HathiTrust Community Metadata Strategy Task Force

Final Report
Submitted to HathiTrust PSC
6.16.2021

Executive Summary

The Community Metadata Task Force has worked from January 2020 to June 2021 to create “a set of guiding principles and potential priorities for metadata enhancement actions and workflows addressing primarily member-contributed bibliographic metadata deployed by HathiTrust in supporting ingestion, management, development, and use of the digital collection.”

The current HathiTrust metadata management model and workflows rely on metadata submitted by contributing institutions, i.e., that metadata editing, enhancement, and correction should be done by each individual member institution rather than by HathiTrust directly. This principle, designed in 2008 when HathiTrust was founded, made sense at the time, as HathiTrust was and is a member-driven organization. However, the metadata landscape has changed since then, and the services that HathiTrust provides have changed drastically.

As of June 2021, HathiTrust curates more than 17 million digitized items from around 60 contributing institutions and its membership has grown to 10 consortia/state systems and 290 individual institutions, at the time of writing. Considering the importance of metadata in HathiTrust’s discovery, access, and preservation services, the Task Force believes that it is time to rethink the metadata workflows as well as its management model.

After meeting with various stakeholders and reviewing previous reports on metadata workflows, the Task Force came up with a set of recommendations for four stakeholders: member institutions, OCLC, HathiTrust, and the HathiTrust Research Center. In addition, the Task Force concluded that it would be ideal for HathiTrust to become an organization that manages its own data, beginning with bibliographic metadata, to accomplish its strategic goals and become a leader in initiatives including Equity, Diversity, and Inclusion (EDI) and Linked Open Data (LOD) in discovery services. The Task Force proposes further that HathiTrust work in three stages of exploration, development, and implementation in metadata life-cycle management systems for the next five years:

- Stage 1 (Exploration - as soon as possible): Explore ways to make metadata corrections in-house that ensures timely services for users around the world.
- Stage 2 (Development - next 2 years): Develop and test new HathiTrust metadata management workflow
- Stage 3 (Implementation - next 3-5 years): Implement new metadata life-cycle management workflows and systems and lead EDI, LOD and other initiatives.

Introduction

The Community Metadata Task Force was formed in January 2020 by the HathiTrust Program Steering Committee (PSC) to “create of a set of guiding principles and potential priorities for metadata enhancement actions and workflows addressing primarily member-contributed bibliographic metadata deployed by HathiTrust in supporting ingest, management, development, and use of the digital collection.” The Task Force charge document focused on the following areas:

- Review HathiTrust's position in the overall metadata ecosystem for libraries and advise PSC and HathiTrust operations on opportunities and policies concerning bibliographic metadata enhancement.
- Identify high priority use cases and issues that result from current limitations around enhancing metadata.
- Review current policies for member-contributed bibliographic metadata, and determine which types of metadata are appropriate candidates for enhancement and correction in HathiTrust.
- Liaise with and educate the member community about issues related to metadata enhancement policies for two-way communication on issues related
  to metadata enhancement of member-contributed metadata.
- Develop a set of guiding principles and, if determined appropriate, draft proposed text for member-contributed metadata policies for PSC review.
- Engage units of HathiTrust and partner organizations potentially affected by metadata policies. Leverage staff membership in the group to consult with HathiTrust Operations to assess the implications of proposed guiding principles and the feasibility of potential policy changes.

This report aims to provide detailed responses to these areas of focus based on the results from an updated environmental scan including reviews of previous reports, use cases, and meetings with stakeholders. It recommends considerations critical to support robust HathiTrust metadata management and services. The Task Group recognizes that these recommendations cannot be achieved without the support of HathiTrust contributing institutions and other stakeholders in the wider community. In addition, several of the recommendations for HathiTrust require a major shift in HathiTrust resource allocation, including but not limited to hiring of additional staff, and new developments on workflows and systems architecture.

Environment scan

As a matter of policy, bibliographic metadata in HathiTrust is contributed by depositing member institutions, i.e., HathiTrust does not create or edit bibliographic metadata. HathiTrust uses this member-contributed metadata for its services such as discovery in its interface, API, print holdings data review, research, etc. As the HathiTrust landscape has grown over the years, the strengths and limits of this approach to metadata have also changed. In order to ensure that members had a complete understanding of this complex landscape and where there was potential for innovation, the Community

Metadata Strategy Task Force reviewed several seminal reports and conducted informational sessions with numerous key stakeholders. This approach ensured that the Task Force had an understanding of topics that had been pursued in the past along with an updated view of the issues.

Metadata in HathiTrust - Past Work

The Task Force reviewed two reports on HathiTrust metadata: the 2016 report "HathiTrust Metadata: An Environmental Scan" and the 2019 report by the Metadata Policy, Strategy, Use and Sharing Advisory Group (MUSAG). The Task Force also invited two MUSAG chairs to hear their experiences. Although these reports were written three years apart, they shared very similar findings as well as recommendations.
Some of the findings include:

1. Working with member-contributed data is challenging, as each member institution has different local metadata practices and different metadata quality;
2. Since institutions have different ways of organizing/binding serials, there are clear limitations on how HathiTrust can work with serial volume (chronology/enumeration) information;
3. Duplication detection and selection criteria for the HathiTrust preferred record should be improved by utilizing OCLC services, other types of identifiers, or additional metadata fields such as subject headings;
4. Current metadata enhancement workflow has gaps in various areas across HathiTrust user services.

Interestingly enough, these findings and some of the recommendations continued to surface at meetings with various stakeholders listed below.

OCLC

As part of the practices established between OCLC and HathiTrust, OCLC has an automated process of creating e-resource records in WorldCat with HathiTrust holdings using contributors' records. As HathiTrust membership, as well as its services and needs for member institutions, have grown significantly since 2008, the Task Force strongly believes that it is time to revisit the established practice with OCLC and address possible areas of collaboration that would improve overall HathiTrust metadata management.

1. **OCLC numbers for print records:** HathiTrust asks member institutions to submit print records for their digitized resources and the majority of these are OCLC records. Zephir uses these print OCLC record numbers as the primary method for clustering bibliographic records submitted by member institutions to display in HathiTrust.
2. **OCLC e-resource records creation:** OCLC uses an automated process to generate e-resource records from the OCLC numbers of print records in HathiTrust when the library sends the files to OCLC. Those e-resource records have HathiTrust as a holdings library.
3. **Corrections to OCLC WorldCat records:** Since HathiTrust is a holdings institution of OCLC, not a member institution, HathiTrust cannot receive updated OCLC records and cannot use them to

update HathiTrust records. Instead, updated OCLC bibliographic records can only be submitted by the contributing institution. This system relies heavily on member institutions that often have limited bandwidth for sending updates.

4. **Metadata sharing**: The current OCLC metadata sharing policy restricts bulk metadata downloads and distribution. In addition, there are restrictions that limit the sharing and use of subject headings. This has frustrated researchers whose work depends on bulk access to this data.
5. **Concordance files:** OCLC regularly provides HathiTrust with the "concordance file," which provides a list of updated and merged OCLC numbers. OCLC numbers in library records as well as in HathiTrust can include out-of-date OCLC numbers due to the regular merging of records in OCLC. During the summer of 2020, HathiTrust began using the concordance file in order to improve Emergency Temporary Access Service (ETAS) and in March, 2021 Zephir began to use it for record matching. For example, if a library submits a record with an old OCLC number, the concordance table will provide related numbers to establish a match with records in various HathiTrust workflows.

Zephir

As a gatekeeper of the HathiTrust metadata, Zephir works with all partner institutions to coordinate the metadata ingestion and remediation processes. The outcomes of the products are delivered to the HathiTrust Digital Library for creating other types of metadata and used for discovery services.

1. **Ownership of metadata**: Zephir's metadata work relies solely on contributing institutions since HathiTrust tries its best to keep contributed metadata in the state in which it was contributed, i.e., Zephir does not directly correct metadata when errors are reported. This impacts discovery services, along with the clustering of the records.
2. **Working with contributing institutions**: There is no requirement for contributing institutions to refresh their submitted metadata. HathiTrust does ask members to refresh their holdings periodically and to correct errors when they are reported but does not have any enforcement mechanisms. This can cause discrepancies between metadata in member databases and that in HathiTrust, i.e., while the metadata in partner institutions is up-to-date, the metadata in HathiTrust is not, and can impact matching/clustering. In addition, metadata can vary greatly between institutions, with some institutions submitting metadata describing digital content even though HathiTrust requires metadata for print resources.
3. **Metadata correction**: Given the circumstances noted above, metadata correction and enhancement is not fast or easy.[1] Zephir and HathiTrust ask institutions to correct and resubmit metadata whenever errors are reported. However, many institutions do not respond to metadata correction requests, furthering the limitations in terms of managing and utilizing metadata that HathiTrust does not own, and that ultimately impacts HathiTrust services overall.
4. **Use of metadata**: Zephir uses different parts of the bibliographic records in order to cluster and ultimately "score" records so that the best possible record is chosen for display in HathiTrust.

[1] The Zephir team has created and submitted recommendations to improve the evaluation of records as Appendix 1.

Since there are limits on the use and reuse of certain fields, some potentially valuable data such as Subject headings are ignored during scoring.

HathiTrust Digital Library

As the HathiTrust Digital Library is the repository of numerous data stores, the Task Force focused on bibliographic metadata, including information stored at the volume or item level, stored in HathiTrust. The Task Force has identified two overarching challenges in evaluating the central position of bibliographic data in the HathiTrust Digital Library:

1. **Uneven metadata quality:** Contributed bibliographic records drive most of the other types of HathiTrust metadata, yet they are uneven in terms of completeness and quality. This is largely due to the variance of cataloging practice both through time and across contributing institutions. This affects title level and item level information, notably the enumeration and chronology information in serials volume information.
2. **Limits on metadata use and sharing:** The richness of the metadata that is contributed is not utilized to its fullest extent. This is largely due to the current practice where one record is elected to represent a given title when various depositors contribute items for that title. The scoring algorithm elevates one record to represent the title, and the other records, and whatever richness they could contribute, is deprecated by default.

HathiTrust Research Center

The HathiTrust Research Center (HTRC) enables computational analysis of the HathiTrust corpus and seeks to help meet the technical challenges researchers face when dealing with massive amounts of digital text. In its development of software tools and data sets for advanced computational access to the HathiTrust Digital Library, HTRC relies heavily on HathiTrust's metadata; it likewise enhances substantial portions of that metadata, and creates substantial additional metadata.

HTRC's efforts in the areas of text mining and non-consumptive research include support for scholars making the fullest possible use of HathiTrust content within the confines of current U.S. copyright law, as well those undertaking larger-scale analysis of the HathiTrust collections, potentially to discover (and even propose remediation of) collection gaps. And it is clear that there is potential for HTRC's past and current work to enhance HathiTrust's use of metadata.

User Engagement Task Force

The User Engagement Task Force shared how they are working to better understand HathiTrust members' needs and challenges. The discussion centered on the task force's upcoming survey. While the task force has usage statistics and anecdotal information, it does not have good data about how people engage with HathiTrust; the survey is intended to fill this gap. The survey will primarily focus on determining what aspects of HathiTrust people are aware of and whether they know how to use the

services. Although metadata will not be a major emphasis of this survey, it is possible that responses will address metadata issues.

Linked Data

Since early 2000, member institutions have been exploring possibilities of utilizing linked data in metadata creation and discovery services in order to improve the discovery of resources and workflow efficiency. The Library of Congress has developed a new ontology, BIBFRAME[2], with the goal of making library data interoperable with other data in the semantic web. OCLC has made all their data available in Schema.org[3] vocabularies in WorldCat with entities represented in URIs, and has been developing entity management initiatives[4]. In addition, the LD4P Group[5] and ShareVDE[6] are now looking at how linked data could benefit the discovery of library resources. The HathiTrust Research Center currently creates and distributes linked data records (including both bibliographic and statistical metadata), describing HathiTrust content in its Extracted Features dataset[7].

HathiTrust Community Week 2020 Review

In response to the COVID-19 pandemic, HathiTrust changed to a fully virtual format for its 2020 Community Week sessions. The Community Metadata Task Force reviewed the sessions for intersections with its current work. Many of the sessions surfaced the same concerns and areas for improvement that the Task Force has seen throughout its survey: inconsistent quality, difficulty utilizing serials information, desire for more subject headings, etc. The sessions also made it clear that the creation of HathiTrust's Emergency Temporary Access Service (ETAS) has made many of these underlying metadata issues more pressing than they had been in the past. HathiTrust uses metadata in various ways to provide or deny access to materials and so areas where information is lacking or inaccurate have had a direct impact on users at a critical time.

The Community Week sessions also surfaced an issue that had been on the margins of some discussions in the Task Force: the current integrated library software (ILS) landscape. Several of the libraries who presented mentioned their use of HathiTrust in the Alma ILS and several Task Force members are at libraries that either already have or are currently migrating to Alma. The Task Force does not intend to make any recommendations that are built around a specific ILS but the prominence of one particular vendor in the space may present opportunities for libraries using shared software in the future.

---

[2] https://www.loc.gov/bibframe/
[3] https://schema.org/
[4] https://www.oclc.org/en/news/releases/2020/20200109-oclc-awarded-mellon-grant-linked-data-management-infrastructure.html
[5] https://wiki.lyrasis.org/pages/viewpage.action?pageId=74515029
[6] https://www.share-vde.org/sharevde/clusters?l=en
[7] https://wiki.htrc.illinois.edu/pages/viewpage.action?pageId=79069329

Recommendations

Based on findings from the environmental scan, the Task Force identified three key stakeholders around which to center recommendations for HathiTrust engagement: member institutions, OCLC, and HathiTrust Research Center. In addition, the Task Force has recommendations that are purely internal to HathiTrust Digital Library. Recommendations are labeled as short-, mid-, and long-term as applicable based on the anticipated amount of time and other resources that would be needed. In addition, appendix 2 shows the combined list of tasks organized by their level of impact and level of difficulty.

Member Institutions

It has been more than twelve years since HathiTrust was founded (2008)[8], and approximately eight years since the Zephir system was inaugurated to ingest metadata (2013)[9]. During that time, there have been many changes to both library staffing models and the systems library use for the description and discovery of resources. These recommendations seek to strike a balance between the critical need for high-quality metadata and the acknowledgement that member institutions are able to support HathiTrust functions at different levels. Therefore, some of the following recommendations suggest mechanisms to provide more mutual support among member institutions.

Short-term recommendations

1. **Increase support for greater member participation in maintaining metadata quality.** HathiTrust should set clearer expectations and requirements, and offer stronger member support, for:
    - Submitting high-quality and current print metadata.
    - Periodic refreshment of metadata records, for example, every two years (which is not uncommon in other programs such as shared print endeavors)
        - Support this effort by providing reports to contributors of the records that they have provided over time to HathiTrust.
    - Participation in metadata corrections, for example keeping contact information current and responding to requests from HathiTrust User Support in a timely manner.
2. **Establish a metadata mentoring program for new contributor institutions:**
    - Compile a centralized list of metadata contacts and the institution's ILS for each existing contributing institution.
    - When new contributor institutions are onboarded, match them with existing members that have the same ILS.
    - This will facilitate and support the goals of improving the quality of contributed records, regularizing updates, and processing bibliographic corrections.

[8] https://www.hathitrust.org/about
[9] https://www.hathitrust.org/zephir

3. **Establish metadata-focused regular communication channels**: ideally channels should be multidirectional, allowing communication from HathiTrust to members, and the reverse, and members to each other.
    - Review current communication channels available in HathiTrust and find ways to communicate with member institutions for any metadata related matters, such as Metadata office hours or listservs.
    - Reinvigorate the community of metadata workers and stakeholders.
    - Share community solutions for metadata scenarios and situations.

Mid-term recommendation

1. **Create a system to support the correction of bibliographic data by volunteer "experts"**, similar to, or perhaps an expansion of, the HathiTrust Bibliographic Corrections Group. As noted above, many libraries do not have the staff or resources to meet all requests for corrections to their records, nor are HathiTrust staff currently able to take on this work. Why not lift the burden from contributing libraries and share it out among a volunteer corps of experienced cataloging professionals? Develop an agreement to use with libraries that wish to participate in a "mutual aid" program for metadata improvement, both those who are willing to have others correct their data and those who wish to work on corrections.

    The Task Force believes that this approach will:
    - Increase accurate, corrected metadata in the HathiTrust catalog.
    - Improve clustering and discoverability and a better user experience.
    - Relieve participating contributors of this work.
    - Create professional development opportunities for the correction experts.

    However the Task Force also acknowledges that there are some challenges as well:
    - Zephir was not designed with any sort of cataloging interface and is unlikely to develop one, so at least initially, changes will have to be made using current ingest routines.
    - A shift from the notion that the canonical "record of record" resides in the contributor's ILS.

Long-term recommendation

1. **Investigate options for merging or amalgamating fields from all cluster members into a preferred record** that is exported from Zephir to HathiTrust, for a richer representative record. This kind of activity (selecting fields from all records in a cluster) already takes place on a small scale, to get the bibliographic system IDs of other cluster members and insert them into the exported preferred record.
    - The Task Force would advocate for adding other fields, such as subjects.
    - Improved merging would render unnecessary the more complicated idea of making all records in a cluster accessible to users, which current technology infrastructure does not support.
    - Investigate and develop new ways to cluster records more accurately (beyond the use of OCLC numbers and other system numbers).

- Research methods and feasibility of “machine learning” algorithms for matching records
- Explore existing tools for validating, analyzing, and matching metadata, developed and employed by other organizations.

OCLC

The Task Force believes that OCLC can play an important role in improving HathiTrust metadata quality and resource sharing. We recognize that heavy reliance on one bibliographic utility carries its own risks and limitations but also acknowledge that the reality is that OCLC plays a substantial role in the current metadata landscape and that the current arrangements need to be revised.

Short-term recommendations

1. **Review the current relationship with OCLC**, especially the following issues:
    - **Metadata sharing:** Clarify terms of re-use / redistribution of metadata records or fields (e.g., subject headings).
    - **Concordance file**: Explore opportunities to improve how to use this file as well as future development of API lookups in OCLC.
    - **Future cooperation**: Begin a discussion about future options for working with HathiTrust in terms of updating or receiving records, using linked-data and next steps with metadata.
    - **Review creation of e-book records:** Review the current processes related to OCLC creating new e-book records from print book records when digitized content is submitted to HathiTrust. Instead of assuming e-book records need to be created from print, what are new ways to approach these records? Many libraries catalogue their digitized collections, so stopping duplicate records would be good. Is there a way to customize the process to do, if needed, instead of always creating new e-book records? This will potentially impact HathiTrust metadata workflows as more and more member institutions will submit born-digital resources that do not have print records.

Mid-term recommendations

1. **Establish a new partnership with OCLC:** Currently there is no clearly defined relationship between HathiTrust and OCLC. Since OCLC works as a shared database for bibliographic records and initiates several research projects that would potentially benefit HathiTrust, the Task Force recommends HathiTrust to consider the following:
    - Identifying the areas where HathiTrust can benefit from OCLC beyond making HathiTrust collection discoverable in the WorldCat.
    - Ingesting OCLC records directly into HathiTrust catalog

    Benefits:
    - HathiTrust would be able to make edits in a centralized, cooperatively developed metadata repository.
    - HathiTrust records would incorporate cooperatively updated metadata improvements.

- HathiTrust database would have the latest MARC data and OCLC number.

Challenges:

- HathiTrust may need to manage its own ILS, or at least a cataloging module, to work with OCLC.
- The OCLC record may not have important local information only available in member contributed metadata.
- The OCLC record is always being updated, and can change in ways that will affect HathiTrust workflows (e.g., incorrectly clustered record, edition/copyright changes)

2. **Identify ways for HathiTrust to work with OCLC to reduce the number of duplicate e-resource records in OCLC.** As noted above, there are instances where the automated process through which OCLC generates e-resource records based on print records submitted to HathiTrust winds up duplicating existing e-resource records in OCLC. If OCLC and HathiTrust were full partners, they may be able to identify a workflow that prevents the duplicates from being generated. In addition, it may provide opportunities for libraries to submit new types of records in cases where libraries already have e-resource records for their collections or digital resources they want to submit to HathiTrust.

Long-term recommendation

1. **Consider the potential benefits of OCLC's Entity Management Infrastructure** in addition to other linked data developments, as OCLC develops services and fee structures.
    - OCLC's linked data infrastructure may be able to help support some of the challenges discussed. For example, there is a possibility that the Entity Management Infrastructure will include the OCLC ID lookups service that could replace the large concordance files being generated for HathiTrust. Discovery of HathiTrust resources could be improved as OCLC's linked data presence grows.
    - OCLC numbers are increasingly being added to Wikipedia and Wikidata which then appear in knowledge graphs in Google search results. This will ensure that WorldCat records have up-to-date HathiTrust links and rights statements and HathiTrust resources discoverable within 2 clicks from a Google search.

HathiTrust Research Center

HathiTrust and its Research Center (HTRC) have a long history of mutually reinforcing policies, funding and management structures, and cross-pollinating professional relationships. These should certainly be maintained and strengthened. But the Task Force feels that even more can be done to strengthen the relationship and deepen the integration of the two organizations, particularly as regards the use and production of metadata. These recommendations reflect three broad questions of HTRC's metadata-related activities:

1. How to enable HTRC to make even more robust use of member-contributed metadata.
2. How to leverage HTRC's metadata enhancements and products, both current and future, for the improvement of HathiTrust collections and services.
3. How to encourage an HT-HTRC whole greater than the sum of its parts.

Our recommendations below reflect these three broad areas of activity.

Benefits and challenges

The “research and development” aspects of HTRC generally reflect not the traditional corporate sense of an R&D division focused on improving the functions of its parent, but rather HTRC’s primary mission: to “enable computational analysis of the HathiTrust corpus” -- that is, to promote research by scholars in the community *outside* of the HathiTrust organization. While the Task Force certainly does not recommend weakening this commitment, we believe that more and better uses of HTRC work by HathiTrust are possible and desirable. Beneficial aspects (and their related challenges) of this approach include:

- An opportunity for HathiTrust to engage and take advantage of metadata work already being done in the course of HTRC business
- An opportunity to further a long-standing commitment to "translational work" -- that is, to “translate” work done by HTRC on behalf of, or in support of, scholarship, into the more infrastructure-focused work of HathiTrust
    - Acknowledging a potentially very broad understanding of “translational work,” for purposes of this report we focus primarily on work that touches on metadata and metadata workflows
- The flow of metadata and metadata practices can and should go both directions:
    - On the one hand, we recommend considering HTRC-produced metadata as a new source for ingest into the HathiTrust metadata environment
    - On the other hand, we will recommend an even more robust use of HathiTrust member-contributed metadata in HTRC work (e.g., easing current proscriptions on the use of subject headings)
- An opportunity and challenge for both organizations to take advantage of slightly different organizational styles, rather than attempting to impose any single style on both. In broad brushstrokes:
    - HTRC operates in nimble, experimental, and research-driven ways, regularly pursuing funding opportunities for innovation and research
    - HathiTrust rightly focuses more on developing and sustaining stable and reliable infrastructures and services for its community

Short-term recommendations

1. **More Robust Use of Member-contributed Metadata:** Support the use of the full wealth of member-contributed metadata in HTRC data products.
    - In particular, enabling reuse of both subject headings and unselected member records in the Zephir system would create a raft of new data for research-oriented users of the HTRC
    - We note that any new OCLC partnership (as recommended above) could change the allowable uses of records, which may end up being characterized as no longer "member-contributed"

2. **Encourage a more permeable metadata membrane between HathiTrust and HTRC (in both directions):**
    - HTRC-created metadata (e.g., machine-generated front-matter metadata) should be made more readily available to HathiTrust for use in the Digital Library.
    - HathiTrust should in turn seek more opportunities to share data and metadata with HTRC for research purposes.

Mid-term recommendations

1. **Leverage metadata to support a more equitable and representative Digital Library**: Metadata-driven gap-filling efforts are an important part of HTRC's current Mellon-funded project, "Scholarly Curated Worksets for Analysis, Reuse & Dissemination" (SCWAReD). We recommend the following in support of these and other gap-filling activities:
    - Use HathiTrust holdings metadata to identify "missing items" in member library collections, and encourage those libraries to digitize and submit them to HathiTrust through existing workflows.
    - Loosen existing expectations for print records such that gap-filling digital objects of non-standard format (e.g., TEI editions), without a corresponding print record but held by a member library, might be ingested into the collection (see above recommendation under OCLC regarding e-resource records).
    - Enable a metadata path for "digital donations" such that gap-filling digitized objects located in a non-member library (or outside of a library altogether, e.g., held by a private collector), might be "donated" to the collection via a member library (or directly to HathiTrust).
2. **Implement HTRC Discoveries into HathiTrust Services**: Some discoveries e.g., the opening of pages of in-copyright works as identified by automated front-matter detection projects, and actively seek similar future opportunities. (This will better support members as they engage in rights review projects, and facilitate more works becoming open to full-text use.)

Long-term recommendations

1. **Make Use of HTRC Metadata Enhancements in the HathiTrust Digital Library**: Explore how and whether some of the HTRC metadata enhancements can be incorporated into the HathiTrust Digital Library, including non-MARC metadata such as automatically-detected genre, image detection, automatically extracted named entities (including their Linked Data URIs), etc. (This may serve to enrich access points for patrons of the HTDL.)
2. **Explore organizational enhancements that encourage an HT+HTRC whole greater than the sum of its parts.** The Task Force recommends a light touch when considering and engaging in the organizational changes described above: in general, we believe that the current organizational structures and relationships between HathiTrust and the Research Center are successful. Our recommendations are focused on building on the strengths of those relationships, not on changing them radically.

HathiTrust Digital Library

The Task Force proposes the following approaches to addressing two broad challenges identified in the above Environment Scan section:

Short-term recommendation

1. **Identify barriers:** Correcting and resubmitting bibliographic metadata is not an easy task for all depositing institutions. The Task Force recommends HathiTrust identify challenges that member institutions experience and ways to help them to contribute and edit their metadata.
2. **Strengthen the bibliographic metadata specifications**: Encourage higher quality metadata on the intake.

Mid-term Recommendations

1. **Devise easier metadata enhancement methods**: As noted elsewhere, the current metadata editing workflow is challenging and the need for streamlining the error correction process is critical. We recommend exploring the creation of a new interface and potentially shifting correction work to others besides contributors. A new workflow would facilitate the improvement of greater portions of metadata already in the HathiTrust Digital Library.
2. **Explore the possibility of merging depositing members' records**: Clusters of records in HathiTrust often have different fields that could be combined to improve the overall record. Merged records may provide the basis of intelligence to begin to deal with challenges presented by enumeration and chronology data, since the metadata in HathiTrust is currently flat MARCXML at the item level that includes information from multiple institutions. This would need to be considered along with utilizing the OCLC Primary Record, as has been described above. The mechanisms to do this might range from a single fabricated merged record, or information merged in a single presentation through use of linked data. Either way, the provenance of each data point could and should be expressed in the merged record.
3. **Use of alternative IDs**: As HathiTrust expands, some new members do not use OCLC records or numbers in their metadata workflow. Workflows being developed need to support this growth. In particular, HathiTrust needs to review what options could be employed as alternative primary IDs for clustering.
4. **Evaluate and revise the scoring algorithm**: As noted elsewhere, records are assigned a score as they are ingested in Zephir. The score affects which record is selected during the export process as the "preferred record" to represent its cluster. The current scoring algorithm only scores on the *presence* of certain metadata, not on the content or quality. It has not been evaluated since its original implementation in 2013. A more discerning assignment of scores would promote selection of records with richer metadata. Revisions could include examining records and adding points for markers of quality such as the ones in this short list - numerous others could be considered for inclusion:
    - Presence of subject fields (currently not taken into account)

- RDA versus AACR2 cataloging
- Records cataloged by Library of Congress, CONSER, or the Program for Cooperative Cataloging
- Encoding Level (Leader/17) for level of cataloging data

Other Considerations

The Task Force also wishes to highlight the following two areas that could impact not only the metadata management but also overall HathiTrust discovery services.

Equity, Diversity, and Inclusion (EDI)

The Community Metadata Strategy Task Force recommends that HathiTrust continue to be an advocate for equity, diversity, and inclusion (EDI), and consider developing goals and initiatives to address EDI concerns in descriptive metadata.

The HathiTrust Statement of Values includes its commitment to EDI. According to the statement, HathiTrust strives to "promote justice and create communities that thrive and advance collective knowledge and understanding" and "develops practices and acts so that our organization, services, and programs fully support these principles." The Community Metadata Strategy Task Force believes that metadata should play a critical role in HathiTrust's EDI efforts through identifying potentially harmful and offensive metadata terms, and leading campaigns and initiatives to update them.

In fact, the Library of Congress (LC) tried to revise the subject heading "Illegal aliens" in 2016 to "Noncitizens" and "Undocumented immigrants" as alternative subjects to "Aliens" and "Illegal aliens" as recommended by the ALA/ALCTS Subject Analysis Committee (SAC) Working Group.[10] Unfortunately, the House of Representatives objected to the change and ordered LC to continue using the terms. However, there have been a lot of discussions and actual conversion works for those headings, especially Aliens and Illegal Aliens, since 2016 as those terms impede efforts towards EDI. HathiTrust has already conducted some research into the existence of these headings in its catalog, so the Task Force recommends that HathiTrust use them as a test case for making changes based on the options featured in the "long-term recommendations" below.

The Community Metadata Strategy Task Force recognizes that some EDI-focused activities in the HathiTrust environment are already underway. For example, the HathiTrust Research Center is engaged in activities in support of these principles. The current Mellon-funded Scholarly Curated Worksets for Analysis, Reuse & Dissemination (SCWAReD) project[11] is focused on making a number of reusable worksets and research models focused on HathiTrust content created by and representative of marginalized and underrepresented groups available for digital scholarship. Specifically the project aims

[10] https://alair.ala.org/handle/11213/9261
[11] https://ischool.illinois.edu/news-events/news/2020/10/htrc-receives-500000-andrew-w-mellon-foundation

both to use HathiTrust metadata for the identification of items for these worksets, and to enable its enhancement for their appropriate description, discovery, and promotion.

Short-term recommendation

1. **Creation of Official EDI Statement:** As a preliminary action, the Community Metadata Strategy Task Force recommends that HathiTrust develop an official statement of support for EDI in metadata beyond what is included in its current value statement, and possibly use additional types of statements to inform users (e.g. disclaimers, sensitivity messages, etc.). Many memory institutions have already created and shared such statements to address institutions' committeements to EDI.[12] This work can be accomplished by a short term task force or by a HathiTrust Metadata team.

Long-term recommendations

1. **Advocate for ethical cataloging practices**: HathiTrust should adopt and employ critical and ethical cataloging practices, such as Cataloguing Code of Ethics[13], where appropriate and/or feasible for member institutions. Because HathiTrust metadata is contributed by member institutions, HathiTrust should promote the importance of EDI-compliant metadata among its member institutions.
2. **Identify action plans for EDI projects**: HathiTrust should start identifying EDI-related metadata projects. The ultimate goal for these projects would be identifying offensive terms in HathiTrust and remediating those terms with alternatives to follow EDI best practices.[14] Many institutions have already implemented EDI initiatives that emulate these principles, such as:
    - Improving description
        - Identifying and removing/replacing offensive LCSH or local terms
        - Advocating for the change of offensive LCSH or local terms
        - Add the alternative terms in 653 as keywords
        - Work with discovery layers to identify offensive terms and display alternate terms
    - Improving representation
        - Connecting and consulting with communities represented in current collections and evaluating description
        - Identifying underrepresented communities and developing associated collections
        - Highlighting and promoting diverse collections

Many memory institutions that address EDI issues in their metadata do so in two ways: changing problematic terms to in records themselves or changing them in display interfaces. While making changes in the catalog records is often more permanent, it has its own challenges. With specific regard

[12] http://cataloginglab.org/list-of-statements-on-bias-in-library-and-archives-description/
[13] https://docs.google.com/document/d/1IBz7nXQPfr3U1P6Xiar9cLAkzoNX_P9fq7eHvzfSlZ0/edit
[14] Please see the ALA/ALCTS Subject Analysis Committee (SAC) Working Group report above.

to HathiTrust, since metadata is contributed and updated by member libraries, if changes are not made at the contributor level they may not stick in the HathiTrust version of the record. Some of these concerns could be addressed through other recommendations made above around changes to updating and correcting metadata within HathiTrust. In addition, many problematic terms are still used and available in LCSH, so many institutions work with their discovery layers to make changes on the user side. The Task Force recommends that HathiTrust pursue both options.

In addition, the Task Force advises that HathTrust consider incorporating subject headings in its metadata scoring algorithm. As noted above, Zephir does not consider subject headings when scoring records to determine the representative record in a cluster. We recommend that HathiTrust work with Zephir to determine the potential for evaluating subject headings so that records with EDI-compliant terms may be weighted more heavily than those without them.

Like many other recommendations in this report, successful implementation relies heavily on HathiTrust taking an active role in maintaining its metadata.

Linked data

Linked data offers HathiTrust users the potential for better discovery of HathiTrust resources on the Web, faster updates to metadata, and easier use and reuse of HathiTrust metadata. The use of URIs in linked data could also help improve the clustering process when records are ingested into Zephir as well as overall metadata management work. In addition, it will also offer an opportunity to redesign the discovery mechanism in HathiTrust Digital Library.

Benefits and challenges

The Task Force recognizes potential benefits that linked data can bring and a list of challenges that HathiTrust faces.

Challenges:

- Currently, few, if any library systems on the market currently can export bibliographic records in BIBFRAME or other linked data formats. This could require that HathiTrust take on linked data transformation and authority reconciliation work for all member-contributed data.
- Sending records to third party vendors for conversion to linked data is likely to be expensive and require new workflows
- Paucity of library applications that can take advantage of linked data would require investment in development work.

Benefits:

- Ease of maintaining accurate metadata
- Expand findability of related works by exploiting various entity relationships
- Expose HathiTrust resources to the web
- Move from record-based to entity-based catalogs

- Enable new forms of research and analysis

Short-term recommendation

1. **Engage with community linked data development work:** The Community Metadata Strategy Task Force recommends that HathiTrust actively engage with current linked data groups and their new developments. While HathiTrust is not involved in creating linked data, groups like OCLC, LD4P and Share VDE, and many individual institutions have been experimenting with linked data in discovery environments. HathiTrust should closely follow and engage with these groups.

Long-term recommendations

1. **Work with HathiTrust Research Center linked data:** One of the premier data products of the HathiTrust Research Center, "Extracted Features," has included automatically generated Linked Data since release 2.0 (2020). During and after the production of this data set, HTRC has engaged in a number of related experimental efforts: visualizations, ontology development, automated record conversion, and automated entity reconciliation. HathiTrust should find a way to leverage the works and findings already accomplished by the HathiTrust Research Center, and start a joint pilot project that tests how linked data can benefit HathiTrust users and member institutions.
2. **Explore ways to facilitate Linked Data in HathiTrust Services:**
    - Make HathiTrust resources searchable on the web: While HathiTrust hosts unique and abundant digital resources contributed by member intuitions around the world, those resources are not searchable/discoverable on the web. There is a way to add microdata or json (or meta tags) on each webpage that allows search engines (SE) to find and index the web pages. As HathiTrust strives to serve all users, multiple approaches should be explored so users can find and discover HathiTrust resources from any search engine.
    - Entity reconciliation: this can be done without converting the entire existing data set and would provide rich contextual information to HathiTrust users. The Task Force recommends identifying appropriate linked data sources and information that is useful for HathiTrust users.
    - Ensure that HathiTrust metadata is in sync with linked data projects like OCLC's Entity Management Infrastructure so that key fields flow in and out. For example, ensure that HathiTrust metadata like URLs, rights information, and other relevant information are in the entity record. In addition, HathiTrust should explore workflows that would consume relevant fields as well, such as related formats, author IDs, subject headings, etc.

The Task Force recognizes that this work will likely require the massive redesign of a new discovery service site redesign and configuration of the system.

Moving Forward

The Task Force acknowledges that the recommendations and considerations outlined in this report cannot be achieved fully in the current metadata management model. We therefore propose that instead of relying solely on contributing insitutitions for metadata enhancement, HathiTrust should transition into an organization that manages its own data and explores other options that will ultimately improve its services.

Why should HathiTrust consider such a significant shift from this fundamental approach? There are several reasons. Much has changed in the library domain and other cultural heritage institutions in the past 10+ years since the HathiTrust Digital Library was originally established, not least within HathiTrust itself. Library budgets for acquisitions and for staffing are continually squeezed and reduced; everyone is urged to do more with less. Concurrently, the amount, range, and format of information resources to which libraries are expected to provide access have exploded, along with fast-changing technologies and products.

Libraries and other information organizations are challenged to keep up with these competing demands and their costs. Many simply cannot allocate the resources to correct records submitted to one of many allied organizations, no matter how strongly they believe in its mission. This leads to outdated and incorrect metadata in the HathiTrust Digital Library, to which contributors have entrusted their digitized materials. While everyone agrees that high quality metadata is a must-have to support the best collocation and discovery of resources, only a few institutions can or will commit to providing it regularly.

HathiTrust services rely on accurate metadata, including the recently developed Emergency Temporary Access Service. The challenges faced by libraries and other organizations leads to a gap between the services that HathTrust is able to provide its users and what the contributing institutions would like to be able to provide and what its users expect.

As soon as possible	Next 2 years	Next 3-5 years
Explore options to make metadata corrections in-house and start doing it on a small scale.	Develop and test new HathiTrust metadata management workflows and infrastructure.	Implement new workflows and provide new services with quality metadata. Lead EDI, Linked data, and other initiatives.

To make this possible, the Task Force recommends that HathiTrust work in three stages of exploration, development, and implementation in metadata life-cycle management systems for the next five years. The appendix 3 includes the three stages in greater detail along with an assessment of required costs and benefits for each stage.

Stage 1 (Exploration): Explore ways to make metadata corrections - As soon as possible

- Purpose - HathiTrust needs to change its metadata correction workflow. The current workflow is a complicated and cumbersome process, with no certain outcome, and this requires immediate action.
- Actions - Find ways to correct metadata in-house, including:
    - Correct and update metadata on a case by case basis;
    - Correct dates and enum/chron in individual records using existing HathiTrust and Zephir protocols;
    - Enlist experienced members of Bibliographic Corrections Group (BCG) to do the work;
    - Recruit volunteers from other member institutions.
- Outcome - Ability to edit/correct metadata.
    - HathiTrust has the ability to make needed metadata corrections in a timely manner, so that researchers and users better find what they need;
    - Contributing institutions are relieved of making “one-off” corrections and record resubmissions;
    - HathiTrust has reduced the number of duplicated tickets for metadata corrections of the same metadata.

Stage 2 (Development): Develop and test new metadata management models - Next 2 years

- Purpose - HathiTrust needs new ways to obtain metadata and keep them up-to-date.
- Actions - Develop a pilot project to identify and test what HathiTrust needs to manage its own data, including:
    - Explore a new partnership with OCLC for possibilities of utilizing OCLC to receive and update records in Zephir and HathiTrust;
    - Rethink modes of deposit and updating of metadata in HathiTrust;
    - Develop new clustering algorithms independent of OCLC numbers;
    - Work with HathiTrust Research Center to identify areas of collaboration in metadata enhancement and services.
- Outcome - HathiTrust has experience with new metadata management models leading to an implementation plan.

Stage 3 (Implementation): Implement new metadata life-cycle management workflows and systems - Next 3-5 years

- Purpose - HathiTrust can be a leader in linked data and EDI initiatives for not only the member institutions, but also all cultural heritage institutions with a new metadata life-cycle management and workflow.
- Actions - Implement new workflows, metadata management model, and systems identified in Stage 2. Determine and executing tasks including:

    - New system architecture to manage the metadata life-cycles and workflows;
    - Re-align or augment Staffing related to metadata life-cycle management;
    - Actions related to utilizing linked data;
    - Continue advocating for EDI initiatives.
- Outcome - HathiTrust can efficiently and effectively manage its metadata and provide best and timely user services to users around the world using the latest technology. Recognized as a leader in support of linked data and EDI initiatives related to metadata.

Conclusion

Metadata is a key to HathiTrust services, including discovery and research support, the cornerstone of the HathiTrust service. However, the current metadata enhancement workflow and metadata life-cycle management model in HathiTrust are far from ideal. The Task Force believes that some significant changes are needed. As the membership grows and the need to provide more robust user services increases, it is clear that HathiTrust needs a new metadata management model that supports present and future services. ~~And~~ HathiTrust has an opportunity to redesign its metadata life-cycle and workflows, and become a leader in critical initiatives in memory institutions.

Given the current role of HathiTrust in the library metadata landscape and discovery services, the potential for and need to become the leading organization continue to grow in its stewardship of metadata such as EDI and linked data work. The Task Force wants HathiTrust to continue to lead developments in the use, stewardship, and sharing of metadata not only for the HathiTrust member institutions, but also all memory institutions and users that use HathiTrust collections. In addition, the Task Force hopes that this new development will make it easy for contributing institutions to submit their metadata and for HathiTrust systems to share that metadata and to keep it up-to-date.

HathiTrust needs to make use of the best descriptive practices and keep EDI values at heart. Discovery and research support are at the forefront of the HathiTrust mission, supported and enhanced by quality metadata. In order to continue to facilitate these core missions, HathiTrust should be part of those current initiatives and a leader of the future thinking projects. The Task Force believes that the current metadata life-cycle management workflow be reviewed, that staff be charged with these efforts, and systems updated or changed, so that HathiTrust is well positioned for the future.

Acknowledgements

The Community Metadata Strategy Task Force would like to thank the following colleagues who shared their metadata expertise relating to HathiTrust.

- Timothy Cole and Stephen Hearn (MUSAG Group)
- Christopher Cox (User Engagement Group)
- Claudia Conrad (Zephir)

- Cynthia Whitacre, Bill Carney, Nathan Putnam (OCLC)
- Heather Christenson (Federal Documents Advisory Committee)

Task Group Members

- Benjamin Bradley, University of Maryland
- Barbara Cormack, HathiTrust staff, California Digital Library
- Graham Dethmers, HathiTrust
- MJ Han, University of Illinois, Chair
- Joseph Hafner, McGill University, PSC Liaison
- Nancy Lin, New York University
- Chris Long, University of Colorado Boulder
- Elizabeth Miraglia, University of California, San Diego
- Andrea Payant, Utah State University
- Michelle Paolillo, Cornell University
- Tim Prettyman, HathiTrust staff, University of Michigan
- Sheila Torres-Blank, Texas State University
- Glen Worthey, HathiTrust Research Center, University of Illinois

Appendices

1. Use Cases for Direct Editing of Zephir Database Records
2. Recommendations in the Grid
3. 5-Year Recommendations for HathiTrust