Did We Scan That Book Twice?: Weeding the Texas Tech Digital Dark Archive

Heidi Winkler, Digital Stewardship Librarian, Texas Tech University Libraries

Introduction
When the Digital Stewardship Librarian was installed in the Library’s Digital Resources Unit in October 2013, one of her first projects was to determine how much overlap existed between seven of Texas Tech University’s digitized book collections and the holdings of the HathiTrust Digital Library, of which Texas Tech had just become a member. Most of these TTU collections had been commissioned by several subject liaison librarians during a period before the creation of the Digital Resources Unit, when they were encouraged to “sponsor” collections of scanned books relevant to their areas; these included sociology, psychology, Texas history, and human sciences. The books had been scanned from the Texas Tech Library stacks and were made accessible through the Library’s CONTENTdm and later our TDL-hosted DSpace instances. Also, because of a number of digital preservation scares over the years, the PDF and TIFF files of these collections were stored on a server in a dark archive. While Texas Tech has made great strides in digital preservation over the years, this dark archive cannot hold a candle to a Trustworthy Repository like HathiTrust. Why duplicate their work? By clearing out space on the dark archive server, the Digital Stewardship Librarian could make room to preserve more TTU-unique files, such as the ever-growing ETD collection.

Physical vs. Digital Weeding
Librarians are taught to weed items from their physical collections when those items are outdated or unnecessary, ugly or worn, or when the materials no longer fit the needs of their patrons. Print copies of academic journals also face weeding when libraries can afford quality electronic copies. We had to develop different criteria for evaluating a collection that was already off the shelves and online:

Weeding the TTU digital collections:

• Not an evaluation of usage or age of the materials. Instead, we focused on rarity: do we hold the only copy of this material?
• Focused on replicated content. For most materials, rather than get into the minutiae over whether we had the 1892 edition and HathiTrust the 1891, for example, we tried to gauge whether the content was essentially same; we weeded those copies.
• Public access would still be maintained. We are continuing to provide open access to the PDFs through DSpace. From the patrons’ perspective, nothing has changed. Instead, we removed the preservation copies – the scanned, uncompressed TIFF files and archived PDFs – from the dark archive. We would prefer to apply this level of digital preservation to rarer digital items unique to TTU.

Methods
• Combined the collections’ respective “manifests” into one long list, creating 391 individual records. These records included novels and plays, textbooks, elementary school readers, records from the Texas House and Senate, and cookbooks.
• Checked each record one-by-one against the Hathi catalog, which is open to all. Checked for content rather than quality of scan or, in the case of novels/plays, exact matching edition.
• In the process of reading these items for deletion, established two new procedures for the maintenance of the archive: an archive change log and a Ready-to-Delete folder. Recorded any deletions in the change log, to track the movement of items and better establish what was done when and why. The Ready-to-Delete folder is a purposeful bottleneck that holds items for one last check before formal deletion and also to keep track of how much is being removed.
• In December 2013, deleted ~105 GB from the dark archive server.

Discussion
We, in fact, had scanned that book twice. To be more specific, we discovered that we scanned at least one textbook twice for two different collections that had been “sponsored” by the same subject librarian. It was just one example of why the careful curation that can only come with a focused digital unit was so vital for the digital collections at TTU. In this way, we have been able to narrow our mission in order to make the most effective use of limited financial and human resources. The HathiTrust’s mission is to preserve and make accessible the broad cultural and academic record, and they do it much better than Texas Tech University ever will. We will soon be exploring how to submit as many of the remaining pieces of these collections as possible to the HathiTrust, in order to help fill out their collections and to finally weed out those items from our dark archive. In this way, the Texas Tech University Library’s Digital Resources Unit and its dark archive can focus on the preservation of those materials truly unique to the university and its research mission.

References:
1 HathiTrust Digital Library. [http://www.hathitrust.org/]

The Weeded Collections

Physical vs. Digital Weeding

Librarians are taught to weed items from their physical collections when those items are outdated or unnecessary, ugly or worn, or when the materials no longer fit the needs of their patrons. Print copies of academic journals also face weeding when libraries can afford quality electronic copies. We had to develop different criteria for evaluating a collection that was already off the shelves and online:

Weeding the TTU digital collections:

• Not an evaluation of usage or age of the materials. Instead, we focused on rarity: do we hold the only copy of this material?
• Focused on replicated content. For most materials, rather than get into the minutiae over whether we had the 1892 edition and HathiTrust the 1891, for example, we tried to gauge whether the content was essentially same; we weeded those copies.
• Public access would still be maintained. We are continuing to provide open access to the PDFs through DSpace. From the patrons’ perspective, nothing has changed. Instead, we removed the preservation copies – the scanned, uncompressed TIFF files and archived PDFs – from the dark archive. We would prefer to apply this level of digital preservation to rarer digital items unique to TTU.

Methods
• Combined the collections’ respective “manifests” into one long list, creating 391 individual records. These records included novels and plays, textbooks, elementary school readers, records from the Texas House and Senate, and cookbooks.
• Checked each record one-by-one against the Hathi catalog, which is open to all. Checked for content rather than quality of scan or, in the case of novels/plays, exact matching edition.
• In the process of reading these items for deletion, established two new procedures for the maintenance of the archive: an archive change log and a Ready-to-Delete folder. Recorded any deletions in the change log, to track the movement of items and better establish what was done when and why. The Ready-to-Delete folder is a purposeful bottleneck that holds items for one last check before formal deletion and also to keep track of how much is being removed.
• In December 2013, deleted ~105 GB from the dark archive server.

Discussion
We, in fact, had scanned that book twice. To be more specific, we discovered that we scanned at least one textbook twice for two different collections that had been “sponsored” by the same subject librarian. It was just one example of why the careful curation that can only come with a focused digital unit was so vital for the digital collections at TTU. In this way, we have been able to narrow our mission in order to make the most effective use of limited financial and human resources. The HathiTrust’s mission is to preserve and make accessible the broad cultural and academic record, and they do it much better than Texas Tech University ever will. We will soon be exploring how to submit as many of the remaining pieces of these collections as possible to the HathiTrust, in order to help fill out their collections and to finally weed out those items from our dark archive. In this way, the Texas Tech University Library’s Digital Resources Unit and its dark archive can focus on the preservation of those materials truly unique to the university and its research mission.

References:
1 HathiTrust Digital Library. [http://www.hathitrust.org/]