Discussion document: recommendations for handling duplicates in HathiTrust

Introduction and summary of findings

As more HathiTrust partners contribute digitized collections, accurately identifying and appropriately handling duplicates will be increasingly important. The HathiTrust Collections Committee (HCC) has drafted for discussion this set of preliminary recommendations for handling duplicate scans, including when and how de-duplication might be performed. Because other committees and groups, both within and outside of the HathiTrust partnership, are examining policy and practice for related issues, this document includes brief summaries of that work as well as recommendations that some outstanding questions be referred back to other groups.

Based on the analysis performed so far, the HCC's basic findings are as follows:

- Primarily due to issues with underlying metadata, current duplicate detection processes are flawed
- Storing duplicates is not yet a very expensive problem
- De-duplication is likely to be an expensive activity and fraught with error
- HathiTrust scans lack quality markers to indicate which is the better duplicate

Based on these findings, the HCC recommends that de-duplication is not something that should be attempted now, but the situation should be reassessed at least annually. Should one or more of the following developments take place, de-duplication may become a more worthwhile activity: the measured rate of duplication increases significantly, new metadata management tools and matching algorithms increase confidence in duplicate detection, or reliable quality measures are developed. Should the HathiTrust partnership decide to initiate de-duplication processes now or in the future, it should do so within the following parameters:

- Scans should not be de-duplicated under these conditions:
    - If they are early printed works
    - If the duplicates result from two different capture methods
    - If there are legal restrictions or uncertainty about whether the retained copy can be broadly shared
    - If the duplicate in question has been integrated into a mashup
- Scans should not be permanently de-duplicated for at least five years

What is a duplicate?

The duplicates problem in the HathiTrust is complex, beginning with formulating an accurate definition of a duplicate. For the purposes of this document, duplicates are bibliographically identical texts, by which we mean "manifestations" in FRBR terms, for which more than one copy exists within HathiTrust. Usually these will be copies that have been digitized and contributed by two or more different HathiTrust partners, although the same copy may also have been digitized more than once by a single partner using different capture methods.

How does HathiTrust identify duplicates?

HathiTrust currently attempts to identify duplicates at the bibliographic (i.e. not volume) level upon ingestion by looking for a matching OCLC number or local system ID of the contributing institution, and groups matched texts together under a single catalog record. For example, the record for Oscar Douglas Skelton's 1911 dissertation, "Socialism: a critical analysis" links to three scans: two made by Google from the University of Michigan and the University of California copies, and one made by the Internet Archive from the same California copy.

Due to the current dependence on system numbers (OCLC or local systems such as library catalog record numbers), not all possible bibliographic duplicates are being correctly identified upon ingestion. For example, Google scanned a copy of "Miniature and Window Gardening" from the New York Public Library and from the University of Wisconsin Libraries, and, although from the same manifestation, they appear under separate catalog records in HathiTrust. Techniques used by HathiTrust to identify duplicates should be improved over time, as new partners and additional content types are added, and additional sources of metadata, such as the HathiTrust print holdings database now under development, become available. However, a major factor impacting the accuracy of duplicate detection will be the quality of the underlying metadata, consistency in application of OCLC numbers, and the like. The new metadata management system under development at the California Digital Library may also offer improvements to the matching and de-duplication algorithms.

Further complications arise for multi-volume sets and for serials. Duplicates are currently identified and clustered for display at the level of the manifestation of the work as a whole. When a set of matched bibliographic records also have multiple volumes and/or multiple scans of one or more volumes, only basic enumeration and chronology normalization is attempted, so that duplicate scans can be detected for reporting purposes. Automated identification of duplication at the volume level is extremely difficult because of differences in binding and in the way libraries record

enumeration and chronology. No action is currently taken to cluster, suppress from view, or remove any duplicates at the volume level.

HathiTrust staff produced a report, generated in early 2011, of suspected duplicates where more than one scanned volume with the same volume information is associated with a single bibliographic record. In many cases, the entries are for single volume works where enumeration and chronology do not apply, as in the Skelton dissertation example above. In others, particularly those with a high number of duplicates per bibliographic record, problems with missing or different enumeration and chronology information become apparent. For example, the duplicate detection report (see excerpt in Appendix A) identified as possible duplicates only the 20 Library of Congress-contributed volumes out of a total of 34 scanned volumes linked to the record for Documents relating to the colonial history of the state of New Jersey. Although the HathiTrust matching programs correctly grouped all 34 volumes under a single bibliographic record, missing enumeration data for the 20 volumes scanned from Library of Congress holdings makes it impossible for the program to properly evaluate these items against the five scanned volumes from Yale and nine from Michigan. In fact, this single record exhibits three variations on the duplicates problem:

- **Missed duplicates**: three different copies of volume 9 were scanned, but they are not grouped together in the duplicate detection report, due to missing and variant forms of enumeration data: Library of Congress's copy does not contain a volume label, Yale's copy uses the form "I.9", and Michigan's copy uses the form "v.9". The program thinks that I.9 and v.9 are different volumes and does not match them as possible duplicates, nor identify that one of the LC copies is also volume 9.
- **False duplicates**: the duplicates that *are* identified as duplicates were identified based on missing enumeration data; the missing volume identification for the 20 volumes provided by Library of Congress makes the program think they are all the same item, when in fact each is a different volume.
- **Correctly identified duplicates**: in the sense that all 34 scans are grouped together under a single bibliographic record, Hathi's matching processes have succeeded, at least to some extent.

In this case the HathiTrust duplicate detection programs have both under-counted and over-counted duplicates. In addition, a basic title keyword search of the HathiTrust catalog reveals at least two, and possibly as many as five additional records that could be the same work, but are cataloged as serials rather than sets, under variant titles, or variant series titles, and were not matched using system identifiers. These may also represent missed duplicates, though whether improved metadata analysis systems will be able to detect if these very different bibliographic records are for the same or different manifestations is a matter for exploration. This problem is further complicated by changes in cataloging practice over time: in this case rules about recording title changes may have been more or less strict at the time the records were created.

How do library digitization partners identify duplicates?

Google, a major source of HathiTrust scans, has been working successfully to improve its clustering processes for identifying duplicate items prior to digitization, as summarized in the October 2010 report by the Google Metadata Working Group (GMWG)[1]. The GMWG performed a systematic evaluation of Google's clustering and measured its accuracy in the 93% - 95% range. The Group attributes most of the clustering failures to metadata: either to errors in the data, or to wide variation in encoding of enumeration, chronology, and other holdings and item information. Although Google's clustering algorithms will limit the number of new duplicates coming from Google, Google scanned many millions of volumes before these algorithms were well tuned. There is little information about pre- or post-digitization de-duplication strategies within other digitization initiatives such as the Internet Archive/Open Content Alliance.

A variation on the duplicates problem arises for Google "**surrogates**": when Google receives a volume it has already scanned from a different library, it rejects it and, if it is in the public domain, returns the already digitized copy, called a surrogate, to the library that sent the duplicate. These surrogates may or may not have been made from a print volume sent by another HathiTrust partner. Surrogates appear in the ingestion queue for HathiTrust, but are not actually being ingested as of July 2011, pending the outcome of discussions about identification, attribution, tracking, and display. Once the Surrogates Working Group has addressed these issues[2], the duplicate principles below may also be applied to surrogates. However, since a surrogate is not a scan of the exact physical copy on the shelf in a library that receives the surrogate scan, it is a fourth, distinct variation on the duplicates problem.

Why should the HathiTrust partnership care about duplicates?

Storage Costs

As of April 2011, based on the duplicate detection processes described above, it is estimated that between 5% and 10% of HathiTrust titles have at least one duplicate item, requiring approximately 32Tb of additional digital storage at an annual cost of $109,000. It is not possible to know by how much this might over-count or under-count actual duplicates, but if it is assumed that these figures are reasonably accurate, the financial impact of retaining duplicates in HathiTrust is not overwhelming ($109,000 distributed equally amongst the 52 current partners is an additional $2100 per partner). However, as the collection grows, this number is very likely to increase, increasing storage costs. It should also be noted that this is a minimum figure, since, as shown above, some items exist in more than two copies in HathiTrust. Current duplicate detection processes indicate that approximately 75,000 bibliographic titles have three or more duplicates.

Usability

Potential for user confusion is among the most important reasons to improve duplicate handling. There is at least anecdotal evidence from user complaints that multiple catalog listings for the same title and lengthy volume lists with scattered duplicates are a real source of frustration. "Catalog users often have difficulty understanding the rationale or the subtle differences between multiple records when searching through a cluster of very similar electronic resource records."[3] If duplicates are not actually removed, efforts must be increased to group duplicates to simplify the public display whenever possible.

What are the risks of de-duplication?

Depending on how they are performed, de-duplication processes may themselves be expensive. If all matching and removal can be automated, costs can likely be controlled. However, if human review is required, or if libraries are to be granted an opportunity to 'rescue' scans prior to removal, either simply to override a decision or to actively move the objects to another repository, workflows become more complex and costs will increase. As much as possible, the HathiTrust partnership should strive to agree on de-duplication procedures that can be automated, including the creation of dark archives or other failsafes to permit recovery in case actions are taken that are later discovered to be in error. These recommendations have been drafted with automation in mind, but in the sections below, additional risks and other considerations are briefly outlined to provide perspective and opportunity for deep discussion prior to implementation. Key risks discussed include removing content that is potentially valuable to certain scholar communities, and basing de-duplication decisions on poor or nonexistent information about scan quality or on metadata that may be faulty or created using very different practices.

Recommendations for handling duplicates in HathiTrust

Unless otherwise noted, the same principles should apply to the removal of duplicates already stored (de-duplication) and to the addition or rejection of new duplicates submitted for ingestion.

1. Acknowledge the value of duplicates to scholars in certain fields

Keeping multiple digital copies of early printed works will be desirable for scholars in certain fields. For example, the English Short Title Catalog[4] gathers records of every known extant copy of works printed between 1473 and 1800, primarily in English and printed in the British Isles and North America. Scholars of early print history are interested in tracking each copy, and will value the opportunity to closely study multiple digital surrogates of a work. As of May 2011, the HathiTrust collection contained just 83,278 pre-1800 titles, accounting for 1.6% of the total collection. Although this number is likely to grow, particularly as more partner libraries contribute locally scanned material, it may never be so large, as a percentage of the overall HathiTrust collection, that duplicate removal will generate significant cost savings. Similar projects, such as the Catálogo Colectivo de Impresos Latinoamericanos hasta 1851 (CCILA)[5], are also engaged in building comprehensive catalogs of all copies of early printed works.

Prior to the widespread adoption of formal copyright laws in the early twentieth century, there was a higher frequency of variant editions—copies of works printed within short time frames by different publishers and often less clearly and consistently identified in title pages. These factors have complicated cataloging activities, so that not only is there a profusion of very similar items, there is a strong likelihood of a higher cataloging error rate. More caution in de-duplication is warranted on these grounds. Works from different time periods may have other characteristics that make a case for retaining duplicates. Many important projects, such as the ESTC, CCILA, Early English Books Online, Eighteenth Century Collections Online, and Digital Evans, have relied heavily on library bibliographic data, and might wish in future to construct links to HathiTrust digital copies. Aggressive de-duplication might, therefore, lessen the value of HathiTrust to these projects.

Recommendation 1.1: the HathiTrust partnership should survey scholars to determine whether there is a generally agreed-upon date before which de-duplication should not be attempted, due the likelihood of a negative impact on scholars of early printed works.

Recommendation 1.2: De-duplication should not be attempted for early printed works. Provided that scans and catalog records for these items meet minimum standards, they should always be retained.

2. Variations in capture method and quality

The HathiTrust corpus contains scans from different providers using different equipment and applying different standards. Google-scanned material is primarily bitonal, with color elements retained when present, but Internet Archive scans are presented in full color. Some readers may find higher-contrast black and white scans easier to read on screen, while others will prefer the full color experience. Quality of Optical Character Recognition (OCR) will vary, as will the accuracy and depth of in-book navigational aids such as the clickable table of contents. These and other differences in the form, accuracy and completeness of a digital scan will have real or perceived impacts on quality that must be taken into account in de-duplication processes.

Recommendation 2.1: Retain scans from different capture methods

Other quality impacts may only be realized as HathiTrust grows, and as the community builds projects and tools around these materials. Teachers building virtual courseware may wish to blend elements from two scans of the same work if, for example, the OCR'd data tables from copy X are superior, but the color photographs in copy Y are better. Crowdsourced applications for correcting OCR or suggesting metadata fixes may benefit from presenting a reviewer with as many versions of a digital facsimile of a difficult original as it can. It may even be possible in future to construct tools to 'digitally heal' torn or otherwise imperfect pages by combining multiple scans of the same physical page.

There may be other long-term quality implications in de-duplication. Google has revealed that it is constantly improving its image processing and OCR software, and does reprocess previously digitized texts. If it is possible to improve OCR as algorithms get smarter, or to better correct for page noise, foreign objects such as fingers and clamps, page curvature or visual flaws, a more conservative approach to de-duplication may provide these applications with better, or at least multiple, inputs.

A Quality, Ingest, and Error Rate Working Group[6] was formed in 2009 to develop a set of quality principles for HathiTrust material. In 2010, IMLS funded the project "Validating Quality in Large-Scale Digitization,[7]" led by professor Paul Conway,

which seeks to identify "methods for detecting and measuring errors and other quality issues" and to assess the impact of these errors on specific end user activities. HathiTrust should charge a working group to synthesize the results of these activities, and translate them into a set of objective quality criteria so that a rating, or quality score, can be assigned, in an automated fashion, to each digital item in the collection. These scores can be used to determine which copy from a set of duplicates is the best copy and worthy of retention. There should be a minimum quality threshold below which no de-duplication actions are taken.

This is not to suggest that poor-quality scans should always be retained when better scans are available. Duplicates that don't meet HathiTrust's own standards for acceptance (e.g., scans that include only the book's covers and not the text) may safely be removed.

Recommendation 2.2: De-duplication should be performed only if the retained copy meets certain objective quality standards.

3. Trust, but verify[8] when matching through metadata

As many (Google Metadata Working Group[9], John Wilkin[10], Jeffrey Nunberg[11]) have recently observed, blending metadata from hundreds or thousands of library contributors reveals both tremendous variations in cataloging practice and, unfortunately, a measurable rate of error. These problems have been visible for decades in OCLC WorldCat, but their implications are also deeply felt in mass digitization projects and collections such as HathiTrust. As discussed above, we know they contribute to failures in duplicate detection (missed duplicates). The more serious condition, however, is false duplicates. If HathiTrust or its contributors incorrectly match catalog records, there is a risk that an automated de-duplication process will remove scans that should be retained. Unless external measures of metadata quality can be leveraged, or until the HathiTrust community can devise a score for metadata quality as it will for scan quality, de-duplication should be approached conservatively. If there is any doubt of a match, the decision must be not to remove a duplicate. HathiTrust must expand and improve matching processes so that elements other than OCLC or system IDs can be used as candidates for identifying matches, but should likewise increase in stringency the requirements for determining that a duplicate has been detected. Multiple points of matching must be present, and should be carefully logged. It may also be possible, in future, to enhance matching by including OCR'd text as a comparison point. This may also help to solve another problem: when the item scanned is not actually the item described in the metadata record.

Recommendation 3.1: Due to known metadata quality problems, HathiTrust should expand the scope of matching activities beyond system numbers, but de-duplication should be carried out conservatively, and should require multiple confirmations of matching elements.

4. Transparency and implications for ownership

When de-duplication actions are taken and items are removed, transparency and tracking will be critical. Ideally, information about any given manifestation's removed duplicates will be tracked, and either displayed on-screen in the catalog record view, or, at minimum, stored within the PREMIS data or other administrative data about the object. Regardless of which digital copy is retained, HathiTrust partner libraries will wish to know if a title in HathiTrust came from or is available in their local library, especially if a print copy can be readily available to a user. This is particularly important in cases where the work itself is still under copyright and not available in full view. This approach should align well with the new cost model that will be implemented starting in 2013[12] and the development of the common holdings database on which it depends.

In addition to documenting de-duplication decisions as described above, if any de-duplication takes place, libraries whose contributed digital copies have been removed must be granted the same rights to digital copies as those libraries whose digital copies were favored for retention. Partnership agreements or contracts that govern ownership of digital copies must be taken into consideration: if these instruments prohibit such extensions of ownership, de-duplication actions cannot be taken. It is possible that the only implications of this will be for partners seeking to leave the HathiTrust; nevertheless, ownership implications for de-duplication must be fully understood prior to any content removal.

Recommendation 4.1: If a partner-contributed copy has been suppressed or discarded in favor of a different scan, this action should be recorded in the metadata and made visible to library curators. Additionally, a library whose digital copy has been discarded must be given joint ownership and/or full exercise of owner's rights in the replacement copy.

Recommendation 4.2: If legal restrictions preclude joint ownership and/or use rights in retained scans, de-duplication should not be undertaken. (items could be suppressed from view)

5. De-duplicate cautiously and reversibly at first: Failsafes

Duplicate removal might actually be carried out in one of several possible ways, or

in a combination of methods. In the most literal implementation of de-duplication, duplicates are actually deleted, the files expunged from HathiTrust servers forever. Alternatively, they might be retained in place, but visibility turned off. They might also be returned to the library or libraries that contributed them. An as-yet-unidentified third party might accept them, either under contract with HathiTrust or individual partner libraries, or as a contribution to the public good. HathiTrust itself may choose to dark archive the duplicates by moving them to less expensive digital storage. At least until the full implications of a massive shared database of digitized texts and the impact of duplicate removal are better understood, HathiTrust should initially implement de-duplication with a failsafe: a fully reversible method, as insurance against possible miscalculations in its de-duplication procedures.

Recommendation 5.1: For at least the first five years of taking de-duplication actions, HathiTrust will implement one or more failsafes, so that it will be possible to restore removed duplicates at some time in the future.

6. Book blending and authenticity

Although not strictly speaking a de-duplication issue, the related issue of book blending, or creating so-called "digital Frankenbooks" is worthy of some attention. Especially for out-of-print works, libraries have developed a tradition of replacing missing or damaged pages with photocopies from a borrowed copy of the same work at another library. It is also not uncommon for microfilm producers to cobble together complete works from copies at multiple institutions. Such aggregation activities may or may not be closely documented in accompanying metadata. In the digital environment, however, special attention should be paid to authenticity and transparency. The technology is perfectly suited to one day support book mashups, particularly in cases where characteristics of the physical text have made it difficult to achieve a good quality scan. If library X was willing for its copy to be disbound for scanning, but the scan is bitonal, it may be very desirable for a user to be able to combine pages from copy X with selected color pages from copy Y, even though copy Y in general has shadows from a tighter binding. Given the growing acceptance of a digital copy as an authentic, reliable copy, if HathiTrust permits book blending, either as part of a production process or an end user process, it must provide complete and accurate documentation of the source of every element. When books are blended, a complete version of each the 'donor' books should be retained in the HathiTrust database. Google, Internet Archive and other contributors should likewise be encouraged not to submit blended books, or at the very least not to do so without providing full documentation. It should always be possible for a researcher to examine the complete original sources, and therefore de-duplication processes must check 'book blending' records to be certain that a user is not relying on a candidate prior to its removal.

Recommendation 6.1: Book blending should be approached cautiously, and any services that permit book blending must be accompanied by complete documentation tracing the origin of pages or page sections to their copies or sources.

Conclusion and next steps

Based on the analysis conducted, the HathiTrust Collections Committee recommends against taking any action to de-duplicate scans at the present time. Should the cost of storing duplicates increase significantly, or should reliable quality measures become available, the case for de-duplication may be more compelling. In the near term, however, the emphasis for the HathiTrust should be to improve the experience for users, and to simplify, as much as possible, labeling and display of duplicates, both in search results and in item view screens.

Version history:
v. 1: May 5, 2011
v. 2: July 12, 2011
v. 3: July 15, 2011
v. 4: July 22, 2011
v. 5: August 11, 2011
v. 6: August 11, 2011
v. 7: September 18, 2011
v. 8: September 22, 2011
v.9: April 6, 2012

Appendix A: Excerpt from duplicate detection report

Entry in the duplicate detection report for “Documents relating to the colonial history of New Jersey”

http://catalog.hathitrust.org/Record/001647401 1880
http://hdl.handle.net/2027/loc.ark:/13960/t0bv7n89z
http://hdl.handle.net/2027/loc.ark:/13960/t0bv7n91j
http://hdl.handle.net/2027/loc.ark:/13960/t0js9tn30
http://hdl.handle.net/2027/loc.ark:/13960/t1dj5kz75
http://hdl.handle.net/2027/loc.ark:/13960/t30293s46
http://hdl.handle.net/2027/loc.ark:/13960/t34174m9x
http://hdl.handle.net/2027/loc.ark:/13960/t3hx1hb90
http://hdl.handle.net/2027/loc.ark:/13960/t5w66mz8f
http://hdl.handle.net/2027/loc.ark:/13960/t6542w71h
http://hdl.handle.net/2027/loc.ark:/13960/t6m048q65
http://hdl.handle.net/2027/loc.ark:/13960/t6m048q9m
http://hdl.handle.net/2027/loc.ark:/13960/t73t9r45n
http://hdl.handle.net/2027/loc.ark:/13960/t7hq43t9g
http://hdl.handle.net/2027/loc.ark:/13960/t7wm1fn65
http://hdl.handle.net/2027/loc.ark:/13960/t81j9k74b
http://hdl.handle.net/2027/loc.ark:/13960/t86h4s40v
http://hdl.handle.net/2027/loc.ark:/13960/t8mc93d7s
http://hdl.handle.net/2027/loc.ark:/13960/t8w95bk0t
http://hdl.handle.net/2027/loc.ark:/13960/t9959q86c
http://hdl.handle.net/2027/loc.ark:/13960/t9b56r573

[1] Julia Lovett, *Google Metadata Working Group Report*, October 19, 2010.
[2] "Surrogates Working Group Charge", n.d., http://www.hathitrust.org/wg_surrogates_charge.
[3] Provider Neutral E-Monograph Guidelines, *Provider Neutral E-Monograph Guidelines: Final Report* (Program for Cooperative Cataloging, Library of Congress, July 30, 2009), http://www.loc.gov/catdir/pcc/bibco/PN-Final-Report.pdf.
[4] British Library, "English Short Title Catalogue", n.d., http://estc.bl.uk/.
[5] "Catálogo Colectivo de Impresos Latinoamericanos", n.d., http://ccila.ucr.edu/.
[6] "Working Groups and Committees", n.d., http://www.hathitrust.org/working_groups.
[7] "Validating Quality in Large-Scale Digitization: Metrics, Measurement, and Use-Cases", n.d., http://www.si.umich.edu/research/project/validating-quality-large-scale-digitization-metrics-measurement-and-use-cases.
[8] "Trust, but verify," *Wikipedia*, n.d., http://en.wikipedia.org/wiki/Trust,_but_verify.
[9] Lovett, *Google Metadata Working Group Report*.
[10] John Wilkin, "Bibliographic Indeterminacy and the Scale of Problems and Opportunities of 'Rights' in Digital Collection Building", February 2011, http://www.clir.org/pubs/ruminations/01wilkin/wilkin.html.
[11] Geoffrey Nunberg, "Google's Book Search: A Disaster for Scholars," *The Chronicle of Higher Education*, August 31, 2009, sec. The Chronicle Review, http://chronicle.com/article/Googles-Book-Search-A/48245/.
[12] "Cost," *HathiTrust*, n.d., http://www.hathitrust.org/cost.