Abstract

The emergence of massive digital library collections presents an opportunity to pose novel, collection-wide questions of our published history, offering new ways to access and use library materials. Studying collections in aggregate may also uncover new knowledge about their individual items, by situating them within a broader field and revealing relationships with other items. Unfortunately, models that learn from massive digital library collections are challenged by bias from duplication and text variants.

The University of Denver, in collaboration with Northeastern University, proposes a content-based study of duplication, text variants, and similarity in digital libraries, applying content-based approaches to learn more about digital collections. This research overcomes a key obstacle in learning from massive digital libraries, biases from repeating text, using the 16 million volume HathiTrust corpus. It will also produce a dataset of stylistic and thematic similarities between books, which will be applicable as a complement to existing retrieval and access methods in all library collections.

Specifically, this project focuses on three questions: 1) how do you accurately disambiguate varying levels of duplication in a digital text collection, 2) how do you identifying a canonical copy of each work from a set of its items, and 3) what are the most similar volumes and authors to each volume in the HathiTrust? The third question will produce recommendations of similar books to each volume in our research corpus; however, given its immense size and the fact that the HathiTrust is a collection of other library collections, the data will be more broadly applicable.

In answering its three questions, this project will produce results that are immediately and freely usable in library collections, while also contributing long-term knowledge to support novel collection curation research.

This project will be performed in two years, from May 2018 to April 2020.

Text Duplication and Similarity in Massive Digital Collections

Massive digital libraries have enabled a class of collection questions that can only be asked of our aggregate published history. It is now possible to look inside library materials at scale, allowing content-based analysis of relationships between text materials. Content-based relationships can improve access in ways that complement metadata-based approaches and human recommendation.

Unfortunately, models seeking to learn from massive digital library materials are frustrated by a key diseconomy of scale: duplication and text variants. We propose a study of text duplication and similarity in massive digital collections. The volume-level relationship data provided by this project will be of immediate value in public and academic libraries, useful in improving access services such as recommendation and retrieval. At the same time, it will tackle a critical hurdle for text mining scholars seeking new historical and cultural insights from massive digital libraries, opening the door to more effective content-based digital library research in the future.

Our research builds from publicly available research data from the HathiTrust Digital Library (HTDL). The HTDL holds scans for over 16 million library holdings (‘About’ 2018), collected in consortial fashion from over a hundred contributing libraries. The scale and provenance are useful here, as the HTDL digitized collection offers more content than each individual institution, and findings learned from that content are likely to overlap with notable portions of other catalogues.

The outcome of this project will be in derived metadata enrichment. The methods being studied and their resultant datasets will produce information about the relationships between volumes in digital libraries. We will use the HTDL as an access point to peer inside 16 million volumes, leveraging the scale to fingerprint the words inside them and to find duplicate, variant, stylistically similar, and thematically similar texts. Those relationships are immediately useful in collections that overlap with the massive HTDL, while the fingerprinting and analysis models will be applicable for additional volumes. The goal is to learn more about the individual volumes by modeling them and their relationships at a collection-wide scale, then taking that learning back to the volume level. This study will focus on three research questions:

RQ1. How do you accurately disambiguate varying levels of duplication in a scanned text collection? Repeating text causes issues for information modeling, but controlling for the issue is surprisingly complex. While scans of identical print runs can be inferred from metadata, the challenge is amplified when considering variant expressions - e.g. aligning works that exist in multiple versions or editions - or variant manifestations - e.g. multi-volume sets or parts of a collected anthology. Scanning errors add further complications.

RQ2. How do you determine a canonical copy of each work in a digital library, the ‘best’ copy for access or analysis? This question applies particularly to massive consortial collections like the HTDL, which may be biased toward the books important enough to be in multiple contributing collections. For retrieval, access, and analysis, however, not all digital volumes are equally valuable. We seek to identify archetypal copies of works by identifying the volume that most resembles the text of all the duplicates, with the smallest amount of deviation from that model.

RQ3. What are the most similar volumes and authors to each volume in the HTDL? Answering this question is the most broadly valuable, yet it necessarily follows the resolution of RQ1. When looking at the variant grades of duplicate content, the problem is one of measuring similarity in the words used within. One step further, beyond different scans, print runs, and editions, it is possible to find volumes which are different works, but merely very similar. Some types of words can reveal stylistic similarities, while other classes of words demonstrate topical or thematic overlaps. In resolving this research question, this project will be able to recommend ‘more like this’ books, information that can improve access and retrieval systems. Content-based analysis is a good complement to reader advisory or metadata-based recommendation, providing an access vector that is good at surfacing lesser-known works.

In the remainder of the proposal, we describe the project background, proposed methods and project activities, broader impact, and planned deliverables.

Statement of National Need

This project addresses two key challenges related to access and use of library collections. First, it seeks to add depth to collection discovery at the item-level, developing methods to recommend books by contextual and thematic similarities. Without regard for the prominence of the work and able to ‘read’ millions of books, this form of algorithmic recommendation can dig deeper into collections to uncover neglected or forgotten items. Secondly, it aims to overcome a key hurdle to realizing the lofty potential of collection-scale learning in digital libraries: identifying clean works and culling duplicated content.

This research infers more about the individual volumes in library collections, producing methods to reduce noise in digital libraries and data on relationships between volumes and authors. For example, at the conclusion of this project, we will know more about Stephen King’s *The Stand* and which literature traverses similar themes.

Digital materials allow libraries to think of their collections beyond their component parts, computationally beholding the collections at scales beyond what people can read. At the aggregate level, this work will greatly improve scholars' ability to research history, culture, language, and style through massive-scale digital libraries. This is an emerging area of knowledge discovery that libraries are well-positioned to steer, and the proposed project will address a significant hurdle to the domain.

Duplication in text collections frustrates unsupervised learning models, which try to uncover latent patterns in the words that co-occur in texts. When a piece of evidence recurs, it likewise duplicates the patterns from that text, making certain word trends seem stronger than others. These types of biases only serve to frustrate content-based curatorial tasks, ones that try to learn something new about a collection or its works from the language used within. For example, in trying to algorithmically infer topics (i.e. topic modeling) in *IMLS Digital Collections and Content*, a federated cultural heritage collection, putting effort into identifying and removing runs of duplicated texts resulted in significantly improved results (Efron, Organisciak, and Fenlon 2011).

This issue was directly studied by Schofield, Thompson, and Mimno in the context of topic modeling (2017). When the proportion of a corpus with duplicates was sufficiently large, or when the amount of duplication grew, the models grew less representative for other texts. Digital libraries suffer from both problems Schofield et al. consider – many texts repeating a few times, as well as some texts duplicated a great many times. Additionally, the longer document lengths welcome new quirks, the duplication biases are unevenly distributed, and the use cases are broader than simply topic modeling; nonetheless, the trend is clear. Going by metadata alone, the HathiTrust 16m volumes are represented by 8.2m records. While that number folds multi-volume sets in addition to known duplicates, the scale of the duplication derived from metadata only shows a fraction of the issue.

The problem of duplication is not only an internal hurdle, but one that affects users. Consider the scope of the problem in the context of the HathiTrust Digital Library. A reader searching the 'natural selection' subject area of the HTDL can retrieve a promising 518 relevant volumes. However, 97 results are various copies, editions, volumes, or remixes of Darwin's Origin of Species. A library that can group these copies in their catalogue search will provide better access to the relevant materials beyond Darwin. A scholar applying text mining to learn overarching themes in natural selection work will be equally challenged by the book's popularity; providing a single archetypal copy for analysis would avoid the bias. In other areas, past scholarship already complements our proposed work to the benefit of library collections, such as research finding that the most effective way to estimate a date of first publication is to use the earliest date of known duplicates in the HTDL (Bamman et al. 2017). Our work will improve this approach and immediately allow libraries to improve publication date metadata.

Project Design

As a research grant, the outcomes of this project are tied to answering the research questions:

- RQ1. How do you accurately disambiguate varying levels of duplication in a scanned book collection?
- RQ2. How do you determine a canonical copy of each work in a digital library, the ‘best’ copy for access or analysis?
- RQ3. What are the most similar volumes and authors to each volume in the HTDL?

The underlying goal driving these research questions is to provide a better understanding of the relationships between works in a large collection. The indicators for whether we adequately meet the outcomes are tied to the evaluation metrics described in the workplan below.

One key assumption underpinning this study is that we are at a point where content-based analysis of library books is tractable, in access and ability. While massive digital libraries now hold the content, legal reasons restrict access to newer, in-copyright portions of those collections. However, this project’s access hurdles have been resolved through the release of the Extracted Features (EF) Dataset from the HathiTrust Research Center (Capitanu et al., 2016). The EF Dataset provides non-consumptive, unreconstructable representations of the volumes in the HathiTrust Digital Library, including page-level counts of words for over 5 billion pages. A jumble of words cannot be read by a person, but still contains the necessary information for fingerprinting the content of books. By releasing quantified information that has already been tokenized, tagged, cleaned, and re-hyphenated, the EF Dataset also makes a project of our scope feasible by greatly reducing the preparatory effort and computational needs of our work.

Beyond access, our preliminary work over a set of 102k works demonstrates feasibility of the approach (Organisciak et al 2017). Scaling to millions of works will present a challenge; however, this can be managed by roughly identifying candidates for comparison and only cleanly measuring relationships between those candidates.

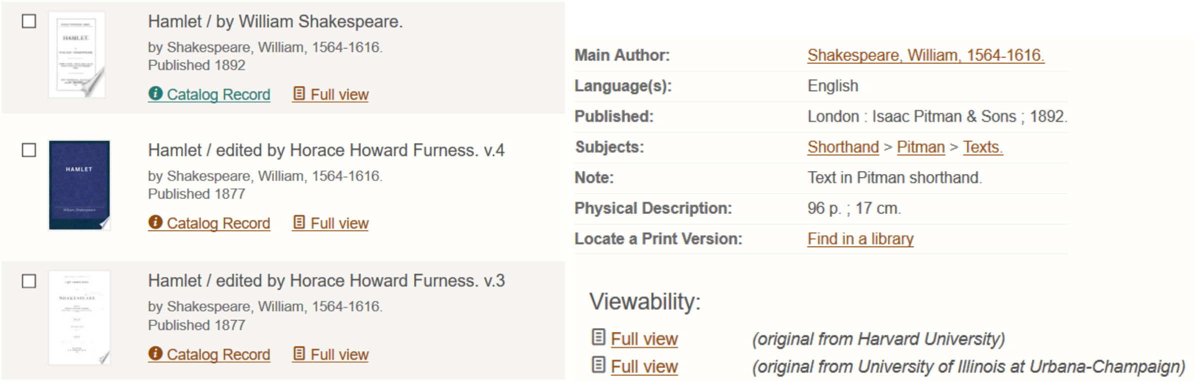
Background

This research takes the FRBR model as its conceptual foundation, specifically the group 1 entities. FRBR – the Functional Requirements for Bibliographic Records – considers library material in the context of *works,* which are realized by one or more *expressions,* which are embodied by one or more *representations,* each of which is exemplified by one or more *items* (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998).

The work may be considered the foundational abstraction that the author or authors imagined, for example *Hamlet*, regardless of how it has been expressed. Expressions are different ways of writing down or otherwise representing the work; e.g. different surviving texts of *Hamlet*. A manifestation is the ‘physical embodiment’ of an expression; e.g. a specific print run of a version of *Hamlet*. Finally, an item is a single exemplar of a given manifestation; e.g. the paper copy (and, for our purposes, its equivalent digitization) of a given expression of Hamlet.

Duplication can exist on various levels: multiple items of the same manifestation, multiple manifestations of the same expression, or multiple expressions of the same work. We aim to capture each level.

Traditionally, metadata-based approaches account for the first type of relationship, using manifestation-based control numbers such as ISBN, ISSN, URN, and OCLC. This is true for the HathiTrust Digital Library, which groups items into manifestation-level catalog records. An item-level search for *Hamlet* by William Shakespeare returns 575 volume scans; grouped by catalog records, there are exactly 500 manifestations (although many other copies are invisible to searchers because they are listed under a title like "Shakespeare's works," volume 12.)



□ Hamlet / by William Shakespeare.
by Shakespeare, William, 1564-1616.
Published 1892
Catalog Record Full view

□ Hamlet / edited by Horace Howard Furness. v.4
by Shakespeare, William, 1564-1616.
Published 1877
Catalog Record Full view

□ Hamlet / edited by Horace Howard Furness. v.3
by Shakespeare, William, 1564-1616.
Published 1877
Catalog Record Full view

Main Author:	Shakespeare, William, 1564-1616.
Language(s):	English
Published:	London : Isaac Pitman & Sons ; 1892.
Subjects:	Shorthand > Pitman > Texts.
Note:	Text in Pitman shorthand.
Physical Description:	96 p. ; 17 cm.
Locate a Print Version:	Find in a library

Viewability:

Full view	(original from Harvard University)
Full view	(original from University of Illinois at Urbana-Champaign)

First item-level results for Hamlet in HTDL (left). Example of manifestation-level catalog record, for 1892 edition printed by Isaac Pitman & Sons, with two items available. (right)

Newer cataloguing standards allow for description of works and expressions, either within an item record or through a separate access point (RDA, 2013 – Ch. 6). There is also earlier precedent for cataloguing relationships along multiple levels of FRBR entities, such as in the Perseus Digital Library (Mimno, Crane, and Jones 2005; Babeu, 2008). However, the tradition of controlling duplication only at the manifestation-level means that for many collections, scaling to a multi-level approach requires a significant manual undertaking. The feasibility of executing FRBR-inspired cataloguing for Perseus was attributed to its relatively small size and well-controlled scope (ibid): features that are uncommon for massive digital libraries like the Internet Archive and HathiTrust Digital Library.

FRBR also accounts for a class of ‘derivative’ relationships. While more faithful derivatives are considered new expressions – such as editions, translations, and revisions – there is an affordance for a relationships to be sufficiently derivative as to be considered new works (Tillett 2003). New, derivative works include adaptations (e.g. *Apocalypse Now*’s take on *Heart of Darkness*), parody imitations (e.g. *Pride and Prejudice and Zombies*), and works with the same style or thematic content. This class of new-but-related works bridges the continuum explored in this study, from duplication to similarity.

Preliminary work

In exploratory work, we approached this problem on a 101k volume subset of the HTDL, modeling lower-dimensional vector representations of each volume, and then applying distance metrics to cluster volumes with their partners (Organisciak et al 2017). Best-copy was determined by identifying the work closest to a centroid of apparent duplicates – an averaged point in vector space.

The preliminary work demonstrated robustness, while revealing key issues that warrant further study. The exploratory approach is not sensitive to deviations from one work per volume, such as multi-volume works and multi-work volumes (e.g. anthologies). In response, we will explore page-level fingerprinting methods. The approach for best copy selection can easily suffer from outlier texts, which we intend to improve and assess with carefully researched evaluation corpora. Finally, the existing methods simply list notably similar works – *candidates* for duplication – without knowing the relationship context. Learning how these measures align with conceptual boundaries and acting on them is a core concern for our forthcoming work.

Audience

The audiences for this project are in library collection services, and in curation-focused text mining scholarship. Domain expertise from both groups is represented among the key personnel on this project. Additionally, the workplan for this project incorporates early dissemination activities, to allow for early and repeated discussion within the library and information science community.

Workplan

The proposed research will be performed in a two-year timeline. The project management and implementation will be led by the Project Director, following the milestones outlined in the accompanying schedule of completion. Year 1 focuses on defining the problem and preparing our methods for a sound evaluation, including annotating corpora for evaluation, training transformations for the book data, and scaling our methods to HTDL sizes. In Year 2 we will complete, evaluate, and document our methods, and prepare maximally useful deliverables, including applying our best approach over the full collection, annotating the models, and generalizing the code assets for easier reuse. We organize the workplan by research question.

RQ1: Identifying Duplicates

The first research question seeks to accurately identify and disambiguate varying levels of duplication in digital libraries, and is a dependency for the next two questions.

The fundamental approach toward answering this question is through similarity measurement between texts. Identifying similarity between texts is an established area of study, with history in information retrieval and text mining. For example, early ranked retrieval search represented documents and queries quantified in Euclidean space, and documents were returned based on the shortest measured distance from the document to the query (Salton 1975).

A number of key issues make this project's focus more complex and novel than simply applying distance measures to quantified word counts. First is the scale. Toward a set of results that not only furthers library scholarship but also can be immediately useful in practice, we focus on one of the largest digital libraries that can be accessed for research. To be tractable at the scale of 16 million volumes, special consideration needs to be given to succinct, dense representation of books. Next is the variety of the source data: the data covers hundreds of languages and subject areas, meaning that the vocabulary of words in the full collection is extremely broad and heterogeneous. There is also the question of data quality: massive digital libraries have been able to grow through digitization and automatic extraction of text through Optical Character Recognition (OCR). In contrast to human transcription, which is cleaner but a labor-intensive bottleneck, OCR produces occasional transcription errors. Such errors are unevenly applied, affected by typographic choice, scan quality, and various aspects of the original text's physical condition.

The duplicate detection portion of this research first requires conceptual modeling of the problem. Developing a framework for discussing the issue and the bounds of the challenge, manually studying example cases, and fleshing out use cases will need to establish a foundation for the subsequent work. As noted, this work builds on FRBR's group 1 entities and will be informed by the current RDA guidelines on encoding works and expressions.

At the same time, an early pilot will be performed, scaling the methods preliminary work (Organisciak et al. 2017) to around 500k volumes. The purpose of this pilot is to highlight edge cases and outliers. Which known manifestation-level duplicates fail to match? When the results deviate from candidates fuzzy-matched by author and title, what is the reason for it? These types of questions will be explored manually alongside the conceptual modeling work, in order to catch unanticipated risks early.

After initial work, success for this project will hinge on developing effective representations of texts, efficient ways to match them, and training breaking points for differentiating different levels of duplication.

Text does not map to quantitative methods naturally, so the preprocessing, preparation, and weighting of the words is important. Some words are more salient than other (e.g. 'locomotive' tells you more about a text than 'and'), and synonyms and homophones complicate the relationship between a word and its meaning. Most seriously, texts are *sparse* when in a common space – there may be thousands or millions of words being counted, but for any particularly page or book, the counts for most of those words will be zero. Huge, low-information representations of books is a problem for computation performance. It also presents a quality problem, because with so many unseen words, the ones which are seen will over-bias the results. Finally, there is the question of which words to keep: because of scanning issues, there are over 75 million unique tokens in the HathiTrust, most of which are infrequent errors and not real words.

The anticipated approach to representing texts will be to 1) cull the term vocabulary to a reasonable size, accommodating multiple languages and still allowing for infrequent but highly representative words; 2) weigh the remaining vocabulary based on a term weighting heuristic, to formalize relative salience of words; 3) train a dimensionality reduction model to learn denser representations of books, ones that combine word counts into fewer, but still comparable dimensions. For example, rather than having four numbers representing how much a book mentions cat/cats/kitten/kittens, a single dimension might represent the amount a text uses animal-related words.

This workflow has multiple points for experimentation. Dimensionality reduction has been used in other contexts for similarity search (Keohn et al 2001; Kanth, Agrawal and Singh 1998) However, the most efficient body of methods, including latent semantic analysis, latent Dirichlet analysis, and auto-encoders, need to be trained. The scale and variability of our source data might make an adequate training intractable, which may require less effective but much more performant methods based on random projections from hashing (Broder 1997, Gionis et al 1999, Aggarwal 2001, Bingham and Mannila 2001, Schmidt 2017).

Once texts are adequately represented, a distance metric such as cosine distance can be employed to measure pairwise similarity between texts. This provides a measure for identifying duplicates. Figure 2, derived on work from Schmidt (2017), shows how a lower cosine distance can uncover duplicate works.

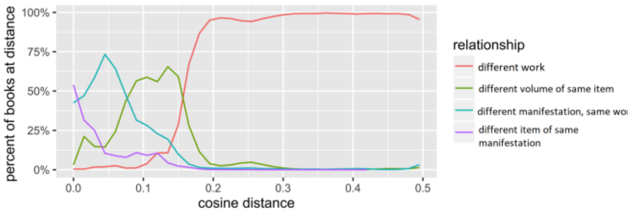


Figure 2: Comparison of Dickens items by distance.

It is expected that the distribution of similarity measures for a target book has a clear drop off. In asking, ‘is *book* Y a duplicate of *book X?’*, this project will seek to find that cut-off point. We will analyze these distributions across the collection, as well as considering the measured similarity for already known, manifestation-level similarity.

The question of specificity – what type of duplicate is *book Y*? – will be pursued after, in Year 2. It is a trickier question that may require page-level methods, a sensible continuation to the above work. We anticipate varying approaches to disambiguating expressions or manifestations. For example, different manifestations may be identifiable by encapsulation – even when paratext like the foreword and front matter is different, the content in the middle of the book is roughly identical, while different expressions may have more consistent deviations. Direct reprints will have content aligned on the page level, which can be measured through the page-level features in HTEF. Still, badly scanned or low quality books may complicate such approaches, and the approach will be informed by the conceptual modeling stage of the study. Additional edge cases will be considered through page-level work in Year 2, like multi-volume and sub-volume (i.e. anthologies) works.

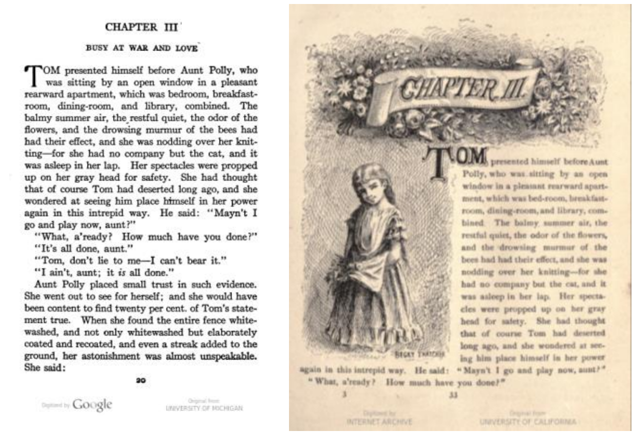
For evaluation data, a manual encoding of relationships for randomly-sampled target volumes will be collected. Evaluation on identified duplicates will be done using precision and recall. The same evaluation dataset will be used for evaluating later work trying to disambiguate levels of duplication.

Additional evaluation will be performed against two forms of metadata: manifestation-level control numbers, and author-title matches. Both types of information are incomplete themselves, but would be expected to group together under the larger umbrella of this study’s results. This evaluation cannot measure how well our methods identify Balzac’s alternately-named (in English) *La Comédie humaine* and *The Human Comedy*, but it should show whether all the translations under those names group together.

RQ2: Selecting a Best Copy

Texts derived from digitization are only as good as their scans, software, and the clarity of their materials. Additionally, for scholars applying text mining to learn about books at large scales, there is a hope for the true content of a book, un-besmirched by extraneous annotations or long forewords. Given a set of duplicate items of *Tom Sawyer*, for example, which one is the most wholly representative of the *expression*?

In preliminary work, we approached this question by taking a centroid of all known duplicate items – an imaginary ‘averaged copy’ – and finding the real item that most resembled it. This approach worked for the fiction-specific test set, with the best copies having the smallest number of unique words, as would be expected of a copy with fewer scanning errors and less commentary.



CHAPTER III

BUSY AT WAR AND LOVE

TOM presented himself before Aunt Polly, who was sitting by an open window in a pleasant rearward apartment, which was bedroom, breakfast-room, dining-room, and library, combined. The balmy summer air, the restful quiet, the odor of the flowers, and the drowsing murmur of the bees had had their effect, and she was nodding over her knitting—for she had no company but the cat, and it was asleep in her lap. Her spectacles were propped up on her gray head for safety. She had thought that of course Tom had deserted long ago, and she wondered at seeing him place himself in her power again in this intrepid way. He said: "Mayn't I go and play now, aunt?"

"What, a'ready? How much have you done?"

"It's all done, aunt."

"Tom, don't lie to me—I can't bear it."

"I ain't, aunt; it *is* all done."

Aunt Polly placed small trust in such evidence. She went out to see for herself; and she would have been content to find twenty per cent. of Tom's statement true. When she found the entire fence whitewashed, and not only whitewashed but elaborately coated and recoated, and even a streak added to the ground, her astonishment was almost unspeakable. She said:

20

CHAPTER III.

TOM presented himself before Aunt Polly, who was sitting by an open window in a pleasant rearward apartment, which was bed-room, breakfast-room, dining-room, and library, combined. The balmy summer air, the restful quiet, the odor of the flowers, and the drowsing murmur of the bees had had their effect, and she was nodding over her knitting—for she had no company but the cat, and it was asleep in her lap. Her spectacles were propped up on her gray head for safety. She had thought that of course Tom had deserted long ago, and she wondered at seeing him place himself in her power again in this intrepid way. He said: "Mayn't I go and play now, aunt?"

"What, a'ready? How much have you done?"

3 33

Matched pages from different manifestations *of Tom Sawyer*. Early work correctly identified the left copy as the cleaner one. Text from the right scan begins "3 33 presented himself before Aunt Polly, who was. sitting…"

However, issues can be anticipated in scaling this approach. First, it is susceptible to skew from outliers, like a single misclassified or particularly messy text. It also does not account for language drift between expressions.

In continuing this work, this project will look to identify the densest clusters of texts among duplicates, using only those for determining an archetypal expression. Other information, like total word count and measures of how many uncommon words there are (indicative of scanning errors) may also be used

Most importantly, continuing this work requires a more thorough evaluation. In this case, an evaluation dataset will be coded, evaluating randomized full-text pages on the quality of the scan.

RQ3: Measuring Similarity

> The final research question and associated digital product aims to identify books that are not the same work, but related through themes and style.

The approach here is a continuation of the measurement described for RQ1. The main difference is that 'similarity' and 'duplication' have different measurement needs. Most word types are important in determining if something is the same work. For similarity to be useful, we want to focus on words that describe topics and themes. For example, proper nouns need to be excluded, because seeing a character named 'Alice' in two books says extremely little of the similarity between the books. This means that a dual representation of each book will be necessary, one full and one thematic.

Consider one demonstrative example from early work. When determining works most similar to a volume of *Sense and Sensibility*, the top results are other Austen works, starting with *Emma* and *Mansfield Park*. The most similar non-Austen book is *Old Manor House* by Charlotte Smith. The recommendation is apt; writing on the influence of Charlotte Smith on Jane Austen, Magee has previous noted the 'verbal echoes' of *Old Manor House* in the language satirized by *Sense and Sensibility* (1975).

Similarity recommendations will be measured against other sources of book relationships: user-based item-to-item recommendations from the website Goodreads, and metadata-based subject and class information. This will capture overlap with known recommendations, but an expected feature of this dataset is that it may dig deeper into the collection to find new synergies. To gain a sense of the quality of newly uncovered recommendations, a small-scale 'odd one out' evaluation will be performed. In this evaluation, a student rater is provided a set of similar book candidates along with one distractor, and has to identify the outlier after a brief review of the books.

Dissemination and Deliverables

Results from this project will be in theoretical and methodological contributions, as well as deliverables that allow the project outcomes to be used immediately.

We aim to describe and publish our research findings in information science venues. Two types of findings will be presented at research conferences. The first will be the results of our year 1 conceptual work, developing a model for describing duplication in digital libraries and studying edge cases and outliers from pilot applications of our deduplication efforts. Additionally, work-in-progress results of the deduplication effort will be presented. Both conference targets are designed with a dual purpose, to solicit community feedback to the research throughout the grant performance period.

The final outcomes of the project will be described in an end of project whitepaper, disseminated openly through the University of Denver institutional repository. While we hope to repurpose the work for peer-review publication after the grant activities, there is a clear and immediate need for a whitepaper: to contextualize the deliverable datasets. The usefulness of the code and datasets that we release is contingent on the strengths and possible weaknesses of the research.

The products of the research will be documented and released with open licenses, as described in the attached Digital Product Plan. The product with the broadest usefulness in the library community is the open similarity dataset. For each work in the HTDL, the dataset will enumerate: topically similar authors and works; items (i.e. a scan of a volume) grouped by related realizations and manifestations, including multi-book and partial book identification; and ranked recommendations for the items most representative of an archetype of the work.

To maximize the reuse potential for the research, code and models are included in the digital product release. Though less useful for libraries with physical holdings, libraries with electronic copies of their materials may be able to plug those materials into the data that we will generate.

In summary, the project findings will be presented at information science conferences and described thoroughly in a summative whitepaper, while the equally valuable datasets and software will be released publicly, with documentation befitting a first-order deliverable.

Risks

The primary risks to the completion of the project stem from the scale and heterogeneity of the data. The HathiTrust Digital Library is large, unwieldy, and sure to produce unanticipated edge cases. We protect against these risks by formalizing a gradual scaling – in scope and complexity – through the workplan. The project also begins a period of conceptual modeling and exploration of edge cases, aiming to plan early for possible risks to completion.

Additional risks relate to the sensitivity of the methods. For example, how do differentiate between scanning errors and variation from versioning? How do we avoid information like running headers skewing measurements of similarity? Disentangling the signals that come from various parts of a book is a challenge, but one that this project has planned for. It builds on data that has already had some cleaning applied to it, and follows a pilot study that demonstrated feasibility.

The experience of the key personnel and advisors on this project was compiled in anticipation of this project's risks. Peter Organisciak and Benjamin Schmidt specialize in the text mining and representation activities performed here, and have worked with the HT Extracted Features Dataset. Krystyna Matusiak has long-time expertise in digitization and digital libraries, and will drive conceptual work with an eye to practice.

National Impact

The stated goal for this project is to broaden access and expand use of the Nation's content and collection. Upon successful completion, we will:

- provide methods for new collection access in digital libraries, by enriching the known item relationships with details of matching manifestations, realizations, and works,
- improve the usefulness of HathiTrust Digital Library content by enabling deduplicated use, with code to allow other digital content to hook into it, and
- distribute a set of inferred similarity relationships between 16 million volumes as well as their authors, deepening item-to-item access in a way that complements the strengths and weaknesses of existing approaches.

The research data used here is the content of the HathiTrust Digital Library, chosen because it is one of the largest and most generalization sources for within-book content. The HathiTrust is a collection of collections, and the results will be immediately applicable to libraries whose collections overlap with the HTDL.

As beneficiaries of this research, library professional will be able to apply it in various ways. Collection services may apply similarity data as supplemental evidence for collection development, identifying related books in a desired subject area. Retrieval may be aided by folding duplicate items by expressions, and user-facing catalogue records can be augmented by the similarity recommendations delivered by the project. The inferred similarity recommendations can also be used as an on-demand supplement to reader's advisory when expert advice is unavailable. Finally, there is a metadata enrichment use: libraries with digital content may re-apply the project's methods, while other libraries can use the inferred relationships for improved cataloguing. One metadata improvement use is for determining date of first publishing for books. Recent work finds that compiling a set of duplicates items and using their first known date serves as an effective heuristic to determining this date (Bamman et al 2017).

The second audience for this research is scholars and researchers. For them, this work will be a stepping stone toward large-scale text mining over massive digital libraries, overcoming the hurdle of duplication while also offering recommendations for the most useful single copies to use if culling duplicate works. The value here will particularly lie in studying the HathiTrust, which is extremely promising as an avenue to new historic and cultural insights, but is also intimidating and unwieldy in its scale and complexity.

To maximize reuse potential and ensure lasting value, the products of this research will be documented and released openly for immediate use, in addition to traditionally published findings. These products include datasets, models, and associated code.

Conclusion

We propose a content-based study of duplication, text variants, and similarity in digital libraries, applying content-based approaches to learn more about digital collections. This research overcomes a key obstacle in learning from massive digital libraries, biases from repeating text, using the 16 million volume HathiTrust corpus. It will also produce a dataset of stylistic and thematic similarities between books, which will be applicable as a complement to existing retrieval and access methods in all library collections.

Schedule of Completion

The Gantt chart on the subsequent page outlines the timeline for this project's activities.

Year 1 focuses on defining the problem and preparing our methods for a sound evaluation, including annotating corpora for evaluation, training transformations for the book data, and scaling our methods to HTDL sizes. In Year 2 we complete, evaluate, and document our methods, and prepare maximally useful deliverables, including applying our best approach over the full collection, annotating the models, and generalizing the code assets for easier reuse.

The schedule of completion is organized into the following sections:

1. Develop methods for identifying duplicates and similar works

   This portion of the schedule describes the activities related directly to the project's research questions. The first year activities are foundational, on which more detailed work builds in Year 2.

2. Conceptual modeling

   The modeling activities is Year 1 help establish the framework for research during the remainder of the project.

3. Evaluation of results

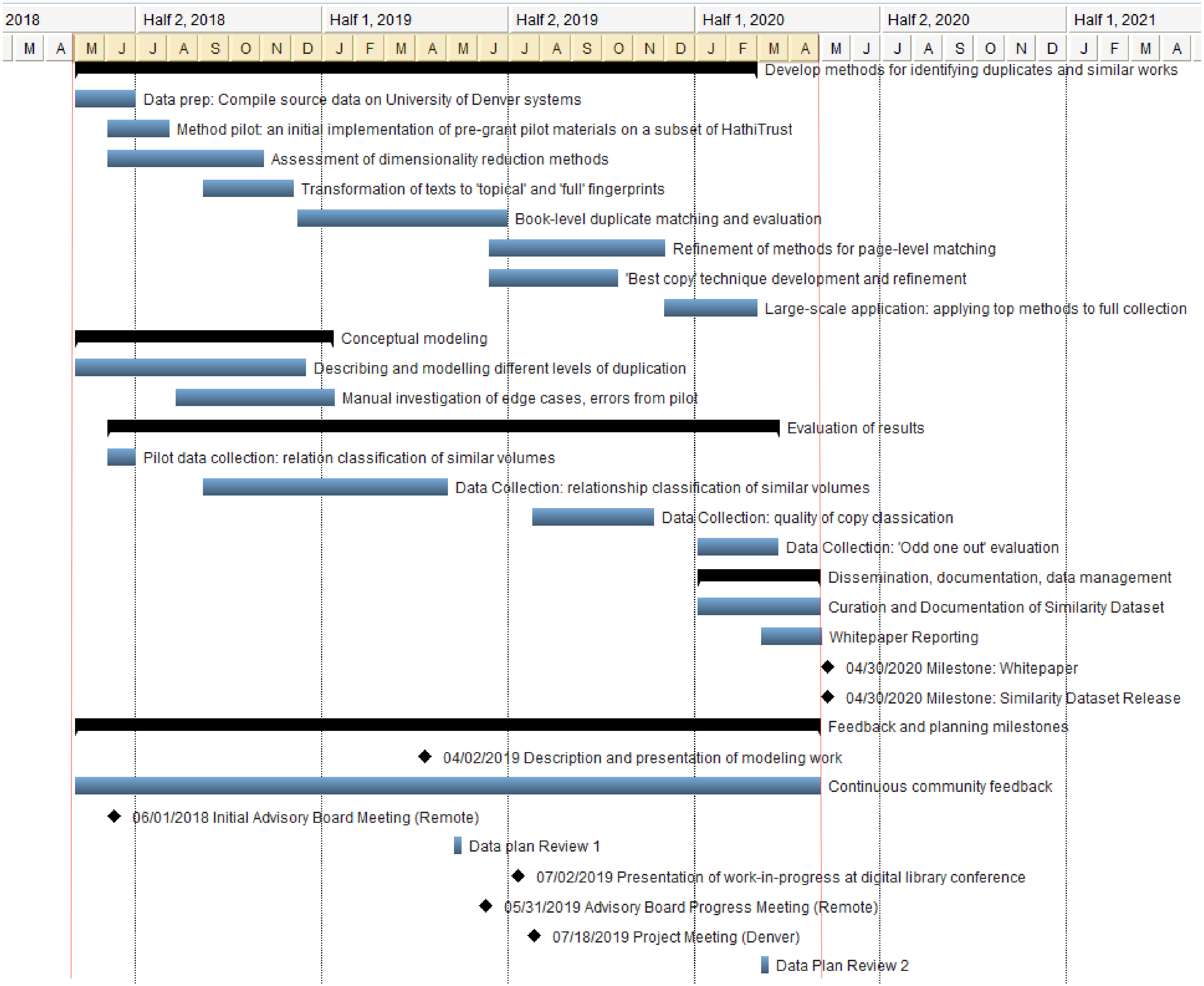
   This is the data collection schedule for evaluation data, which will be performed by student workers at the University of Denver.

4. Dissemination, documentation, data management

   These are the activities related to disseminating results, maximizing the reuse value of the research, and releasing digital products.

5. Feedback and planning milestones

   These are the necessary activities for planning and development, tracking work alongside the work plan, and soliciting feedback from advisors and community members.


2018
Half 2, 2018
Half 1, 2019
Half 2, 2019
Half 1, 2020
Half 2, 2020
Half 1, 2021
M A M J J A S O N D J F M A M J J A S O N D J F M A M J J A S O N D J F M A
Develop methods for identifying duplicates and similar works
Data prep: Compile source data on University of Denver systems
Method pilot: an initial implementation of pre-grant pilot materials on a subset of HathiTrust
Assessment of dimensionality reduction methods
Transformation of texts to 'topical' and 'full' fingerprints
Book-level duplicate matching and evaluation
Refinement of methods for page-level matching
'Best copy' technique development and refinement
Large-scale application: applying top methods to full collection
Conceptual modeling
Describing and modelling different levels of duplication
Manual investigation of edge cases, errors from pilot
Evaluation of results
Pilot data collection: relation classification of similar volumes
Data Collection: relationship classification of similar volumes
Data Collection: quality of copy classication
Data Collection: 'Odd one out' evaluation
Dissemination, documentation, data management
Curation and Documentation of Similarity Dataset
Whitepaper Reporting
04/30/2020 Milestone: Whitepaper
04/30/2020 Milestone: Similarity Dataset Release
Feedback and planning milestones
04/02/2019 Description and presentation of modeling work
Continuous community feedback
06/01/2018 Initial Advisory Board Meeting (Remote)
Data plan Review 1
07/02/2019 Presentation of work-in-progress at digital library conference
05/31/2019 Advisory Board Progress Meeting (Remote)
07/18/2019 Project Meeting (Denver)
Data Plan Review 2

DIGITAL PRODUCT FORM

Introduction

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (i.e., digital content, resources, assets, software, and datasets). The products you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. However, applying these principles to the development and management of digital products can be challenging. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions

□ Please check here if you have reviewed Parts I, II, III, and IV below and you have determined that your proposal does NOT involve the creation of digital products (i.e., digital content, resources, assets, software, or datasets). You must still submit this Digital Product Form with your proposal even if you check this box, because this Digital Product Form is a Required Document.

If you ARE creating digital products, you must provide answers to the questions in Part I. In addition, you must also complete at least one of the subsequent sections. If you intend to create or collect digital content, resources, or assets, complete Part II. If you intend to develop software, complete Part III. If you intend to create a dataset, complete Part IV.

Part I: Intellectual Property Rights and Permissions

A.1 What will be the intellectual property status of the digital products (content, resources, assets, software, or datasets) you intend to create? Who will hold the copyright(s)? How will you explain property rights and permissions to potential users (for example, by assigning a non-restrictive license such as BSD, GNU, MIT, or Creative Commons to the product)? Explain and justify your licensing selections.

The intellectual property of the software will be owned by the University of Denver and released under a permissive MIT license. The data generated will be distributed with a Creative Commons Attribution License, which allows *any* use of the data as long as it is attributed. Assigning an especially permissive license will reduce ambiguity for users and will work to maximize the usefulness of the data.

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

The organization will maintain ownership of the digital products but will not assert conditions on access and use.

A.3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

We do not anticipate privacy concerns with any products.

Part II: Projects Creating or Collecting Digital Content, Resources, or Assets

A. Creating or Collecting New Digital Content, Resources, or Assets

A.1 Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and format you will use.

See *Part IV: Projects Creating Datasets*.

A.2 List the equipment, software, and supplies that you will use to create the content, resources, or assets, or the name of the service provider that will perform the work.

n/a

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to use, along with the relevant information about the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

n/a

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

n/a

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

n/a

C. Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata. Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

n/a

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

n/a

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

n/a

D. Access and Use

D.1 Describe how you will make the digital content, resources, or assets available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

D.2 Provide the name(s) and URL(s) (Uniform Resource Locator) for any examples of previous digital content, resources, or assets your organization has created.

n/a

Part III. Projects Developing Software

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) it will serve.

The created software will be the code assets associated with this project's research activities: for transforming books from the HTRC Extracted Features Dataset to a common geometric space, classifying levels of duplication, and encoding similarity to other volumes and authors.

A.2 List other existing software that wholly or partially performs the same functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

None exactly. General purpose fingerprinting approaches exists, such as *datasketch* (https://github.com/ekzhu/datasketch). Our software will reflect the specific considerations required for analysis of digital library collections.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software and explain why you chose them.

Project code will be written in Python, drawing on the SciPy Stack for data analysis (e.g. Numpy, SciPy, Matplotlib, Pandas, Scikit-Learn) and the HTRC Extracted Features Reader. The SciPy stack is a robust set of tools that is well adopted in LIS communities, easily interpretable, and includes a robust tool for documenting code called Jupyter.

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

See B1.

B.3 Describe any underlying additional software or system dependencies necessary to run the software you intend to create.

The resulting code should be fully useable with the Anaconda distribution, which bundles Python with a set of tools, including the SciPy Stack. Given that Anaconda can be run on Window, Linux, and Mac OS, our software will be able to also.

B.4 Describe the processes you will use for development, documentation, and for maintaining and updating documentation for users of the software.

Code will be documented inline during development. During the digital product assessments in the middle and at the end of the grant, tutorial-style documentation will be written.

B.5 Provide the name(s) and URL(s) for examples of any previous software your organization has created.

- HT+Bookworm: https://github.com/Bookworm-project
- HTRC Feature Reader - https://pypi.python.org/pypi/htrc-feature-reader

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under open-source licenses to maximize access and promote reuse. What ownership rights will your organization assert over the software you intend to create, and what conditions will you impose on its access and use? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain and justify any prohibitive terms or conditions of use or access and detail how you will notify potential users about relevant terms and conditions.

The University of Denver will maintain ownership of the software, with no restrictions on its use. The source code will use the MIT License.

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

Software will be available as source code on Github. We don't anticipate its use as a library; however, if community feedback suggests otherwise, it will be also be packaged to the PyPi and Anaconda code repositories.

C.3 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository: Github

URL: https://github.com/organisciak

Part IV: Projects Creating Datasets

A.1 Identify the type of data you plan to collect or generate, and the purpose or intended use to which you expect it to be put. Describe the method(s) you will use and the approximate dates or intervals at which you will collect or generate it.

1. Book similarity dataset and author similarity dataset.

   These datasets will be products derived from the methods at the completion of the project. They will be distributed in a compressed JSON format. The book similarity dataset will describe each of 16 million volumes through brief metadata – author, title, year, OCLC number, HathiTrust number, ISBN, classification number – then list all recurrences of the work in other volumes and forms, as well as listing the most similar non-duplicate works. The author dataset will be provided at the same at an author level.

2. Similarity models

   Part of this project will be determining the ideal manner to represent books computationally with the same vocabulary and dimensionality. While the form of this data cannot be entirely anticipated until the research is completed, we expect the models to be matrices, distributed in a CSV format and comprising under 1 GB of data. This data will be generate over the full course of the grant.

3. Training and evaluation data

   The project will generate two types of evaluation data. The first will be an encoding of the relationships amongst most similar books, for a sample of target books. The second will be an assessment of the 'cleanest' scanned copies for a sample of duplicate book clusters. Both will be small datasets, distributed alongside the software described below. This data will be collected from the middle of Y1 into Y2, and should remain static once collection.

A.2 Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

IRB is not applicable to this data.

A.3 Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

No.

A.4 If you will collect additional documentation, such as consent agreements, along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

No.

A.5 What methods will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

The data will be generated using the code described in Part II and methods described in the Narrative.
There will be no technical requirements for understanding or processing the dataset – packaged using standard formats with associated documentation, a person will be able to read them as they see fit. To use the models, the associated software will be useful.

Formats

- JSON – A project specific schema will be developed and documented alongside the data release
- CSV – tabular evaluation data will be release in basic comma separated form
- MM – models will be released in the Matrix Market format, the NIST standard for disseminating matrices

A.6 What documentation (e.g., data documentation, codebooks) will you capture or create along with the dataset(s)? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

Data documentation will be maintained throughout the course of the grant, with two scheduled assessments against this plan. It will be distributed packaged with the datasets. The data is not expected to change after the project lifecycle, nor is its associated documentation.

A.7 What is your plan for archiving, managing, and disseminating data after the completion of the award-funded project?

Digital products will be released on institutional services. Similarity datasets and models will be included on the University of Denver institutional repository, and evaluation data will be bundled with the software on a hosted code repository (Github). The project's outputs will produce derived, descriptive metadata. We will align our dataset structure with the RDA recommendations for describing works and expressions in MARC; however, a more succinct hierarchically structure will be used in JSON.

Data will be posted freely, with an open access license, for download. It will be platform or language agnostic.

Examples of past datasets from authors:

- IDF Dataset: https://www.ideals.illinois.edu/handle/2142/89691
- EF Dataset: http://dx.doi.org/10.13012/J8X63JT3

A.8 Identify where you will deposit the dataset(s):

Name of repository: Digital Commons at DU

URL: https://digitalcommons.du.edu/

A.9 When and how frequently will you review this data management plan? How will the implementation be monitored?

This data management plan will be reviewed at the start of year 2, as part of the report to the advisory board, and near the conclusion of the project. These activities are included in the project plan and are tied to a successful completion of this project.