HathiTrust Program Steering Committee's Response to the Collection Committee's *HathiTrust Monographic Duplication and Uniqueness: 2017 Report and Recommendations*

September 2017

The HathiTrust Collections Committee (HCC) reviewed the 2012 HCC Duplicates Report[1] (*HCC DR 2012*, hereafter) and submitted its findings to the HathiTrust Program Steering Committee (PSC) in February 2017. The PSC shared and discussed its comments with the Collections Committee in March 2017. The final report (May 2017) reaffirms the conclusions in the *HCC DR 2012*, and also extends our understanding of the types of duplication and relative uniqueness in the corpus by means of this duplicates taxonomy:

> **Type 1: Multiple Scan Submissions Produced with the Same Scanning Method from a Single Item. These are considered "true duplicates" that current ingest methods largely prevent.**
>
> **Type 2: Multiple Scan Submissions from a Single Item**
> **These occur where a single HathiTrust 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 event.**
>
> **Type 3: Multiple Distinct Items by a Single Provider**
> **These occur where 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 HathiTrust bibliographic record.**
>
> **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 well as different scan events and/or methods.**

The report raises a number of questions about the value and usefulness of duplicates (or what initially appears to be duplicates) to scholars. As the report notes, Type 1 duplicates, "true duplicates" providing no additional value, are being successfully screened out. And at the same time, the use cases illustrated in the report demonstrate the value of stewarding the other types of duplicates. If such duplicates are of value, might we be more intentional about it as a collection development driver? Are there areas of the collection that might be targeted for efforts needed to fully benefit from non-Type 1 duplication? In doing so, what would be the costs weighed against the benefits? Assuming we can effectively screen out the Type 1 duplicates, then we can plan to steward the valued duplicates? So the key issue is how to achieve greater deliberateness in ingesting, describing, and supporting useful duplicates in the corpus. PSC will want to explore the cost/benefit of accepting these other duplicate types that, with better metadata, might improve the user and researcher experience. We also want to affirm that there is sufficient precision in HathiTrust's ingest mechanisms to insure that the preponderance of duplication is of the valuable sort.

In offering its responses to the specific recommendations below, the PSC acknowledges the considerable complexity and inherent constraints of the data collection task that the Committee undertook to arrive at its analysis and recommendations. While an expanded methodology may produce some variance in the numerical findings, the PSC does not regard such potential differences as significant enough to change the recommendations or policy implications. In looking ahead at next steps, the PSC notes that in many cases agreed-to metadata and other enhancements resulting from these recommendations would benefit from drawing on expertise from a number of other HathiTrust groups: HathiTrust Operations, the Metadata Policy, Strategy, Use and Sharing Advisory Group (MUSAG); Zephir; and the Quality Assurance and Standards Working Group (QASWG). Any resulting implementation of these recommendations would likely be the responsibility of HathiTrust Operations.

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

Response to Specific Recommendations

Regarding Recommendation A, PSC agrees that Type 1 scans are the only category of duplication in the corpus that is truly redundant. While eliminating and preventing them is desirable, the report suggests that they are believed to be rare or non-existent in the corpus. Therefore, it's unclear whether their numbers warrant significant effort. We will aim for improved identification of duplicate types in general, by drawing upon QASWG and MUSAG to define metadata that can better identify the varying duplication types as a basis for future duplicate remediation decisions.

Type 2 duplicates, particularly those resulting from different scanning methods (as opposed to different scan events) have been identified in both the 2012 and 2017 reports as valuable for some purposes (see Recommendation D); Recommendation B suggests that their use cases be subjected to further study to better understand their value to users. Until this is better understood, and in light of their relatively small numbers and the difficulty of establishing priority among them, we agree that there should be no attempt to eliminate such duplicates from the corpus at present. We will ask the Collections Committee to investigate further such use cases and/or "survey" to clarify the value of this duplicate type for users.

With respect to prospective ingest of Type 2 duplicates, Appendix A notes the desire among some digitizing partners to allow new Type 2 duplicates to be ingested under certain conditions, which current ingest routines do not support (see the report, Appendix A, item 2, Example 2). While PSC agrees in general with the desire to minimize additional Type 2 duplication prospectively, we believe there is a need to develop a more explicit policy about whether certain Type 2 duplicates can be accepted on a going-forward basis and under what conditions (e.g., library scans of items previously scanned by Google).

PSC also agrees with Recommendation B, that the HathiTrust partnership should continue outreach to both the HathiTrust 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. We will encourage exploration of HTRC's computational capacities to do the kind of text mining and quality analytics among duplicates done at UC-Berkeley that is described in the report's section III. A. This seems to be a ripe opportunity of intersecting concerns (HTRC, QASWG, MUSAG).

The PSC also agrees with recommendation C, which affirms the *HCC DR 2012* that de-duplication should not be attempted for early printed works. Provided that scans and catalog records for these items meet minimum standards, these scans should always be retained. To further note, this recommendation appears principally aimed at Type 3 and 4 duplicates, i.e., those representing different physical items or manifestations. The 2012 report recommended that the HathiTrust partnership survey scholars to determine whether there is a generally agreed-upon date before which de-duplication should not be conducted; however such a survey was not undertaken. HathiTrust should solicit feedback scholars on this point as part of the user research suggested in Recommendation B, or refer the question of a cutoff date to the Collections Committee. Until such research is undertaken, we provisionally recommend that 1850 be considered a cutoff date earlier than which duplicates should never be removed.

PSC very much supports Recommendation E, that leveraging Type 3-4 scans, and possibly Type 2, is a valuable and important initiative. We support consideration of efforts to enhance the metadata and user experience to allow bibliographically identical but unique HathiTrust monograph scans to be more accurately identified, and more easily leveraged for close and distant reading. As noted above, we ask MUSAG and QASWG to advise on the appropriate metadata requirements.

PSC would like to consider more use case examples to support Recommendation F, which considers the value of mashups to help create a best HathiTrust copy composed of best page scans of a given manifestation.

PSC fully supports Recommendation G, which identifies metadata enrichment as key to leveraging the identification and discernment of valued duplicates. PSC will refer this to MUSAG and QASWG for further investigation.

Summary

In summary, PSC recommends making it clear that, following the further study and metadata work recommended in the report, that the membership be consulted to affirm that the value of such duplication warrants the costs of effort and infrastructure entailed.

In accepting the report, PSC is grateful to the Collections Committee not only for the thoughtful analyses and recommendations it includes, but especially for framing the issue in terms of the use cases that HathiTrust aims to support.