The HathiTrust Digital Library's Potential for Musicology Research

Article begins on next page

International Journal on Digital Libraries manuscript No.
(will be inserted by the editor)

The HathiTrust Digital Library's Potential for Musicology Research

J. Stephen Downie · Sayan Bhattacharyya · Francesca Giannetti · Eleanor Dickson · Peter Organisciak

Received: date / Accepted: date

Abstract The HathiTrust Digital Library (HTDL) is one of the largest digital libraries in the world, containing over fourteen million volumes from the collections of major academic and research libraries. In this paper, we discuss the HTDL's potential for musicology research by providing a bibliometric analysis of the collection as a whole, and of the music materials in particular. A series of case studies illustrates the kinds of musicological research that may be conducted using the HTDL. We highlight several opportunities for improvement, and discuss promising future directions for new knowledge creation through the processing and analysis of large amounts of retrospective data. The HTDL presents significant new opportunities to the study of music that will continue to expand as data, metadata and collection enhancements are introduced.

Keywords musicology · distant reading · digital humanities · digital libraries

J. Stephen Downie
University of Illinois at Urbana-Champaign
E-mail: jdownie@illinois.edu

Sayan Bhattacharyya
Price Lab for Digital Humanities
University of Pennsylvania
E-mail: sayan@illinois.edu

Francesca Giannetti
Rutgers, The State University of New Jersey
E-mail: francesca.giannetti@rutgers.edu

Eleanor Dickson
University of Illinois at Urbana-Champaign
E-mail: dicksone@illinois.edu

Peter Organisciak
University of Illinois at Urbana-Champaign
E-mail: organis2@illinois.edu

1 Introduction

A comprehensive digital library for musicology would ideally include all of the formats and sources typically used by musicologists: text sources; printed music, including performance editions and scholarly collected works; scores, ranging from full scores and vocal scores to individual parts; and audio recordings of performances. It should allow text-based searching as well as searching by note values, and, ideally, it should even support searching by pitch, that is, "query by humming." While digitization of cultural heritage materials, including music, has led to a proliferation of library, archives, and museum content available online, no such ideal digital library for musicology yet exists. Current digital collections for musicology are often built around an individual composer,[1] genre, repertory, or format,[2] or on the basis of a specific characteristic, such as the location where the materials were published.[3] In this paper, we discuss the potential for musicology research of a large-scale, general-purpose digital library, the HathiTrust Digital Library (HTDL).

At over 14 million volumes, the HTDL is one of the largest digital libraries in the world (York, 2012). We describe the history, coverage, and features of the HTDL collection. We then explore the HTDL as a re-

[1] E.g., the Bach digital database portal, designed to provide Bach researchers with "solid information on the works of Johann Sebastian Bach and other composers from the Bach family and their whereabouts," http://www.bach-digital.de/content/infos.xml.

[2] E.g., Alexander Street's Classical Scores Library, http://alexanderstreet.com/products/classical-scores-library-package.

[3] E.g., a virtual library of nineteenth century California sheet music (from between the years 1852 and 1900), http://people.ischool.berkeley.edu/~mkduggan/neh.html.

source for musicologists through a bibliometric analysis of the music-related material in the HTDL and its characteristics, including its languages, genres and formats, and its chronological characteristics,. We illustrate several case studies showing how the HTDL could be used to study music and music history, and we present challenges and opportunities for domain-specific research using an unspecialized repository such as the HathiTrust's. We conclude with recommendations regarding how to develop the HTDL into an even more useful resource for musicology research.

2 The HathiTrust Digital Library

2.1 Overview

The HathiTrust is a "partnership of major research institutions and libraries working to ensure that the cultural record is preserved and accessible long into the future."[4] The HathiTrust Digital Library consists of digitized volumes contributed by partner libraries, as opposed to being built around a particular physical library's collection. The HathiTrust has roots in the Google Books initiative, begun in 2004, which aimed to digitize all the books in the world. In 2008, the universities in the Big Ten Academic Alliance (formerly the Committee on Institutional Collaboration) and the University of California system, all of them partners in Google Books, formed the HathiTrust as a not-for-profit consortium. Since then, membership in the HathiTrust has grown to over one hundred and ten partner institutions that continue to deposit digitized materials into the repository.

As of August 2016, there are 14.68 million volumes in the HTDL, comprising more than 4.9 billion individual scanned pages (Table 1). For each scanned page image in the HTDL, there is an associated plain-text file generated from it via an Optical Character Recognition (OCR) process, as well as, in some cases, page coordinate OCR information. Each volume in the collection is described by a Metadata Encoding and Transmission Standard (METS) file that includes bibliographic metadata derived from the MAchine Readable Cataloging (MARC) record for the volume that is provided by the contributing partner institution. The HTDL collection of metadata, images, and text consists of approximately 658 terabytes of information. The content of the collection encompasses in-copyright material as well as material in the public domain.

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

2.2 HTDL Functionality

The HTDL interface allows users to search both the bibliographic metadata and the full text of volumes in the repository in order to discover items. In the case of volumes that are restricted from the user's view due to copyright or privacy concerns, the user can see a list of pages on which their search term was found, but the page images and full-text OCR are not displayed. In the case of volumes that are not restricted and are open to the user's "full view," both the scanned page images and the OCRed text are available for reading. This functionality allows users to contextualize their search results and enables researchers to quickly discover how a particular search term is being used in a work. It can assist a user in disambiguating between variant meanings of a word or phrase and in determining if the volume is actually of interest to the user. Additionally, the possibility of search within the OCRed full text in the HTDL redresses, to some extent, the notorious difficulty of locating individual works within collected editions — a task that ordinarily requires the use of additional reference tools, such as bibliographies and thematic catalogs.

The user is able to view, for restricted as well as unrestricted volumes, an abbreviated MARC catalog record for the item. Although the catalog record display in the HTDL interface does not show all the fields of the MARC record (notes fields, for example, are not shown), researchers can request custom datasets from the HathiTrust that include full metadata records, including the entire MARC record. Additionally, users can formulate and facet their search queries in the HTDL interface on information derived from the MARC bibliographic metadata. As the MARC records from contributing institutions are processed and indexed by the HathiTrust, their content is associated with the search and display fields in the HTDL interface. For example, the metadata for the 'Subject' search facet in the HTDL interface is drawn from the contents of MARC 6XX subject fields as well as other selected fields such as the 043 ('Geographic') and 752 ('Hierarchical Place Name') fields.

2.3 Copyright

Because the HTDL contains digitized material spanning over 500 years of publication up to the present day, a portion of the library's contents are restricted from public view by copyright law. The rules that surround legal dissemination of digitized library material complicate user access, especially for the international audience. Nevertheless, the ability to perform full text

Table 1 Size of the HathiTrust Digital Library as of June 2016

Total Volumes	14,685,004
Titles	7,742,501
Pages	5,139,715,400
Disk Size	658 TB

searches, which are otherwise impossible in standard library catalogs, on the entire HathiTrust Digital Library allows users to discover previously unfindable material, and also to track the location of volumes and to read the physical copies. The material in the HTDL that is open and in the public domain can be distinguished into two groups: material that is in public domain worldwide, and material that has only been confirmed to be in public domain in the United States. In order to abide by various complex international copyright laws, the HTDL uses 1873 as the cut-off date to determine what can be shown in "Full view" to those accessing the library's Web interface from outside the United States. The ratio of public domain material to restricted material is 24:76 for users outside the United States and 38:62 for users inside. [5]

For the purposes of this paper, we have grouped the items in the HTDL into the following three broad sets by rights status.

- *World Public Domain*: Each volume in this set is an item that is either internationally open or published with a Creative Commons zero license.[6]
- *U.S. Public Domain*: Each volume in this set is an item that is either in the World Public Domain or has been determined to be in the public domain in the United States although not in the World Public Domain; or is published with a Creative Commons zero license.
- Restricted: Each volume in this set is an item restricted from full view (usually for reasons of copyright).

It can be seen from the above definitions that the set *U.S. Public Domain* is a superset of the set *World Public Domain*, and that the sets *Restricted* and *U.S. Public Domain* are disjoint sets. Figure 1 conceptually illustrates these relationships.

3 HTDL Coverage

The large quantity of digitized text in the HTDL makes it unique in its scale and coverage as a resource for research. Given such a scale, investigating the library's coverage can help reveal information about the contours of the collection. While overview-oriented studies of the HTDL collection, such as that by Wilkin (2011), can characterize it as a whole by the different kinds of resources contained within it and by their attributes (such as language, chronology, etc.), little work has been done so far to explore specific subject areas of the collection.

The metadata we explore in this paper is a snapshot of bibliographic metadata obtained from the HathiTrust in April 2016. It was obtained serialized as MARC-in-JSON format that had undergone post-processing after having been received from contributing institutions. It was then converted by the HathiTrust Research Center into Metadata Object Description Schema (MODS) format.[7] This metadata snapshot describes 14.6 million items. It is an earlier, and therefore smaller, set of metadata than is reflected by the more up-to-date numbers seen in Table 1. These volumes can be represented by 7.8 million bibliographic records, as there is a one-to-many relationship between MODS records and item records. Some tables in this paper describe items (labeled *volumes*) while others describe MODS bibliographic records (labeled *records*). Because there is considerable overlap in volumes classified using the Library of Congress Classification system and the Dewey Decimal system, we have chosen to focus on Library of Congress classifications for our exploration.[8]

[5] See "What are the different copyright statuses of items in HathiTrust, and what do they mean?" https://www.hathitrust.org/help_copyright#RightsCodes.

[6] See http://creativecommons.org/publicdomain/zero/1.0/.

[7] We first converted the records we obtained from the HathiTrust into MARCXML (see: http://www.loc.gov/standards/marcxml/) using a Perl module (see: http://search.cpan.org/~gmcharlt/MARC-File-MiJ/lib/MARC/File/MiJ.pm). Then we used a Library of Congress XSLT stylesheet to convert the records to the Metadata Object Description Schema (MODS) format. The stylesheet was enhanced locally by consolidating information encoded in multiple MARC data and control fields, to reduce data loss and retain more detail about the conceptual characterizations of the items. Finally, a locally developed XSLT style sheet was used to transform the records into Structured Query Language (SQL) *insert* statements for populating the customized MODS database tables.

[8] Of the 14.6 million item records that we examined, 6,672,311 (46%) had Library of Congress Classification numbers and 2,461,361 (16.8%) had Dewey Decimal Classification numbers. 2,165,140 had both, leaving only 296,221 (2% of the total items) that had a Dewey Decimal number but not a Library of Congress classification number. Of the volumes with a recorded classification authority (approximately 50% of the total records), only a very small number had any authority

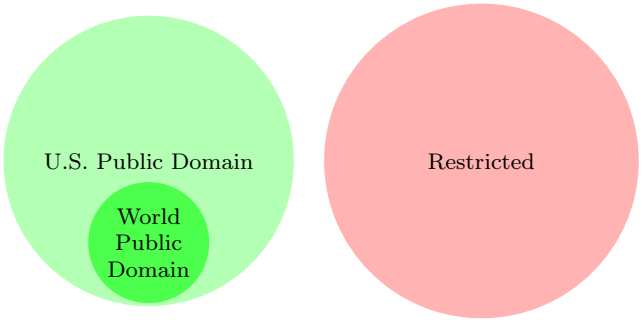

Fig. 1 Venn diagram illustrating the set relationship between the three rights statuses (not to scale)

3.1 Contributing Institutions

The HathiTrust Digital Library collection is drawn from partner libraries that include many of the largest and most prominent research libraries in the United States. Two of the original partners, the University of Michigan and the University of California, are still the best-represented among the HTDL's contributing partner institutions. Over time, various other partner libraries have also been contributing new material, so that the relative proportions of the contributions made by different HTDL partners have gradually changed.

Figure 2 shows the proportions of the content in the HTDL received from each of the top five contributors.

3.2 Language

The HTDL contains volumes in no less than 455 languages, from Adygei to Zuni. The proportions of the most prominent languages found in the HTDL are shown in Figure 3. English, not surprisingly, is the predominant language, reflecting the HTDL's existence as a primarily American project so far.

3.3 Subjects

The topical subject headings applied to volumes in the HTDL help reveal the nature of the collection. As {fig:subcounts} illustrates, volumes cataloged under such topics as *history*, *periodicals*, and *politics and government* predominate.

other than Library of Congress or Dewey Decimal. Classification authorities were not ascertainable for the remaining volumes, primarily because the HTDL does not retain local call numbers that can help in determining classification information. Although not retained, the local call number for a volume can, however, be retrieved if needed — via the holding record for the volume, using the contributing library's local system number, which *is* stored by the HTDL.

3.4 Chronology

As we mentioned earlier, the MARC bibliographic records undergo post-processing after they are received from contributing institutions. The addition of a non-standard MARC field 974, which includes normalized date information pulled from other places in the MARC record, such as 260$c (date of publication) and 863-865 (enumeration chronology), constitutes a step in this post-processing. Given the issues of cataloger uncertainty, re-publication of works, and differences in local cataloging practice, all of which contributes to the occurrence of ambiguous dates in bibliographic records,[9] the HathiTrust's best estimation of date of publication for each item is the publication date in the 974 field of the HathiTrust MARC bibliographic record, which is translated directly into the MODS database that was used in this analysis. The date information in field 974 subfield y is also incomplete; nearly 8% of records in the HTDL do not have a date recorded in that subfield. Nevertheless, (5) shows the general chronological spread of the HTDL collection by analyzing these normalized publication dates by century. The concentration of the collection in the twentieth century is notable from the figure.[10]

[9] An important source of inconsistency is the significant variation that is found in the format of the date field across records, originating from variations in the local standards used by contributing institutions for their own bibliographic records. For example, some bibliographic records use wildcard characters in the date field, which are not consistent with each other as, sometimes, different wildcard characters have been used. Ranges of years often appear, and their formats, too, are frequently different. (For example, "1904-1924," "between 1920-1950," etc.). Other sources of variation include the fact that some records use the character 'u', for 'unknown', in place of a digit — as in '18uu' to denote an year which is not precisely known but is from the nineteenth century — other records may use '–' (for example, '198–').

[10] See https://www.hathitrust.org/visualizations_dates for up-to-date chonological information.

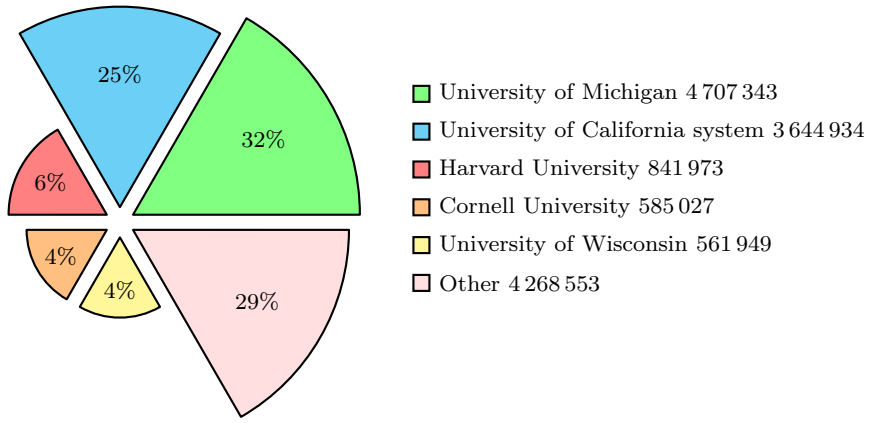

Fig. 2 Institutional representation within the HTDL collection, all works

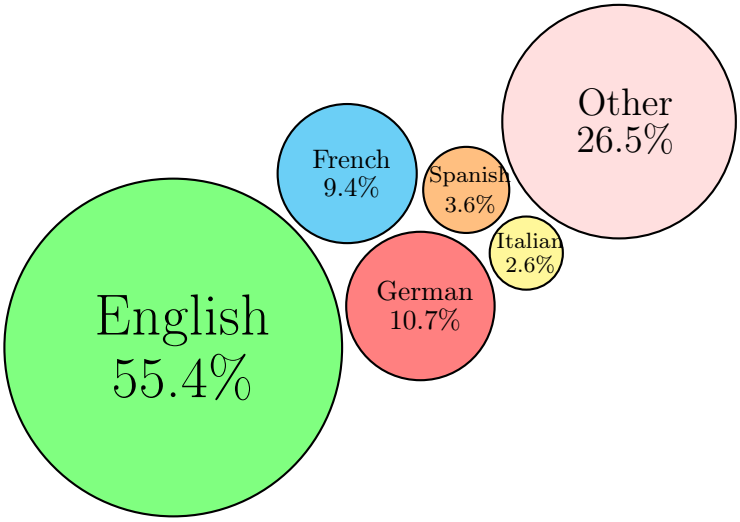

Fig. 3 Preponderance by language within the HTDL collection, all works

3.5 Genre and Form

Genre and form information is held in various places within a MARC record. The MARC-to-MODS conversion process that was applied to our snapshot of HTDL metadata pulls together genre information from several fields of MARC records, including that which was encoded in MARC control field 008, informed by leader positions 06 (type of record) and 07 (bibliographic level), as well as fields 655$a (genre term subject heading), 047 (form of musical composition), 336 (physical description content type) and 880$6 (alternate graphic representation linkage).[11]

The *format* facet available on the HTDL interface allows users to limit their search by material type. This information about form, which is distinct from genre, is derived primarily from the 00x control fields in the MARC metadata, and only a small, controlled set of terms can be used. *Electronic resource*, *remote*, and *print* are some of the more commonly occurring and self-evident terms describing the material digitized and deposited in the HTDL. As we discuss later, form terms can help in the discovery of some specialized material, including music.

Table 2 shows the most frequently applied genre terms. (It is to be noted that multiple genre terms may apply to a single bibliographic record.)

4 Coverage of Music Material in the HTDL

For items in the HTDL classified by a Library of Congress classification number, a total of 114,167 (0.8%) belong to music-related class M (Music) and its subclasses, M (Music), ML (Literature on music), or MT (Instruction and study). By way of comparison with a non-digital library, as of 2014 the library of the University of Illinois at Urbana-Champaign, one of the largest collections in

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

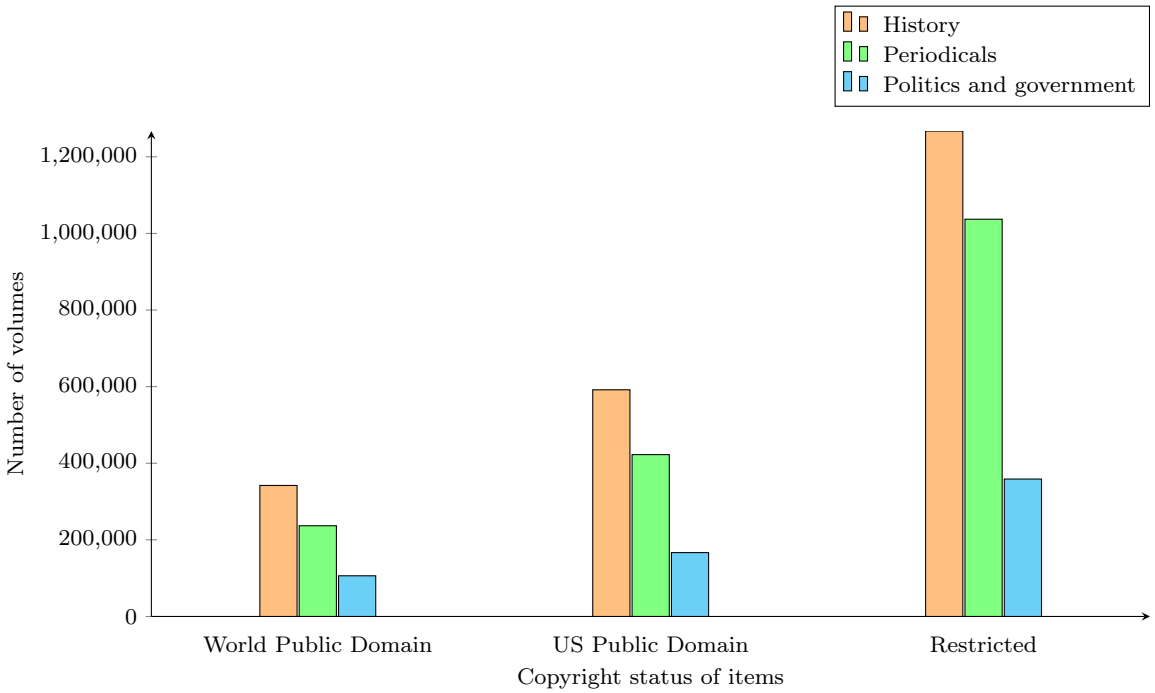

Fig. 4 Top three most frequently occurring subject headings in the HTDL—'History', 'Periodicals', 'Politics and government'—and the number of bibliographic records to which they are applied

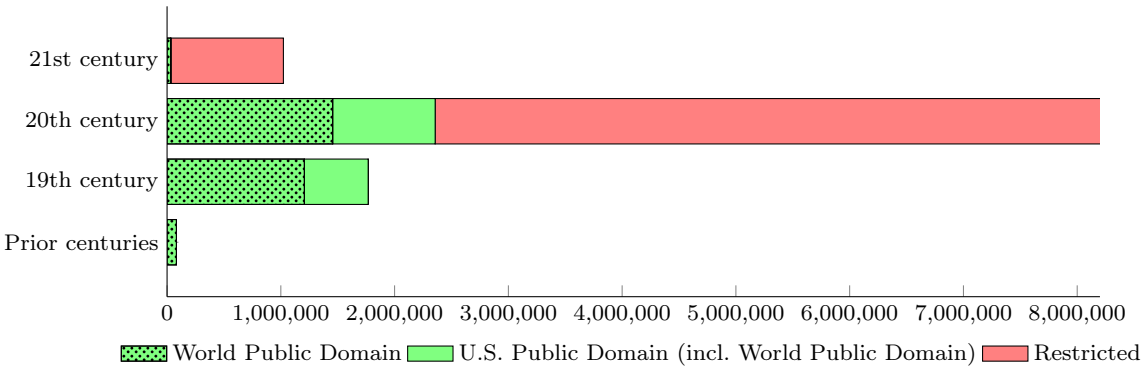

Fig. 5 Publication dates of volumes in the HTDL, by century and rights status

Table 2 Top five most frequently occurring genre terms in the HTDL

Public Domain Worldwide	No. records	U.S. Public Domain	No. records	Restricted	No. records
not fiction	852512	not fiction	1204758	not fiction	2902227
periodical	201761	periodical	360287	bibliography	1640081
bibliography	154250	bibliography	214219	periodical	834296
government publication	143626	government publication	148834	biography	249473
biography	68134	biography	93103	government publication	185366

the United States, held 113,470 volumes (1.0% of all items in the University's library) that were cataloged as music scores by either Library of Congress Classification M call numbers or Dewey Decimal Classification 78x call numbers.

4.1 Contributing Institutions

Figure 6 shows the top-ranked contributors of material in the HTDL belonging to Library of Congress Class M. The libraries of the University of Michigan and the University of California system predominate as contributors.

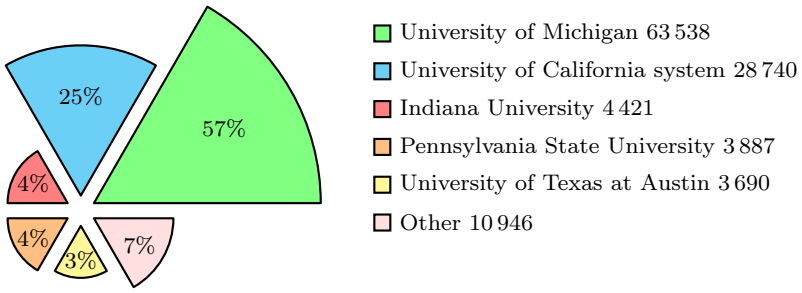

Fig. 6 Institutional representation within the HTDL collection, music-catalogued materials

4.2 Languages

Table 3 shows the number of titles in each Library of Congress subclass, along with the number of titles that also have a language specified in the record. As with the entire HTDL, English is by far the most prevalent language among material in the library belonging to the Library of Congress 'Music' Class. Comparing the music-classified material in HTDL to the overall HTDL collection shows that certain languages have a greater footprint among music-classified material than they have in the overall HTDL collection.[12] German-language material makes up a greater-than-expected proportion of music-class content, as seen in Table 4. This difference may be due to the preponderance of German composers in Western music and to the considerable quantity of research that musicologists have consequently dedicated to them, as well as to the strong tradition of German-language publication in the field of music.[13] Similarly, Italian-language materials also make up a slightly higher-than-expected proportion of music-class content, presumably due to the relatively large Italian influence on western music.

4.3 Subjects

Analyzing the subject headings in the HTDL allows us to further characterize the music-related material it contains. For example, the string 'music' appears in the MARC 655 subject fields of 106,418 of the volumes in the HTDL either as a standalone word or as part of another term such as 'musician' or 'musicology.' Table 5 and Table 6 show the most frequently occurring subject headings where the string "music" appears in either the subject or the title. Given that only 43% (45,387) of these volumes are in the Library of Congress Classification M class, subject headings are likely to improve discoverability of additional music-related resources. As we will discuss later in this paper, many of these subject terms indicate the volume's genre.

The subject headings for the 45,387 bibliographic records containing the string "music" within the title also help reveal information that is potentially useful for musicologists. Table 6 shows the four most frequent subject categories (by number of volumes) from among these records (which have "music" in the title).[14]

4.4 Chronology

As with our analysis of the general HTDL collection, we have pulled normalized date information from the HathiTrust-specific MARC field 974 subfield y to avoid ambiguous, incomplete, or otherwise difficult to analyze date information located elsewhere in the bibliographic record. Date information for 7.6% of the items cataloged as music is unavailable through this field. Still, as was the case with the general HTDL collection, looking at publication dates in aggregate helps contextualize the music-related material in the HTDL. Figure 9 shows that the vast majority of items cataloged as music in the HTDL were published in the 20th century. As a result, the majority of the items cataloged under Library of Congress class M are restricted.

4.5 Genre

We investigated the genre information in the 66,678 bibliographic records for items in the HTDL belonging to Library of Congress Classification class M. Some of

[12] The footprints of languages in the overall HTDL collection are shown in Figure 3.

[13] Consistent with the relative decline of German as a language of international scholarship starting in the 1920s, the proportion of German-language material in the music class in the HTDL collection declines from about 61% for texts that are considered to be in public domain for researchers in the United States (which are mostly pre-1923 publications) to about 30% for texts that are in-copyright (which are mostly post-1923).

[14] Each distinct Library of Congress Subject Heading in the HathiTrust metadata explored in this paper is counted separately, so that "Piano Music," "Vocal Music," etc. are categories distinct from the category "Music."

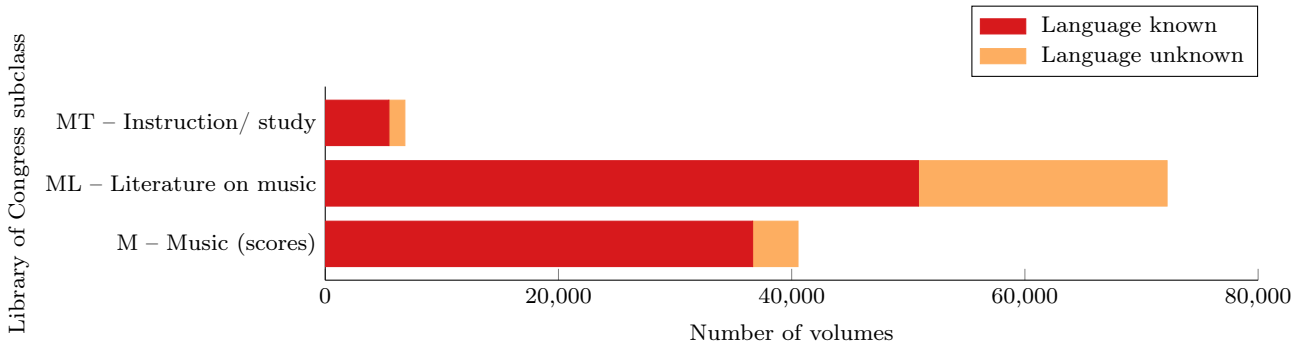

Fig. 7 Number of volumes in Library of Congress classification class 'M' (music)

Table 3 Counts of volumes in Library of Congress Classification class 'M' (music), and of those among them which have language specified

Library of Congress Subclass	Volumes	Language known
M – Music (score)	40,542	36,650
ML – Literature on music	72,192	50,867
MT – Instruction/ study	6,828	5,461
Total	119,572	92,978

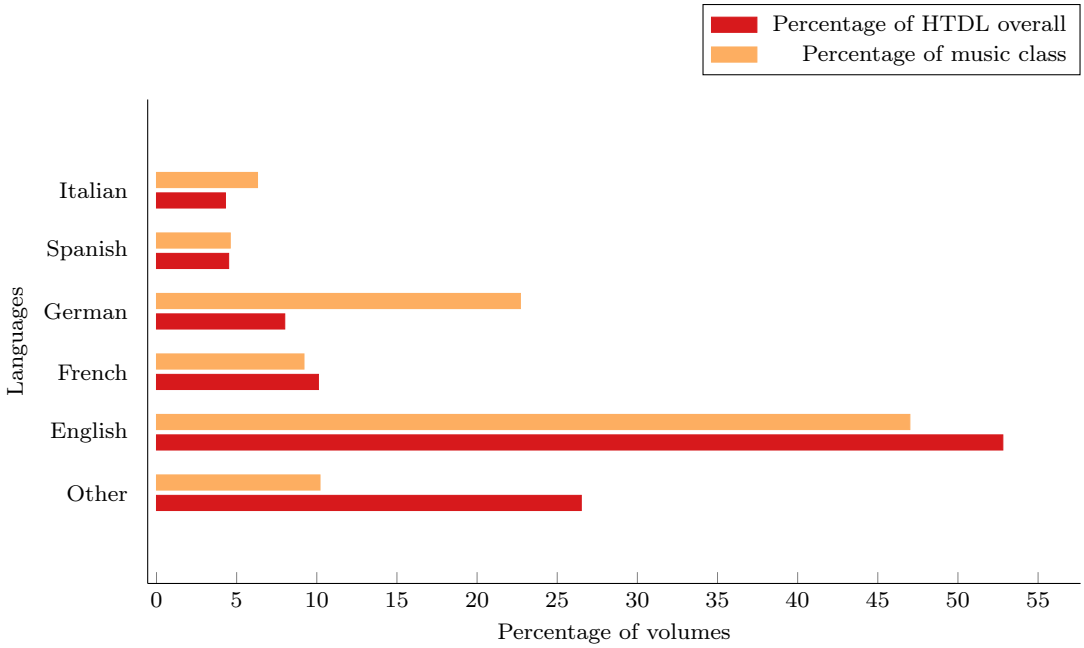

Fig. 8 Proportions of (in red) all HTDL material and (in orange) of HTDL material specifically classified as music that are in some of the most highly represented languages in the HTDL collection, as compared to other languages

Table 4 Preponderance by language within that part of the HTDL collection that is specifically classified as music

Language	Percent classified as music
English	47.0%
French	9.2%
German	22.7%
Spanish	4.6%
Italian	6.3%
Others	10.2%

Table 5 Four of the most frequently-applied Library of Congress topical subject terms containing "music" and the number of items for each topic

World Public Domain	No. of volumes	U.S. Public Domain	No. of volumes	Restricted	No. of volumes
Music	7992	Music	16269	Music	50252
Piano music	1281	Vocal music	2934	Piano music	6628
Vocal music	879	Piano music	2739	Vocal music	5572
Musicians	683	Musicians	1258	Instrumental music	5401

Table 6 Top five most frequently occurring Library of Congress 655 subject headings in volumes where the string "music" appears in the title

World Public Domain	No. of volumes	U.S. Public Domain	No. of volumes	Restricted	No. of volumes
Music	5616	Music	8528	Music	27823
Bibliography	3354	Bibliography	3431	History and Criticism	9352
History and criticism	1014	History and criticism	1680	Periodicals	5972
Instruction and study	795	Periodicals	948	Instruction and study	3743
Periodicals	544	Instruction and study	947	Bibliography	2791

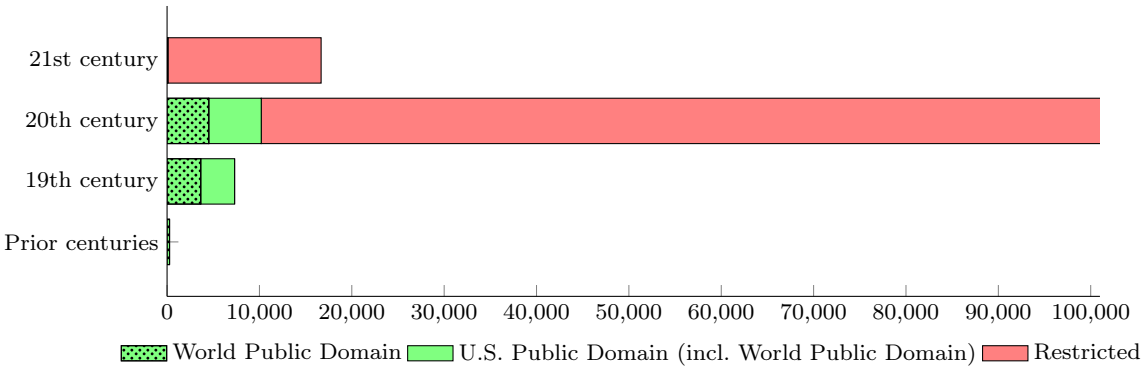

Fig. 9 Publication dates of volumes classified as music in the HTDL, by century and rights status

the more interesting and commonly-applied genre terms are displayed in Table 7.

As in any catalog, bibliographic metadata, by itself, is often insufficient for discovering materials of interest. For example, Table 7 shows that there are only 43 bibliographic records with the genre term "tune-books."[15] However, a keyword search of items in the musical score format available in the HathiTrust Digital Library user interface returns over 2,000 items that have the string "tune book" in the full text, with many of them having "tune book" in the title itself. Finding fewer bibliographic records than there are items in the digital library is not unexpected, as there is a one-to-many relationship between bibliographic records and digitized items. Nevertheless, that our searches should uncover such a large gap is an unmistakable symptom of the dearth of genre-specific metadata in the HTDL, and consequently indicates the potential advantage to be gained in some situations by preferring full-text search over the collection.

4.6 Format

Musical scores have a distinct physical format and constitute a critical object of study in musicology. According to MARC standards, "Music" and "Printed music" are among the permitted terms in MARC leader position 06 and MARC control field 008 of a bibliographic record. These designations for music-related material benefit researchers seeking to find musical scores. The HTDL "Advanced Catalog Search" allows users to limit their search by formats "Music" and "Musical score." Such a search in the HTDL yields 103,744 results, of which 21,405 are available for full view as of April 2016.[17]

[15] Tune books constituted a genre of early American music publication that had a pedagogical aim, and they usually contained an instructional preface. That so few tune books are classified as such by subject in the HTDL is probably due to their having often been cataloged by libraries as 'hymns.'

[16] The preponderance of 'Biography' as a genre term in the M range may be an indication of the magnitude of attention that musicological scholarship, in particular, has devoted to the figure of the composer. However, 'Biography' tends to be one of most frequent genre terms within the HTDL as a whole, as Table 2 attests.

[17] 'Music' and 'Musical Scores' are overlapping categories for a small number items in the HTDL, which is an artifact of how

Table 7 Genre terms prevalent within Library of Congress Classification music class M

Genre term	World Public Domain records	U.S. Public Domain records	Restricted records
Biography[16]	770	1804	9721
Government publication	115	130	1563
Discography	1	4	528
Catalog	34	73	500
Festschrift	10	21	316
Dictionary	84	135	296
Tune-books	28	28	15
Encyclopedia	2	2	24
Handbook	2	2	13

Items in this category include everything from full opera scores to items with individual instrumental or vocal parts for performers. They are a subset of the material in the HTDL useful for research on the performance cues and dynamic, tempo or expressive markings of particular instrument or vocal parts, such as references to the human voice in instrumental expression marks, and vice versa.

By way of comparison, as of 2014 the physical collection in the Music and Performing Arts Library at the University of Illinois, Urbana-Champaign included 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)

Musical parts in the HTDL are presented as a single PDF document, with the different parts—for example, those in a string quartet with violin 1, violin 2, viola, and cello—following each other, though not necessarily in the correct score order. If one uses the HTDL interface to search for "parts" in the subject field and limits the format to "music" or "musical score" only, then 26,925 items are returned, of which 2,811 are available through 'Full view' in the United States. Table 8 shows some of the most commonly used genre-specific subject headings for items with parts. The number of scores in the HTDL that have parts is not as high as that in the International Music Score Library Project (IMSLP), but the HTDL can still be a useful point of reference for performers and for researchers of music notation.

5 Case Studies

Our analysis of the music-related material in the HTDL so far has attempted to characterize the library bibliometrically as a resource for musicologists. In the following case studies, we will now focus on some particular kinds of musicological research that can be conducted using the HTDL. These case studies illustrate the potential for finding musical editions, researching popular genres of music, finding musicological source readings, as well as conducting research in music notation and audience studies.

5.1 Discovery

Many of the tables and figures in this paper address the volume of materials to be found in the HTDL. However, the HTDL is arguably also just as useful for research about "edge cases," oddities, and "needles in haystacks." In music studies, these include forgotten works, obscure or less prominent creators, lesser known performers and historically significant rare recordings. For instance, the HTDL contains several scores of historical significance, such as the first edition of the Brahms Sonata op. 120, no. 1, published by Simrock in 1895 (1895), 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 (1892).

The HTDL's full-text search functionality is particularly useful for discovering smaller works held within larger ones, a problem which, while also similarly challenging for subject areas like literature and for publications like serials, is frequently encountered by researchers of music. An aria from an opera may be published in an edition of that opera or in an anthology of arias from multiple operas. A keyword search of bibliographic records can locate the aria only if the full contents have been listed in a "notes" field, which is not typically the case for operas, whereas keyword searching in full text improves discoverability. "Ombra mai fu" from Handel's *Serse* is an illustrative example. A full-text search for the title of this aria returns over two thousand results, but a catalog search of the title retrieves only a handful of matches. Thus, the ability to search the full-text OCR of volumes in the HTDL

these metadata are generated from the MARC bibliographic records.

Table 8 Top genre subject headings in the HTDL for items with parts

Genre	Items
String quartets	2,040
Part songs, sacred	1,060
Violin and piano music	887
Motets	816
Sonatas (Violin and piano)	812

obviates the need for searchers to know which larger work the smaller work belongs to.

5.2 Popular Genres

The HTDL has utility as a massive union catalog that can view both the metadata and the contents of public-domain as well as copyright-protected volumes — even though human readers cannot see inside the latter. Full-text searching across volumes facilitates retrieving mentions of music and musicians of the twentieth century, as the following examples illustrate. A search for "The Beatles" returns 2,055 items available to read in "Full view" out of a total of 76,944 items. Original formats returned by this search include books, but also newspapers, journals, legislative hearings, and music industry statistics, with dates of publication ranging back to 1960, the date of the band's formation. A researcher in jazz studies interested in compiling every mention across all discographies contained in the HTDL of a jazz standard, such as the frequently recorded "St. Louis Blues," would be able to do so. A search for "St. Louis Blues" reveals 1,160 items in the HTDL, 139 of which are available in "Full view." Not only does this type of searching have obvious benefits for bibliographers, it can also provide a helpful starting point for scholars working on social, cultural, legal or institutional histories of music.

5.3 Source Readings

Within the rather large category of works included under music history and criticism (the 'ML' subclass), there are, of course, multiple areas of study such as ethnomusicology, cultural musicology and music iconography that either did not exist or did not become codified as disciplines until the twentieth century. For scholars in these areas, the HTDL may not be a resource of first resort since copyright restrictions would in most cases pose a severe challenge to viewing research products such as monographs and disciplinary journals. For example, searching the HTDL catalog with the subject term "ethnomusicology" returns 380 items overall, but only nine of them are available for unrestricted ("Full view") access outside the USA. This is a far lower proportion of material available for "Full view" compared with the proportions for the four most frequent Library of Congress subject terms that mention "music," as shown in Table 5 and Table 6. However, the many source readings about musical performance as well as impressions of music in the eighteenth century in the HTDL will be of great potential interest for historiographers of music.[18] These works are valuable sources of information on eighteenth-century performance practice and contemporaneous impressions of music.

The HTDL also does contain some literature by Westerners (and non-Westerners) on non-Western music. Representative examples include writings on indigenous Brazilian music by Jean de Léry (1534-1611), Western African music by Mungo Park (1771–1806), Arab music by Guillaume-André Villoteau (1759–1839), South Asian music by S. M. Tagore (1840-1914) and Ernest Clements (b. 1873), and Japanese music by Sir Francis Taylor Piggott (1852-1925).

5.4 Music Notation

The OCRed data associated with HTDL resources presents research possibilities to scholars of music notation. However, at present there is a great deal of variation in the accuracy of OCRed data from musical scores in the HTDL. The staves and notes of musical scores are invisible to full-text search within volumes in the HTDL because Optical Music Recognition (OMR) technologies have not yet been applied to them. However, the OCR process does detect dynamic, tempo and other performance markings, textual content such as lyrics, and titles of movements and works. As the OCR processing that is applied to material in the HTDL does not systematically capture text within scores, gaps occur in the OCR-ed text — most notably in vocal literature, in which melismas and line breaks routinely chop up words. Nevertheless, dexterous use of the HTDL's

[18] Important early histories and reference works related to music and contained in the HTDL collection include Jean-Jacques Rousseau's *Dictionnaire de Musique* (1768), Sir John Hawkins's *General History of the Science and Practice of Music* (1776), and Charles Burney's *A General History of Music: From the Earliest Ages to the Present Period* (1789).

advanced full-text search interface can enable scholars to carry out useful preliminary work. Researchers exploring Beethoven's notational practice over the course of his career or in relation to his contemporaries may be able to find productive paths of inquiry through advanced full-text search of the HTDL. For example, it has been claimed that Beethoven made more liberal use of soft dynamics in his late period (Sheer, 1998), which contributed to a more reflective, introspective quality to his music; this claim can be investigated by means of a search of the full text for the character string "ppp" (*pianississimo*), with the catalog fields for "author" and "format" restricting results only to Beethoven's musical scores. Similarly, researchers may be able to explore Beethoven's idiosyncratic and conflative use of the signs ***sf*** (*sforzando*), ***sfp*** (*sforzando piano*), and ***f*** (*forte*), as well as their occasional standardization by many of his editors. For instance, November mentions a bar in the autograph manuscript of the second movement of *String Quartet in E-flat, Op. 74*, in which Beethoven placed an ***sf*** marking in the first violin while the lower voices have ***sfp***. Editors have standardized this to either ***sf*** or ***sfp*** in all voices, although there is reason to believe that Beethoven intended the first violin part to be foregrounded (November, 2004). Thus, a comprehensive study of Beethoven editions could clarify the history of interpretations of Beethoven by shedding light on the range of meanings ascribed over time to his autograph manuscripts.

5.5 Audience Studies

A musicologist with an interest in audience or reception studies may find that the HTDL uncovers documents of significant evidentiary value from both inside and outside the Library of Congress 'M' class. One example can be found in the early nineteenth-century Italian opera libretto, Felice Romani's *I due Figaro*, which was based upon a French play by Honoré-Antoine Richaud-Martelly entitled *Les deux Figaro*. A prefatory note in an edition of the libretto published for the musical setting by Giovanni Antonio Speranza suggests that the French source play was so well known as not to require further description (Romani, 1840). Richaud-Martelly's play vanished from the stage long ago; it is Beaumarchais, author of *Le Barbier de Seville* and *Le Mariage de Figaro*, who is remembered today. The HTDL simplifies the task of assessing the extent to which theatres throughout France and the rest of Europe favored Martelly's work in the period after its *début*. A keyword search of the HTDL for "deux Figaro" finds hundreds of results, including periodicals, bibliographies, repertory lists, theatre histories, and even memoirs and published letters, which in turn can point to translations, records of productions, advertisements, reviews and first-person narratives and impressions of performances. Meanwhile, a keyword search for the name of the Italian opera ("due Figaro") – based upon the French source – returns similar document types for the musical settings and productions of the composers Michele Carafa (1820), Giovanni Panizza (1824), Saverio Mercadante (1835), and G. A. Speranza (1839). Thus, while searching the HTDL cannot replace a thorough and knowledgeable exploration of the relevant available sources in print and electronic formats, even a simple keyword search across the HTDL can yield useful, and occasionally surprising, research-advancing discoveries.

6 Challenges and Opportunities

Overall, the HTDL would benefit from improvements that would make it more suitable for musicological research. These changes include new recommendations for capturing page images, metadata enhancement, improved OCR and OMR processes, targeted collection development, and building the capacity to include audio formats.

6.1 Page Images

Some volumes of music in the HTDL have scanning errors, which is, in general, a well-documented problem in the repository (York and Hagedorn, 2015). For example, we found that a set of string quartets by Schumann show only the bound cover of the item (Schumann, n.d.), and a quartet by Dvořák that had obviously been repaired and supplemented with replacement photocopies, making the scans unsuitable for performance (Dvořák, 1890). Riley and Fujinaga describe best practices for image capture of scores, which are different from those for standard text-based documents because of several reasons: music notation is visual in nature, accidentals and note connections include fine detail, and music engraving practices show considerable variation (2003). As York and Hagedorn note, the HTDL has a system for requesting new scans — of either single pages or entire volumes — when poor quality is noted. Encouraging contributing libraries to follow Riley and Fujinaga's guidelines for recapturing music would lead to both improved quality in the HTDL of these resources.

A recent initiative of interest for the music content of the HTDL is the Workset Creation Through Image Analysis of Document Pages (WCTIADP) project, (Biggers, Audenaert, and Houston, 2015) one of the

prototyping projects of Workset Creation for Scholarly Analysis (WCSA), a HathiTrust Research Center initiative funded by the Andrew W. Mellon Foundation (Jett, 2015). Using the HTDL as dataset, the WCTIADP project aims to develop a software application that will use the visual characteristics of digitized printed pages to identify documents containing three kinds of visually distinctive materials — poetry, music, and illustrations — of interest to humanities researchers. This approach can help automatically identify those regions of a page that additionally contain musical notation even though the page as a whole does contain standard text elsewhere. This can potentially enable researchers to execute queries that will run exclusively against those regions within pages that only contain music.

6.2 Metadata Enhancement

Duffy notes that, compared with searches of Google Books, searches of the HTDL tend to be more accurate, as a result of the latter's more nuanced use of library bibliographic records (2013). However, the metadata in the HTDL do have limitations. Since contributing partners supply bibliographic metadata for volumes they ingest into the repository, the pathway by which an item arrives in the HTDL has an impact on the quality of its metadata. Items that arrived at the HTDL via an intermediary such as Google Books may have less detailed metadata than those contributed directly by member libraries. Additionally, Duffy observes some inconsistencies and ambiguities in the metadata (2013). For example, the subject terms for a string quartet by Malipiero in the HTDL that was imported from Google Books was found to misleadingly indicate that the scanned item was a score with parts, even though it was really a miniature score (1921). Because scores are complex items, both the metadata and the OCRed text associated with them are arguably more error-prone than is the case with those HTDL volumes that are text-only.

Efforts can be made to improve or "clean up" existing metadata, as well as to include new metadata in the HTDL. New display functionalities, along with metadata schemas such as those based on Functional Requirements for Bibliographic Records (FRBR) (Tillett, 2003), could also be used to link related items in the HTDL together — for example, scores to recordings or to music-related texts. Hagedorn et al have experimented with improving HTDL metadata through the use of the algorithmic approach of topic modeling. Conducting a two-part study using records from the HTDL about art, art history, and architecture, they found that topic modeling could help users discover more relevant materials than standard library metadata by itself (2011). Additionally, researchers have explored how to add a new metadata element, the gender of the author, to items in the HTDL. Researchers with the HathiTrust Research Center, discussed in further detail below, have used data sources such as census data and the Virtual International Authority File (VIAF) (Tillett, 2004) to perform a preliminary identification of gender for approximately 80% of the authors in the HTDL public domain holdings (Peng, Chen, Plale, and Kowalczyk, 2014). Other useful metadata enhancement with potential for further work would be to provide the country of origin of authors, composers, and performers, and to display — and allow searching based on — the period or date of creation would be yet another However, as Newcomer et al have noted, such a change would be difficult to implement as the use by music catalogers of the MARC fields where period of composition would be recorded has not always been consistent.

6.3 OCR and OMR

Music presents a special challenge to digital libraries on account of the relative difficulty of machine reading of musical notes. In her study investigating the music-related content in various large-scale repositories, Dougan found that, since non-text matter such as musical notation does not lend itself well to the standard OCR process for recognizing text in optical scans, text-based repositories like Google Books, Open Content Alliance (OCA), and the HTDL do not favor including scores (2010). Specialized tools for recognizing musical notes, which we discuss later in this paper, have so far only been in the early stages of development. For the purpose of discovering materials that are not findable through searching traditional bibliographic data, the full-text search that OCR enables is, of course, as valuable to musicologists as to scholars in any other field (Duffy, 2013). For example, full-text search of the text data is likely to be useful to a musicologist studying printed vocal music, as lyrics are not usually included as part of the metadata in catalog records.

Researchers from the Single Interface for Music Score Searching and Analysis (SIMSSA) project are already exploring ways to improve discovery and access of sheet music. This project is developing techniques for identifying musical scores within digitized books, and will create improved Optical Music Recognition (OMR) technology similar to Optical Character Recognition (OCR) for text. OMR will be used to further process the musical scores that SIMSSA researchers identify and locate within digitized books in existing collections such as Google Books, the Internet Archive, and the HTDL,

so that those scores can be made searchable. The ultimate goal of the project is to create a central place to search digitized scores (Fujinaga, Hankinson, and Cumming, 2014; Motuz, 2013). The HTDL can potentially begin enabling researchers to search by means of musical values such as pitch in addition to doing so by means of textual criteria such as words, by leveraging any improvement to the OMR process that the SIMSSA project may generate.

6.4 Collection Development

There are several areas where focused collection development in the HTDL would improve the library as a resource for musicologists. First, attempts could be made to have more opera librettos in the HTDL. Academic libraries usually did not collect the opera libretto in the past, as it was considered merely a secondary artistic product. Although searching by the subject heading for opera librettos in the HTDL does return a reasonable number of matches for these slender print volumes, we found, based on searches for the texts of seventeenth-century operas as well as for works by prolific librettists like Pietro Metastasio, Carlo Goldoni and Lorenzo da Ponte, that there still exists room to improve the number of opera librettos available. The texts of operettas, zarzuelas, and works of musical theater tend to occur even more rarely in the HTDL's collection. "Folk music" scores of almost any era, region, or ethnicity are likewise difficult to come by in full text in the HTDL collection, although some ballads and folk songs from the United States, Latin America, and various European ethnicities can be found. The antecedents of jazz, including spirituals, ragtime, and blues, constitute another area that could benefit from strengthening. Finally, literature on the intersection of music with race, ethnicity, anthropology, politics and technology—primarily twentieth and twenty-first century concerns—is another potential area in which further growth can be sought in the future, even though "Full view" access to most materials in this area would doubtless be restricted for some time to come as a result of copyright law.

6.5 Sound Recordings

A collecting area relevant to musicologists but not yet included in the HTDL is sound recordings. The HTDL, as of this writing, only accommodates digitized textual material. This gap is a significant deficiency for musicology research because the aural is, obviously, a critical modality in the study of music. Users are able to facet their search by the format "sound recording" in the HTDL interface, but th items returned by the search consist merely of the digitized textual material accompanying recordings, such as program notes, synopses and/or librettos. Figure 10 shows an item appearing in the catalog as a sound recording of Verdi's *Falstaff*. Clicking on "Full view" pulls up, for browsing or searching, the digitized version of the text accompanying this sound recording. Incorporating sound recordings, as well as building new discovery tools for audio, would eventually allow researchers to search by sound, which will obviously be a very desirable improvement.

The challenge of incorporating audio recordings into the HTDL has recently been taken up for consideration. Providing an overview of the HTDL in the context of Google Books and the Open Content Alliance (all three of which have overlapping content), Christenson states that the HTDL aims to eventually include other media types beyond the text-based materials with which it originally started, as the mission of academic libraries is to build, preserve, and provide access to collections that meet the research needs of their users not only for now but also for the future (2011). This will be a challenging endeavor because, while preservation and access are difficult enough for text and image collections, they tend to be more so for audiovisual collections—whether digitized or not—as recording media, especially magnetic carriers, become increasingly unstable with age, and playback hardware becomes obsolescent.

Creating ingest and validation procedures for audio presents further difficulties. Beers and Parker describe the challenges faced during a 2009 pilot project at the University of Michigan for digitizing unique audio items of high research value for inclusion in the HTDL (2011). They found that the existing routines for ingest and validation of non-text media, having been built exclusively with images in mind, could not be adapted easily to audio items. They also mention the paucity of audio digitization standards and the insufficiency of digital audio metadata standards. However, the publication of guidelines for audio preservation by the International Association of Sound and Audiovisual Archives (Bradley, 2009) and by the Association for Recorded Sound Collections (Brylawski, Lerman, Pike, and Smith, 2015) has probably improved the situation. Beers and Parker conclude that while over time the HTDL may evolve to incorporate audio, the preservation of digital audio would still require resources different from, and at a larger scale than, the resources needed for the preservation of digital images (2011).

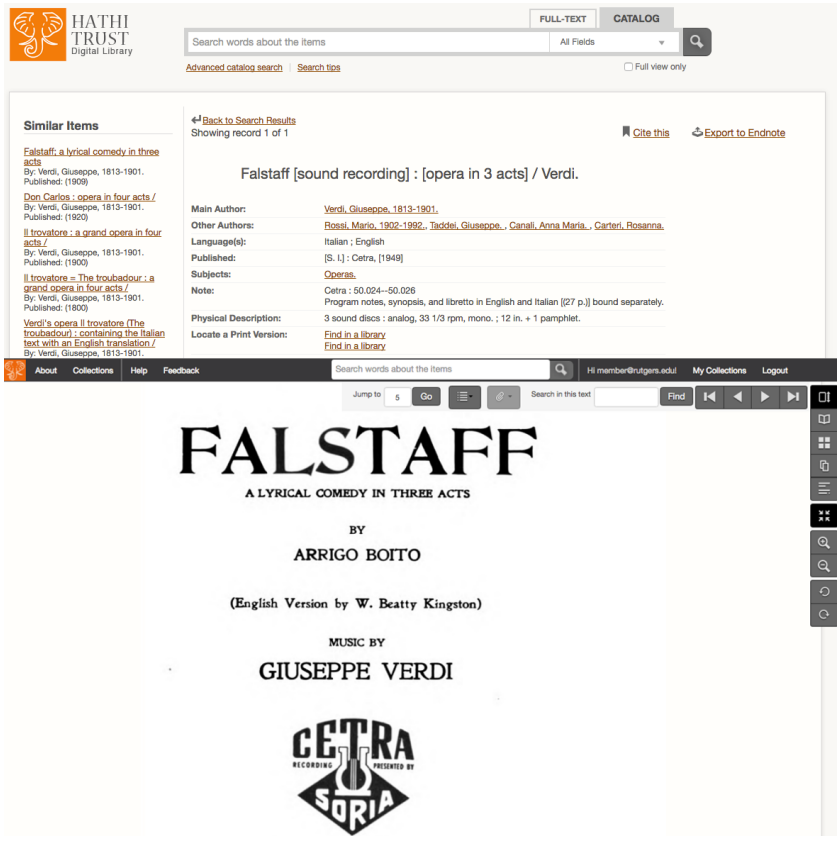

Fig. 10 A portion of the HTDL bibliographic record for a sound recording of Verdi's *Falstaff* (top); a page image from the accompanying liner notes (below)

7 Future Directions

Analysis of the digitized content of the HTDL is an opportunity to create new knowledge for musicology. Large digital libraries such as the HTDL are stewards of materials across a long span of time, making them, as Rumsey suggests, an ideal laboratory in which to process retrospective data and to research how dynamic systems evolve and change over time (2016). For instance, the analysis of a large quantity of text within the collection from the era in which a composer lived can enable the tracing of evolutionary patterns within a composer's musical and aesthetic views and of their relations to those of the composer's predecessors or contemporaries. Solomon has recently shown that there are parallels between the development of Beethoven's thoughts about aesthetics and the thoughts of E.T.A. Hoffmann and Karl Marx, Beethoven's near contemporaries (2008). It is meticulous close reading of texts that informs Solomon's research; however, in the future, algorithmic methods of automated text analysis involving computational "distant reading," to use a term popularized by Moretti, are likely to help researchers in carrying out similar investigations on a much larger scale, by analyzing large quantities of text. Bringing statistically based algorithms to bear on corpora that consist of works by many different writers, including composers and musicians, may allow comparative studies of temporal patterns of similarities and contrasts among writing to be carried out across the textual record, contained in the HTDL, of entire swathes of cultural life. The scale of the HTDL makes it possible for the shifting boundaries of a culturally and historically determined musical "mainstream" to become visible. Researchers in literary studies such as Underwood and Sellers are already utilizing the HTDL to trace the shifting outline of this "mainstream" in specific languages, particularly English. Applied to musicology, taking a "distant" approach to reading music-related materials can produce insights about their history and cultural context, and even about musical forms. As an additional benefit, the scale and characteristics of the HTDL will also permit useful inferences about the values and biases of the collecting and digitizing practices of its constituent member libraries, and thus of the régime of tastes and needs — legible only in aggregated form — of the constituencies with which they were implicated.

The HTDL is useful for cultural studies of music. Ratliff has recently drawn attention to the synoptic or synaural quality of our present cultural conjuncture, in which all surviving music from all of historical time is, for the first time, simultaneously present and available (2016). The HTDL embodies this trend in the domain of printed scholarly materials. The extensive quantity of scanned musical scores in the HTDL can potentially enable their automated analysis at such large a scale as has not hitherto been possible. For example, researchers are now starting to develop tools for analyzing digitized scores which can be used to investigate potential similarities between the musical styles of different composers over long periods of historical time, revealing how musical themes have been reused, developed, and borrowed. Just as a digital library with an extensive holding of digitized text allows for the automated discovery of patterns in text, a digital library with an extensive holding of music in the form of digitized scores can potentially enable similar discoveries of patterns in music, especially with the help of the OMR tools that projects such as Single Interface for Music Score Searching and Analysis (SIMSSA), mentioned earlier, are developing.

Composers sometimes use the same melodic motif in more than one composition, and musical "quotation" of a melodic phrase from another composer in a work is not uncommon either. Some composers are known to recycle entire arias and choruses with new texts; a well-studied example of this is Johann Sebastian Bach's *Christmas Oratorio*, written in 1734, for which Bach recycled, using a process that musicologists describe as the "parody technique," material from several of his earlier works — most notably his cantata *Laßt uns sorgen, laßt uns wachen*, BWV 213, from fifteen months earlier (Rathey, 2016). Better OMR will enable the discovery and enumeration of such re-occurrences, re-use and borrowings of musical phrases and larger chunks, which will make it possible, in principle, to trace chains of influence and of borrowing both within the same composer's work and among composers. Collections of musical scores can be created from the HTDL's holdings and subjected to automated analysis, yielding potentially significant new discoveries since a digital library as large as the HTDL contains scores even by composers hitherto neglected or less studied. It is worth noting that researchers in literary studies, focused on text, have already started to undertake similar investigations using the HTDL's resources to detect influence and borrowing among writers.[19]

The HathiTrust Research Center (HTRC), the research arm of the HathiTrust, currently provides tools and services to facilitate large-scale "distant reading" of the HathiTrust Digital Library (Downie et al, 2016). Founded in 2008, the HTRC continues to explore methods to allow for large-scale "distant reading" by means of computational analysis. The HTRC allows users to create custom corpora from within the HTDL and run text analysis algorithms against them.[20] Such methods open the door to so-called non-consumptive or non-expressive research[21] of even those materials that are not open to "Full view" in the HTDL interface (Sag, 2012). Non-consumptive research creates avenues for researchers to explore large-scale resources such as the HTDL within the constraints of copyright restrictions that otherwise prohibit making the actual text accessible for direct, human-scale, reading or display of substantial portions of the text (Zeng, Ruan, Crowell, Prakash, and Plale, 2014). Even for out-of-copyright texts for which human-scale reading of the actual text is permissible, distant reading is still useful for getting a sense of the large-scale, aggregate-level, properties of the text data. In that context, the ability to move back and forth easily between close reading and distant reading is ideal — for musical scores as much as for text. The HTRC is currently developing such a tool for text through the HathiTrust+Bookworm (HT+BW) initiative.[22]

8 Conclusion

The HathiTrust Digital Library is a resource for scholarship, unique in the quantity and range of digitized material from the world's research libraries it gathers and preserves. As a digital library, the HTDL may not currently be able to fully support the needs of all researchers in a specialized field like musicology since its general-purpose functionalities cannot fully accommodate the special characteristics of certain kinds of items, such as musical scores and audio recordings, that

[19] This was the focus of a collaborative project carried out under the auspices of the HTRC in 2015; details can be found at: https://www.hathitrust.org/htrc_acs_awards_spring2015.

[20] For simple examples of how comparison and contrast between two corpora created from the HTDL collection can be performed by using the algorithmic tools provided by the HathiTrust Research Center, see: 'Workset Builder and Portal of the HathiTrust Research Center'. HathiTrust Research Center UnCamp. Ann Arbor, Michigan. 30-31 March 2015, http://bit.ly/1NF7QLi.

[21] Sag notes: 'The HathiTrust aims to develop and facilitate the development of data mining and analysis of its digital collection. This activity would have qualified as "non-consumptive research" under the now defunct Amended Settlement Agreement [ASA]. "Non-consumptive research" as defined in the ASA is a form of nonexpressive use...' (2012).

[22] See: https://htrcbookworm.wordpress.com/.

are important to the field. However, in the future, advances such as the incorporation of audio content, Optical Music Recognition, and enhanced metadata will make the HTDL much more useful to musicologists. Meanwhile, the sheer scale of the HTDL's collection already enables significant text-based musicological research through the use of computational text analysis. Non-consumptive techniques, which make the content available for statistical and algorithmic analysis at aggregate levels while disabling human-scale consumption of the material, will make even the copyright-restricted content of the HTDL usable for many research purposes in the near future. In addition, ongoing expansion of the HTDL beyond the set of institutions currently contributing to it is likely to lead to more variation and diversity within its collection. Overall, while currently the HTDL falls short of being a comprehensive digital library suitable for *every* possible kind of musicological research, it already presents significant new opportunities to the field.

Acknowledgements We gratefully acknowledge the contributions of Kirstin Dougan and Colleen Fallaw, who were co-authors of a preliminary version of this paper, which appeared in the *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*. We wish to thank Janina Sarol for her help in generating the data for this paper. Xiao Hu provided valuable suggestions and critique. Mike Furlough of the HathiTrust, as well as Tim Cole of the HathiTrust Research Center, helpfully answered our queries. This work was made possible with generous support from the HathiTrust Foundation.

References

Beers S, Parker B (2011) Hathi Trust and the challenge of digital audio. IASA Journal 36

Biggers K, Audenaert N, Houston NM (2015) VisualPage: Workset creation through image analysis of document pages. Tech. rep.

Bradley K (2009) Guidelines on the production and preservation of digital audio objects. IASA Technical Committee

Brahms J (1895) Sonata, op. 120, no. 1, F moll = Fa mineur = F minor: Clarinetta & Piano: Viola & Piano. N. Simrock, London

Brylawski S, Lerman M, Pike R, Smith K (2015) ARSC guide to audio preservation. Tech. rep., Council on Library and Information Resources

Christenson H (2011) HathiTrust: A research library at web scale. Library Resources & Technical Services 55(2)

Dougan K (2010) Music to our eyes: Google Books, Google Scholar, and the Open Content Alliance. portal: Libraries and the Academy 10(1):75–93

Downie JS, Furlough M, McDonald RH, Namachchivaya B, Plale BA, Unsworth J (2016) The HathiTrust Research Center: Exploring the full-text frontier. EDUCAUSE Review URL http://er.educause.edu/articles/2016/5/the-hathitrust-research-center-exploring-the-full-text-frontier

Duffy EP (2013) Searching HathiTrust: Old concepts in a new context. Partnership: The Canadian Journal of Library and Information Practice and Research 8(1)

Dvořák A (1890) Quartett, op. 34. N. Simrock?, Berlin?

Fujinaga I, Hankinson A, Cumming J (2014) Introduction to SIMSSA (Single Interface for Music Score Searching and Analysis). In: Proceedings of the International Workshop on Digital Libraries for Musicology, London,UK

Hagedorn K, Kargela M, Noh Y, Newman D (2011) A new way to find: Testing the use of clustering topics in digital libraries. D-Lib Magazine 17(9/10)

Jett J (2015) Modeling worksets in the HathiTrust Research Center. Tech. Rep. CIRSS Technical Report WCSA0715, URL https://www.ideals.illinois.edu/handle/2142/78149

Malipiero GF (1921) Rispetti e strambotti per quartetto d'archi. J. & W. Chester, London

Moretti F (2013) Distant Reading. Verso, London

Motuz C (2013) CIRMMT Workshop, September 7th, 2013, Part I : Introduction

Newcomer NL, Belford R, Kulczak D, Szeto K, Matthews J, Shaw M (2013) Music discovery requirements: A guide to optimizing interfaces. Notes 69(3):494–524

November N (2004) Editing Beethoven's middle-period quartets: Performers, scholars and sources in dialogue. Ad Parnassum: A journal of eighteenth- and nineteenth-century instrumental music 12(24)

Peng Z, Chen M, Plale B, Kowalczyk S (2014) Author gender metadata augmentation of HathiTrust digital library. In: Annual Meeting of The Association for Information Science & Technology (ASIS&T). Seattle, Washington. Oct 31-Nov 4, 2014

Rathey M (2016) Bach's major vocal works: Music, drama, liturgy. Yale University Press, New Haven, Connecticut

Ratliff B (2016) Every song ever: Twenty ways to listen in an age of musical plenty. Farrar, Straus and Giroux, New York

Riley J, Fujinaga I (2003) Recommended best practices for digital image capture of musical scores. OCLC systems and services 19(2)

Romani F (1840) I due Figaro, ossia Il soggetto di una commedia: da rappresentarsi nel Ducale Teatro di Parma la Primavera del 1840. Filippo Carmignani, Parma, Italy

Root GF (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. S. Brainard's Sons, Chicago

Rumsey A (2016) When we are no more: How digital memory is shaping our future. Bloomsbury, London

Sag M (2012) Orphan works as grist for the data mill. Berkeley Technology Law Journal 27

Schumann R (n.d.) Drei Quartette für 2 Violinen, Viola, Violoncell: Op. 41 / von Robert Schumann; Revised by von Fried. Hermann. Retrieved from http://catalog.hathitrust.org/Record/100143217 on May 4, 2016

Sheer M (1998) Dynamics in Beethoven's late instrumental works: A new profile. The Journal of Musicology 16(3)

Solomon M (2008) Reason and imagination: Beethoven's aesthetic evolution. In: Historical musicology: Sources, methods, interpretations, University of Rochester Press, Rochester, NY

Tillett BB (2003) What Is FRBR? A Conceptual Model for the Bibliographic Universe. Library of Congress; Cataloging Distribution Service, Washington,DC

Tillett BB (2004) Authority control: State of the art and new perspectives. Cataloging & Classification Quarterly 38(3/4)

Underwood T, Sellers J (2015) How quickly do literary standards change?

Wilkin J (2011) HathiTrust and print storage: Building around a digital core. In: Committee on Institutional Cooperation Center for Library Initiatives Annu. Conf., East Lansing, MI

York J (2012) HathiTrust: The elephant in the library. Library issues 32(3)

York J, Hagedorn K (2015) Quality in HathiTrust. In: https://www.hathitrust.org/print/1975

Zeng J, Ruan G, Crowell A, Prakash A, Plale B (2014) Cloud computing data capsules for non-consumptive use of texts. In: 5th Workshop on Scientific Cloud Computing (ScienceCloud), Vancouver, Canada