Building Digital Collections through Partnerships: Brigham Young University's Harold B. Lee Library and the Internet Archive

Elizabeth Smart; Scholarly Communications Librarian, Harold B. Lee Library, Brigham Young University; Provo, UT, USA

Abstract

Mass digitization requires extensive budgetary, staff and technical resources. Brigham Young University's Harold B. Lee Library has partnered with the Internet Archive to share goods and services in an arrangement that maximizes the assets of each institution. While the Harold B. Lee Library supports a large staff of student employees and extensive library collections, the Internet Archive specializes in large-scale image processing, publishing and hosting. By combining these strengths, the Harold B. Lee Library has scanned ten times as much printed material in the last 18 months as it had scanned in the previous ten years combined.

The efficiency of the Internet Archive digitization process presents challenges in keeping a full scanning queue, but the Library continues to make careful selections by relying on subject specialist expertise and patron interest.

Background

Since June 2009, Brigham Young University's Harold B. Lee Library (HBLL) has worked in partnership with the Internet Archive (IA) to scan library holdings for public access in the Internet Archive's digital library (http://www.archive.org). Founded in 1996 as a nonprofit organization, the Internet Archive aims to provide researchers, historians, scholars and the public with persistent access to digitized historical collections [1]. While many libraries send microfilm, monographs and serials to IA Scanning Centers for digital imaging on the Internet Archive's Scribe scanning stations, HBLL is one of very few academic libraries working with the Internet Archive in a "no-cash model" partnership. This partnership represents an exchange of goods and services: the Internet Archive provides HBLL with Scribes, relevant software, post-scan processing and the digital repository for scanned materials. In exchange, HBLL provides the staff to operate the Scribes for a minimum of 40 hours per week as well as the content for scanning.

Collaborative Workflow

In order to maintain scanning productivity of 25,000 – 30,000 pages per week, HBLL has established an "all hands" approach for selecting content, preparing materials and digital imaging that involves multiple departments across the library. Participating staff include:

- Subject librarians
- Curators
- Archivists
- Book repair personnel
- Catalogers
- Copyright staff
- Digital imaging specialists

The key to success has been distributing elements of the workflow across the library and integrating these tasks with existing library processes. The Digital Imaging Lab coordinates the effort. This distributed model encourages solutions from specialists working close to the problem. For example, discovering a volume with missing pages before it arrives in the Digital Imaging Lab puts the responsibility on the subject selector (or his/her student employee) to find a complete copy, allowing the employees in the Lab to focus on digitization.

Selecting Materials

Materials are selected for digitization through a variety of methods.

Subject Librarians, Curators and Archivists

Subject specialists select materials for digitization that support curriculum, research and patron interest. These content selectors also consider the uniqueness of material, the demand for material, the condition of the physical materials and the prevalence of existing digital copies. Examples of materials selected by subject librarians, curators and archivists include university yearbooks, university course catalogs, Victorian serials, local histories and papyrology reference sources.

Much of this content is organized within distinct Internet Archive sub-collections, which allow users to search within specific groups of content. However, the number of sub-collections HBLL can define is not unlimited. Subject specialists will be required to define categories that are broad enough to encompass a range of content, but narrow enough to be meaningful to users.

Generally, HBLL scans materials that are in the public domain. In those instances where materials are under copyright, HBLL requests a release from copyright holders and documents that release with a form developed by BYU's Copyright Licensing Office.

Book Repair

Brittle books in the public domain that are routed to Book Repair through the circulation desk are considered for digitization as an alternative to boxing or physical reformatting (copying).

Patron-Driven Selection

HBLL is in the process of establishing two patron-driven scanning streams. The first content stream is derived from scanning all printed materials requested in the L. Tom Perry Special Collections Reading Room. One of HBLL's three IA Scribes will be dedicated to digitizing these materials.

The second stream of patron-driven content will capture books that patrons check out from the circulating collection that are clearly in the public domain (published before 1923). These materials will be routed to the Digital Imaging Lab upon returning to the library.

Material Preparation

As materials are pulled from the shelves for digitization, we ask that content selectors review the materials in the same way they would if they were preparing materials for microfilming or binding by:

- Flagging missing pages and foldouts
- Selecting the best copy for scanning
- Confirming that materials are in the public domain, or that we have permission from the copyright holder to digitize

Concurrent with preparing materials for scanning, content selectors create a pick list representing all items selected for digitization. The pick list spreadsheet includes these standard fields recommended by the Internet Archive:

- Title
- Author
- Publication date
- Catalog ID number

HBLL requires an additional field to identify the disposition of the materials after digitization (return to stacks; withdraw; auxiliary storage, move to Special Collections, etc.). As we scan more fragile materials, we anticipate adding a field to identify whether digital preservation measures are required for specific digital objects.

Digitization

After the materials have been selected and prepared for scanning, the HBLL Digital Imaging Lab uses the pick list to identify the catalog record for each item to be scanned and transfers the catalog record into the Internet Archive digital library.

Using the Scribe workstation, student employees in the Digital Imaging Lab work in two-hour shifts to digitize the items selected. We have found that limiting shifts on the Scribe to two hours keeps students more engaged in the scanning process. Limiting time at the Scribe also gives each student a chance to participate in other areas of the process, from quality assurance to republishing.

After digitizing, materials are returned to the circulation desk for re-shelving. Materials to be withdrawn (at the request of the content selector) are routed through traditional withdrawal procedures.

Cataloging

After digitizing and republishing are complete, the Digital Imaging Lab forwards each pick list to the Special Collections and Formats Catalog Department. This department creates a master and an institutional provider-neutral e-monograph OCLC record for each digitized book added to the Internet Archive. The institutional e-monograph records are added to the HBLL catalog.

These changes make newly digitized IA assets visible in the HBLL catalog. By extension, the materials are also visible in HBLL's installation of the Ex Libris search product, Primo, known to our users as ScholarSearch (http://lib.byu.edu/).

Productivity

Scribe workstations are generally in operation on weekdays from 8:00 a.m. to 9:00 p.m. Since June 2009, HBLL has scanned more than 4,600 items. These items represent more than 1.5 million pages, with an average daily output of 5,000 pages. Total downloads of these materials exceeds 127,000. The most popular items are from our family history collection (e.g. parish registers, pedigrees) and University Archives (e.g. yearbooks). As the partnership continues in its second year, HBLL is installing two additional Scribes for a total of three workstations.

Though production rates consistently exceed HBLL's weekly goals, the Digital Imaging Lab continues to develop methods to further increase output on the Scribes. One successful approach has been monitoring productivity on a shift-by-shift basis. As each Scribe operator begins a shift, he or she is given a specific page count to complete. The Digital Imaging Lab supervisor monitors progress throughout the day and adjusts page-count goals for each shift. Breaking down expectations into shift-level goals gives each operator a stake in achieving the week's goal.

The Digital Imaging Lab is developing a second approach to increasing productivity and accuracy by pinpointing specific re-training needs. The Lab will use the detailed quality assurance reports provided by the Internet Archive to identify the most frequent scanning errors. Additional training will focus on these common errors, and can be targeted to specific operators.

Success in Special Collections

Though HBLL has scanned a significant quantity of materials from its circulating collection, the positive impact of managing a Scribe in-house is felt most deeply in Special Collections. Operating a workstation that quickly scans materials and facilitates their use online within 24 – 48 hours mitigates concerns about sending rare materials to a third party for scanning, or concerns about keeping vault materials off the shelf for an extended period. Keeping scanning operations on-site protects the Library's collection of rare and high-value materials from the extra handling required for packing and shipping to an off-site location.

Another benefit is being able to fulfill inter-library loan requests for Special Collections materials through digitization. HBLL has also scanned (with permission) ILL materials requested from other libraries to support campus curriculum.

Library exhibits comprised of materials from Special Collections have a deep online presence because items on display are represented by full-length digital surrogates on exhibit websites.

HBLL's current IA scanning strategy excels in reformatting print monographs and serials. A potential future challenge will be making adjustments in metadata creation to expand HBLL's IA

scanning operations to include manuscript and photograph collections.

References

[1] Internet Archive. “About the Internet Archive”17 March 2011. http://www.archive.org/about/about.php.

Author Biography

Elizabeth Smart is the Scholarly Communications Librarian at Brigham Young University’s Harold B. Lee Library. She earned an MLS from the University of Maryland and also holds an MA in American Studies from Utah State University and a BA in Humanities from Brigham Young University.

Thanks to Robert Miller, Director of Books at the Internet Archive; Marisa Snyder, Digital Imaging Lab Supervisor at the Harold B. Lee Library; and the original Internet Archive Project Team at the Harold B. Lee Library.

