This paper was presented at the panel "(1-46) Creating East European and Eurasian Digital Repositories: Archiving for Preservation and Access," 40th Annual AAASS National Convention, held on November 20–23, 2008, in Philadelphia, Pennsylvania, USA.

Long-term Preservation versus Innovation: The Challenge for Producers of Digital Content
Kevin S. Hawkins

Introduction

Once upon a time, there was a clear division of labor between publishers and libraries in questions of who developed and preserved content. Publishers produced books and other printed matter, and libraries preserved them.

In this traditional model, the print codex is a discrete, fixed medium. Ideally, publishers offer hardback binding and use acid-free paper that won't yellow or tear, but in any case, they have no role in preserving the end product once it is safely distributed to libraries. Libraries, on the other hand, are experts at maintaining their print collections. They know optimal climate conditions and conservation techniques, and they have clear privileges in copyright law allowing them to make preservation copies in certain circumstances. But more importantly, each item in the collection is discrete and fixed, limiting the scope of any preservation work.

However, times have changed for producers of content now that so much of it is distributed online. All sorts of people—not just "publishers" but also libraries and individuals—are putting material online. The reality of the Internet's architecture is such that the content distributor, rather than a library, ends up maintaining the content. Furthermore, digital preservation requires more regular human intervention than preservation of print material. For these widely known reasons, preservation of digital content is not as straightforward as preservation of print content. What is discussed less often, however, are the constraints on innovation that the needs of digital preservation present to producers and distributors of online content.

Preservation versus innovation in data formats

The basic issues in preservation of digital data are familiar: choosing open, standard file formats that are not tied to particular software and that preserve as much information as possible; refreshing data by transferring to newer or more durable media; and migration of data to better or more current file formats. No library wants to be stuck with the digital equivalent of Betamax tapes.

Some digital repositories take a tiered approach to preservation of digital objects, with varying levels of commitment to maintaining files depending on the format of the file submitted by the content creator.[1] Distributors of digital content may take a similar approach with those who create or curate their content: for example, the Scholarly Publishing Office of the University of Michigan University Library (where I work) pledges to maintain all image and text content stored and delivered in its digital library architecture, but any content outside of this architecture—multimedia works, custom databases, and accessory information about a publication—is only hosted with the understanding that it may not be possible to maintain this content indefinitely.

Such a tiered approach creates a constraint on innovation. As long as digital repositories and online publishers coerce content creators into using standard file formats

---

[1] See, for example, "Deep Blue Preservation and Format Support Policy" (http://deepblue.lib.umich.edu/about/deepbluepreservation.jsp).

and default access systems, these content creators temper their innovative uses of technology in order to make digital preservation manageable.

Preservation versus innovation in access systems

However, it's not just decisions about the content itself that are important: care must also be taken in choice or design of an access system. How will the content be presented to the user for browsing and searching? Individual webpages, even those with valid code, are difficult to update en masse and prone to editing mistakes. A CMS or digital library platform, on the other hand, dynamically generates pages and offers interactivity not possible through individual pages. It's important not only to choose an access system with promise of longevity (a developer community, open-source code that works across platforms, and institutional support are all good signs), but you also need to make sure that your implementation of this infrastructure will be maintainable. If you customize the software, how easily will these customizations carry over to new versions of the software? While a knowledgeable programmer working with well-designed software can minimize future migration difficulties, you usually can't predict whether and how the software's developers will re-architect it for future versions.

In short, if you design your own access system, you need to be fully prepared to maintain it. If you use someone else's software, the more you customize it—whether for all your content or just for a portion of it—the more you will need to maintain the code through future upgrades. In either case, the inescapable need to maintain an access system is a significant constraint on customization and, more broadly, on innovation. Everything from custom skins to new features may cause problems in the future.

The tension between innovation and preservation

How significant is this tension between the desire to innovate and the constraints on customization based on the need to maintain the content? Simply put, you have to make a tradeoff. The less responsible the content distributor feels for maintenance of the content, the less likely they will allow innovations desired by those creating or curating that content. But the more responsible the content distributor feels for maintenance, the more likely that the content will, in fact, be preserved intact for future generations, subject to "weeding" of digital content based on use and perceived future value.

Who will preserve your content? Since today's publishers have to maintain their own content, we can't simply say that publishers don't care about preservation. Nor would it be fair to say that libraries *only* care about preservation: they understand the need to offer innovative services. If a library closes, its online collection will likely be absorbed by another institution. If a publisher is purchased by another publisher, its online collection will likely be preserved as well. But if a publisher simply goes out of business or decides that no more revenue can be generated from its older content, we can only hope that an arrangement will be made for a library to take over maintenance of this content. Better yet, these arrangements can be made ahead of time: two notable partnerships are e-Depot[2] and CLOCKSS.[3] Still, any broad preservation effort like these will inevitably have a limited preservation scope in order to minimize the need for human intervention in the preservation process.

---

[2] www.kb.nl/e-depot
[3] http://www.clockss.org/

Who offers content creators the opportunity to innovate? Libraries understand economics of scale, whether it be in technical services procedures or in managing inexhaustible user demand. But publishers also make compromises in order to keep publishing costs down. The most innovative content distributors end up being those with the greatest resources, not just to pay for the work of innovation now but also to preserve it in the future.

Weighing the options

Distributing content on your own and not through an institution of some sort gives you maximum flexibility in innovation but puts your content at the greatest risk of not being preserved unless you make arrangements for its stewardship in the future. While there are admirable projects like the Internet Archive that attempt to address such at-risk content, they are limited in their ability to migrate content to new file formats and essentially end up offering the lowest tier of a preservation pledge: preservation of a bitstream.

Distributing content through a library will ensure long-term preservation but limit innovation; however, the commitment of preservation and the constraints on innovation vary depending on the resources of and support for the institution.

Distributing content through a publisher may provide more opportunities for innovation, but there is a risk that content will be lost, especially if there is no preservation arrangement, but possibly also if there is one.