New Google Duplicate Detection and Return Procedure
Impacts on HathiTrust Bibliographic and Item Level Metadata

J. Rothman, October 27, 2009

DRAFT

Introduction -- New Google duplicate detection and return procedures

In August 2009, Google implemented new duplicate detection and return procedures for items received for scanning from library partners. Under the new procedures, Google's intention is to avoid scanning any item more than once. When Google identifies a volume as unique, the return process remains essentially unchanged. When Google identifies a volume received for scanning as a duplicate, the volume is returned without scanning. If it is also identified as a public domain item, the files associated with the previously scanned volume that Google identifies as the same work are made available via GRIN. In this case, there is a code indicating that this is a duplicate and the GRIN entry for the duplicate contains two sets of identifying information: the institution code and item ID associated with actual source volume that was digitized, and the institution code and item ID for the volume that was submitted but rejected for scanning as a duplicate.

This change raises questions and issues for HathiTrust in a number of areas spread across our systems and processes. The goal of this document is to examine the impacts of this change on our handling of bibliographic and item-level descriptive metadata in HathiTrust.

Background – bibliographic metadata in HathiTrust

Prior to ingest of an object in HathiTrust, relevant bibliographic and item level metadata is loaded into an Aleph ILS (Mirlyn). For each object in the HathiTrust repository, we maintain an Aleph item record containing the relevant HathiTrust object ID. The object ID is composed of a namespace identifier (which is associated with a specific source institution) plus a unique (within that namespace) item identifier, e.g. uc1.b4174938 or mdp.39015004214865. Each item record is associated with a bibliographic record, thus associating the HathiTrust object with bibliographic metadata that describes it. Prior to ingest, the related item record is "inactive" (i.e. suppressed from public display, not included in output files of HathiTrust content, etc.). After ingest, the IDs for newly ingested objects are supplied to a process which uses item level (enumeration/chronology) and/or bibliographic (publication date) data in Aleph to make an initial rights determination. Once the rights have been determined and set in the rights database, the HathiTrust object IDs in the item records are "activated", making these objects visible to public access systems and allowing their inclusion in metadata output files, etc..

The HT object IDs and associated bibliographic/item metadata are used by HathiTrust in a number of different contexts and in a variety of ways. These include:

- VuFind catalog displays
- pageturner displays
- automated rights determination (as noted above)
- indexing for full text search
- API outputs
- OAI files and tab-delimited data files provided for use by the library community
- the OCLC e-Content Synchronization process (which will be based on the tab-delimited files)
- the planned HathiTrust/OCLC public catalog (which will, in turn, be based on the records created by the e-Content Synchronization process, augmented in real-time with item-level metadata from the SOLR indexes on which the VuFind catalog is built).

Partners and IDs

Prior to the advent of these Google designated duplicate (GDD) materials, we have been able to assume that for each HathiTrust object there is a single associated HathiTrust partner institution and that the partner institution has two roles relative to the object: the HT partner is both the Source of the Original Print item scanned by Google (SOPG) and the Contributor Of the Digital Item to HathiTrust (CDIH). Up to this point those two roles have been tightly linked and, for a single item, could not be split between different institutions. Given the tight linkage between those roles, we have been able to assume that the namespace code part of the HathiTrust Object ID for any item indicates both the Source and the Contributor for that item and that the identifier portion of the HathiTrust Object ID is the same identifier that was supplied to Google by the Source institution. This has allowed us to make the following operational assumptions:

1. The HT namespace ID code in any HT object identifier can be used to infer the source institution (SOPG).
2. The SOPG will always be a HathiTrust member.
3. There is a direct relationship between an HT object ID and a Google volume ID (i.e. an HT object ID plus a relatively small table linking HT namespace IDs with Google institution codes provides all of the information needed to generate a Google volume ID that links to the equivalent volume via the Google API).

Introduction of the GDD materials invalidates the assumption that SOPG and CDIH institution will always be the same. For GDD items, those two roles may belong to two different institutions and the SOPG is not necessarily a HT partner. This introduces a new level of indirection in identifiers – the identifier from the CDIH item that was rejected as a duplicate becomes, in effect, an alias for the SOPG identifier for the item that was actually scanned. It is also possible, of course, that multiple items, potentially from multiple CDIH partners, will be returned as duplicates for a single original. For each of those items, the files returned would be identical, the only difference being that each would have its own distinct CDIH identifier – in other words, multiple "aliases" associated with a single object in the HathiTrust repository.

This added complexity in potential partner roles and the meaning of associated identifiers faces us with an important decision regarding which identifier we will use to build the HT object IDs for the GDD items. While that decision is, arguably, outside the scope of this paper, it has significant ramifications that run throughout the topics that **are** in scope and, therefore, I am treating it as in scope for the purposes of this discussion. As I see it, there are two options:

1. Construct the HathiTrust object ID using the first CDIH identifier under which the item is returned in GRIN.
    a. Pros
        i. The HT Object ID always represents an item that was contributed by a HathiTrust partner institution.
    b. Cons
        i. It can no longer be assumed that the HT Object ID directly represents the print original that was scanned to create the object stored in the repository, so the source institution can no longer be consistently inferred from the object ID.

ii. In the case of a second/subsequent GDD associated with a given print original, the GRIN entry would not contain any direct link to the existing HT Object ID created when the first GDD associated with that original was ingested.
iii. Relationships between IDs that need to be managed and recorded are more complex. Under this scenario there are potentially three types of IDs to contend with:
    1. The ID associated with the original print item that was scanned.
    2. The CDIH ID associated with the first rejected GDD, which becomes the HathiTrust Object ID.
    3. Additional CDIH IDs (if any) associated with second and subsequent GDD returns of the same original item.

    ID #1 is the constant from the Google perspective, and it is needed to connect data from multiple GRIN entries if the same item is returned as a GDD more than once. ID#2 is an alias for ID#1 and is used as our local constant (HT Object ID). ID#3 is, for HT purposes, an alias to ID#2, but linkage must be established via ID#1 at the time of ingest.
iv. A direct link between the HT object ID and the equivalent Google volume ID can no longer be assumed.

2. Construct the HathiTrust object ID using the SOPG identifier:
    a. Pros:
        i. The HT object ID continues to consistently represent the print original that was scanned to create the object in the HT repository.
        ii. There continues to be a consistent direct relationship between the HT object ID and the Google volume ID.
        iii. Relationships between IDs are less complex:
            1. The SOPG is always the same as the HT object ID.
            2. CDIH IDs for GDD items are never HT object IDs – they always function as aliases to the HT object ID/SOPG.
    b. Cons:
        i. We will need to create and maintain namespaces to represent Google library partners who are not HathiTrust partners. (Note – assuming that we will want to identify the source of the print original when that source is not a HathiTrust member, we'll need some kind of codes to represent these institutions regardless of our choice about HT object IDs.)
        ii. It seems conceivable that some non-HT partners could object to being identified as the source of original in Hathi public interfaces. (Note – assuming, again, that we will want to identify the source even when it is not a HathiTrust partner, this issue exists regardless of our choice about HT object IDs.)

The sections that follow explore the effects each of these two approaches will have in various Hathi-Trust related contexts where bibliographic and item-level metadata are used.

Ingest and Item Activation

HathiTrust partners provide bibliographic data, including their original item IDs, for loading into Aleph prior to ingest. This data is loaded suppressed from public display. The ingest process includes checks for the presence of bibliographic/item data (keyed on the item ID) which must be passed before ingest is

allowed to occur. After items are ingested into the repository, their HT object IDs are provided to a process which performs initial automated rights determination and updates the rights database. Finally, there is a process that "activates" the items by unsuppressing them in Aleph. The two approaches for creation of HT object IDs for GDD items would have the following effects:

1. Under the "first CDIH" approach, the HathiTrust object ID would (in the case of the first contributing partner) continue to match the data originally supplied in the bibliographic data. No changes to the ingest check on bib. availability, the initial rights determination, or the "activation" process would be required.
2. Under the SOPG approach, special handling for GDD items would need to be added to the check for availability of bibliographic data (i.e. the check would need to be keyed on the ID from the HathiTrust contributing partner, which would not be the same as the intended HathiTrust object ID). The same would be true of the initial rights determination process. When the list of IDs is supplied for activation processing, it would need to include both the CDIH partner's ID and the HT object ID (based on the SOPG ID). The activation process would need to match based on the CDIH ID and then insert the correct HT object ID into the call number 2 field of the item record (rather than assuming that the ID stored in the record's barcode field is the HT object ID as it does now).
3. Under either approach, there will be special handling needed for second and subsequent duplicates associated with the same original. Assuming that we do not want to display multiple items that link to the same digital object, these items would not be included in the lists for rights determination or activation. We may want to set up a periodic clean-up process to report on these long-term inactive items and, possibly, remove them. We will also need to keep a record of the disposition/relationship of these items, preferably as part of a new supplementary storage database (SSD) that would be available to any of our systems via real-time queries.

Source of original print identification in catalog displays

The current VuFind-based HathiTrust public catalog displays the source of the original print item "original from…" using the HathiTrust namespace identifier portion of the object ID to generate the appropriate display constant. Version 1 of the planned OCLC-based HathiTrust public catalog will display the same data in a very similar manner since the item-level displays will be generated in real time from the SOLR data underlying the VuFind catalog. The two approaches for creation of HT object IDs for "duplicates" would have the following effects:

1. Under the "first CDIH" option, the HathiTrust namespace ID in the object ID can no longer be relied on to infer the source of the original print item. In order to maintain the display of this data, there will be two options:
    a. Put information identifying the source of the original item into the Mirlyn item record in addition to the HT object ID currently stored there. This is not an attractive prospect given the ways that our use of the Mirlyn item record to store HT data already strains the data model.
    b. Store entries associating the HT object IDs with the associated SOPGs in an external supplementary storage database. Display routines would need to be modified to perform real-time queries on this data in order to display the "original from" text.
2. Under the SOPG option, the namespace ID included in the object identifier will continue to reflect the source of the print original on a consistent basis. With the exception of adding non-HT google partners to the namespace/text constant table(s), no change would be needed in order to maintain display of this information in the current manner.

Pageturner displays

The pageturner pulls basic bibliographic data about the item to be displayed from Mirlyn. I don't believe that the HT object ID approach which is chosen will affect the pageturner's use of bibliographic data one way or the other. Note: Display of the appropriate watermark indicating the source of the original print item in pageturner displays is, presumably, based on the namespace identifier in the HT object ID and, thus, IS affected by this decision. That impact is, however, outside the scope of this paper since the data used to generate the watermark is not drawn from Mirlyn item (or bibliographic) records.

Source of original facet/filter in catalog displays

The current VuFind-based HathiTrust catalog provides a facet allowing users to filter by the source of the original print item. Availability of an equivalent feature in version 1 of the planned OCLC-based HathiTrust catalog is considered desirable, but is still under negotiation. The two approaches for creation of HT object IDs for GDD items would have the following effects:

1. Under the "first CDIH" approach, the data needed to provide this facet would not be available in the Aleph records for "duplicate" items. Either the feature would need to be discontinued or data from the external SDD would need to be merged with the data extracted from Aleph prior to SOLR indexing.
2. Under the SOPG option, the namespace ID in the HT object ID would continue to reflect the source of the print original and, thus, would be available for indexing as a facet.

Contributing institution identification in catalog displays

As noted previously, with GDD materials it is no longer possible to assume that the SOPG institution is the same as the CDIH institution, nor that there is only one CDIH institution associated with a given object. This raises a question on whether to include "contributing institution" data in public catalog displays, a decision that is outside the scope of this paper. If, however, we decide to display that information, then the two approaches for creation of HT object IDs would have the following effects:

1. Under the "first CDIH" option, the namespace ID included in the HathiTrust object ID could, unsurprisingly, serve to identify the "first contributing partner". Information about second and subsequent contributing partners, however, would not be stored in the item record. Checking for the existence of additional contributing partners would require a real-time query to the SSD when building the display. This would be in addition to the query required to retrieve information on the SOPG institution from the SSD for these records.
2. Under the SOPG option, information about the HathiTrust contributing partners would not be stored in the item records. A real-time query against the SSD when building the display would be required in order to display such data.

Contributing institution facet/filter in catalog displays

If the event that there is a desire to facet based on contributing partners, the effects are as follows:

1. Under the "first CDIH" approach, the namespace ID identifying the first contributing partner would be available in the item record, but data identifying second and subsequent contributing partners would not. Data from the external SSD would need to be merged with the data extracted from Aleph prior to SOLR indexing in order to make this facet useful.

2. Under the SOPG option, information about contributing partners would not be stored in the Aleph record. Again, data from the external SSD would need to be merged with the data extracted from Aleph prior to SOLR indexing in order to make this facet useful.

Shared bibliographic data

1. Tab-delimited files
   The "source" field in the tab-delimited files contains the HathiTrust namespace identifier code from the HT object ID.
   a. Under the "first CDIH" approach, either:
      i. the "source" field would contain the code for the source institution in non-duplicate cases, but would contain the code for the first contributing partner institution in duplicate cases; or
      ii. Assuming the above would be unacceptable, the programs which create these records would need to be modified to use data from the SSD to populate the "source" field.
   b. Under the SOPG option, the "source" field would continue to consistently represent the source of the print original.

   In neither case would the tab-delimited files contain complete, reliable information identifying HathiTrust contributing partner institutions. To include that data we would need to add new fields to the tab-delimited files and more complex logic to the programs which write those files.

   Data formats and program logic aside, I would have serious reservations about adding contributing partner data to these files. It seems likely that any potential information value this data carries would be outweighed by the potential confusion between source institution and contributing institution which this could introduce for users of this data.

2. OAI files
   Each record in the OAI files of HathiTrust content which are made available for harvesting contains a handle that links to the HathiTrust pageturner. While I do not believe that we have ever asserted or promised that the namespace identifier portion of the object ID in the handle indicates the source of the original print, it has been valid to assume that up to this point.
   a. Under the "first CDIH" approach it would no longer be possible to assume that the namespace code indicated source of print.
   b. Under the SOPG option, the namespace ID would continue to consistently indicate the source of original print.

Conclusion

There are clearly pros and cons to each of the two options for assigning HathiTrust object IDs to GDD items**.** This decision is bound to have significant consequences in areas that are not explored within the scope of this document, but which need to be understood prior to making a choice. When looking at this question from the bibliographic/item metadata perspective of this document, however, I see clear advantages in using the ID associated with the original print item that was actually digitized by Google. I, therefore, recommend the following:

1. Adopt the SOPG approach to assigning HT object IDs. While this will require some one-time modifications to ingest procedures and the item activation routine, the benefits of maintaining

availability of “source institution” identification and consistency/predictability in the contents of the object ID more than justifies that initial effort by reducing complexity in all of the downstream uses of the object ID stored in the Mirlyn item record.

2. Do not add any new data elements to Aleph records. Our current use of the Mirlyn item record already strains the data model in problematic ways.
3. Create an external supplementary storage database (SSD) to record relationships between HathiTrust object IDs and contributing partner item IDs (where they differ). Make this data available via API, both internally and externally. I also strongly recommend that we choose the platform and design for this SSD with a careful eye toward extensibility, allowing easy rational expansion when other elements which do not fit easily into current record structures are identified.
4. Do not include contributing partner data in tab-delimited files. It is likely to be of limited utility in public-facing interfaces and is potentially confusing to users. Where needed, it can be drawn from the SSD via API.