3-13-2019

Characterizing Same Work Relationships in Large-Scale Digital Libraries

Peter Organisciak
University of Denver

Summer Shetenhelm
University of Denver

Danielle Francisco Albuquerque Vasques
University of Denver

Krystyna K. Matusiak
University of Denver

Characterizing Same Work Relationships in Large-Scale Digital Libraries

Peter Organisciak [0000-0002-9058-2280], Summer Shetenhelm, Danielle Francisco Albuquerque Vasques, and Krystyna Matusiak

University of Denver
1999 E Evans Ave, Denver, CO, USA
peter.organisciak@du.edu

Abstract. As digital libraries grow, they are prompting new consideration into same-work relationships. They provide unique opportunities for resource discovery, but their scale and aggregated models lead to challenges presented by duplicates and variants. Addressing this problem is complicated by metadata inconsistencies as well as structural/content differences. Following from work in algorithmically identifying duplicate works in the HathiTrust Digital Library, we present some cases that complicate our existing language for work entity relationships. These serve to contextualize the complexities of same-work alignment in digital libraries, ground future discussion around content similarity, and inform methods to better identify duplicates in large-scale digital libraries.

Keywords: digital libraries, entity alignment, text mining, text duplication.

1 Introduction

As the next generation of digital libraries (DLs) has grown to massive scales by aggregating materials from multiple collections, they are prompting new consideration into same-work relationships. This new environment of large-scale DLs provides unique opportunities for resource discovery but also poses challenges in distinguishing between duplicates and finding the most representative edition. Reconciling duplicate volumes from multiple sources is important for practical information organization and retrieval. It also allows for better downstream uses of digitized print collections, such as improving the quality of metadata, reconciling better OCR text quality, and effectively utilizing text mining for higher-level concepts and trends. However, determining the relationships between DL volumes is complicated by metadata inconsistencies as well as structural/content differences, such as variant editions, anthologies, derivatives, and aggregated works. Following from ongoing work in content-based alignment of digital works, this paper characterizes the scope of the issue.

We present a set of cases that complicate our existing language for work entity relationships. These three cases serve to demonstrate the complexities of same work alignment in massive DLs and ground the discussion around various levels of content similarity. These case studies are derived from a manual review of algorithmically matched duplicate works. Our content-based algorithm was intentionally preliminary, and our expectation was that the evaluation of its results would surface areas where content-based analysis is challenged. Instead, the results often reveal metadata failures, where inconsistencies in practice or the standards themselves are insufficient for ascertaining relationships between items contributed from multiple libraries.

Large-scale DLs like the 16.7 million work HathiTrust Digital Library present great potential value in their scale and coverage but are also quite complex. They carry the inconsistencies in bibliographic data from the print environment but add another layer of complexity by presenting users with multiple digitized versions of the same edition. The versions may differ in the symbolic representation because of the quality of OCR and scanning technology. What criteria could end-users apply to distinguish between different versions?

The remainder of the paper introduces early work in algorithmically identifying duplicate works in the HathiTrust Digital Library at various levels of granularity, from variant existing versions of content (same work, different expression) to variant presentations of the same content (same expression, different manifestation) to simply different copies from the same print run (same manifestation, different item). We introduce issues related to this task through demonstrative cases. Same-work relationships are a pertinent issue in large-scale DLs, and these case studies provide a language for better exploring that issue.

2 Background

Why should we identify same-work relationships? Library practice has traditionally focused on preserving duplicate relationship information only at the level of the artifact; e.g. same OCLC or ISBN number. To understand the need for identifying more nuanced same-work relationships, it is important to understand the context of recent large-scale DLs.

Recent projects have grown digital collections in consortial fashion. Europeana and the Digital Public Library of America aggregate cultural heritage metadata from institutions throughout Europe and the United States, and Internet Archive and Google Books have scanned holdings from libraries around the world. One of the largest DL projects is the HathiTrust Digital Library, a collection of 16.7 million scanned volumes [1]. The HathiTrust aggregates metadata and full-text materials not only from multiple libraries, but from multiple digitization initiatives.

Large DLs present a great deal of value in information science and other fields, due to their broad coverage, accessibility, and malleable digital representations. However, they also hold many duplicates and variants, and identifying them is difficult due to metadata inconsistencies and differing representations of the work. Identifying same-work duplicate relationships may a) improve access and retrieval, b) assist with the normalization of metadata, c) improve OCR capabilities, and d) better discover trends about history and culture via text mining and topic modelling.

For traditional access and retrieval, a massive searchable DL makes it easier to connect information-seeking users to the published record. However, culturally prominent

– and well-represented – works may overwhelm search results. As such, a retrieval system knowing exact or near-duplicate works could fold same results and provide room for other results. As per Xu and Smith, "...removing duplicates improves efficiency and user satisfaction," [2]. Such information can also aid people seeking to compare multiple versions of a work, perhaps contrasting forewords to a novel or evaluating translations of a foreign book.

For cataloguing practice, identifying matching works allows metadata to be compared and normalized. Inconsistencies as fundamental as title differences become apparent when matching works in large-scale DLs. More complete metadata can be created by aligning complementary records [3], like the date of first publication [4]. Identifying duplicates can address coverage and text quality issues as well. Our ongoing work seeks to overcome issues of poor OCR text by determining the 'cleanest' copy of multiple options. Alternately, Xu and Smith propose 'consensus decoding' of aligned documents for improved OCR [2]. Another area of improved access may be in completing incomplete multi-volume sets; e.g. if a scanning project at one library missed two volumes of a set, they may be found elsewhere in the collection.

The most revolutionary potential presented by large-scale DLs is in text mining the collection, to learn more about aggregate trends in history, language, and culture. However, redundancy confounds any language model's efficacy, and removing it is in the interest of many text mining uses. A model that tries to learn from words in a collection like the HathiTrust's will be have a skewed view of reality from repeating texts [5]. Removing duplicates helps mitigate the potential for skewed results that do not accurately reflect the reality of a certain topic or time period.

A note about language: we uses materials from the HathiTrust, and adopt their term *volume* to refer to a scanned copy of a single physical library item. In discussing various levels of similarity, we adopt the FRBR class 1 entities. Here, *work* refers to an abstract concept that denotes intellectual content of a distinct creation; e.g. *Hamlet* as concept. Works are realized through *expressions*, the "distinct combination of signs conveying intellectual or artistic content" [6], such as various versions of Hamlet. Expressions are embodied by *manifestations*, e.g. a physical or e- book. An individual copy of a manifestation **–** i.e. the specific bound set of pages for a book **--** is an *item*.

3 Related Work

Much of the literature regarding using the FRBR model leverages catalog records rather than content. Bennett, Lavoie & O'Neill used FRBR to examine how works were shown in WorldCat [7], finding in their sample of 996 works that 20% contained more than one manifestation, and only 1% had eight or more. More recently, the Global Library Manifestation Identifier (GLIMIR) project clusters WorldCat records into work-clusters using a heuristic-based metadata matching approach [8]. Our exploration of work relationships uses content rather than metadata, allowing us to consider duplicates with divergent metadata as well as works that exist in varying forms.

A survey on translating FRBR to practice found a desire for verification of the model with real data applied in different communities. Participants also expressed a desire for

“...tools that facilitate the FRBRization processes,”, albeit with appropriately broad testing and research. Our ongoing work pursues both of these goals.

FRBR and its related standards are being consolidated through the IFLA Library Reference Model (LRM) [6]. LRM maintains the FRBR entities and refines them further by defining attributes and relationships. In recognition of end-users’ needs and tasks, LRM introduces a new attribute of a work – representative expression [6, p.41]. This attribute can be particularly useful in the large-scale DL environment where users are faced with duplicates, multiple editions, and aggregated expressions. It is essential for users to distinguish between editions and to identify that a certain group of editions relates to the same work. In some cases, users may want to identify and select an expression that is the most representative of a work. LRM still has a focus on forms and relationships established in the print environment but states that extension can be developed for DL needs. It suggests a *digital item* entity as an intermediary between the *manifestation* and *item* entities [6].

4 Approach

This work uses a sample of 102k literature books from the HathiTrust Digital Library [9], all of which are in the public domain and easily accessible for manual review. From the sample, we subsampled 20 'target' volumes, asked a basic algorithm to rank other volumes in the set by their similarity to the target volume, and then manually reviewed the top results to determine their relationship to the target. The purpose was in uncovering quirks and complexities that a mature algorithm needs to account for.

The basic algorithm used word frequencies from the HathiTrust Research Center's Extracted Features Dataset [10], term weighted with TF-IDF and modelled using Latent Semantic Analysis (LSA) [11]. LSA has the effect of reducing the term count representation of a text to a smaller set of dimensions, each signaling a latent set of co-occurrence patterns. It has the effect of smoothing between trivial differences in word usage (e.g. 'car' and 'automobile' may be represented together), allowing better comparisons between texts than by directly comparing words. Pairwise comparisons of texts' LSA representations were performed with cosine similarity.

The evaluation surfaced many areas of complexity for thinking through the conceptual boundaries and technical complexities of same work relationships. In this paper, we do not report on the entire evaluation; rather, we aim to tell the story of three cases that provide a face to some of the issues we see.

5 Case Studies

Nested Whole/Part Work-to-Work relationships. A difficulty in considering manifestations of a work is when they do not neatly align with the physical or organizational form; e.g. a work that manifests across a set of physical volumes. FRBR and LRM allow for these types of whole/part relationships, where a work may exist as a subset of another work.

Such whole/part relationships can nest, as in *The Works of Charles Dickens* (1897). Our evaluation's target text was volume 31 of 36 for this work (*w1*), which was entirely comprised of *Christmas Stories* (*w1.1*). This is a typical whole/part work-to-work relationship; however, *w1.1* itself is comprised of 21 smaller stories across two volumes (15 in v.1, or v.31 of the greater work, and 6 in v.2/v.32). That means that a story like *A Message from the Sea* may be published on its own, but also exists as part of *Christmas Stories*, which may be published alone or as part of an even larger "*Works of*" work. Our basic algorithm found instances of each circumstance.

LRM affords consideration into aggregate manifestations, which may allow for better classification of these relationships in the future. However, this particular example was complicated in further ways that do not appear in the current HathiTrust metadata. *w1.1* in this version of *w1* is a facsimile of an earlier manifestation, represented in a special kind of manifestation-to-manifestation relationship. The algorithm found numerous exact duplicates, with the same appearance but different titles, years, control numbers, and volume numbering. These cataloguing quirks may be understandable given the hodgepodge nature of *w1* - the title page of our target volume says "Vol. XXXI" followed shortly by "Vol. I".

Identical Manifestations with differing catalogue metadata, control numbers. Why not match same-work relationships by metadata, particularly by title and author? Currently, HathiTrust identifies duplicate manifestations in their holdings by grouping items by OCLC or institutional control numbers [12]. This metadata-based approach is also used by OCLC for their FRBR work matching algorithm, and later the GLIMIR project to allow fuzzier matching. It is effective but not complete, due to inconsistencies in practice and record completeness. A telling example was observed with *Writings of Samuel Richardson, vol. 19* (1902, *w1*), which includes *vol. 6* of the book *The History of Sir Charles Grandison (w1.1)*. The matching algorithm surfaced four volumes which manual review confirmed were identical, manifestations. However, between the four matches and the target text, there were four different titles, five different OCLC control numbers, and even variation in the volume numbering. Once the metadata was classified as *w1.1*, the other four times it was classified as *w1*, though with inconsistent titles.

HathiTrust Digital Library volumes are linked by the catalogue records from the contributing institution, and records are grouped together in the interface using OCLC numbers, roughly intended to keep identical manifestations together. Considering the coverage of catalogue records for our five identical volumes, only two records have complete coverage of their multi-volume set: *Writings of Samuel Richardson* (OCLC:12097044) and *The history of Sir Charles Grandison, Bart* (OCLC:6359966). *The novels of Samuel Richardson* (OCLC:3451879), *The novels* (OCLC:68137659), and *Writings of Samuel Richardson* (No OCLC) are incomplete.

Between just these five HathiTrust records, there are 126 scans of physically bound books, but limited ability to infer their relationships - perhaps guessable through title heuristics but otherwise requiring manual or automated content review.

Same Works with Itinerant Expressions. One complicated history seen in our evaluation was *The Life and Adventures of Robinson Crusoe,* by Daniel Defoe. This book has many small variants, non-canonical tweaks, and additions by publishers. It shows the breadth of how expressions may deviate beyond more common instances like editions, translations, and abridgements. This work was frequently adjusted by publishers who saw the story as more interesting to new readers than the style that Defoe wrote it in [13]. Consider these three observed variants on the opening line:

- About the year 1632, I was born in the ancient city of York, of respectable parents.
- I was born at York, in the year 1632, of a reputable family.
- I was born in the year 1632, in the city of York, of a good family, though not of that country, my father being a foreigner of Bremen, who settled first at Hull.

There are 1198 English-language manifestations of the first book listed in one bibliography [13], ranging in date from 1719 to 1979. After the first edition was published, publishers found "means to increase the size and price of Robinson Crusoe," [13] including adding images, new introductions, biographies, and information about Alexander Selkirk, a Scottish sailor who spent four years alone on Juan Fernandez Island. Abridgment was common practice for broadening the audience in *Robinson Crusoe*'s time, such as removing mention of cannibals or "leaving out the dull parts" [14, 13]. The copyright frequently changed hands, further exasperating the issue.

A case such as *Robinson Crusoe* illustrates some of the hurdles that content-based methods for identifying same-work relationships must overcome: an evolving and branching text. It also challenges assessment of representative expressions in LRM.

6 Conclusion

The scale and biases introduced by multi-institutional aggregations in large-scale DLs have prompted new consideration of same-work relationships, not only at the traditionally considered level of exact duplicates (i.e. same manifestation) but in tracking a work across different iterations or configurations. This type of alignment can improve traditional access and retrieval, as well as providing important evidence for improving metadata, correcting OCR, and mining history and culture in aggregate.

Conceptual models provide a language for thinking through these types of relationships, but the reality is complicated, as this paper shows through a selection of lucid examples. In future work, we are developing content-based methods to match relationships in large DLs, informed by the complex real relationships described in this paper. Where duplicates are usually identified by metadata, as with the HathiTrust [12], we aim to create a more thorough inventory of same-work relationships.

Same-work relationships are not exclusively useful to DLs. As cataloguing standards transition to FRBR-based Resource Description and Access (RDA), there is now the ability to encode higher-level relationships between works and expressions. Leveraging these new fields is difficult due to the effort required in identifying duplicate relationships, punctuated by the complexities discussed in this paper. As we produce data on the massive HathiTrust collection, it will be possible to align with other collections, aiding other libraries in meeting modern cataloguing standards.

7 Acknowledgements

This project is supported by IMLS grant #LG-86-18-0061-18.

References

1. "About". HathiTrust Digital Library. https://www.hathitrust.org/about, last accessed 2018/9/6.
2. Xu, S., & Smith, D. Retrieving and combining repeated passages to improve OCR. In Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries (pp. 269-272). IEEE Press. (2017).
3. Hillmann, D. I., Dushay, N., & Phipps, J. Improving metadata quality: augmentation and recombination. http://hdl.handle.net/1813/7897. (2004).
4. Bamman, D., Carney M., Gillick J.,Jon Gillick, Hennesy C., Sridhar V. Estimating the date of first publication in a large-scale digital library. In Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries (JCDL '17). IEEE Press, Piscataway, NJ, USA, 149-158 (2017).
5. Schofield, A., Thompson, L., & Mimno, D. . Quantifying the Effects of Text Duplication on Semantic Models. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2737-2747 (2017).
6. Riva, P., Le Boeuf, P., & Žumer, M. IFLA Library Reference Model. International Federation of Library Associations (IFLA). Available at: https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf (2017).
7. Bennett, R., Lavoie, B. F., & O'Neill, E. T. The concept of a work in WorldCat: an application of FRBR. Library Collections, Acquisitions, and Technical Services, 27(1), 45-59. https://doi.org/10.1080/14649055.2003.10765895 (2003).
8. Thornburg, G. A candid look at collected works: challenges of clustering aggregates in GLIMIR and FRBR. Information technology and libraries, 33(3), 53-64. https://doi.org/10.6017/ital.v33i3.5377 (2014).
9. Underwood, T., Capitanu, B., Organisciak, P., Bhattacharyya, S., Auvil, L., Fallaw, C., Downie, J.S. Word Frequencies in English-Language Literature, 1700-1922 (0.2) [Dataset]. HathiTrust Research Center. http://dx.doi.org/10.13012/J8JW8BSJ. (2015).
10. Organisciak, P., Capitanu, B., Underwood, T. & Downie, J. S. Access to Billions of Pages for Large-Scale Text Analysis. In iConference 2017 Proceedings, Vol. 2 (pp. 66-76). https://doi.org/10.9776/17014. (2017).
11. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407 (1990).
12. HathiTrust Collection Committee. Discussion document: recommendations for handling duplicates in HathiTrust. Available at: https://www.hathitrust.org/documents/hathitrust-collections-duplicates-report-201204.pdf (2012).
13. Lovett, R. W. & Lovett, C. C. Robinson Crusoe: a bibliographical checklist of English language editions (1719-1979) (No. 30). Greenwood Pub Group (1991).
14. Howell, J. Eighteenth-Century Abridgements of Robinson Crusoe. The Library, 15(3), 292-343. https://doi.org/10.1093/library/15.3.292 (2014).