The HathiTrust Corpus: A Digital Library for Musicology Research?

J. Stephen Downie
Graduate School of Library & Information Science

Kirstin Dougan
University Library

Sayan Bhattacharyya
Graduate School of Library & Information Science

Colleen Fallaw
Graduate School of Library & Information Science

University of Illinois at Urbana-Champaign
{jdownie, dougan, sayan, mfall3}@illinois.edu

ABSTRACT
The HathiTrust Digital Library (HTDL) consists of digitized print materials contributed from the collections of some of the foremost research libraries of the world. The HTDL contains over 11 million volumes comprising approximately 3.9 billion pages. In this paper, we describe an exploratory bibliometric study to examine and characterize music-related content in the HTDL. Our study provides an overview of the music-related content in the HTDL as seen through the lenses of format, genre, language, and chronology. We seek to determine in what ways, if any, the materials in HTDL could be considered to form a unique music digital library for use by musicology scholars and students. We also suggest ways in which the music-related content of the HTDL holdings could be made more useful to users with musicological needs and interests.

Categories and Subject Descriptors
H.3.7 [Information. Storage and Retrieval]: Digital Libraries -- *Collections*

General Terms
Measurement, Documentation, Reliability.

Keywords
Music Digital libraries, HathiTrust, Musicology, Informetrics

1. INTRODUCTION
An ideal all-encompassing digital library for musicology would include all of the formats and sources typically used by musicologists and would have features currently unavailable in traditional library tools. It would include text sources; printed music ranging from performance editions to scholarly collected works editions and in formats ranging from full scores, to vocal scores, to individual parts; and audio recordings of performances, among other materials. At a minimum, it would allow text-based searching and musical searching (both by note values and even perhaps by pitch, that is, "query by humming"). While no such library yet exists, there are repositories that attempt to address one or more of these areas of need.

DLfM '14, September 12, 2014, London, UK.
ACM 978-1-4503-3002-2/14/09…$15.00.
http://dx.doi.org/10.1145/2660168.2660173

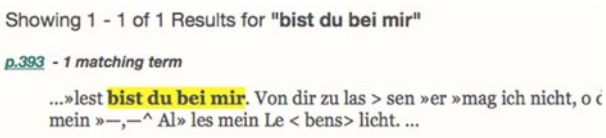

Figure 1: Search-in-page OCR results for Fraktur page, with errors

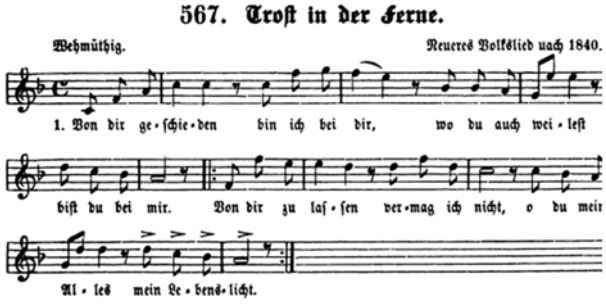

Figure 2: Original image of Fraktur score

The HathiTrust (HT) is a "…partnership of major research institutions and libraries working to ensure that the cultural record is preserved and accessible long into the future. There are more than ninety partners in HathiTrust, and membership is open to institutions worldwide." The HathiTrust Digital Library (HTDL) is the digital preservation repository and access platform of the HathiTrust, providing the digital preservation of, and access to, both public-domain and in-copyright content from a variety of sources, including Google, the Internet Archive, Microsoft, and some of the world's foremost research libraries that are the HT's partner institutions. There are over 11 million volumes in HTDL, comprising more than 3.9 billion individual scanned pages. For each volume there is a Metadata Encoding and Transmission Standard (METS) file that includes bibliographic metadata derived from MARC records provided by the scanning institutions. For each scanned page image there is text generated via an optical character recognition (OCR) process. The HTDL collection of metadata, images and text consists of approximately 500 terabytes of information.

Table 1: Partner library contribution distribution for HTDL resources

Institution	% of collection
University of Michigan	42%
University of California	31%
University of Wisconsin	5%
Cornell University	4%
New York Public Library	3%
Others	15%

Table 2: Partner library contribution distribution for HTDL resources for Library of Congress subclasses M, ML and MT

Institution	% of resources
University of Michigan	65%
University of California	28%
Other	7%

The HathiTrust Digital Library[1] collections are drawn from many of the biggest and best research libraries in the United States. Materials from two institutions, the University of Michigan and the University of California, are predominant in the collection at this point in time as both of these were sites for the original GoogleBooks scanning project on which the HT corpus is based. Over time, each of the partner libraries has been contributing new materials, so that the relative levels of contributions of the different institutions are slowly but clearly equalizing.

Table 1 shows the proportions of the content of the HTDL collection contributed by the top-ranked contributors, as of April 2014, for all content types. Table 2 shows the top-ranked contributors for materials catalogued with the music-related Library of Congress subclasses 'ML' ('Literature on Music'), 'M' ('Music') and 'MT' ('Music Instruction and Study') respectively. Again, the predominance of the University of Michigan and the University of California libraries is worth noting.

As stated earlier, the HTDL contains the digitized text of more than 11 million volumes. This makes the HTDL a new and unique resource for scholarship. However, at this time, relatively little is known about the precise nature of the subject area coverage for the myriad topics of interest contained in its holdings. Some overview-oriented studies have been done to characterize the HTDL collection as a whole by the different types of resources contained within it and by their attributes (such as language, chronology, etc.), such as that by Wilkin [1], but not much work has yet been done to explore specific subject areas within the HTDL collection.

Music, the content area of this survey, presents some special problems and challenges. In a study investigating the music-related content of Google Scholar, Google Books, and the Open Content Alliance (OCA), Dougan determined that while there was more music content in these tools than anticipated, the inherent different-ness of printed music (music scores), in particular, means that music content is not yet a prime candidate for text-based repositories like Google Books, OCA, and HathiTrust [2]. The OCR process that works for recognizing text in scans does not work for music notes (although there are tools that can do this). Duffy points out that the full text search enabled by OCR is of huge benefit to scholars because they can find things that are not captured in traditional bibliographic data [3]. Those studying printed vocal music would therefore greatly benefit from full text search, as lyrics are not included in catalog records. Duffy also describes how the HTDL's use of the bibliographic records of libraries means that it has more accurate metadata than does Google Books (which creates its own records), causing HTDL's collection to be better for searching.

Table 3: Items in the HTDL with music-related Library of Congress subclasses

Library of Congress Class M		Language Specified
Subclass M – Music (score)	23,667	17,848
Subclass ML – Literature on music	39,016	38,932
Subclass MT – Instruction/ study	3,994	3,891
Total	66,678	60,666

The MARC records from contributing institutions are processed and indexed, so that the content can be associated with search and display fields in the Digital Library interface. For example, when performing a catalog search or advanced catalog search, the Subject option includes the content of MARC 6XX subject-related fields, along with other selected fields, such as 043 (Geographic) or 752 (Added Entry-Hierarchical Place Name). The catalog record display in the HathiTrust Digital Library Interface does not currently show all of the information from the MARC record, such as notes fields, but researchers can request custom datasets from HathiTrust that include full metadata records.

In this paper, we provide a brief survey of the music-related material in the HTDL with a special eye toward those materials that may be of interest to students and scholars of musicology. We begin with a description of our data and metadata sources, upon which we based our bibliometric analyses. We provide summary descriptions of the types of material and the source languages of the material in the collection. Next, we analyze the distribution of subjects contained within the M, ML and MT-classed material found in the HTDL. We also describe the chronological distribution of the materials and their genre and format characteristics. We conclude with some comments on how the HTDL could possibly better position itself in the future for use as a digital library for musicology.

The HTDL contains both public domain and in-copyright materials. About 30% is in the public domain and 70% still under copyright. Only material that is in the public domain is available for "Full view." Full view allows scholars the ability to read closely, page by page, the materials as presented via page images. That 70% of the materials are *not* available for viewing is a problem that one must take into account when discussing the utility of the HTDL as a digital library for musicology. In this paper, all measurements are with respect to data that is available for "Full View" unless otherwise noted.

One benefit of the full view is that one can search within the text of a specific page, the results of which are displayed as highlighted words in the OCRed text. This is especially useful when reading items in challenging fonts such as Fraktur, which was commonly used in older German-language publications.

[1] See http://hathitrust.org/about

Table 4: Distribution of resources in the HTDL by language

Language	Percent
English	49%
German	10%
French	7%
Spanish	4%
Russian	4%
Chinese	4%
Japanese	3%
Italian	3%
Arabic	2%
Latin	1%
Other	13%

Table 5: Distribution of music-related resources by language in the HTDL

Language	Percent
English	40%
German	17%
French	8%
Italian	4%
Urdu	4%
Russian	3%
Spanish	3%
Latin	3%
Arabic	2%
Chinese	2%
Other	16%

However, as noted earlier, the OCR data is not corrected, so there are mistakes and extra characters. See Figures 1 and 2.

2. DATA

2.1 Exploration via Metadata Records

6,067,835 MARC records (August, 2013) were obtained serialized as MARC-in-JSON from the HTDL. Each record had undergone post-processing after having been submitted by the contributing institutions. The records were converted first into MARCXML[2] using a Perl module[3] and then into Metadata Object Description Schema (MODS) format via a Library of Congress (LC) XSLT stylesheet enhanced locally to reduce lossiness by decoding more information encoded in MARC fields and combinations of fields. This transformation consolidated information from multiple MARC data and control fields to facilitate characterization at a conceptual level. Finally, a locally developed XSLT stylesheet was used to transform the records into Structured Query Language (SQL) "insert" statements for a custom MODS database schema. 26% of our 6.07 million records represent material in the public domain. However, of those records that have a Library of Congress subclass of M, ML, or MT only a scant 8% represent public domain materials.

Of the 6,067,835 bibliographic records we examined, there were 2,484,032 (43%) with a Library of Congress classification, and 1,010,893 with a Dewey Decimal Number. Of those, 929,576 have both, leaving 81,317 (only 1.3% of the total records) with a Dewey Decimal Number but not a Library of Congress classification. Hence, we used Library of Congress classifications for our exploration. For 3% of the titles with an LC classification, the classification is music-related (subclasses M, ML, or MT). See Table 3 for subtotal counts of the 66,678 music-related records found in our sample.

By way of comparison, the library system of the University of Illinois at Urbana-Champaign holds 113,470 volumes cataloged as music scores (either with Library of Congress M call numbers or Dewey Decimal Classification call numbers), held at multiple different locations: the Music and Performing Arts Library, in the Sousa Archives and Center for American Music, and in the library's high-density storage facility.

2.2 Languages

There currently are music resources in the HTDL in 188 languages, from Adygei to Zuni. Table 3 shows the number of titles found in each Library of Congress subclass, along with the number of titles that also have a language specified in the record.

While 97% of records in the HTDL overall specify at least one language, only 91% of records in the combined M, ML, and MT subclasses specify a language, largely because some musical scores in subclass M are cataloged as having no language content. Tables 4 and 5 summarize more detailed information about language distribution.

It is interesting to note the significantly larger proportion of German-language materials in the music classes, 17% vs. 10% in the entire HTDL corpus. This is likely due to the strong contributions by German musicologists to the discipline as wel as the predominance of German composers in Western art music and the frequency with which they are studied by musicologists.

2.3 Subjects

The subject headings of items in the collection also provide important information that may be helpful in characterizing the contents of the collection. 57,256 of the titles in the HTDL collection contain, in MARC 6xx subject fields, "music" (not necessarily as a standalone word — the search was carried out for just the string (sequence of characters) "music", and hence any word that contains "music" as part of itself, such as "music," "musician," "musicology," etc., would all be included in the topics returned by the search.) While 34,668 of these (62%) are in the LC Class 'M' as well, this also uncovers additional resources. Many of these subject terms also function as genre indications, which are discussed later in this paper. Table 6 shows the most frequently represented topics among items in HTDL that contain the character sequence "music" in their title.

There are 39,025 bibliographic records in the collection that have the Library of Congress subclass ML, 38% of which (14,935 records) have at least one subject heading that contains "music" as a string. From among these 14,935 records, we identified five frequent subject categories, as shown in Table 7. These five categories constituted 1690 records (11% of subclass 'ML').

As the number of records containing more than one of these five categories (47) is quite low, overlap introducing skew into a category-wide chronological characterization of these records

[2] See http://www.loc.gov/standards/marcxml/

[3] See http://search.cpan.org/~gmcharlt/MARC-File-MiJ/lib/MARC/File/MiJ.pm

Table 6: The most frequently represented topics among items in HTDL that contain the character sequence "music" in the MARC 6xx subject fields

Topic	# of titles
Music	24,565
Piano music	5,045
Folk music	2,026
Popular music	2,023
Organ music	1,930
Orchestral music	1,861
Church music	1,388
Songs and music	1,295
Music theory	1,217

Table 7: Five common categories in subclass ML ('Literature on Music'), with the word "music" in at least one of the subject headings

Music--History and criticism	432
Music--Philosophy and aesthetics	406
Music--Periodicals	315
Music--20th century History and criticism	310
Musicians--Correspondence	227

does not appear to be a serious concern. It should be noted that music-related periodicals (which are included in the Library of Congress subclass 'ML') present a special case: in the data informing Table 7, if the periodical is currently ongoing, then we have not included it among the numbers we have reported in the table.

2.4 Chronology

For musical scores, the date of publication can often be very misleading, as a score would be listed by the year of actual publication of that particular edition of the score, rather than the original year of composition of the music. For this reason, we attempted to characterize the collection chronologically only for the Library of Congress subclass ML ('Literature on Music'), and not for the M ('Music') or MT ('Music Instruction and Study') subclasses. Figures 3 and 4 show our findings.

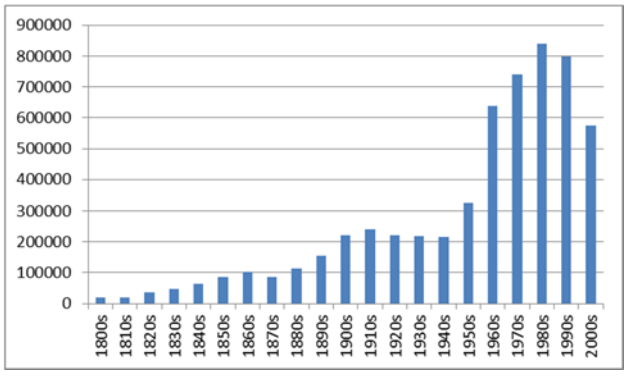

Figure 3: Resources in HTDL by decade

2.5 Genre and Format

We investigated the genre information in the HTDL bibliographical records for material in the collection that belonged to at least one of the pertinent Library of Congress subclasses (M, ML, MT). The total number of records for the set is 66,678.

Table 8 shows some of the more interesting genres and their frequencies. These come from the MODS database described earlier, which gets information about genre from various places in the MARC records.[4] Primarily it is comprised of information encoded in control field 008, informed by leader positions 6 and 7. Other places include data fields 655$a, 047, 366, and 880$6. The categories can overlap, as the same record may belong to multiple genres and therefore have multiple genre subject headings applied.

As in any catalog, searchers cannot rely on headings alone to point them to appropriate material. For example, Table 8 shows that there appear to be only 36 items with the genre heading "tune-books". However, a keyword search of items in the musical score format available in full view finds over 1,900 items that have "tune book" in the text, many with tune book in the title. In many cases the subject heading applied to these items is "hymns."

Selecting "Music" and "Musical Score" together from the HTDL's "Advanced Catalog Search" page (by selecting both these fields in the 'Original Format' box, or selecting just "Music" from the "Original Format" box) returns 82,892 records, of which 9,778 are available for full view (as of June 25, 2014). Both "Music" and "Musical Score" categories appear to contain musical scores.

Scores as a separate physical format, of course, constitute a critical component of musicological study and have their own types of genres indicating what type of musical composition they are. This is often reflected in the subject headings applied to a work in its bibliographic record. Searching the HTDL via the public interface can reveal general characteristics of the scores in the collection. HTDL, as in regular library catalogs, combines all formats of printed music in class M and they are cataloged as "musical scores" or "printed music." This includes everything from full opera scores to items that include individual instrumental or vocal parts for performers. These make a good subset for examination. For example, if one searches for "parts"

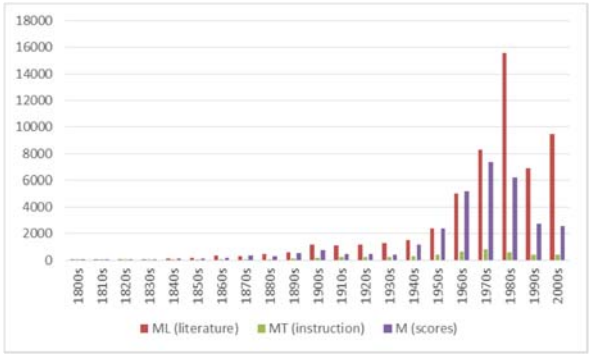

Figure 4: Resources in HTDL for Library of Congress subclasses M, ML, & MT by decade

[4] See http://www.loc.gov/standards/mods/v3/mods2marc-mapping.html#genre

Table 8: Some common genres, with the character sequence "music" in at least one of the subject headings

Genre	Record
Biography	7880
Government publication	1071
Discography	310
Catalog	285
Festschrift	249
Dictionary	125
Tune-books	36
Encyclopedia	17
Handbook	8

Table 9: Top genre subject headings for items with parts

Genre	Records
String quartets	115
Violin and piano music	92
Violin and piano music, Arranged	87
Violin and piano music, Scores and parts	77
Violin and piano music, Arranged Scores and parts	71
String quartets, Parts	67
Sonatas (Violin and piano)	61
Sonatas (Violin and piano) Scores and parts	52

in the subject field and limits the format to "music" or "musical score" only, then 22,648 items are returned, 1,491 of which are available in full view. Table 9 shows the top genres among the items available in full view. This table omits genres that contain 'part' in other contexts, such as in "Part songs, English".

By way of comparison, the physical collection in the Music and Performing Arts Library at the University of Illinois at Urbana-Champaign includes the following among its top headings for items with parts:

- String quartets (1245)
- Violin and piano music (673)
- Sonatas (Violin and piano) (474)
- Piano trios (444)
- Flute and piano music (392)

Figure 5: The HTDL catalog entry for a "sound recording" of Verdi's *Falstaff* from the HTDL collection

Parts in the HTDL are scanned into one PDF, with parts — for example in a string quartet with violin 1, violin 2, viola, and cello — following each other, but not necessarily in the correct score order. As one would need to print the items for use, this is generally not a problem. There are scanning errors with some items, which is a well-documented problem. For example, a set of string quartets by Schumann was found to show only the bound cover of the item [4], and a Dvorak quartet was found to be from a copy that had obviously been repaired and supplemented with replacement photocopies, making the item unsuitable for performance [5]. Riley and Fujinaga describe best practices for image capture of scores, and these standards differ from what is required by standard text-based documents, given that music notation is visual in nature, that accidentals and note connections include fine detail, and that music engraving practices show considerable variation [6]. This makes it unlikely that a collection like the HTDL which contains a wide array of materials can support musical scores as fully as required. However, there are many items of historical significance in the HTDL. (Some examples are: the first edition of the Brahms Sonata op. 120, no. 1, published by Simrock in 1895 [7], and George Root's *Our National War Songs; A Complete Collection of Grand Old War Songs, Battle Songs, National Hymns*, published by S. Brainard's Sons in 1892 [8].) So, the benefits gained from scanning may, in this case, outweigh the problems with scanning.

In addition, there are examples of "complete works" editions of composers available through "Full View" in HTDL, such as:

- *Oeuvres complètes* / Jean-Philippe Rameau; ed. C. Saint-Saëns (Durand), http://catalog.hathitrust.org/Record/000077075
- *Mozart's Neue Ausgabe sämtlicher Werke* (Barenreiter), http://catalog.hathitrust.org/Record/007943559
- *Werke* / Palestrina, ed. Theodor de Witt (Breitkopf & Härtel), http://catalog.hathitrust.org/Record/001834605
- *Werke* / Handel (Breitkopf & Härtel), http://catalog.hathitrust.org/Record/011538541
- *Werken* / Josquin des Prez, ed. A Smijers (Vereeniging voor Nederlandsche Muziekgeschiedenis), http://catalog.hathitrust.org/Record/008322384

The ability to do a full-text search on the OCRed text of these items is valuable, as it can be notoriously difficult to locate individual works within such collected editions, and additional reference tools are usually needed to do so when dealing with the

Figure 6: Digitized version of text corresponding to the catalog entry shown in Figure 5

physical objects. The “smaller works within larger ones” issue frequently encountered in music further complicates this. For example, the famous aria “Ombra mai fu” from Handel’s opera *Serse* could be found either in a full edition of the opera or in an anthology of arias. Keyword searching in full text will reveal this, but a keyword search of bibiliographic records will reveal this only if the full contents have been listed in a notes field, but that is typically not listed for operas. So, had there not been OCRed full text, searchers would have needed to know which opera the aria is from in order to have been able to locate the aria.

As noted by Duffy, there are occasional problems with the metadata in HTDL [3]. Although the HTDL uses original library metadata for anything ingested first to itself, items that came to the HTDL from a library through the Google Books project may have less detailed metadata, as the Google Books project maintains, on average, less metadata than other digital libraries do. For example, in one item imported from Google Books, a string quartet by Malipiero, the subject terms indicate that it is a score with parts but the item scanned is in fact a miniature score [9].

Certain music-related items in the libraries contributing to the HTDL are not in the form of typical natural-language textual media, but are documents representing past performance (sound recordings) or intended future performance (musical scores). However, it is important to note that the HTDL collection does not currently accommodate such non-printed text media as sound recordings. Notwithstanding this shortcoming, the sound recordings represented in the collection do often contain accompanying program notes, synopses and librettos, which have been digitized, and these appear in the HTDL’s catalog as "sound recordings" — even though they are, in reality, digitized text only. An example consisting of Verdi’s *Falstaff* is shown in Figure 5.

Clicking on “Full view” brings up, in a browsable and searchable form, the digitized version of the text that accompanies this “sound recording” (Figure 6).

3. FUTURE DIRECTIONS and RECOMMENDATIONS

Researchers have already recognized the potential of the music holdings within the HTDL. The Single Interface for Music Score Searching and Analysis (SIMSSA) project is working to develop techniques for identifying musical scores within digitized books. The project will locate musical scores within the digitized books in existing collections such as the HTDL, Google Books and the Internet Archive. The project will also create improved Optical Music Recognition (OMR) technology (similar to Optical Character Recognition for text). SIMSSA will then use OMR to further process the music scores that they have identified and located, in order to make them searchable, the ultimate goal being to create a central place to search digitized scores. Having OMR as well as OCR data for printed music in the HTDL would be very valuable to researchers, as this will enable them to search by musical values such as pitch, as well as by words [10].

Another recent initiative of interest in relation to music and the HTDL is the Workset Creation Through Image Analysis of Document Pages (WCTIADP) project, which is one of the prototyping projects of the Workset Creation for Scholarly Analysis, a HathiTrust Research Center initiative funded by the Andrew W. Mellon Foundation. Using the HTDL as its dataset, the WCTIADP project aims to “develop a software application that uses the visual characteristics of digitized printed pages to identify documents that contain three types of visually distinctive materials of interest to humanities researchers: poetry, music, and illustrations” [11].

Hagedorn et al. have experimented with improving HTDL metadata with topic modeling algorithms. They conducted a two-part study using records from the HTDL about art, art history, and architecture. They found that topic modeling can help users discover more relevant materials than if they had relied solely on standard library metadata [12].

A range of possibilities are afforded researchers because of full-text search capabilities in HTDL. They now have the capability to directly search the musical literature for references to works and composers, and even to identify trends in scholarship. For example, an examination of the application of subject headings would show how music literature has changed over time. In addition the HathiTrust Research Center can provide additional support to researchers who want to do more extensive research using the HathiTrust holdings.

Improvements to HTDL would make it of even greater use to music scholars. These include incorporation of audio formats, standardization of how scores (with and without performing parts) are scanned, and enhancement of the metadata. Incorporating audio formats as well as more printed music would allow scholars to use another important source for the study of music, sound. In addition, efforts should be made to improve or “clean up” existing metadata and include new metadata. New display mechanisms (along with FRBR or other metadata schemas) could also be developed to link related items in the HTDL, whether they are scores, recordings, or writings about the music or composer. One new metadata element already being addressed is that of author gender. Researchers with the HathiTrust Research Center have used various data sources such as census data and the Virtual International Authority File (VIAF) to make a preliminary identification of gender for approximately 80% of the authors in the HTDL public domain holdings.[5]

Another new area for potential metadata enhancement is that of date of composition. Many musical works are published in multiple forms and editions after their original composition, but the date of publication is currently the only date normally captured in cataloging. Having the date of composition to search on would be beneficial to researchers and performers alike. Providing the country of origin of authors and composers would be another valuable metadata enhancement that has great potential for further work.

A large body of scanned musical scores also provides opportunities for researchers to perform analysis that has previously been challenging, if not impossible, to do by hand. For example, researchers can now develop programs to analyze digitized scores to illuminate similarities between composers’ musical styles and to reveal how musical themes were reused, developed, and borrowed over large periods of time. Just as a digital library with extensive textual holdings allows for the automated discovery of patterns in text, a digital library with an extensive holding of music in the form of digitized scores could potentially enable similar discoveries, especially with the OMR tools being developed by projects like SIMSSA. For example, composers sometimes use the same melodic line in more than one

[5] See http://www.hathitrust.org/updates_august2013

composition — as Jan Swafford notes, "the melodic line, virtually intact, of the opening piano soliloquy of the Fourth Piano Concerto" of Beethoven also occurs in Beethoven's Fifth symphony [13]. Better OMR will enable the discovery and enumeration of similar melodic phrases that show up not only in different works by the same composer, but also in works by different composers. Once melodic phrases can be so identified, chains of influence and borrowing among composers can potentially be traced. As a large digital library is likely to contain even scores by relatively obscure composers, many new discoveries can potentially be made in this way once OMR technology advances sufficiently.

A final recommended improvement is to harness lyrics captured in the OCR process. At this time many (but not all) items undergo the OCR process, whether they are text-based items or not — which means that printed music with text (lyrics) can in some cases be searched by words.[6] In addition to the image files, researchers can view the OCR text files if they exist. However, it would be even more beneficial if the lyrics were displayed in the record for the score and if a field for lyrics (which does not currently exist) could be a search limiter, so that researchers could search only on lyrics if they chose to do so.

4. CONCLUDING DISCUSSION

The nearly ten-fold disparity in counts between the 82,892 items mentioned that are contained in the HTDL and the 9,778 that can be seen in "Full View" is a matter of crucial concern, not only for musicology, but also for other scholarly domains: much of the material still remains in copyright (or undetermined copyright, which amounts to the same thing in practice). For all content still in copyright, the content can neither be displayed for scholars nor distributed to them. While it is possible to search for copyrighted materials (and this can be helpful in many cases), the lack of access to many of the actual scores or texts is a critical shortcoming that limits the HTDL's utility as a general musicological resource. Work is being done to provide computational non-consumptive ("distant reading") access to materials in the collection that cannot be seen in "Full View" [14]. Easy ways to move back and forth between distant reading and close reading of a score (or text) will also be needed. This will begin to be addressed in the forthcoming "HathiTrust + Bookworm" ("HT+BW") initiative [15]

The presence of music-related content in the HTDL corpus presents an ongoing challenge that has begun to be addressed only recently. In providing an overview of the HTDL in the context of Google Books and the Open Content Alliance (all three of which have content overlap), Christenson states that the mission of academic libraries is to build, preserve, and provide access to collections that meet the research needs of their users now and in the future [16]. She explains that although the HTDL began with text-based materials, the eventual goal is to expand to other media types (p. 95). The lack of audio functionality in the HTDL is, therefore, an impediment to its fulfilling its role as a comprehensive music digital library for musicology at the present time. Audio is evidently a critical component to the study of music, but the content of the HTDL corpus currently consists only of digitized printed text.

While preservation and access are difficult enough to provide for text and image collections, they are inherently more difficult for media collections (whether they are digitized or not), given the increasing instability of (magnetic) carriers and eventual obsolescence of playback hardware. Beers and Parker describe the challenges faced during a 2009 pilot project at the University of Michigan to digitize unique audio items of high research value for inclusion in the HTDL [17]. They found that there were few audio digitization standards and that there was also a lack of digital audio metadata standards (p. 40), and they reported that "creating ingest and validations for this workflow also presented difficulties, as the current routines were built completely around images…and could not be easily adapted for audio" (p 41). They concluded that, while over time the HTDL may evolve to incorporate audio, "large scale digital audio preservation requires different resources than large scale digital image preservation" (p. 44).

To conclude, the HTDL is a new and unique resource for scholarship. The inclusion of audio, OMR, and metadata will all make the HTDL more useful to musicologists. Non-consumptive access techniques, making the content available for aggregate-level statistical and algorithmic analysis but not human-scale comsumption, will make material that is under copyright be useful for research purposes. Expanding the range of contributors beyond the current set of contributing institutions is likely to lead to more varied content. Thus, overall, as a source for some music materials, the Hathi Trust Digital Library still falls short of being a comprehensive digital library for musicology, but it does provide new opportunities.

5. ACKNOWLEDGMENTS

The authors gratefully acknowledge the assistance of Thomas G. Habing, who developed the schema of our SQL metadata database and the associated XSLT style sheets used for data ingest.

This research was supported in part by the HathiTrust Research Center and a grant awarded by the Andrew W. Mellon Foundation; however, any opinions, findings, and conclusions or recommendations expressed here are those of the author(s) and do not necessarily reflect the views of our sponsors.

6. REFERENCES

[1] Wilkin, J. 2011. HathiTrust and Print Storage: Building around a digital core. Presented at: *Committee on Institutional Cooperation (CIC) — Center for Library Initiatives (CLI) Conference*, May 2011. Retrieved from http://www.hathitrust.org/documents/HathiTrust-CIC-201105.ppt on July 1, 2014.

[2] Dougan, Kirstin. 2010. Music to our Eyes: Google Books, Google Scholar, and the Open Content Alliance. *portal: Libraries & The Academy* 10, 1: 75-93.

[3] Duffy, Eamon P. 2013. Searching HathiTrust: Old Concepts in a New Context. *Partnership: The Canadian Journal of Library & Information Practice & Research* 81: 1-13.

[4] Schumann, Robert. *Drei Quartette für 2 Violinen, Viola, Violoncell : Op. 41 / von Robert Schumann* ; Revised by

[6] For example, see http://catalog.hathitrust.org/Record/100002074, which was returned with a search for Joplin Rag, text that appears on the back cover of this sheet music by another composer.

von Friedr. Hermann. Retrieved from http://catalog.hathitrust.org/Record/100143217 on July 1, 2014.

[5] Dvořák, Antonín. *Quartett, op. 34.* Retrieved from *http://catalog.hathitrust.org/Record/100025795* on July 1, 2014.

[6] Riley, Jenn, and Ichiro Fujinaga. 2003. Recommended best practices for digital image capture of musical scores. *OCLC Systems & Services* 19, 2: 62-69.

[7] Brahms, Johannes. Sonata, op. 120, no. 1, F moll = Fa mineur = F minor : Clarinetta & Piano : Viola & Piano. London : N. Simrock, 1895. Retrieved from http://catalog.hathitrust.org/Record/000904321 on July 1, 2014.

[8] Root, George. 1892. *Our national war songs; a complete collection of grand old war songs, battle songs, national hymns, memorial hymns, Decoration Day songs, quartettes, etc., with accompaniment for piano or organ.* Chicago : S. Brainard's Sons. Retrieved from http://catalog.hathitrust.org/Record/008722111 on July 1, 2014.

[9] Malipiero, Gian Francesco. Rispetti e strambotti per quartetto d'archi. London: J. & W. Chester. Retrieved from http://catalog.hathitrust.org/Record/001091611 on July 1, 2014.

[10] Motuz, Catherine. 2013. CIRMMT Workshop, September 7th, 2013, Part I : Introduction, *SIMSSA Blog*, September 12, 2013. Retrieved from http://simssa.ca/node/79 on August 7, 2014.

[11] Biggers, Keith. 2014. Workset Creation through Image Analysis of Document Pages, *Workset Creation for Scholarly Analysis*, Retrieved from http://worksets.htrc.illinois.edu/worksets/?page_id=108 on August 7, 2014.

[12] Hagedorn, Kat, Michael Kargela, Youn Noh, and David Newman. 2011. A New Way to Find: Testing the Use of Clustering Topics in Digital Libraries. *D-Lib Magazine* 17, No. 9/10: 1-10.

[13] Swafford, Jan. 2014. *Beethoven: Anguish and Triumph: A Biography.* New York: Houghton Mifflin Harcourt.

[14] Zeng, Jiaan, Guangchen Ruan, Alexander Crowell, Atul Prakash and Beth Plale. 2014. Cloud Computing Data Capsules for Non-Consumptive Use of Texts, *5th Workshop on Scientific Cloud Computing (ScienceCloud)*, Vancouver, Canada, June, 2014.

[15] HathiTrust Digital Library. 2014. HathiTrust Research Center Awarded Grant from National Endowment for the Humanities. July 29, 2014. Retrieved from http://www.hathitrust.org/neh_implementation_grant_award_2014 on August 9, 2014.

[16] Christenson, Heather. 2011. HathiTrust: A Research Library at Web Scale. *Library Resources & Technical Services* 55, no. 2: 93-102. p. 93.

[17] Beers, Shane and Bria Parker. 2011. HathiTrust and the Challenge of Digital Audio. *IASA Journal* (36): 38-46.