the social sciences community may be more advanced than the physical and engineering sciences or the biomedical community.[6,7]

The custodial role requires:
1) deep and long-term commitment;
2) full integration of technology into database management; and
3) leadership in standards establishment.

It includes not only the technologically sophisticated active preservation activities outlined above, but also, perhaps, the preservation of the technological context in which the original document may have been created.

The resources required should not be underestimated: a National Archives (USA) study has estimated that the cost of each migration of a record to a new format, in terms of dollars, person-power and technical expertise, may be comparable to that of creating the record.[9] On a per record basis, active archiving in the digital world may involve costs two to three orders of magnitude greater than passive custodial storage of print media.[8]

The distributed model of scientific literature preservation, e.g., OAIS created by NASA,[10] as well as ICPSR for social science research data,[7] implies a collaborative effort. This effort includes cost-sharing, among various custodians, i.e., commercial publishers, professional societies, academic and governmental institutions, etc., to the end that formats are interchangeable among databases, and allowing transfer of custodial responsibility where necessary.

References

1. T. Strang, *Symposium 2003--Preservation of Electronic Records: New Knowledge and Decision Making*, Canadian Conservation Institute, Ottawa, Canada, 2003; published on CD-ROM; http://www.cci.ca/symposium.
2. *Handbook for preservation management of digital materials*, The British Library, London, UK, 2001; http://www.dpconline.org/graphics/handbook; see also D. Woodyard and H. Shenton, in Ref. 1.
3. T. Eastwood, in Ref. 1.
4. T. Au Yeung, in Ref. 1.
5. It has been opined by professionals in the archiving community (Refs. 2-4) that migration as understood in this context corresponds to unauthorized reproduction when practiced by other than the copyright holder, under prevailing legislation. This is another example of where the legal environment has not caught up with the technological environment.
6. C. Humphrey, in Ref. 1.
7. http://www.icpsr.umich.edu.
8. F. Frey, in Ref. 1.
9. S. Puglia, unpublished result cited by P. Adelstein, in Ref. 1.
10. http://ssdoo.gstc.nasa.gov/nost/isous.

Biography

Mel Sahyun received a Ph.D. in chemistry from UCLA and has had a career spanning government, industry and academia, in the course of which he has published extensively. After working in the area of environmental science for the US Public Health Service, he joined the Corporate Research Laboratories of 3M, where he specialized in silver halide and photoconductor-based technologices. From 1996-2003 Dr. Sahyun was a member of the chemistry faculty at the University of Wisconsin-Eau Claire, with responsibilities in the department's Chemistry in Business program. Since 1997 he has been Editor of IS&T's *Journal of Imaging Science and Technology*, after several years service as Associate Editor. He is a Fellow and Senior Member of IS&T and has served as a Vice-President of the Society.