

UC Davis

Library Staff Publications

Title

Final report on the Material Selection Pilot from the Google Books Partners Working Group on Material Selection and Collection Development

Permalink

<https://escholarship.org/uc/item/6dv310kk>

Authors

Majors, Rice

Strieb, Karla

Ewing, Renata

Publication Date

2026-02-01

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Final report on the Material Selection Pilot from the Google Books Partners Working Group on Material Selection and Collection Development

February 2026

Rice Majors, Karla Strieb, Renata Ewing

Pilot participants

University of California Davis Library

Jeff Brainerd

Rice Majors, co-chair

Sage Koller

Cornell University Libraries

Michelle Paolillo

Mira Basara

Ohio State University Libraries

Tony Maniaci

Gene Springs

University of Michigan Libraries

Simon Hunt

Geoff Stoller

Rutgers University Libraries

Jeff Carroll

Christopher Sterback

Cathy Pecoraro

California Digital Libraries

Renata Ewing

Paul Fogel

HathiTrust

Kristina Hall

Heather Weltin

Big Ten Academic Alliance

Karla Strieb, co-chair

Cite this report:

Majors, R., Strieb, K., & Ewing, R. (2026, February.) *Final report on the Material Selection Pilot from the Google Books Partners Working Group on Material Selection and Collection Development*. <https://escholarship.org/uc/item/6dv310kk>

1. Executive Summary

Executive Summary: Material Selection Pilot Report

Overview

After two decades of mass digitization, Google Books partner libraries—which provide **95% of HathiTrust content**—are shifting from passive "shelf-clearing" to **intentional collection building**. The 2023–2025 Material Selection Pilot tested whether five U.S. institutions could coordinate to fill specific scholarly gaps in the digital corpus, specifically targeting "Bees & Insect Pollinators" and "Puerto Rico."

Key Findings & Challenges

The pilot proved that coordinated digitization is operationally feasible, contributing **2,248 targeted volumes**. However, it revealed significant systemic barriers:

- **Logistics over Strategy:** Current workflows prioritize "convenience" (e.g., offsite storage items) over high-value, high-use materials in branch libraries or special collections.
- **The "Black Box" Model:** Google's "waterfall" assignment system prevents libraries from seeing each other's candidate lists, meaning one library's logistical hurdle creates a permanent gap that another partner could easily fill.
- **Content Gaps:** Reliance on Google's pre-generated lists has created a "missing middle" of scholarship, particularly for items published after 1964.

Strategic Recommendations

To transition from passive scanning to a responsive digital library, the report recommends:

- **For Google:** Refresh candidate lists to include post-1964 materials and provide transparency across partner lists to allow for "divide and conquer" coordination.
- **For HathiTrust:** Develop analytics to identify "negative space" (what is missing from the corpus) and lead community-wide "Calls for Scanning."
- **For Leadership:** Library Deans must advocate for collective action to prioritize high-impact research content over logistical ease.

Conclusion

The pilot demonstrated the community's capacity for strategic selection. By addressing data-sharing and infrastructure gaps, the partnership can move from a "completist" scanning operation to an active force for preserving the scholarly record. A **Phase 2 Pilot is proposed for late 2026** to further refine these coordinated workflows.

2. Background and Context

As the mass digitization of print books has matured, and as the COVID-19 pandemic sharpened our ability to see the value of mass digitization, Google Books partner libraries (especially those who are also member libraries of HathiTrust) have recognized a shift in needs. With Google Books scanning accounting for approximately 95% of the content in the HathiTrust Digital Library, the partnership has moved through two distinct phases:

1. **Shelf-Clearing Phase (began 2004):** Initially, mass digitization workflows focused on scanning all volumes on the shelf at a small number of partner libraries to build the base corpus.
2. **Candidate Selection Phase (2011 - now):** a larger group of U.S. libraries select materials from candidate lists approved by Google (U.S. public domain materials, the “renewal period” of U.S. copyright (1931-1963), and various types of government documents), built from snapshots of the libraries’ holdings.

However, this current model has limitations:

- **Passive Engagement Risks:** Relying solely on candidate lists creates “holes” in the corpus. Branch library collections and any closed access or non-circulating collections are deprioritized, and high-use items—which are frequently checked out—are rarely on the shelf to be pulled for scanning.
- **From Volume to Value:** Libraries need to manage assembling shipments with reasonable efficiency, but this can make digital collection building deeply suboptimal. High value items may take years to move into the workflow, waiting in line behind arbitrarily selected low-priority titles.
- **Logistical Divergence:** The libraries’ understanding is that Google assigns items only to a single library’s candidate list. Partner libraries operate under different logistical constraints but are blind to each other’s candidate lists, meaning that a candidate item that is logistically inconvenient at one library can’t be contributed by another partner whose copy might be readily available for shipment.
- **HathiTrust Integration:** HathiTrust stewards the digital corpus for most U.S. scan partners – and Google scanning provides the vast majority of content for the corpus. To intentionally build the digital corpus inherently requires that coordinated digital collection building by partner libraries harness Google scanning efforts.

3. The Material Selection Pilot

3.1 Scope and Participation

The impetus for this project formed out of discussions at the 2023 Google Summit in Ann Arbor, Michigan. The pilot was developed by the Material Selection and Collection Development Working Group of the partner libraries. From the beginning, the group included Google partners from the U.S. and a handful of libraries in the UK and EU. However, as we scoped out a pilot,

we realized that the digitization workflows of non-U.S. Google partners differed significantly from U.S. based partners. The UK and EU partners exclusively send pre-20th century volumes for Google digitization and they do not work from Google candidate lists. Also, unlike U.S. based partners, very few UK and EU partners contribute their scans to HathiTrust. Ultimately, we decided that the pilot should focus on the interested U.S. based libraries who are contributing members of HathiTrust.

Once the pilot project was scoped, all Google partner libraries were invited to participate. Ultimately, librarians from five HathiTrust contributing institutions agreed to participate and met regularly from August 2023 to November 2025 to collaborate. This included library colleagues from Cornell University, University of Michigan, Ohio State University, Rutgers University, and the University of California, Davis. Kristina Hall, HathiTrust's Rights and Access Specialist, as well as Paul Fogel and Renata Ewing, both from the California Digital Library and Karla Strieb from the Big Ten Academic Alliance, also joined the meetings to provide information and support while learning alongside the participants.

The group agreed upon three overall pilot goals. These included, to:

1. Discover challenges and pathways to **optimize existing workflows** for the intentional contribution of materials going to Google for scanning
2. Demonstrate and learn about the **appetite and capacity** among partner libraries for more intentional contribution of materials
3. Show a **positive impact on users'** ability to engage with digitized materials, especially in research and scholarship; and to begin to develop ways of assessing such impact

To help realize goals, the group decided to coordinate the intentional contribution of two topic areas during 2024 and 2025 and to collaborate and support each other with operational realities and best practices as much as possible. For example, a team member shared information on how to mash-up a Google candidate list with Alma data.

3.2 The Identification of two focal topics

In scoping the pilot, the Working Group decided upon two prompts to test the pilot's proposition and workflow(s) as part of the initial scoping of the pilot.

- Bees & insect pollinators
- Puerto Rico

To investigate ways that materials could be selected for the pilot, two types of prompts were selected. "Bees & Insect Pollinators" maps well to Library of Congress call number ranges, and participating libraries could more easily locate materials within their collection using the suggested call numbers. The prompt "Puerto Rico" was conceived broadly and included many subjects, including the literature, history, culture, and economics of Puerto Rico. It required the team to grapple with a prompt that was not clearly defined and did not map easily to Library of Congress call numbers.

There were additional reasons these prompts were chosen. For instance, “Bees & Insect Pollinators” were of particular interest to Cornell and UC Davis (and Rutgers, once they joined) because they are land grant institutions and have agricultural programs.

Both selected prompts shared the following characteristics:

- Books and serials on these topics published before 1964 remain **highly relevant** to contemporary research and scholarship.
- These topics would likely include publications that are likely to be government documents that are (or could be) in the **public domain**.
- These materials are likely to be **held in the “main” library** on a campus which should help to streamline operational processes for paging/sending/etc.
- The topics are broad enough that any scan partner library is likely to have some **materials to contribute** for the pilot.
- Together, the prompts encompass a **broad array of disciplines** across the academy.

The pilot participants shared methods they used to identify the materials in their collections that were sent in support of the two prompts. This included sharing subject headings, keywords, and Library of Congress call number ranges used to search their Integrated Library System (e.g., Alma or Sierra). As well as strategies for comparing these findings to Google’s candidate lists.

Library of Congress call number ranges from our November 2024 meeting

Topic	Range	Notes
Bees and beekeeping	QL461-569.999 AND keyword="bee*"	Too many titles about other insects if no keyword limit. This range is about insects.
Bees and beekeeping	SF521-539.999	This seems to be the apiculture section, cleaner than perhaps some others.
Pollinators	QK926-926.999	This is for pollen/pollination. Could perhaps expand QL bees/beekeeping searches to pick up other insect pollinators.
Puerto Rico	E183.8-184.999 AND keyword="puerto rico"	These don't seem to be Puerto Rico-dedicated call numbers. Note that E184X seems to be diaspora, i.e., "Puerto Ricans."
Puerto Rico	F1950-1999.999 AND keyword="puerto rico"	These don't seem to be Puerto Rico-dedicated call numbers.

Puerto Rico	HC154-157.999 AND keyword="puerto rico"	These don't seem to be Puerto Rico-dedicated call numbers.
Puerto Rico	PQ7420-7442.999	This is likely the dedicated range for Puerto Rican literature.

3.3 Candidate List Characteristics

The pilot participants analyzed the efficacy of Google's candidate lists. Typically for Google scanning, the candidate lists are created by Google largely without input from library partners. Candidate lists are currently restricted by publication date based on assumed copyright status. Initially, Google scanned everything regardless of publication date or copyright status. Starting in 2011, Google has chosen not to digitize library volumes assumed to be in copyright (at the time, those published after 1923 that are not U.S. federal government documents). Google later expanded their candidate lists to include volumes they deemed likely to be public domain but in need of individual copyright review. This included the U.S. renewal period (i.e., published in the U.S. before 1964), U.S. state and local government documents, and international government documents. Taken together, these practices have left the HathiTrust corpus with many gaps in volumes published since 1964, and in particular, a dearth of volumes published since 2011.

Google also allows individual publishers to opt out of having their publications digitized by the Google Books Project which has exacerbated the gaps in HathiTrust's digital collection. Google started publisher opt outs at the beginning of the Google Books Project, so long ago that participating publishers may have lost sight of their "opt out" decision. But more recently, the risk/value proposition of digitization has also potentially changed for publishers with the new potential for scans to advance AI training. The full impact of publisher opt-outs and their alignment with current publisher preferences is of concern to all library scan partners generally, and more specifically is generally a significant barrier to intentional digital collection building.

One potential mitigation strategy is for partners to complement the Google Books Project by digitizing volumes themselves and contributing them to HathiTrust. This allows HathiTrust member libraries more control over what they scan and contribute. But local scanning is cost and resource prohibitive, and at this time only a few HathiTrust members are contributing volumes they've digitized themselves. Overall, the number of volumes digitized and contributed by members is extremely small compared to Google's mass digitization efforts, but they are selected according to the library's criteria, not Google's. However, even selecting volumes for local digitization becomes difficult when Google partner libraries lack visibility into each others' candidate lists.

4. Outcomes

The pilot demonstrated that coordinated digitization is possible ("We were able to do it!"). The effort was deemed "very manageable," though the yield for most individual partners was modest. As of the pilot conclusion, partners had contributed **2,248 volumes** intentionally selected for the two prompts. All five pilot partners were able to discover and contribute volumes from their candidate lists in response to the prompts. Hathitrust was able to verify that their collection quickly reflected the volumes contributed through the pilot.

5. Key Learning & Challenges

5.1 Community Coordination and Selection Criteria

A significant finding from the pilot was the lack of infrastructure for partner libraries to identify and agree on strategic focuses for coordinated collection building. There are currently no obvious or objective criteria for selecting collective prompts for partner libraries to use, nor is there a governance entity empowered with providing a community venue to make these decisions for the Google Books partnership (or for HathiTrust).

Anecdotally, there was considerable collective interest among participating libraries to respond to requests for high-value content. The pilot validated that the community can respond to community-determined prompts. There remains a gap around how to identify and promulgate calls for targeted collection-building.

5.2 "List" vs. "Non-List" Workflows

The working group's pre-pilot conversations engaged a broad cross-section of partner libraries to reveal a major structural divide between:

1. **Candidate List Libraries:** Largely bound by Google's pre-generated candidate lists. Their ability to "select" is limited to prioritizing within the candidate list except as authorized by Google.
2. **Non-List Libraries:** Have greater freedom to push distinct collections but face higher logistical hurdles in getting items approved and shipped. Yet, these libraries are also constrained by their jurisdictional copyright laws and consequently are sending older materials.

5.3 The Collection-level "Gap Analysis" Problem

Partner libraries need access to better analytics to see what the collective corpus *lacks* versus what an individual library *has*.

HathiTrust is currently investing in creating new capacity to analyze its corpus. Identifying what is *not* digitized remains even harder than identifying what *is*. Both the Working Group and the group of pilot libraries discussed the need for tools to analyze the "negative space" of the corpus— anecdotally, for non-English languages and specific subject domains (e.g., [TRAIL](#) technical reports).

For example, we hypothesize but cannot prove that the following issues exist – or exist as we describe them:

- Based on our understanding of the value of our print collections, we theorize that applying a “completist” approach to selected topics or groupings of content would afford new research outputs and potentially new research methodologies, new educational outcomes, and/or new clinical care outcomes. This is a difficult hypothesis to test not least because currently we can’t confidently identify all the material on a topic among the Google partner libraries to know that we’ve scanned what we collectively own.
- Gaps from earlier phases of digitization linger in the corpus without easy remediation. For instance, during the Shelf-Clearing Phase, anything from a multi-volume set that was checked out by a user (which we assume strongly correlates to “the most wanted parts of the collection”) was not scanned, and we assume these “holes” have never been addressed by subsequent scanning.
- Convenience workflows introduce unintentional deprioritization of digitization for some topics. For instance, partner libraries prioritizing logistical convenience for decades means that materials most often shelved in branch libraries are underrepresented in the corpus, as they have commonly been de-prioritized. A lightweight review of the branch libraries of the U.S.-based partner libraries showed some trends in what disciplines were most likely to be supported by branch libraries, and we assume that some topics (e.g., veterinary medicine) are therefore systematically underrepresented.
- In another example, for partner libraries prioritizing shipping materials from offsite storage facilities, there could be an unfortunate tendency to ship materials moved to storage because of age and non-usage. Prioritizing those materials unintentionally moves low use, older materials to the front of the digitization line while leaving behind high use and likely higher value materials. During the pandemic, users commonly noted the apparent preferencing of older, low use items in the digit collection.
- Current candidate lists make it difficult to assess the rarity of items at the community level. “Last copies” remain a common concern both for current and future access. Partner libraries have no simple visibility into how rarely-held the materials on their candidate lists may be, either rare among libraries in general or rare among partner libraries. Partners share an understanding that candidate lists are built on a “waterfall” model of partner libraries – add a book to library 1’s list if possible, else add it to library 2’s list, etc. (A library can mark something as “will not send” to indicate the item is too fragile, too valuable, etc., to send, which we understand will cause it to be pushed to another library’s candidate list.) We also assume Google does some load-balancing among partner libraries. Therefore, the presence of something on the list of a lower-on-the-waterfall library does not necessarily tell us anything about how uniquely it is held. Any effort to focus on prioritizing unique items is stymied by this situation.

5.4 Logistics Drive Strategy

Ideally, libraries want the opportunity for (intentional) selection to drive digitization. The pilot supports our suspicion that, in our current reality, logistics drive digitization.

Partner libraries place an appropriate (even necessary?) priority on logistical convenience in assembling shipments. Library facilities with a combination of easy access to space for shipment assembly, loading docks for pickup and delivery, and staff capacity are preferred. Issues such as binding tightness, fold-outs, shipping constraints (e.g., not having access to a loading dock or available work space at a branch library), and condition checks often override the potential contribution of high-value material. Finding ways to moderate this will be essential to incorporate intentional selection.

In addition to issues involving open collections in branch locations, **Special Collections** pose perhaps more difficult challenges. The inclusion of Special Collections was identified by the pilot group as critical, yet these materials often require workflows that differ from standard mass digitization. In addition to concerns about security and handling rare materials, these materials are often stored in smaller workspaces that don't easily allow for batch work of any kind.

6. Recommendations

The pilot successfully demonstrated that scan partner libraries are interested in coordinating their digitization activities and are able to modify their local selection practices to advance targeted collection building. These successes and the learning that occurred through our work together suggest that further action is necessary and in fact essential.

Based on our pilot results and working group discussions, we submit the following recommendations:

6.1 Strategic "Dean-Level" Advocacy

We believe that our findings merit high-level attention by Library Deans and Directors. Library leaders must undertake collective action to realize the potential of intentional selection for scanning. Leadership must understand that the goal needs to be bigger than someday "finishing" mass digitization; rather, the goal should be to leverage large scale digitization to build a responsive, comprehensive digital library. Our efforts could provide substantially more near term value of digitized collections for users if we can select high-impact content for nearer term access.

6.2 Recommendations for Google

Responsive, comprehensive digital library building is a shared goal with Google. This pilot demonstrates that supporting both selection and collective coordination by library scan partners can advance a more intentional digital library development process. But many of Google's practices still reflect assumptions and needs from 20 years ago, and are overdue to be revisited.

Google's current system prioritizes the prevention of duplicative scanning. This remains important but additional functionality is needed to allow libraries to "divide and conquer" in managing local workflows to aggregate and prioritize their available materials. This, in turn, would allow them to reflect priorities the community identifies for coordinated attention. To advance this objective Google could:

- **Expand range of publication dates on candidate lists:** Expand the publication ranges included in candidate lists removing post-1964 date restrictions to allow access to newer materials (where copyright allows or for discovery, search, and preservation purposes).
- **Engage with its U.S. partners to assess needs and opportunities** for modifying current processes for identifying candidates for digitization. The current inability of partner libraries to see across each other's candidate lists is a critical barrier to coordinated selection and workflow design and management for a collection-building process. For instance:
 - **Offer mechanisms to identify items that are uniquely or rarely held** by partner libraries (based on Google's matching). This could be accomplished simply by adding a column in the candidate list with the number of Google partners that hold a given title/volume in common. Our mass digitization efforts need to better support broader preservation and access priorities, especially for scarcely held items. Digitization is one step of collection stewardship and management.
 - **Create flexibility regarding waterfalls:** To our knowledge, currently items are allocated to a single partner's candidate list via a black box process. Libraries need the ability to see each other's candidate lists, access them in ways that allow for analysis and sharing among the libraries, and move items across each other's candidate lists based on things like local priorities, logistics, and staffing.
 - **Publish the order of the waterfall** to partners so partners can understand who is upstream and downstream. Keep this list up to date as it changes. This allows for more collaborative decision making among partners.

6.3 Recommendations for HathiTrust

In the U.S., the library scan partner community almost entirely contributes its scans to build the HathiTrust corpus. HathiTrust is both an essential partner and a natural locus for coordinated action for North American scan partners and augment their digitization efforts with those of other HathiTrust members. The participation of HathiTrust staff in discussions throughout the pilot process have helped the group understand the opportunities for action for HathiTrust and demonstrated how essential the engagement of HathiTrust will be to any collective efforts to practice intentional collection building.

- **Analytics for gaps:** Prioritize the development of analytics capabilities that allow partners to see "item-level gaps" in the collective library. Insight is needed at the local collection level, but also across the corpus.

- **Organize programmatic priorities:** HathiTrust is uniquely positioned to organize the community by identifying and implementing issue-specific "Calls for Scanning" based on programmatic priorities determined by the community.

6.4 Further Operational Coordination across U.S. Partner Libraries

The gaps, barriers, and recommendations above will take time for Google and partner libraries to address. While infrastructures are being addressed, the pilot group believes that the community would benefit from continuing its learning by organizing a Phase 2 of the pilot.

With the prospect of changes in the publication years accepted by Google for digitization and the resulting enlargement of partner libraries' candidate lists, a "phase 2" pilot beginning in 2026 would be particularly well-timed. Any changes in candidate lists may provide 2024-2025 pilot participants more opportunity to build collections in the focus areas and will allow additional scan partner libraries to explore the implications of incorporating intentional selection into their Google digitization workflows.

Key elements of a Phase 2 pilot:

- Launch and duration: Begin either in Q3 of 2026 or Q1 of 2027 and operate across 12 months of shipments.
- Continue with the current prompts to offer comparability of the two pilots.
 - That said, the most attractive Phase 2 would aim to satisfy identified gaps in the HathiTrust Digital Library. This would likely realize the best return on investment, and be the most attractive to partners/members broadly.
- Refresh participation - phase one participants can continue or withdraw, any new scan partners would be welcomed to join.
- Proactively work with HathiTrust to better track the effectiveness of the pilot in building the corpus.
- Proactively work with Google to build a shared understanding of the way candidate materials are identified and allocated to candidate lists among partner libraries.

7. Conclusion

The pilot proved that partner libraries have the will and the capacity to coordinate digitization efforts, contributing over 2,200 targeted volumes with minimal impact from normal logistical constraints. However, without improved data tools from HathiTrust and Google along with more flexible ingestion criteria from Google, intentional collection development is inhibited by the lack of venue for selection of collective priorities and remains needlessly labor-intensive. By addressing the recommendations above, the partnership libraries can transition from a passive scanning operation—which risks leaving enduring holes in the scholarly record—to an active, strategic collection-building force.