About the HathiTrust Copyright Search Pilot Study in Germany

Joyce Ray and Melissa Levine

Abstract: The protocol for determining author death dates for digitized books published in Germany (1873–1933) and held in the HathiTrust Digital Library was tested by students in a project seminar. The paper discusses the feasibility of searching without the digital scans of the books, the usefulness of the researched resources and conclusions about the overall process.

Key words: Copyright search; HathiTrust; diligent search

1 Introduction

HathiTrust’s partnership comprises one of the largest digital libraries in existence. It was founded in 2008 by a group of research libraries in the United States and has expanded to include academic and research institutions in the United States, Canada and Europe. To date, there are more than 90 partner institutions.1

HathiTrust has grown rapidly since its initial launch, ingesting millions of volumes digitized from the collections of partner institutions and through partnerships with Google, Internet Archive, Microsoft, and other sources. It has grown rapidly with content continually digitized locally by partner libraries. Today, HathiTrust comprises over 11 million volumes, over 3,7 million of which have been identified as public domain, either in the United States or worldwide. The intention is to build a comprehensive archive of published literature from around the world and to develop shared strategies for managing and developing the partners’ digital and print holdings. The stated mission of HathiTrust is “to contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge.” While HathiTrust first serves “the members (faculty, students, and users) of its partner libraries, the materials in HathiTrust are available to all to the extent permitted by law and contracts, providing the published record as a public good to users around the world.”

Unfortunately, that small phrase “the extent permitted by law and contracts,” is not just a pro-forma caveat. It has had a profound impact on HathiTrust’s ability to go beyond its own partner institutions to serve the larger public good. Copyright law is extremely complicated, especially when the laws of different countries must be considered.

For books published outside the U.S., the copyright laws of the country of publication are considered in making a copyright status determination. There are approximately 587,000 titles in German in HathiTrust (and there may be multiple volumes associated with each title). This represents the largest group of non-English books in the collection. About half of these are viewable in the United States, having been published prior to 1923. But for many of them, further research is required to identify those that may be made viewable anywhere in the world.3 In many cases only the catalog records can be made publicly avail-

1 The organization’s name incorporates the Hindi word ‘hathi’ which means ‘elephant’, because like an elephant, HathiTrust never forgets. A helpful link to a recording of the correct pronunciation is provided on the HathiTrust web site at http://www.hathitrust.org/about (last downloaded 2015-01-08).


3 Ironically, there are German books published after 1923 that are in the public domain in Germany – and could be viewed there with
able in the digital library at this time, even if a work has passed into the public domain in Germany and could be made available as a full-text scan. Note that full-text search is offered even for limited-view items.

Under German law, like that of the U.S., a work passes into the public domain 70 years after the death of the last copyright holder. It is likely, however, that many books that might be entering the public domain under German copyright law were destroyed in World War II. Physical copies in the collections of North American research libraries may no longer exist in Germany, or at least the number of copies may be very limited.

Through his long residence in Michigan (a fact known to even casual acquaintances) and his long-time involvement in copyright issues and digital libraries, Michael Seadle was of course keenly aware of this situation. As Director of the Berlin School of Library and Information Science (IBI, by its German initials) and Dean of the Faculty of Philosophy I at Humboldt-Universität zu Berlin, he was interested in exploring how to advance the process of identifying German public domain works in the HathiTrust digital library so that they could be made available to readers and scholars in Germany and elsewhere. He was also interested in opportunities for Humboldt students to work on real-world problems. With the concurrence of the University of Michigan, he initiated a pilot study at IBI to test the HathiTrust protocol for identifying author death dates for German works.

2 The study

The study was conducted by students enrolled in a “Copyright Review Project Seminar” during Humboldt-Universität’s summer term, April–July 2013. The two authors of this paper directed the seminar and served as project advisors, respectively. Project Seminars are a uniquely German approach to education, in which students are charged with organizing and conducting work leading to a product, such as a research paper, with minimal guidance from their professor. Five students participated throughout the project and earned credit for the course, and four of them went on to turn the project seminar report into an article that was subsequently accepted for publication in the September 2014 issue of D-Lib Magazine.4

The students (and their professor) considerably enhanced their knowledge of German copyright law in the course of the project, including implications of a law passed in 2013 requiring a “diligent search” for living authors, or their death dates, before content holders may make digital copies of works publicly available, either because the works have been determined to be in the public domain or because they have been determined to be orphan works for which the copyright holders cannot be located.5 The German law was passed in accordance with a European Commission Directive on Orphan Works issued in 2012.6 While the issue of orphan works was not the focus of the project seminar, the students were intrigued by the concept of the “diligent search,” which was not defined in the Directive or the German law but was described in the context of specific resources to be used in conducting such a search.7

The students carefully followed the HathiTrust search protocol in their work, with at least two students attempting to verify all information located and everyone working to verify information in at least two sources. This was based on the methodology of the HathiTrust Copyright Review Management System (CRMS), which requires two reviewers to check each item. The seminar itself provided new insights to the CRMS team because they had not previously used students for copyright research and reviews; the seminar was an opportunity to examine the implications of using students to do this type of research. Because of this concern, the Humboldt students made extraordinary efforts to document their work, creating a master spreadsheet to record information about each search and a “search diary” maintained by each student for all the searches they performed.
The search began with a list of about 125 volumes of unknown copyright status in the HathiTrust digital library, entitled “German editions of Greek and Latin works 1873–1933.” HathiTrust supplied the catalog records for each volume.8

The search protocol required students to examine an exact physical copy of the work to be sure they were examining the same work as the one in the HathiTrust collection. The University of Michigan could make digital scans of the works available only to HathiTrust partners (and only under very limited, secured conditions) because their copyright status had not been determined. Since Humboldt-Universität was not a partner – and, indeed, only one European university is a partner as of this writing – students located a physical copy of each work in Berlin for examination and compared it with the catalog record rather than the digital scan. This led to some complexities, as the catalog records were not always exactly consistent with the physical volumes. Fortunately, most of the works turned out to be available in the Jacob-und-Wilhelm-Grimm-Zentrum, Humboldt-Universität’s main library. Two students independently examined a physical copy of each volume to ensure it was the same as the work described on the catalog record. In order to document their search, students took pictures of the title pages and other relevant pages of each volume with their cell phones to send to the University of Michigan for verification that the works were identical to the digital scans. This was a cumbersome but necessary step under the circumstances.

Because the title list represented works from the classics, the original authors (e.g. Seneca) were long dead and any existing copyright would belong to the more modern editor or translator. The underlying works are long in the public domain, but the new elements in the more recent editor or translator. The underlying works are long in the public domain, but the new elements in the more recent

students quickly observed that many of the works were multi-volume but had the same editor, substantially reducing the number of names to be searched. In addition, many of the catalog records were duplicates, since the same work may have been scanned by more than one institution, with each institution producing its own catalog record. In addition, some editors edited more than one title, further reducing the number of names to be searched. Finally, some of the HathiTrust catalog records were found to already contain the editor’s death dates. Since the catalog record was considered reliable by HathiTrust, these titles required no further searching. In the end, only twelve names had to be searched, but it can be said that each name was searched comprehensively and with useful results.

The twelve names that remained to be searched after the initial review were searched thoroughly as students attempted to use sources identified in the new German copyright law in the context of “diligent search”. The students made several significant discoveries relating to the diligent search provisions in the course of the project. They found that only a few of the identified sources were useful for the particular names they were searching, and some that might have been useful were not available online and were restricted to internal staff use. However, students were sometimes able to verify information through other sources, such as Google (a resource not mentioned in the law), after having initially found the information in one source identified in the legislation, thus providing a second source in accordance with the HathiTrust protocol. The students also found that knowledge of German was often necessary in conducting searches, especially in the case of common names that might be associated with a number of different titles by different authors or editors with the same name.

3 Recommendations on search strategies

Students made several important observations useful to future researchers looking for biographical information about authors, including:

- VIAF (the Virtual International Authority File) should be the starting place for author information, as it is the most comprehensive resource and is publicly available online. However, it may be necessary to look for variations in the spelling of the person’s name and/or the title of the work, so knowledge of the particular language is helpful in guessing at possible variations in spelling.

- Information for the same person may be found in more than one VIAF entry, since different national libraries may have reported on the same person (for example, if the person was born in one country but died in another, both may have reported the same information to VIAF). In such a case, the biographical data can be

8 The list was compiled by Gregory Crane, Humboldt Professor of Digital Humanities at Leipzig University. See the HathiTrust catalog records for Professor Crane’s list at http://babel.hathitrust.org/cgi/m?b7a=listis;c=161080990;sort=title_a;pn=1 (last downloaded 2015-01-08). Please note that the current HathiTrust records may not exactly match the title list of books searched due to subsequent changes to the records at HathiTrust.
regarded as verified, since each national library represents a single source according to the HathiTrust protocol.

- Similarly, other national VIAF entries should be searched if no information is found in the country of publication, since other national libraries may have reported it.
- Sources identified in the German law on orphan works were of limited value when access to the source was limited to internal use by professional staff of organizations, such as collection-holding institutions or collecting societies (which generally represent the interests of authors, in the manner of authors' guilds).
- Google searches could be useful as an additional source to verify information found in another source or for additional information that might be verified elsewhere.
- In Germany, the records of local Bürgerämter (roughly, "records offices") could be useful; however, research in these archives would be quite time-consuming as they can be presumed to be analog-only for records from the 20th century and earlier.
- University archives could also be useful sources of biographical data for persons known to have been on the faculty of a particular university, but again this research would be painstaking for pre-digital data.

4 Conclusions

In addition to the observations and recommendations of the seminar students regarding research for biographical data on German copyright holders, the authors made several useful observations about some of the issues relating to the HathiTrust and its processes:

- If access to digital texts in HathiTrust must remain restricted to consortium partners for security reasons, Humboldt-Universität and other European universities should consider joining the consortium in order to provide access to this important resource for students and faculty.

- It would be useful to follow up on the pilot study with a more extensive search of German works in the HathiTrust digital library (in anticipation of this possibility, the seminar students began a glossary of terms, included in the project report, to facilitate future searches). As the pilot study has demonstrated, it is critical that researchers have a good working knowledge of the language in question in order to fully exploit available resources. At the same time, the research could be greatly facilitated if HathiTrust were able to do some preliminary work, such as eliminating duplicate records and identifying records that already contain death dates. It seems likely that some of this work could be automated.

- HathiTrust could consider under what circumstances students could be accepted as reliable researchers (e.g., what documentation of the search process is necessary and what checks are required).

From the perspective of CRMS, the staff was able to experiment with working on German-language materials with the benefit of reviewers with native language skills and access to language and subject-specific research materials. The team was able to rely on the research product because it was well documented and followed the substance of the CRMS methodology. The project was able to bridge the physical distance through an effective partnership relying heavily on email and Skype. Further, the project convinced the CRMS team that students like those who conducted the pilot study – comparable to graduate level students in the US – could be deployed to reliably assist with the necessary research and documentation.

Unfortunately, the inherent efficiencies that would be gained by providing secured access to the digital scans were not possible for this project, since Humboldt-Universität was not a HathiTrust partner. The process that was implemented was reliable but inefficient. At this time, it is workable for relatively small bodies of material. Further, it still requires a check by CRMS staff of the research results and manual input of the findings by CRMS staff into the CRMS interface. That said, the CRMS team is thinking about the lessons learned in this project to inform ideas for a framework for similar projects that might involve other non-English language projects, refining the research and information needs for projects that might consider the copyright status of discrete bodies of material of particular scholarly interest (especially where the body of material falls under a single legal regime), what criteria might be useful for such projects and the level of coordination and oversight that might be required. An immediate outcome of the project is that CRMS is now working with the creator of the titles list for this project9 to make rights determinations of an expanded list of volumes published in Germany of interest in the classics. Note that this work is almost exclusively on books and monographs, as the complexity of copyrights relating to serial publications has not yet been tackled.

9 See note 8.
As a result of the pilot study, the University of Michigan’s CRMS project made nearly all of the individual titles of the provided list of German editions of Greek and Latin works publicly available through HathiTrust after verification. The outcome was useful as a learning experience for students, for HathiTrust and the University of Michigan as a test of its search protocol, and, most significantly, for the public that now has access to these works online.

Joyce Ray  
Digital Curation Certificate Program in Museum Studies  
Johns Hopkins University  
1717 Massachusetts Ave., NW  
Washington, DC 20036  
USA  
jray16@jhu.edu

Melissa Levine  
Copyright Office  
Library Administration  
University of Michigan  
818 Hatcher South  
Ann Arbor, MI 48109  
USA  
mslevine@umich.edu