eBooks: The Preservation Challenge

by **Amy Kirchhoff** (Archive Service Product Manager, JSTOR & Portico)

Shifting from Print to Electronic

Narrow shelves full of books, some new and sparkling, some old and musty, have long been the retreat of undergraduates frantically finishing papers, graduate students searching for the perfect argument in support of their theses, and faculty performing literature reviews. eBooks, however, are starting to make inroads in the purchasing patterns of libraries and individuals. By December 2010, eBooks made up "9 to 10 percent of trade-book sales,"[1] and in the last week of December "about 3 million to 5 million e-readers were activated."[2] By May 2011, **Amazon** was selling "more e-books for the **Kindle** than … print books — by a ratio of 105 **Kindle** books to 100 print books."[3]

As with mass market eBook growth, scholarly eBook publications have seen a measurable increase in sales in 2011, with the percentage of sales from eBooks at one university press going from 1.6 percent in 2010 to 11.3 percent in February 2011 (perhaps attributable to the number of eBook readers given as gifts in the 2010 holiday season).[4] Public libraries are also seeing a dramatic increase in eBook lending: "according to the **New York Public Library**, which has the highest circulating eBook library in the U.S., eBook loans are up 36 percent compared to the same time last year [June 2010]."[5] The academic community has been licensing and becoming dependent on eBooks for years, since before the debut of the first e-reader — the **Sony LIBRIé** — in 2004.

Those narrow shelves of print books are preserved for the long-term due to the conservatorship of a few dedicated libraries and the general ownership of many libraries. Librarians and archivist know much about both the challenges of and solutions for preserving traditional books — for centuries, if need be. What is not so clear is if we even understand the problems entailed in, much less have any solutions for, preserving eBooks for the long haul.

Many individuals, publishers, and libraries have copies of eBooks today, but simply knowing that many copies of electronic content exist does not protect digital content. Long-term protection arises from constant care and attention to the preserved content. Today's eBooks are often tied to a specific piece of software or hardware just to read them or they reside only on the publisher's servers. Even if an individual or library owns the bytes that compose the eBook, it is impossible to move those bytes from one platform to another (and, most libraries and individuals are likely to have licensed eBooks and do not actually own them). To preserve access to eBooks, the intellectual content of the book must be unpacked from its reliance on particular hardware and software and then that content must be securely stowed away and maintained by one or more preservation agencies (such as third party organizations dedicated to preserving digital content, national libraries, or cooperative digital preservation efforts among libraries).

Within the scholarly community, an early expression of the need for robust preservation solutions for digital content was *Urgent Action Needed to Preserve Scholarly Electronic Journals,* a statement endorsed by the **Association of Research Libraries**, the **Association of College and Research Libraries**, and others in 2005.[6] At that time, the consensus of the academic community was that e-journal content was the genre of electronic scholarly publication most in need of preservation. Following this call to action, a variety of reliable long-term preservation arrangements for e-journals emerged, including the e-journal preservation service offered by **Portico**. Since 2005, however, more and more scholarly content has been published in electronic form, including digitized collections, grey materials, research output, government documents, and, of course, eBooks — addressing eBook preservation is a logical next step for the academic community. Library reliance on this material is increasing as the number of published eBooks is growing exponentially.

eBook Specific Preservation Challenges

Digital preservation (whether of e-journals, eBooks, or anything else) is the series of management policies and activities necessary to ensure the enduring usability, authenticity, discoverability, and accessibility of content over the very long term. The key goals of digital preservation include:

- Usability — the intellectual content of the item must remain usable via the delivery mechanism of current technology;
- Authenticity — the provenance of the content must be proven along with its authenticity as a replica of the original;
- Discoverability — the content must have logical bibliographic metadata so that the content can be found by end users through time; and
- Accessibility — the content must be available for use by the appropriate community.

At a base level, one published digital object looks like any other. Every object consists of some metadata and some files:

Some metadata	+	some files	=	Digital Song
some metadata	+	some files	=	Digital Slide
some metadata	+	some files	=	Digital journal article
some metadata	+	some files	=	Digital Book

While eBooks are built from the same building blocks as all digital content, they do present some unique preservation challenges. Three particularly thorny challenges are highlighted below: versions, digital rights management, and metadata.

Books have a history of publication complexity. They have different editions, translations, publishers, publishing runs, sizes, and even different covers. As an exemplar, consider *Anna Karenina*. There are hundreds, maybe thousands, of manifestations of this work: the original manuscripts, the original serial publications in *The Russian Messenger*, the first version published in book form, the many subsequent print editions, the many language translations, the 15+ **Kindle** eBook versions, the 15+ **Nook** eBook versions, the two **Project Gutenberg** eBook versions, and more. In the electronic world, these existing issues are complicated by the ease with which it is possible to make updates or issue retractions on digital content, such that there may be multiple versions of each manifestation. Managing this complexity will be one of the unique challenges of eBook preservation.

Digital Rights Management (DRM) is another challenge for eBook preservation. DRM is technology, often embedded in a file or device, which enforces the rules of use defined by the provider of the content. DRM is particularly prevalent with eBooks, where it is common for books purchased by individuals to be tightly tied to that individual (e.g., it is often difficult to share or lend one's eBook with a friend) or to a particular device (e.g., books purchased for one appliance or application can only be read on that appliance or application). eBooks sold or licensed to public and academic libraries are also wrapped in DRM which limits the number of times the book can be borrowed, the number of users who may borrow it at one time, or even the locations at which it can be read. The purpose of DRM (which is to limit access and replication) increases the complexity of preserving access for the long-term.

Another challenge of eBook preservation is the proliferation of bibliographic metadata at many different levels of the publication. Metadata is neither simple nor straightforward — a publication does not have only an author but an editor, a translator, and so on. eBooks have all the traditional challenges of bibliographic metadata, plus a number of unique considerations. For example, many eBooks within the academic community are delivered a chapter at a time and thus there is chapter-level metadata to be preserved (and perhaps a representation of the book as a whole and as individual chapters must be preserved). In addition, many books, especially within the scholarly community, are part of a series and thus must include metadata placing them within the context of the series or they are one volume in a multi-volume set, where the entire set is the "book." Managing

continued on page 00

this hierarchy of metadata in such a way that preserved eBooks can be accurately delivered in the future is a challenge that differentiates eBooks from e-journals.

Portico's eBook Preservation Solution

Portico is a not-for-profit digital preservation service providing a permanent archive of electronic journals, books, and other scholarly content. **Portico** launched in 2005 with an e-journal preservation service. In 2009, **Portico** ingested the first eBooks into the **Portico** archive as part of an aggregated e-journal and eBook preservation service and fulfilled its first eBook post-cancellation access request in 2010. In 2011, **Portico** began to offer a separate eBook preservation service in order to allow libraries and publishers to select the preservation services best suited to their particular needs. The **Portico** eBook preservation service is modeled after the **Portico** e-journal preservation service; libraries and publishers both contribute to defray the costs of preservation. Publishers commit their current and future eBook holdings to **Portico** for preservation. eBook content is made accessible to all institutions participating in the eBook service in the case of a trigger event: cessation of a publisher's operations, discontinuation of a title by a publisher, removal of back issues or a portion of a title by a publisher, or catastrophic and sustained failure of a publisher's delivery platform. In addition, publishers have the option to designate **Portico** as one of their post-cancellation access (also known as perpetual access) methods to eBooks.

The preservation actions **Portico** takes with eBooks match those of both the **Portico** e-journal and d-collection preservation services. To meet our rigorous definition of preservation — the series of management policies and activities necessary to ensure the enduring usability, authenticity, discoverability, and accessibility of content over the very long term — **Portico** is guided by the following principles:

- Preservation metadata describing the technical and bibliographic natures of the content preserved is gathered as the content is being processed into the archive.
- Preservation must be practical (for example, migration of files to new formats is only done when it is necessary and is not preemptively performed without valid archive management reasons.)
- The **Portico** archive is self-describing and contains sufficient information and documentation to make it possible for a third-party to understand and manage the archive.
- The **Portico** archive is a dark archive, but transparency to participants is required. To that end, **Portico** provides audit privileges to participants and regularly reports on content in the archive.
- The preserved content is replicated to multiple on-line and off-line locations on multiple continents.
- The preserved content is regularly checked for bit rot and corruption and any problems are immediately corrected.
- The hardware on which and machine rooms in which the preserved content is located must be maintained to industry standards.
- **Portico** receives accreditation — **Portico** was certified as a trusted, reliable digital preservation solution by the **Center for Research Libraries (CRL)** in 2010.

As of June 2011, **Portico** has over 5,000 eBooks preserved from four publishers and over 100,000 eBooks committed to the archive from twelve publishers.

Conclusion

Given the dramatic increase in publication and sales of eBooks and the growing reliance of the academic community on eBooks, the moment has arrived to address the preservation needs of eBooks. The preservation of eBooks may be met in numerous ways, including preservation through community supported independent archives such as **Portico**, national preservation efforts, or cooperative efforts among like-minded institutions. While eBooks have many unique challenges, if the community begins to preserve the entirety of eBooks right now, those challenges can be addressed over time.

Endnotes

1. **Julie Bosman**, "Christmas Gifts May Help E-Books Take Root," *New York Times* (December 24, 2010). Retrieved online June 19, 2011: *http://www.nytimes.com/2010/12/24/books/24publishing.html?sq=ebook%20sales%20growth&st=cse&adxnnl=1&scp=1&adxnnlx=1304450152-aI+T0nqS3a0m21fXCL2dKg*.

2. **Alex Knapp**, "What Do Amazon's E-Book Sales Mean for the Future of Books?" *Forbes* (May 19, 2011). Retrieved online June 19, 2011: *http://blogs.forbes.com/alexknapp/2011/05/19/what-do-amazons-e-book-sales-mean-for-the-future-of-books/*.

3. **Steve Kolowich**, "The E-Reader Effect," *Inside Higher Ed* (June 1, 2011). Retrieved online June 19, 2011: *http://www.insidehighered.com/news/2011/06/01/e_books_becoming_a_greater_priority_of_university_presses_in_the_age_of_ipad_and_kindle*.

4. **Bob Minzesheimer** and **Carol Memmott**, "Week after Holidays, E-book Sales Outdo Print," *USA Today* (January 5, 2011). Retrieved online June 19, 2011: *http://www.usatoday.com/life/books/news/2011-01-05-1Aebooksales05_ST_N.htm*.

5. **John R. Quain**, "The Real Force Behind Ebook Sales: Heaving Bosoms," *FoxNews.com* (June 14, 2011). Retrieved online June 19, 2011: *http://www.foxnews.com/scitech/2011/06/14/force-behind-kindle-nook-ebook-sales-heaving-bosoms/#ixzz1PSEZpdP6*.

6. **Donald J. Waters** (ed.), Urgent Action Needed to Preserve Scholarly Electronic Journals (October 15, 2005). Retrieved online June 19, 2011: *www.diglib.org/pubs/waters051015.pdf*.