November 2009

One Book, Twelve Hundred Downloads: How the LYRASIS Mass Digitization Collaborative Is Bringing the University of Scranton's Out-of-Print Books Back to Life

Kristen Yarmey-Tylutki

University of Scranton, yarmeyk2@scranton.edu

Follow this and additional works at: <https://docs.lib.purdue.edu/atg>

Part of the [Library and Information Science Commons](#)

Recommended Citation

Yarmey-Tylutki, Kristen (2009) "One Book, Twelve Hundred Downloads: How the LYRASIS Mass Digitization Collaborative Is Bringing the University of Scranton's Out-of-Print Books Back to Life," *Against the Grain*: Vol. 21: Iss. 5, Article 9.
DOI: <https://doi.org/10.7771/2380-176X.2645>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

One Book, Twelve Hundred Downloads: How the LYRASILS Mass Digitization Collaborative Is Bringing The University of Scranton's Out-of-Print Books Back to Life

by **Kristen Yarmey-Tylutki** (Digital Services Librarian, The University of Scranton Weinberg Memorial Library)
<yarmeyk2@scranton.edu>

In June 2008, **PALINET**, the Mid-Atlantic regional consortium of cultural heritage organizations, announced that it had won a \$1 million grant from the **Alfred P. Sloan Foundation** to create a collaborative digitization service for its members. **PALINET** partnered with **Internet Archive**, a nonprofit organization that aims to provide "permanent access... to historical collections that exist in digital format,"¹ and set a goal of digitizing 60,000 books. Just over a year later, **PALINET** (now **LYRASILS**, after a merger with **SOLINET**) is well on its way to meeting that goal, with 33 institutions signed on to participate in its Mass Digitization Collaborative and over 6,200 items digitized.

A participant from the beginning, **The University of Scranton Weinberg Memorial Library** witnessed the growth of the Collaborative from a thirteen member mini-pilot to its current production phase. Along the way, the Library digitized 169 volumes with **Internet Archive** and has watched their download counts grow by the day — giving these books, most of which are out-of-print, new life in a digital world.

The **Weinberg Memorial Library** is a mid-sized academic library, serving a population of approximately 5,000 students. In order to provide digital access to Special Collections materials, the Library's Digital Services department scans photographs, slides, negatives, and documents in-house. In the summer of 2008, we were exploring the possibility of purchasing a book scanner and training our staff to digitize part of our book

collection, but had not found it to be cost-effective. The **PALINET** announcement, then, caught our attention; the Mass Digitization Collaborative seemed a solution for providing digital access to our Special Collections books in an affordable way.²

At the end of August 2008, with an invitation from **PALINET**'s Digital Services Manager, **Laurie Gemmill**, the **Weinberg Memorial Library** joined the Mass Digitization Collaborative's mini-pilot. **PALINET**'s goal in running this pilot was to test out their plans on a small group of diverse participants and solicit feedback on all aspects of the project before opening the Collaborative to all members; on our part, the **Weinberg** team wanted to participate at this early stage in order to contribute to the refining of the Collaborative's structure and workflow.

Throughout the next month, the thirteen pilot participants, including academic and public libraries of various sizes, a historical society, and a museum, met five times (via conference call) to discuss the program's procedures and report on progress. We were each tasked with selecting five to ten books or microfilm reels for digitization at one of **Internet Archive**'s regional scanning centers in Indiana, New Jersey, or New York.

Selection was a more difficult task than anticipated. **PALINET** encouraged pilot participants to digitize "unique items of regional interest"³ and to select either items in the public domain or items for which permission to digitize

could be obtained from the copyright holder, since the digital images would be available to the public on **Internet Archive**'s Website. **The University of Scranton**'s selection team (made up of the Special Collections Librarian, the University Archivist, and myself as the Digital Services Librarian) further narrowed down our options by ruling out any books that were already available to the public in digital format. In examining our collection, we were surprised to discover that several of the books that we had thought were relatively unique had already been digitized by other institutions and were publicly accessible via either **Internet Archive** or **Google Books**. We eventually chose six books on the history of Northeastern Pennsylvania, because of their appeal to our University community as well as our local Scranton community. Three were out-of-print; all six were in the public domain.

Since **Internet Archive** pulls descriptive metadata from library catalogs, our next step was to open a Z39.50 connection in order to share our MARC records. This process was time-consuming and was not completed for approximately three weeks; thankfully, the connection only had to be set up once. (Note: Since the **LYRASILS** pilot, **Internet Archive** has changed this process so that only large collections utilize a Z39.50 connection; smaller collections provide metadata via a customized spreadsheet tool called "WonderFetch.")

With books selected and our Z39.50 connection functioning, we were finally ready to digitize. While Collaborative participants can ship books to an **Internet Archive** regional scanning center in whatever manner they see fit, **The University of Scranton** chose to send our Special Collections Librarian to personally drop off our first batch of books at the **Princeton Theological Seminary** site in Princeton, New Jersey, in order to view the facility and observe the digitization process. **Internet Archive** employs "scribe" scanning stations, where books are placed on a cradle under a glass platen. Two digital single lens-reflex cameras photograph the recto and verso pages of the book, and then the scanner operator lifts the glass platen with a foot pedal and turns the page by hand. To keep file sizes reasonable, **Internet Archive** captures full color images not in TIFF but in JPEG2000 format. An ISO standard, JPEG2000 is used by the **Library of Congress**⁴ and the **Harvard University Library**,⁵ among others, but is still somewhat controversial among digital preservation experts who are concerned about the ease of future data migration for a format that has not yet been widely adopted.⁶ Books are scanned at between 300 and 500 dpi, depending on the

continued on page 20

Dr. Jekyll ...
from page 16

surprised. He sometimes sells rarer books that are in terrible condition, but he describes them as such. He's received a few negative and neutral evaluations, but he would have adjusted the price or refunded the purchase if the buyer had contacted him first. Even a careful examination can sometimes miss faults or hidden problems, especially in media. Hyde has decided to honor the principle that the customer is always right to the extent that he's reshipped books for free when the buyer gave the wrong address and given refunds when customers mistakenly ordered the wrong item.

To conclude, Mr. Hyde has sold a reason-

able number of books to libraries of all types — academic, public, school, and special. To keep his anonymity, Dr. Jekyll is hesitating whether to tell one of his best friends that her library bought one of Hyde's books. On a final note, Mr. Hyde would like me to tell you his vendor name so that you might order some of his solid academic titles, but Dr. Jekyll convinced him that discretion is the wiser choice. 🍄

Endnotes

1. **Holley, R. P.**, et. al., The effect of the Internet on the out-of-print book market: Implications for libraries. *Library Collections, Acquisitions, and Technical Services* v. 29 no. 2 (2005) p. 118-39.

One Book, Twelve Hundred ... from page 18

size of the book, but larger format foldouts are scanned at lower resolution, also depending on size (for example, an 18" x 24" foldout would be scanned at around 182 dpi). Foldouts larger than 18" x 24" are scanned with a multiple shot approach and then stitched together digitally. Text content on each page is captured via optical character recognition (OCR) using ABBYY 8.0.⁷

About fourteen business days after we dropped off our books, **Internet Archive** completed the digitization process and posted our books to their Website. Users and contributors can browse the books directly on the Web using a flip book viewer but can also download the books in PDF, JPEG2000, and DjVu formats. While no available format is specifically optimized for print-on-demand, public domain books on **Internet Archive** can be printed from **Espresso Book Machines** via **On Demand Books'** proprietary **EspressNet** software.⁸ As **Internet Archive** has expressed interest in print-on-demand in the past,⁹ other options for convenient printing will likely emerge in the future.

To judge the quality of digitization, I conducted a page-by-page review of all six books, looking at both JPEG2000 and PDF images. While compression artifacts were clearly visible on the PDF versions, the JPEG2000 images were much more readable and showed little evidence of compression. The quality of the OCR text was excellent, making the content of our books full-text searchable for the first time. My only quality concern was for foldouts. Because **Internet Archive** scans foldouts at a lower resolution than ordinary book pages, small text on some larger foldouts became difficult or at times impossible to read. However, outside of this single concern, the digital versions of the books were of high quality. Considering the affordable cost of digitization, as well as **Internet Archive's** intent to host and maintain a digital copy of each book in their collection in perpetuity, we felt that the digital product was a good return on investment, making our participation in the Collaborative worthwhile.

In October 2008, then, when the **PALINET** Mass Digitization Collaborative shifted into production mode, **The University of Scranton** went along for the ride.

Since joining the production phase, we have digitized another 163 books, including a series of *Pennsylvania Department of Mines* annual anthracite mining reports from 1873 through 1921 (all of which are out-of-print) as well as the University's own published undergraduate and graduate course catalogs.¹⁰ In the meantime, we have carefully monitored the number of downloads for each book from **Internet Archive**. **Internet Archive's** reporting mechanism provides only a count of how many times each book was downloaded, so we are unable to tell how many unique IP addresses have downloaded each book, or track when or from where each book was downloaded, but we can at least use this single statistic to get a ballpark idea of how much our digital books are being used.

While some of our catalogs have only been downloaded a handful of times, almost all of our book download counts greatly exceeded our expectations. Forty-nine of our books have been downloaded over fifty times each in the past few months, and we have excitedly watched as two of the local history books we digitized in our first pilot batch have each racked up over 1,000 downloads. Most impressively, the out-of-print *Prominent Men: Scranton and Vicinity, Wilkes-Barre and Vicinity; Pittston, Hazleton, Carbondale, Montrose and Vicinity, Pennsylvania*, written by **Dwight J. Stoddard** and published in 1906 by **Scranton's Tribune Publishing Company**, has been downloaded 1,288 times as of September 2009, less than one year after digitization. Comparatively, this same book in physical format is used only a few times each year by researchers in our University Archives and Special Collections.

Indeed, as the **Weinberg Memorial Library** contributes more books to **Internet Archive** via the Mass Digitization Collaborative, we are effectively increasing our user base from a small cluster of patrons who can visit our Archives and Special Collections in Scranton, during regular business hours, to the millions of visitors to **Internet Archive**, who access the site from anywhere, at anytime. Out-of-print books that on a regular day sit on our shelves, protected but untouched, are revitalized and reinvigorated, useful once more to a digital community.

If your library would like to join **The University of Scranton** and the 32 other members of the Mass Digitization Collaborative in

bringing out-of-print books back to life, attend a **LYRISIS** information session or request a quote at <http://www.lyrasis.org/massdig>. Both **LYRISIS** members and non-members are welcome to participate. 🍁

Endnotes

1. **Internet Archive**. "About the Internet Archive." (2009): <http://www.archive.org/about/about.php>.
2. As fees may change after this article's publication, please contact **LYRISIS** at <http://www.lyrasis.org/massdig> for pricing information.
3. **PALINET**. "Mini Pilot Participant Process Guidelines," draft of August 19, 2008. This document was distributed in PDF format to participants in the Mass Digitization Collaborative mini-pilot.
4. **Library of Congress**. "Meeting the Challenge: Setting Standards (JPEG2000)." (2008): http://www.digitalpreservation.gov/library/challenge/formats2_challenge.html.
5. **Harvard University Library**. "Highlights from FY 2006 — Enhancements to Library and User Services." (2006): <http://publications.hul.harvard.edu/ar0506/highlights-from-fy-2006-enhancements-to-library-and-user-services.html>.
6. **Lowe, David and Michael J. Bennett**. "A Status Report on JPEG 2000 Implementation for Still Images: The UConn Survey." *UConn Libraries Published Works Paper* 19 (2009). http://digitalcommons.uconn.edu/libr_pubs/19.
7. **LYRISIS**. "Internet Archive Book Digitization and Quality Assurance Processes," draft of March 31, 2009. This document is available in PDF format to participants in the **LYRISIS** Mass Digitization Collaborative. See also: **Bullock, Dave**. "The Internet Archive Keeps Book-Scanning Free." *Wired* (2008): http://www.wired.com/entertainment/theweb/multimedia/2008/03/gallery_internet_archive.
8. **On Demand Books**. "FAQ." (2009): <http://www.ondemandbooks.com/faq.htm>.
9. One of **Internet Archive's** past projects was the Bookmobile, a traveling van that used print-on-demand to pass out free books across the country. See: **Internet Archive**. "The Bookmobile." (2008): <http://www.archive.org/texts/bookmobile.php>.
10. The **University of Scranton's** digitized books may be viewed on **Internet Archive** at http://www.archive.org/details/university_scranton.

Rumors from page 12

Learned from the latest *Charleston Report* that **Bloomsbury** has made selected titles available to library users via the digital platform **Exact Editions**. Libraries are able to subscribe to individual bookshelves, including the Book Group, a children's shelf, a sports shelf, one devoted to Shakespeare and a reference books shelf. Libraries will be charged £100 per 100,000 library users for accessing each shelf, with a minimum fee of

£250. The books are readable from the **Exact Editions** Website. This was initially reported in **Bookseller.com**.

<http://www.thebookseller.com/news/82925-bloomsbury-unveils-digital-shelf-for-libraries.html>

And, did you know that the awesome **Daryl Rayner** <darylrayner@gmail.com> is one of the people behind **Exact Editions**? **Daryl** used to write **Rumours from Paddington** for *Against the Grain* many years ago and we still miss her!

www.exacteditions.com/

Had lunch the other day with one of my most favorite people, **Anne Kabler Robichaux**, librarian emeritus at the **Medical University of South Carolina**. **Anne** and her husband **Ken** had just returned from Dallas for a "roast" of **Bill Leazer**. Remember him? **Bill** used to be with **Majors** and then was hired by **EBSCO** as a consultant. **Bill** was turning 80 which was the reason for the roast but it's hard to believe! Congrats, **Bill**, and **HAPPY BIRTHDAY!**

And speaking of **EBSCO**, **EBSCO**

continued on page 32