HathiTrust Monographic Duplication and Uniqueness:

2017 Report and Recommendations from the HathiTrust Collections Committee

I. Executive Summary

The HathiTrust Collections Committee (HCC) recently reviewed the 2011 (revised 2012) HCC duplicates report (*HCC Duplicates 2012)*[1] and analyzed a sample of item-level metadata related to monograph duplication in the HT corpus. The current HCC has reached similar conclusions to those expressed in the 2012 report, and the present report expands and extends the earlier recommendations.

Monograph scans in the corpus are not simple surrogates of the print items of origin. Most overlap in the corpus represents things that are valuable in their own right and potentially useful in their ability to stand alongside other bibliographically identical, but functionally non-identical entities. Multiple overlapping scans of given manifestations, whether produced from different physical items or different scanning methods, have intrinsic value that merits their continued presence in the corpus. This report details HCC's recent analysis and recommends actions and areas of focus to leverage the overlap with the HT corpus in support of various research needs.

Truly duplicative entities consist only of those scans produced from the same physical item through the same scanning method. Such ingests are meant to be avoided under current ingest methods and should continue to be avoided in the future. Such duplicates likely represent only a very small proportion of bibliographic overlap in the corpus. HCC has reviewed current practices for detecting and preventing duplication (see Appendix A), and believes HT should uphold those procedures to continue to minimize true duplication within the corpus.

Section II of this report suggests a new taxonomy of manifestations and items that can help frame further discussion of overlap. The use of such a taxonomy will allow HT to take into account the potential value of scans of different physical items representing the same manifestation, and the value of different scanning events applied to the same physical item. With a focus on the researcher experience, Section III reaffirms the need for technical and metadata enhancements to support and leverage overlap in the corpus to fuller effect. Recommendations to the PSC comprise Section IV.

Appendix A details current HT technical practices related to overlap, and Appendix B provides an overview of HCC's analysis of the present overlap within the corpus.

II. Uniqueness and Duplication: A Taxonomy

This report uses the following taxonomy when referring to bibliographically identical entities in the corpus. For our purposes, an OCLC number match is the standard for calling given entities bibliographically identical. Three of the four types outlined below are usually both valued and valuable, and should not be considered candidates for removal from the HT corpus. Scan Types 1, 2, and 3 are difficult to reliably identify and isolate, and it is predicted that fewer than

[1] https://www.hathitrust.org/documents/hathitrust-collections-duplicates-report-201204.pdf.

one-third of scans that share the same OCLC number with another scan are Type 2 (see Appendix B), a category that would thus represent about 5.3% of the overall HT monograph corpus, roughly 534,000 such items. Only Type 1 scans are essentially redundant, i.e., produced from the same item with the same scanning method, and given the current state of metadata and tools available, HCC does not think it is prudent to pursue machine-based detection and removal of these duplicates.

Type 1: Multiple Scan Submissions Produced with the Same Scanning Method from a Single Item

Type 1 occurrences within the corpus, should any be found, would be rare. No such entities should be deposited in the corpus under normal circumstances given current ingest practices, and no examples were found in examining the sample dataset addressed in Appendix B below. Any that are presently in the corpus are of little value. HCC refers to such entities as the “true duplicates.” See Appendix A.

Type 2: Multiple Scan Submissions from a Single Item

These occur where a single Hathi bibliographic record consists of more than one scan, each originating from the same physical item by the same library provider, but with each scan provided by a different scan provider and/or scan event. These scan events occurred at different times, and the most common instances of this are cases in which scans by Google, Internet Archive, and/or a local library provider of the same physical item have all been linked to a single bibliographic record in the Hathi collection. A common feature of these Type 2 scan groupings is the inclusion of one bitonal and one color scan. There are processes in place to prevent such occurrences (see the discussion in Appendix A, section 1), and the standing protocol when such Type 2 cases occur—detected by the intake of an item with a bibliographic record that matches a previously accessioned item—is for both copies to be stored but only one to be provided on the HathiTrust platform. This situation is discussed from an ingest standpoint in Appendix A, sections 1 and 2. Such scans are unique since they are produced by different scanning methods.

Example:
H. P. Liddon, *Sermons Preached before the University of Oxford* (1865), Hathi Record 0006663238, https://catalog.hathitrust.org/Record/0006663238

1. University of California scan by Google: https://hdl.handle.net/2027/uc1.b3478115
2. University of California scan by Internet Archive: https://hdl.handle.net/2027/uc2.ark:/13960/t3nv9k824

Type 3: Multiple Distinct Items by a Single Provider

These occur when a submitting library owns multiple physical items of the same manifestation of a given bibliographic entity, and each has been scanned and correctly linked to the same Hathi bibliographic record. Such scans, while bibliographically identical, are still distinct in that each represents a unique physical object and each scan may contain additional unique information such as a bookplate, marginalia, library circulation tracking, or other extratextual additions that can aid researchers studying the physical provenance of the scanned item. Although rights and access for a given item are drawn from the base bibliographic record, situations in which subsequent scans have been linked to that record, or cases where a library provider has contributed material via two different channels (*e.g.,* Internet Archive or Google Books, at a restricted access, but subsequently directly with an unrestricted access) means that rights and access policies may need to be reconciled across all items linked to a record by the appropriate rights and access advisory committee. Type 3 scenarios are discussed from an ingest standpoint in Appendix A, section 3.

Example:
Agnes Hunt, *The Provincial Committees of Safety of the American Revolution* (1904), Hathi Record 315825, https://catalog.hathitrust.org/Record/315825.

The unique bookplate information, owner signatures, and other marginalia should be noted on each of these copies.

1. Harvard University scan by Google of local call number **US 2818.16 A**: https://hdl.handle.net/2027/hvd.32044046730180
2. Harvard University scan by Google of local call number **US 2818.16 B**: https://hdl.handle.net/2027/hvd.32044011464211
3. Harvard University scan by Google of local call number **HW U23C A:** https://hdl.handle.net/2027/hvd.hwu23c

Type 4: Single Bibliographic Items by Multiple Providers
This situation is very common in the corpus, representing the natural overlap of collections held by various libraries. As with Type 3, each scan is representative of the same manifestation but a different item in a different institution. Thus, such scans can each potentially contain unique extratextual information, and may come with unique access rights linked to the provider library. They also represent different scan events and/or methods, with the resulting advantage of enabling researchers access to multiple scans to aid activities like text extraction and quality assessment. This situation is covered by Appendix A, section 3.

III. Leveraging Overlaps

As HCC began sampling data and discussing duplicates in 2016, it was clear that the value of duplicates outweighed the unreliable methods and cost of any weeding effort (See Appendix B). *HCC Duplicates 2012* included some preliminary recommendations for managing duplicates and guidelines on when and how deduplication efforts could be performed. The recommendations outlined in *HCC Duplicates 2012* immediately acknowledged the value of duplicates and their potential benefits for scholars. HCC reaffirms the earlier recommendation. Type 2-4 entities in the Hathi corpus should be viewed as an asset and not a liability, although further analysis needs to be done to determine—based on user needs—the degree to which these assets need to be surfaced to leverage their use.

There are many reasons to consider multiple digital versions of the same title or even the same physical book as an asset. This was already alluded to in *HCC Duplicates 2012*, in which the importance of each unique printed object of the same title, and therefore the multiple digital scans produced by those multiple printed objects, was identified as an asset to scholars of early modern books. This advantage is not confined to researchers using pre-1900 material, however, and could be extended across the Hathi copus.

A. Value of Type 2-4 Overlaps for Distant Reading, Close Reading, and the User Experience

Type 2-4 entities in the Hathi corpus can enrich scholarship and usability, *e.g.,* as in *HCC Duplicates 2012*, which pointed to the importance of multiple capture methods and the possibility of mashups to support close reading. The simple ability to quickly access an alternative scan of a single page would be very valuable in support of such use cases.

More recent scholarship, particularly research using text mining and other machine-assisted forms of reading, have also pointed to the use of multiple scans of the same object to correct for OCR errors and to identify marginalia and other unique non-printed information that may be found in distinct items representing the same manifestation. For example, a recent text mining project by a Berkeley economics researcher highlights the value of Type 2-4 entities in Hathi's current corpus. This research sought to answer the following question, "How have California universities contributed to the state's growth, economic mobility, and gender equality over the past 150 years, and what is their role today?" The information resources identified for this text-mining project from the Hathi corpus included California university registers, directories, high school teacher directories, and doctor directories. This text mining project included photography, OCR, natural language processing, and algorithmic inference. Duplicate surrogates for the various documents allowed the computer program to determine which version of the duplicates represented the highest quality, and to thereby extract needed information accurately.[2] Other scholars working with the HathiTrust Research Center have expressed a similar need to effectively identify duplicate titles in order to ensure accurate counts in their analyses, necessitating a user interface for machine- and human-readers alike to enable identification of such items.

There are currently barriers, however, to leveraging the strength of these duplicates by both human- and machine-based research uses of Hathi. Improved metadata to identify the presence and types of duplicates in the corpus would be essential to harnessing this strength. Making this metadata available, particularly on the main record page where currently linked scans are presented without identification other than the provider source, would also be very helpful.

B. User Display and User Metadata Enhancement

Keeping in mind the benefits to researcher access to the various types of overlaps, any metadata enhancement should concentrate on easing the ability of scholars to understand from the main bibliographic record page how many physical objects are represented by the listed providers

[2] See Zachary Bleemer, "UC Berkeley Researcher Mines HathiTrust Volumes for Cliometric History of Postsecondary Education in California," *California Digital Library Blog* (16 September 2016), http://www.cdlib.org/cdlinfo/2016/09/16/uc-berkeley-researcher-mines-hathitrust-volumes-for-cliometric-history-of-postsecondary-education-in-california/

and links, and some vital information about the type of scan available (*e.g.* color versus bitonal, etc.). Currently, a researcher can indirectly surmise that there are multiple scans of different physical objects by the presence of more than one contributing institution listed on the main record page. However, when presented with multiple links by the same provider, the researcher must consult the scanned pages of each item in order to establish whether they represent two different physical objects or not. This problem is further compounded when scans of the same manifestation are associated with different bibliographic records resulting from a lack of shared OCLC numbers. Users should not be expected to understand the peculiarities of MARC records and HathiTrust record handling which result in such behavior.

This difficulty is magnified in cases of multi-volume, multi-scan records such as the following entry for a government document, an *Annual Report of the Commissioner of Patents* (1838-1976): https://catalog.hathitrust.org/Record/002138126, where a full 429 scanned documents are linked to a single bibliographic entity. Because the entries in HathiTrust are presented by year or volume in the first instance, and by provider secondarily, and because providers are the primary means of identifying duplicates, the researcher must sort through a long list of links to draw conclusions about any duplication or the number of physical objects represented for the item in question.

One might envision as a solution different ways of grouping and displaying item links on the main record page, or exposure of further metadata elements such as the Hathi ID or the scan provider, to enable identification of duplicates.

IV. Recommendations to PSC

A. Affirm that Type 1 scans are the major truly redundant monographic entity duplicates in the corpus. Type 2 scans are valuable for establishing best-quality scans, but current methods for identifying them and handling them at point of ingest should be sustained and even honed. Future efforts should look to limit increases in Type 1 and 2 items in the corpus, and future metadata and analysis tool development might focus on enabling the removal of any existing Type 1 items.

B. HathiTrust partnership should continue outreach to the HT Scholar and HTRC user communities to gather more examples of researchers leveraging the duplication types above in the future. Such information can be used to expand the known use cases and seek UX enhancements to support those uses.

C. Affirming *HCC Duplicates 2012*: 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.

D. Acknowledge the inherent value of Type 2-4 scans within the corpus overall, especially in relation to physically distinct modifications post-publication, scan (and therefore potentially OCR) quality, processing anomalies, and inherent capture differences. Affirming and expanding upon *HCC Duplicates 2012* recommendation *"Retain Scans from different capture methods."*

E. Leveraging Type 2-4 scans is a valuable and important initiative. Enhance the metadata and user experience to allow bibliographically identical but unique HT monograph scans to more freely and easily be leveraged for close and distant reading.

F. Reconsider the value of mashups to help create a best HT copy composed of best page scans of a given manifestation. Consider UI and metadata enhancements to allow for bulk and user-contributed identification of the best pages and best editions for given uses. This will support close readers and could also be especially valuable for HTRC's general use cases, where the best work-level entity is easily understood as valuable.

G. Metadata enrichment is key to leveraging overlaps. Prioritize the following:
   a. Institutional provider provides the best marker of Type 2 overlap: metadata related to provider should be available more easily as a way to filter or sort multiple scans of a bibliographic item.
   b. Expose metadata that will help researchers easily see when two or more scans are of the same physical object (Type 2), or of the same bibliographic manifestation but different physical object (Types 3 and 4).
   c. Ensure that the relevant metadata regarding institutional and scan providers is provided in the HathiTrust API so that machine-based research applications can similarly identify overlaps.

Version history:

v. 1: 14 October 2016
v. 2: 6 January 2017
v. 3: 3 February 2017
v. 4: 27 February 2017 (sent to PSC for circulation and comments)
v. 5: 5 May 2017 (revised in response to PSC)
v. 6: 16 May 2017 (final)

Appendix A. Duplicate Handling Practices at Ingest in HathiTrust

Current practice divides duplicates into three categories. Each is handled differently, and they overlap in obvious ways with the taxonomy suggested elsewhere in this document. For clarity, this section does not assume an adoption of the new construct but a reflection of current practice.

1. Scans of the same print item digitized by different entities, where HT does not know that these are from the same item because different identifiers are used. However, these are still associated with the same bibliographic record.

   Example: University of California sent items for digitization both to Google and to Internet Archive. When submitting these to HathiTrust, both scans were associated with the same bibliographic record, so a user can visually identify that they are copies of the same print items. See https://catalog.hathitrust.org/Record/006585502 for an example.

2. Scans of the same print item, digitized by different entities, and HT knows these are duplicates because the same identifier is used by the contributor.

   Example: Google scanned some University of Michigan items, and then University of Michigan scanned these same items through their local digital conversion unit. The same barcode is used to identify these two different scans, but because it is impossible to have two items in the HathiTrust repository that share the same identifier, only one of these scans can be present in the repository at any time. HT has processes in place that will "blacklist" specific identifiers in Google so that HT does not download a Google scan that overrides the UM scan.

   Example: Some partners have indicated that they wish to preserve both the Google scan and their own scan. HT has not yet ingested any of this content. Theoretically, both scans can exist at the same time in the repository if one of the identifiers is assigned to a different namespace. HT could still know that these are scans of the same print item because the barcode after the namespace is identical.

3. Scans of different print items that are of the same manifestation, and HT knows they are duplicates because they are associated with the same bibliographic record.

   Zephir loading processes have several ways of grouping related records together. First, records submitted by one contributor are checked against other records submitted by that same contributor, and they are matched together based on identical record system numbers. Second, records submitted by multiple contributors are checked against other records from all contributors, and they are matched together based on identical OCLC numbers. In both cases, items listed on these matching records are grouped together on the same HathiTrust cluster record. The record at https://catalog.hathitrust.org/Record/000398723 is an example of both of these types of matching, as it includes multiple copies from one contributor (University of Wisconsin-Madison) and also copies from multiple contributors.

   If, however, a contributor submits records that contain different system identifiers for associated items, those items will not be grouped together in the HathiTrust interface. This can occur by accident when a contributor submits a corrected record for only one or some of the volumes on a record and misses others.

Appendix B. HathiTrust Collections Committee Analysis and Conclusions: Identifying Duplicates and Uniqueness

HCC Duplicates 2012 estimated that between 5 and 10 percent of the titles in the Hathi corpus have at least one duplicate item associated with that record, and that 75,000 titles have three or more duplicates attached to that record. The previous committee also noted, however, the difficulty of precisely identifying duplicates in the corpus given the reliance on matched OCLC or local catalog numbers to link incoming scanned material and cluster together multiple scans of items all referring to the same manifestation. It also noted the difficulties of differentiating multi-volume records, which by definition contain multiple scans linked to a single record, from single-volume duplicates. Thus, as *HCC Duplicates 2012* report noted, there is the risk of both undercounting and overcounting duplicates in the system using the metadata at hand.

Nevertheless, HCC considered it worthwhile to revisit the question of the scale of duplicates in the HathiTrust corpus using targeted samples of the full set of metadata records provided by the Zephir team.

Sample Parameters

In preparing an extract of records consisting of a single-volume work with suspected duplicates, Zephir noted that there were 10,063,130 "monograph volumes", *i.e.,* scans of monograph, or non-serial, items in the corpus, representing 7,302,396 monograph titles. Of these 7.3 million, 1,189,018 titles, representing 3,949,752 scans, were determined to consist of multiple "volumes", *i.e.,* multiple scans attached to a single bibliographic record, and were therefore candidates to for investigation as to whether the presence of multiple "volumes" in reality represented multiple scans.

From this set of 3.9 million, the Zephir team determined that 1,780,721 of these scans were candidates to be duplicates because of the lack of enumeration/chronology (*i.e.*, item description information) values for those records -- a sign that these were likely not multi-volume works despite having multiple scans attached to a single bibliographic record. This would mean that approximately 17-18% (1,780,721 of 10,063,130) of the HathiTrust corpus of single-volume monographs consist of single scans and one or more of their duplicates.

On this basis, the Zephir team created two smaller extracts from the 1,780,721 suspected single-volume monographs that had been identified, one from title clusters with more than 10 scanned items attached to a single bibliographic record, and one from title clusters with more than 1 scanned items attached to a single bibliographic record. These extracts consisted of all scans linked to a single title, *i.e.,* sharing the same CID (clustering identifier) number, derived from its OCLC number, where that single title could be identified as a book that was also likely a single-volume title. Single-volume status was determined using the lack of Enum/Chron information in the record. The first extract, of clusters of greater than 10, yielded 5,479 records, representing 424 titles, 0.3% of the suspected 1,780,721 records found in clusters. The second extract, a sampling of clusters greater than 1, contained 8,059 records, representing 3,002 titles, or 0.45% of the suspected 1,780,721 records found in clusters. This sample was derived from Zephir metadata with a goal of including representation and coverage of both resource publication dates and source institutions who have contributed content and metadata to HathiTrust. The results below derive from these two small samples of suspected duplicate monograph clusters.

Analysis

The analysis of these two extracts sought to address three questions:

1. How successful had the extract parameters been in excluding multi-volume monographs? That is, what proportion of each extract consisted of true single-volume monographs linked to multiple scans and therefore potential duplicates?

2. Once those multi-volume titles had been identified and excluded, to what extent were the resulting multi-item titles a result of contributions from more than one source, *i.e.*, more than one provider institution or library? These correspond to the Type 4 instances above, believed to be very common in Hathi by nature of overlapping library collections.

3. Similarly, excluding multi-volume titles, what proportion of the bibliographic records with multiple attached scans involved multiple contributions by a single library (Types 1-3)? Were they multiple scans of the same exact physical item,*i.e.,* Type 1 or 2? Or do they represent multiple scans of the same manifestation but distinct physical items (Type 3)?

Methods

These extracts had already been created using the absence of Enum/Chron information as a strong indicator of single-volume status. However, as the Zephir team suggested, the absence of Enum/Chron information was not a sure marker of single-volume status since a scan contributor could have simply failed to supply that information. This left two options for further identifying multi-volume titles within the extracts: look for a second publication date attached to a single title (an indication of a subsequent volume publication), or look for volume information (e.g. “v. 2”) in the description field. A glance at the second date field indicated that such information was often not present, so it was decided for this analysis to identify as multi-volume any titles in the extract that had the character string “v. “or “v.” (i.e. with or without a space character) in its description field. Having identified remaining suspected multi-volume titles, each extract was then sorted using the Pandas Python library to group titles (by CID) and then sources (by source name). Frequency counts were then generated in Python for multi-source titles, and the results double-checked using SPSS statistical software.

Results

The result was a count for each extract of multi-volume single-provider titles, multi-volume multi-provider titles (a potential provider of true duplicates), single-volume single-provider titles, and single-volume multi-provider titles. These can be summarized as follow:

Extract	**Count of Multi-Volume Single-Provider Titles**	**Count of Multi-Volume Multi-Provider Titles**	**Count of Single-Volume Single-Provider Titles**	**Count of Single-Volume Multi-Provider Titles**
Clusters of >10 (424 titles)	166	1	64	193

Clusters of >1 (3002 titles)	30	31	512	2429

With regard to the first question above, these results show that although multi-volume titles were still well-represented in the extract consisting of clusters of greater than 10, once the cutoff was loosened to great than 1 cluster, the number of multi-volume titles became negligible. This confirms that the Zephir team was able to identify single-volume monographs successfully, with just over 2% (61 of 3002 titles) of the overall >1 Cluster sample extract consisting of multi-volume works. We can therefore rely in particular on this second larger extract, consisting of records with clustered scans of more than one, as a means to identify the nature of single-volume monograph duplicates in the overall corpus.

Thus, it is clear that by far the greater proportion of titles in the Hathi corpus with more than one scan attached to its record consist of cases where multiple libraries contributed scans (2,429 of 3,002, or 81%) to the same bibliographic record. Extending this out to the larger corpus, this would mean 81% of the 1,780,721 suspected records in clusters (itself 17-18% of the single-monograph corpus) consist of multiple scans by multiple providers, or 1,442,384 scans. This would be about 14% of the suspected 10 million "monograph volumes" identified by Zephir.

Turning to questions 2 and 3, it is a lot more difficult to identify Type 1, 2, and 3 scenarios, and we should recognize that such scenarios might arise any of the four cases described in the table above, whether multi-volume or not, or multi-provider or not. For example, a record with multiple library providers might also include multiple scans linked to the same library. Moreover, the metadata provided does not make it possible to identify Type 1, 2, or 3 scenarios from the sample without examining the scans directly. That process, is further hindered by the restricted viewing access of some records.

Nevertheless a rough sense of how frequent Types 2 and 3 might be obtained by identifying cases where a single library had provided more than one scan attached to a single bibliographic record, and then investigating a random subsample of this selection. No Type 1 ("true") duplicates were found in the course of this analysis.

In all, of 8,059 records in the >1 Cluster sample set, 3,007 records (or 37%) involved a single library providing more than one scan attached to a single CID (i.e. HT bibliographic record). A random sample of 15 was selected from this group to identify whether the multiple linked scans were scans of the same physical item (Type 2) or scans of different physical items but the same bibliographic entity (Type 3 or Type 4):

- 5 records could not be examined because they were not full-text searchable
- 5 records had multiple scans of distinct physical items provided by a single library (Type 3).
- 2 records had multiple contributions (none of the same physical item) from multiple libraries (a mix of Type 3 and Type 4).
- 3 records contained multiple distinct scans by the same provider of the same physical object (Type 2). For a representative example, see *Sermons Preached Before the University of Oxford* (1869) by H. Liddon (https://catalog.hathitrust.org/Record/0006663238), available from the University of California in a black and white scan by Google

(https://babel.hathitrust.org/cgi/pt?id=uc1.b3478115;view=1up;seq=7) and in a colorized scan of the same physical object by the Internet Archive (https://babel.hathitrust.org/cgi/pt?id=uc2.ark:/13960/t3nv9k824;view=1up;seq=7).

Assuming the records that were restricted from view followed the same pattern as those that could be investigated, that would indicate that 30% of the 8,059 sampled record provided by Zephir, itself a sample of the 1.78 million believed single-volume monographs with multiple attached scans, contain Type 2 duplicates; at least 50% would be Type 3 duplicates; and another 20% would be a mix of Type 3 and Type 4 duplicates, meaning multiple scans of distinct physical objects from multiple library providers.

This limited examination of a subset of the Zephir team's extract is too small a sample size to come to a firm conclusion, but would suggest that approximately 534,000 scans (30% of the 1.78 million that are suspected single-volume, multiple-scan records provided by the Zephir team) could potentially involve Type 2 scenarios in which scans of the same physical were supplied by a single library provider.

Conclusion

More recent analysis by the Zephir team suggests that initial estimates of the proportion of the Hathi corpus involving duplicates was an underestimate if we define duplicates to include all single-volume monographs that have multiple scans attached to a single title/record (Types 2, 3, and 4 combined). An extract of suspected items in the corpus falling into this category by the Zephir team suggests that closer to 17-18% of the HT corpus falls into those categories.

But if duplication is defined more narrowly as falling under the Type 2 scenario, then *HCC Duplicates 2012* may have been not far from the mark. A small subsample of the Zephir extract indicates that about 30% of the 1.78 million records, involve Type 2 instances, scans of the same physical object. This is much closer to the original estimate of the earlier report.