Google Designated Duplicates: Implications for HathiTrust End User Display

Heather Christenson, California Digital Library
2/11/2010 Version 3.0

Introduction

When a Google library partner institution sends a print volume to Google for digitization, Google evaluates whether the volume has already been digitized, and will reject the volume if it has already been digitized from another institution. The institution that offered the print volume may be entitled to a copy of the already-digitized volume. In August 2009 Google began returning digital copies of public domain volumes to a given library partner that were digitized from the print collection of a different library partner. The HathiTrust partners have concerns about how successful Google has been in determining exactly what constitutes a duplicate volume; that topic is being explored via other channels. However, the introduction of these "Google designated duplicates" (GDDs) will need to be addressed within the HathiTrust repository for a variety of reasons. The goal of this document is to explore the impact of this change on HathiTrust's methods of clearly and accurately identifying the provenance of digital and print GDD volumes within end user displays. Broader goals for repository data management are outside of the scope of this paper.

Vocabulary

The following abbreviations[1] and terms are used in this document:
GDD: Google designated duplicate
SOPG: Source of Print volume scanned by Google
CDIH: Contributor of the Digital Item to HathiTrust
Volume: Print or digital version of one bound volume (book, journal, etc)

Potential "Google return" situations

The return of GDDs could cause the following conditions to occur. These conditions may be reflected in the user interface.

Case 1: A volume from a Hathi partner is returned to originating Hathi partner. Examples:
UC original returned to UC alone
UM original returned to UM alone

Case 2: A volume from a Hathi partner is returned to non-originating Hathi partner. Examples:
UM original returned to UC
UM original returned to UC, IU

Case 3: A volume from a Hathi partner is returned to both originating Hathi partner and non-originating Hathi partners(s). Examples:
UM original returned to UM, UC
UM original returned to UM, UC, IU

Case 4: A Volume from non-HT institution is returned to non-originating Hathi partners (s). Examples:
Stanford original returned to UC
Stanford original returned to UC, UM

[1] Abbreviations GDD, SOPG & CDIH are from *New Google Duplicate Detection and Return Procedure Impacts on HathiTrust Bibliographic and Item Level Metadata*, Jon Rothman, October 27,2009

Case 5: Different SOPG copies of the same volume could be returned to different Hathi partners[2] Example: Stanford original returned to UC, and UC original returned to UM

End user experience

In his recent paper[3], Jon Rothman proposed two potential approaches for identifying partner roles associated with the GDDs within HathiTrust systems and processes via object identifiers. The HathiTrust end user displays draw upon these object identifiers, and providing clarity for the end user of HathiTrust services will be an important consideration in choosing an approach. For the purposes of this discussion, the two approaches are can be summed up in a greatly simplified manner:

1. A volume in the HathiTrust is identified as originating from the first CDIH who contributed it. When the end user display identifies the source of the print original, it shows the CDIH. However, the CDIH can no longer be relied on to infer the SOPG. If a display of the SOPG is required, it will have to be achieved via a lookup to an external database.
2. A volume in the HathiTrust is identified as originating from the SOPG. When the end user display identifies the source of the print original, it shows the SOPG. If a display of the CDIH is required, it will have to be achieved via a lookup to an external database.

Currently, there are three situations in which the end user could encounter GDDs: 1) results display (including facets) in VuFind, 2) "item page" in VuFind, and 3) Pageturner display. In the future, the source of original print item may be worth displaying within full text search results & faceting, and in the mobile interface (whether faceting, results display, or another feature). Each of these end user situations was considered, with the following questions in mind:

- How is the institutional provenance currently displayed?
- What are the assumptions driving the display?
- What are the pros and cons of each of the two potential object identifier approaches, for the end user?

Results display in VuFind

The current VuFind results display presents the user with a facet, "Original location", and may provide details: "multiple copies", "Full-text", or "Search-only (no full-text), and an icon indicating full-text items.

Although usability testing would shed more light on this, the current VuFind results display reflects a few assumptions about user behavior:

A. Users may perceive that limiting by "Original location" facet to their own institution will lead them to the exact print volume displayed in digital form
B. Users may perceive that limiting by "Original location" facet to their own institution will lead them to a volume that they have access to locally, in print
C. Users may perceive that limiting by "Original location" facet to their own institution will deliver an accurate count of the number of items "owned" by that institution "How much of our stuff is in there?" There may be administrative or political reasons for their choices

In situations A and B, the user would be best served by approach 2, since "Original location" would be the SOPG and they would be successful. Situation C would be better served by approach 1, since "Original location" would reflect the CDIH. However, since it is possible that multiple institutions could be entitled

[2] This case is somewhat of an edge case for books that were digitized before Google began rejecting duplicates. Each volume is assigned a "quality score" by Google, and the volume with the highest score is selected for return to the non-digitizing library partner. The highest score volume is distributed to all library partners who are entitled to a copy. It could be the case that with incremental processing improvements, another copy of the same volume could be updated and grab the higher "quality score." In this case copies from different SOPGs could go to different CDIHs.

[3] "New Google Duplicate Detection and Return Procedure Impacts on HathiTrust Bibliographic and Item Level Metadata" Jon Rothman, October 27,2009. Apologies to Jon for the gross oversimplification.

to the same volume (Case 3), work would need to be done to create facets that bring together the SOPG and the CDIH in such a way that when a SOPG volume is attributed to multiple CDIHs, and each "Original location" facet would bring up that volume.

Item page in VuFind

The current VuFind item page display presents the user with a view status "Full-text", or "Search-only (no full-text), and an "Original from" statement.
A number of assumptions may be driving item page display:

A. The user may arrive at this display from a non-HathiTrust context. The display must be immediately understandable no matter where the user came from.
B. It is useful to the end user to be able to choose between volumes when there is more than one available.
C. Users want to choose the volume originating from their own institution
D. Users may want to compare full text volumes originating from several different institutions

If approach 1 is used, "Original from" will be a false statement for some volumes. For those volumes, it will be confusing for end user to reconcile "Original from" with the true SOPG. If multiple links are displayed, the user is presented with an even more confusing choice.

If approach 2 is used, "Original from" will be accurate for the end user, but users who arrive from a non-HathiTrust context will likely not be able to distinguish between Hathi partner and non-Hathi.

It is worth pointing out that there may be possible ways to bring together the SPOG and CDIH information in a blended statement, e.g. "Digitized from Stanford and provided to UW" which could provide multiple elements of context to the end user.

Pageturner display

The current Pageturner display presents a watermark on the page images with two statements: "Digitized by…" and "Original from…".

Assumptions driving page turner watermark display:

- The user may arrive at this display from a non-HathiTrust context. The display must be immediately understandable no matter where the user came from.
- The user needs to understand provenance of the original print item.

If approach 1 is used, "Original from" will be a false statement for some volumes. For those volumes, it will be confusing to end user to reconcile "Original from" with visible marks on the image identifying the true SOPG.

If approach 2 is used, "Original from" will be accurate for the end user, but users who arrive from a non-HathiTrust context will likely not be able to distinguish between Hathi partner and non-Hathi.

Again, there may be possible ways to bring together the SPOG and CDIH information in a blended statement that could provide multiple elements of context to the end user.

Discussion

Regardless of what data we are tracking within the repository, we may choose to display or not display any piece of information we've accounted for. Making a decision about which identifier approach to use, and to what extent we are willing to do development of new architecture requires us to make a decisions about principles for the discovery interface.

There are potential approaches to display that take a more "all or nothing" bent, and are useful to consider in order to identify our principles:

- What if we dispensed with institutional provenance altogether, and put the HathiTrust name in instead? “HathiTrust, Digitized by Google”
- What if we simply left “Original from…” off?
- What if we just changed the text to “Contributed by…” instead of “Original from…”?
- What if we removed the watermark entirely and provide a link to “for more information about this digital object” page or other mechanism such as hover box?

Considering these approaches raises some questions we need to answer, which may require more research and discussion.
Related to user behavior:

1. Is it a goal to lead the user to the exact print volume represented in the display?
2. Is it a goal to lead the user to a print item that is the same manifestation as the volume represented in the display?
3. What use cases would require the user to understand the provenance of the original print volume?

Related to HathiTrust as an organization:

4. Is it our goal to identify the CDIH? Why?
5. If there are multiple CDIHs, is it our goal to identify them all? Why? If so, is the order of CDIH display important?
6. Are there reasons to display the SOPG when it is not a Hathi partner?
7. Are there concerns related to non-HT SOPG institutions’ perception regarding the presence and display of “our copies of their copies” within the HathiTrust?
8. Are there concerns that identifying SOPG outside of HathiTrust would enable “free riders” and could freely link to all their volumes that the HathiTrust manages and displays and pays for? The number of GDDs may increase in the future.
9. How do we accommodate the change that is involved when a non-HT SOPG joins the HathiTrust? How does this play out in the UI?
10. Are there concerns regarding liability, if a SOPG in-copyright work is designated as full view based on the CDIH bib record?
11. Is it our goal to have the user recognize HathiTrust as a trusted “brand”? What qualities do we want to be associated with our name?
12. Are there aspects of HathiTrust partner Google contracts will need to be considered in end user displays, and what range of variation needs to be accommodated? Contractual limitations and variations in contracts may need to be addressed. For example the current UC contract says “…will identify the works, in a statement on a web page or other access point to be mutually agreed by the parties, as ‘Digitized by Google’ or in a substantially similar manner.” What do other partner contracts say?

Recommendations

Near term

1. Take approach 2, and use the SOPG within the identifier.
2. Display the SOPG, and find a way to also display the CDIH, whether by using an external database lookup or otherwise. Explore ways to bring together the SPOG and CDIH information in a blended statement.

Longer term
Work through the questions above, with attention to:

3. More research, both usability and literature review, to determine our target audience and what those end users really care about in terms of provenance statements and filtering.
4. More broadly, determination of who our stakeholders are for the end user display, and what their needs are.
5. Ways to separate out administrative, partnership and political needs from user needs.