Archiving Supplemental Materials

David S. H. Rosenthal
Victoria A. Reich

Introduction

It has long been considered important that institutions other than the publisher preserve academic journals. Libraries fulfilled this role when the publishing medium was paper. Shortly after journals started their transition to the Web in the mid-90s, The Andrew W. Mellon Foundation started studying how they should be preserved [1]. These studies bore fruit; now institutions other than the publisher routinely preserve e-journals. There are two architectures in use: centralized and distributed. The LOCKSS Program represents the distributed approach [2], but some years of experience in operating systems with both architectures shows that the differences are matters of detail. The two approaches share many major issues, and in particular those caused by the importance of preserving supplemental materials [3].

We examine this problem from the perspective that the same institutions that already preserve the primary journal content should use the same technologies to preserve supplemental materials. We have relevant experience; the LOCKSS program currently preserves supplemental materials in this way. Setting up these institutions, providing them with viable business models, and developing the necessary technologies has over the last decade has proven to be a major effort. It is unrealistic to believe that a similar but separate effort could be undertaken on behalf of supplemental materials, which by their nature are normally regarded as less valuable than the primary content.

Sustainability

It has become clear that the single most difficult issue in digital preservation in general, and in e-journal preservation in particular is “economic sustainability.” To be blunt, on a cost-per-byte basis it costs too much. Even the most cost-effective approach known, that of the Internet Archive, is too expensive to meet its goals. The vastly more expensive techniques being used for e-journals are similarly inadequate [4]. The LOCKSS program has been cash-flow neutral for some years, but that may not be enough for long-term sustainability. As for the sustainability of Portico, the other major e-journal preservation system, early this year an audit reported:

> “the ongoing business viability of Portico as a service is not yet assured, judging from financial information disclosed to date.”[5]

Examination of the latest tax returns available (2008 [6]) for Ithaka, the parent organization, shows that Portico was at that time short of cash-flow neutrality by a substantial margin.

Adding a broad range of supplemental materials to the task of preserving primary e-journal content, however it is done, will inevitably add costs and thus make the problem of economic sustainability worse. How significant are these costs likely to be?

Ingest

The question thus becomes "how to minimize these additional costs?" Experience has shown that the dominant cost in e-journal preservation is the ingest process [7 p18]. This is not unexpected. Long-term studies of the cost of storing data at the San Diego Supercomputer Center [8] have shown that it decreases over time, although not as fast as the famous "Innovator's Dilemma" [9] exponential drop in the cost per byte of disk storage would lead one to expect. The much-feared costs of format migration have not in practice been incurred, since formats have not been going obsolete at anything like the rate predicted by Jeff Rothenberg in the mid-90s [10]. Ingest costs cannot be delayed to take advantage of the time value of money [11].

The reason why ingest is the dominant cost is that it consumes staff time, which is expensive in absolute terms and tends to increase over time. In order to ingest an e-journal's content the archive must first obtain permission from the copyright owner to do so. This takes negotiation between the archive and the publisher, and often involves lawyers on both sides. Second, the preservation system must be adapted to the peculiarities of each journal. Third, the ingest process must be carefully monitored to ensure that routine problems such as intermittent network and publisher outages, or unannounced changes in the publishers' systems, do not interfere.

It is relatively cost-effective to ingest content from major publishers such as Elsevier. Although the negotiations are time-consuming, this cost is amortized across a large number of journals. Their systems are well engineered, consistent across many journals, stable, and well documented [12]. But the content of the major publishers is at low risk of being lost, so the value of archiving the large amount of content obtained in this way is low.

The content at high risk of loss comes from smaller publishers. Although negotiations with smaller publishers are typically easy, each results in only a small amount of content. Their systems are more diverse, less well documented, and less stable than those of major publishers. Thus the efforts involved in adapting the preservation system to the journal, in monitoring the ingest process, and in handling the more frequent exceptions detected, are all much greater both absolutely and on a per-byte basis.

Thus we see that the two major staff time sinks in the ingest process are diversity and low volume. Supplemental materials are by their nature more diverse and lower volume than the primary content, and can thus be expected to be more expensive both on a per-item and a per-byte basis.

This expectation is reinforced by preliminary results from a survey of data preservation costs under the auspices of the UK's JISC. This found that ingest costs including the process of obtaining permission to do it were over half the total:

> "the cost of archiving activities (archival storage and preservation planning and actions) is consistently a very small proportion of the overall costs and significantly lower than the costs of acquisition/ingest or access activities...As an example the respective activity staff costs for the Archeology Data Service are Access (c.31%), Outreach/Acquisition/Ingest (c.55%), Archiving (c.15%)"[13].

Best Practices

The application of best practices, or even better, standards to the process of publishing supplemental materials can reduce diversity and aggregate the materials into larger, uniform collections. Doing so will not merely reduce the cost of preservation but also the cost of publishing, finding, and accessing them. In particular, these best practices should be aimed at reducing staff time, since this is the biggest cost component in each of these tasks.

In what areas are best practices likely to have the greatest impact in reducing staff time? We identify four:

- Intellectual Property
- Location and Structure
- Technical Metadata
- Bibliographic Metadata

Intellectual Property

The existing intellectual property constraints, both informal and formal, on sharing of data are diverse, unclear and in flux. This complicates the archive's task, which requires clarity as to the permission that the archive has to keep copies of the publisher's intellectual property, and about the terms under which the data is to be preserved and in future accessed.

As regards informal constraints, surveys of authors' attitudes to sharing data [14,15,16] uniformly report a great diversity among fields, funders, and practitioners. In some fields, such as astronomy, sharing is the norm, albeit modulated in some cases by delays allowing those researchers who captured the data "right of first publication." In others, particularly those with the potential of valuable patents, sharing is the rare exception. In these fields it is to be expected that any supplemental materials actually published will be of low value; they will have been sanitized to ensure they do not compromise the potential for commercial exploitation.

As regards formal constraints, the legal situation is complex. Some supplemental materials are copyrighted, but it isn't clear that the same copyright terms apply to them as to the text of the article to which they are attached. Some data are just facts, so are not subject to copyright. Some data represents a compilation of facts, so may be subject to copyright or may in the European Union (but not elsewhere) be subject to "database right."

Furthermore, despite the efforts of Science Commons [17], there is no widely used equivalent of the Creative Commons (CC) license for copyrighted data [18]. The reason is not hard to understand; the CC license is grounded in well-established copyright law. Because the legal framework surrounding data is much less clear, it has been much harder to establish a strong means for allowing the right to use data while providing a guarantee of credit that is the most frequent desire of researchers. The recent release of the Open Data Commons "BY" license for databases [19] is a step in the right direction.

Similarly, there is no equivalent of the machine-readable means for labeling content with the appropriate CC license [18]. Lacking such means, ingest programs that collect data from supplemental materials are on shaky legal ground.

Best practice efforts are thus urgently needed in two related areas:

1) An analog of the CC "attribution" license, allowing researchers to grant general permission to use their data provided credit is given. This would satisfy a substantial proportion of researchers. Additional license versions could be added later to satisfy other groups of researchers.
2) A machine-readable form of this license, similar to the RDF form of the CC license, allowing automated harvesting of data from supplemental materials.

Location and Structure

There are two basic forms in which e-journal content can be ingested:

1) "Source" content, in which the publisher packages up the content it wishes the archive to preserve in some form different from that in which the content was originally delivered to readers and then transmits it, often via FTP, to the archive for processing and preservation. Source is in practice something of a misnomer. Sometimes, the "source" content includes the actual source (e.g., SGML markup) but it almost always includes exactly the same rendered form of the content that was delivered to some readers (e.g., PDF).
2) "Presentation" content, in which the archive behaves exactly as a reader would, accessing the e-journals web site and ingesting the same HTML, CSS, PDF, JavaScript, and other formats that the reader's browser would interpret.

Note that for primary e-journal content this distinction has some long-term relevance. If a “source” publisher supplies actual source, for example SGML markup, then the archive will contain information not normally available from a “presentation” publisher. Conversely, the archive of a “presentation” publisher will contain information, for example the CSS and JavaScript implementing the e-journal’s look-and-feel, not normally available from a “source” publisher.

But for supplemental materials, and especially data, this distinction is unlikely to have long-term relevance. This data is frequently neither source to be processed by the publisher’s web infrastructure into a form usable by a web browser nor is it a presentation to be interpreted by a web browser. It is normally raw input to some other program, in particular one different from that used by the original authors.

Thus the only important difference is whether the archive collects the data in the same way that readers would (presentation) or in some form packaged by the publisher (source):

- In the presentation case, the archive’s ingest web crawler must be able to identify those links in an article pointing to supplemental materials and any associated metadata that should be preserved along with the article. For example, a recent article in *Science* [20] illustrates AAAS’ approach to supplemental materials. A link in the *Article Views* sidebar *Supporting Online Materials* points to a landing page [21] describing and linking to a single PDF file with a Materials & Methods section, figures, and tables. The ingest web crawler needs to know that it should follow this chain of links.
- In the source case, the archive’s ingest process must be able to identify in the package form created by the publisher (often a tar or zip archive), the relationship between the article text, any associated components, the supplemental materials, and any associated metadata. For example, a recent article in the *Journal of Monetary Economics* [22] as it appears on the Web has a link near the end of the paper’s text to a single PDF file with supplementary material. In the packaged source format Elsevier uses [12], this PDF appears in the same directory as the PDF, XML, and raw ASCII of the primary article. There is no XML or raw ASCII for the supplement.

Best practices for making these connections that were robust enough to enable similar automatic processing across a range of e-journal publishing technologies would be useful. Weaker best practices would have little effect, either on preservation or other tasks.

The LOCKSS software currently ingests supplemental materials in both presentation and source forms, but only from major publishing platforms such as HighWire Press [23] and Elsevier [12]. In both cases there are one-time and continuing per-publisher costs involved in doing so, but they are not large. We would expect these costs to increase as smaller publishers and smaller publishing platforms increase their use of supplemental materials; effective best practices would reduce the expected increase.

Technical Metadata

We have argued elsewhere [4] that the advent of the Web triggered a switch from documents as private to applications to documents published for many applications, and that this effectively turned document formats into network protocols, which are almost immune from the backwards-incompatible changes that cause format obsolescence. We have also argued [4,18] that the increasing importance of open source has similar effects for similar reasons.

A corollary of these arguments is that the technical metadata provided by the Web (Mime type, magic numbers, etc.) is adequate, since it clearly enables web browsers, including open source browsers, to render the content.

Some of these arguments are weaker when applied to supplemental materials in the form of data. Although it is not the private property of a particular application, it is also less “published” and more dependent on metadata other than the basic web metadata. These considerations raise the importance of publishing supplemental materials in forms that can be accessed by open source tools, such as XML with public DTDs. Best practices codifying this would be useful both for preservation and the kinds of data-mining activities championed by, for example, Peter Murray-Rust [24].

In addition, standards for representing the technical and scientific metadata that supplemental materials need in addition to the basic web technical metadata would be very useful, although the benefits would not accrue primarily to preservation but rather to the eventual users of the preserved materials.

Bibliographic Metadata

There are standards for the attribute names [25] and to a lesser extent for the formats [26] and vocabularies for the bibliographic metadata describing the articles in journals. They are not as well observed in practice as one might hope, but they are useful. Extending them to cope with supplemental materials would be useful, as would best practices stressing the importance of conformance to metadata standards, and tools verifying such conformance.

It is noteworthy that while Elsevier’s source format [12] includes the most comprehensive bibliographic (and technical) metadata about primary articles of any publisher we have worked with, it includes no metadata about supplementary materials except an MD5 digest of the file. It is not even possible to discover from the supplied metadata whether or not an article has supplementary material.

Conclusion

We have identified that one goal of codifying best practices, or even standardization, with respect to supplemental materials should be to reduce the cost of ingest by eliminating tasks needing human intervention. Suggested areas with the potential to do so are:

- An analog of the CC "attribution" license, allowing researchers to grant general permission to use their data provided credit is given.
- A machine-readable form of this license, similar to the XML form of the CC license, allowing automated harvesting of data from supplemental materials.
- Uniform means for connecting articles, their supplemental materials, and the metadata for the supplemental materials, both in e-journal web sites and in the packaged formats used by "source" publishers.
- Standard representations of the metadata needed by supplemental materials in addition to the basic web metadata.
- Publishing data in supplemental materials in forms that can be processed using open source tools.
- Extensions to existing metadata standards and practices to allow for detailed description of supplemental materials.

=========

David S.H. Rosenthal (dshr@stanford.edu) is Chief Scientist and Vicky Reich (vreich@stanford.edu) is the Director at LOCKSS (www.lockss.org). David's blog on digital preservation is available at: blog.dshr.org/.

References

1. Cantara, Linda (ed). Archiving Electronic Journals. Digital Library Federation, 2003. www.diglib.org/preserve/ejp.htm

2. LOCKSS Program www.lockss.org/

3. Roundtable on Best Practices for Supplemental Journal Article Materials January 22, 2010 Washington, DC. Co-sponsored by NFAIS and NISO. www.niso.org/topics/tl/supplementary/4. Rosenthal, David S. H. How Are We Ensuring The Longevity Of Digital Documents. Coalition for Networked Information, April 2009. blog.dshr.org/2009/04/spring-cni-plenary-remix.html

5. Report on Portico Audit Findings. Center for Research Libraries, "January, 2010. www.crl.edu/archiving-preservation/digital-archives/certification-and-assessment-digital-repositories/portico

6. Ithaka tax returns available from: www2.guidestar.org/Home.aspx

7. Eakin, Loraine et al. A Selective Literature Review on Digital Preservation Sustainability. Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2009. brtf.sdsc.edu/biblio/Cost_Literature_Review.pdf

8. Moore, Richard L. et al. Disk and Tape Storage Cost Models. Archiving 2007, Arlington, VA, May 2007, pp. 29-32. www.imaging.org/IST/store/epub.cfm?abstrid=34413

9. Christensen, Clayton M. The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail. Harvard Business School Press, June 1997. ISBN: 978-0060521998

10. Rothenberg, Jeff. Ensuring the Longevity of Digital Documents. Scientific American, 272(1), 1995.

11. Rosenthal, David S. H. Format Obsolescence: the Prostate Cancer of Preservation. dshr's blog, May 2007. blog.dshr.org/2007/05/format-obsolescence-prostate-cancer-of.html

12 ScienceDirect OnSite Product Specifications. Elsevier, March 2006. info.sciencedirect.com/implementation/implementing/sdos/

13. Beagrie, Neal et al. Keeping Research Data Safe. JISC, 2010. www.beagrie.com/KRDS2_selectioncriteria.pdf

14. Key Perspectives, Ltd. Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long term Viability. Digital Curation Centre, January 18, 2010. www.dcc.ac.uk/sites/default/files/documents/publications/SCARP-Synthesis.pdf

15. Harley, Diane et al. Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines. Center for Studies in Higher Education, UC Berkeley, January 2010. cshe.berkeley.edu/publications/publications.php?id=351

16. Nelson, Bryn. Data Sharing: Empty Archives. Nature, 461, pp. 160-163, 2009. doi:dx.doi.org/10.1038/461160a

17. Scholar's Copyright Project. Science Commons. sciencecommons.org/projects/publishing/

18. Rosenthal, David S. H. Format Obsolescence: Scenarios. dshr's blot, April 28, 2007. blog.dshr.org/2007/04/format-obsolescence-scenarios.html

19. ODC Attribution License (ODC-By). Open Data Commons, June 24, 2010. www.opendatacommons.org/licenses/by/1.0/

20. Gibson Daniel G. et al. Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. Science, 329 (5987), pp. 52–56, July 2, 2010. doi:dx.doi.org/10.1126/science.1190719

21. Science magazine supporting material example: www.sciencemag.org/cgi/content/full/sci;science.1190719/DC1

22. Balia, Turan G. and Robert F. Engle. The Intertemporal Capital Asset Pricing Model with Dynamic Conditional Correlations. Journal of Monetary Economics 57, (4), pp. 377-504, May 2010. doi:dx.doi.org/10.1016/j.jmoneco.2010.03.002

23. HighWire is Hosting. HighWire Press. highwire.stanford.edu/about/

24. Murray-Rust, Peter. Data-driven Science - A Scientist's View. NSF/JISC Repositories Workshop, April 10, 2007. www.sis.pitt.edu/~repwkshop/papers/murray.html

25. Dublin Core Metadata Initiative. dublincore.org/

26. Term Name: date *in* DCMI Metadata Terms. DCMI Usage Board, January 14, 2008. dublincore.org/documents/dcmi-terms/#terms-date
E.g: "Recommended best practice is to use an encoding scheme [for dates], such as the W3CDTF profile of ISO 8601."