Building large-scale collections of genre fiction
HTRC ACS Final Report

Laure Thompson and David Mimno

Summary: We collected a set of more than 3000 distinct volume-length works of speculative fiction based on title/author queries. This process was largely automated, but required considerable manual oversight. A large number of small and specific issues were not amenable to automation, and we do not expect this situation to improve in the near future. The resulting collection was sufficient to support unsupervised machine learning methods, offering a new approach to a collection that has been difficult to study at scale due to copyright. While we were able to bring automated tools to bear on late 20th century material, we did find that our ability to cover works declined as dates of publication approach the present. We also identified a large number of highly influential works that are not available in HathiTrust. With these caveats, we believe that HathiTrust can be a powerful tool for studying in-copyright genre fiction.

Introduction

The HathiTrust Digital Library presents an unprecedented opportunity to build and access collections of contemporary literature. But not every published work is currently in the collection. This problem could be even more troubling if works are not "missing at random": are certain genres or subject areas less likely to appear in the collection? The goal of this project is to determine what is and is not in HathiTrust from the perspective of a specific genre, speculative fiction. In this project we addressed two main issues. First, does our target genre exist in HathiTrust? The academic libraries that make up the HathiTrust source libraries collect speculative fiction. But, like all popular fiction, it has not been a primary focus of academic library holdings and even less of a focus for digitization. Second, if it does exist, can we find it? We prefer to build a collection at the level of works, not specific editions. As a result, our queries are in terms of titles and authors, with no additional information. So what happens when we combine noisy, limited information with noisy, limited catalogs?

Part I: How many speculative fiction works can we find using the HathiTrust catalogue?

We began this project by focusing on testing how many speculative fiction works we could find through the HathiTrust catalogue. Within this initial process, we first built a manually curated list of works represented by title-author pairs. Then, with the support of HTRC, we queried the HathiTrust catalogue database for volume-level matches for each title-author pair. Then, with these results, we analyze what we found and what we didn't.

To build a curated list of speculative fiction works, we use Worlds Without End (WWE), an extensive fan-built database of speculative fiction. We select all works published from 1900 to 2010. We determine publication dates based on WWE. If WWE lists multiple publication dates, we use the earliest for our selection criteria. This gives us a list of 18,809 works written by 3,718 authors (counting collaborations as distinct authors).

Note that WWE includes works of all lengths and organizations. We choose the types that are likely to be published as independent works: novels, novellas, novelettes, collections, anthologies, and omnibuses. We exclude the additional WWE types of short story, non-fiction, graphic novel, light novel, and miscellaneous. Restricting by book type makes sense because not all fictional works are published as stand-alone works. As a result, we do not expect to find matches for all works within our curated list (i.e. novella and novelette).

Book Type	**# Works**
Novel	13585
Collection	1713
Anthology	1428
Novelette	1003
Novella	654
Omnibus	426

In addition to book type, WWE provides rich work-level metadata including genre, sub-genre, awards nominations and wins, inclusion in notable book lists (e.g. David Pringle's Science Fiction: The 100 Best Novels and NPR's top 100 science fiction and fantasy survey), and series information. The genre tags are a combination of the genres Science-Fiction, Fantasy, and Horror. Most works are labeled as a single genre, but many works (especially collections and anthologies) are a combination of the three. WWE contains many more Science-Fiction and Fantasy works than Horror.

Genre	**# Works**
Science-Fiction	15033
Fantasy	10897
Horror	2496
Science-Fiction / Fantasy	808
Fantasy / Horror	565
Science-Fiction / Fantasy / Horror	115
Science-Fiction / Horror	97

With our large working set of author-title pairs, we can query the HathiTrust catalogue to see what we can find automatically. With the help of HTRC, we constructed a query to search for the speculative fiction works of interest by matching author and title information. Given the large variability of the author field within the catalogue, we had to modify our initial script to better account for middle names and initials, as well as accented characters (and their variable representations). Note that our process does not directly account for pseudonyms. We find that catalogue records vary on the inclusion of pen names versus legal names. Additionally, our method does not directly account for alternative titles. This will cause us to miss a number of volumes, but this is a case for content-based analysis which we will return to later.

Ultimately, the script produced 4,920 work-volume pairs. After hand-checking these matches, we found that 4,382 are good matches and 538 are mismatches. Of the mismatches, 245 volumes had a corresponding good match, while the other 293 do not match any work within our working set.

We find that the mismatches for which a true match exists (and is found), that there are two general types: (1) distinct works with similar titles and (2) overlapping content for compilations versus contained works. There are two reasons for the first type. In the first case, works in the same series can have highly overlapping titles leading to mismatches. For example, within Frank Herbert's *Dune* series includes novels *Dune*, *Dune Messiah*, and *Chapterhouse: Dune*. Novels by the co-authors Brian Herbert, Frank Herbert's son, and Kevin J. Anderson which are also set within the *Dune* universe also produce mismatches because of similar titles such as *Dune: House Atreides* and *Paul of Dune*. In the second case, it is possible for the same author and title to refer to multiple different works. Typically, these instances involve a work of short fiction that is later expanded into a novel, such as was the case for *Ender's Game*. **Our overall result is that while automated title/author matches are very good, they are not so good that they do not require manual checks.**

In the case of "partial" matches, where the computer-found volume contains or is contained within the matched title-author pair, there is also large title overlap. This can arise for a few reasons. The title of a short work might be directly used for the compilation work (e.g. *Identity Theft* versus *Identity Theft: And Other Stories*) or be very similar for descriptive or series-related reasons (e.g. *Jungle Book* versus *Jungle Books*).

Speculative fiction, and genre fiction in general, is often characterized by *series* of novels, and these lead to a critical case where volumes may be misidentified or skipped entirely. In such cases, volumes are catalogued under the series title (e.g. *Lord of the Rings*) rather than their volume- /work- level title (e.g. *The Two Towers*). This means that unless our working set includes the series title as a possible work, we could miss all of these volumes despite their being within HathiTrust. To highlight why this is critical we will discuss the search results for J. R. R. Tolkien's *Lord of the Rings* trilogy. The overwhelming majority of these novels are cataloged under the series title rather than their original titles. This observation may be due to the fact that this series is known for being reprinted as three-volume sets. All the same, we only find two copies of *The Fellowship of the Ring* automatically and no copies for the other two novels in the trilogy. However, the series-level search identifies nine additional copies of *The Fellowship of*

the Ring, eleven copies of *The Two Towers,* and eight copies of *The Return of the King*. **This points to the value of performing series-level searches as well.** One unfortunate problem with these serial records is that not all volume-level records come with their series number. This makes it impossible to match these volumes automatically and instead a content-based approach must be used.

For mismatches for which a true match does not exist, we find many works that are related to the ones of interest. This highlights the diversity of HathiTrust's contents. For example, the automatic search finds non-English volumes of works of interest whether it be the non-English original or a translation of an English work. There are also a number of adaptations including screenplays, operas, stage productions, and graphic novels. Many of these "bad" matches can be automatically eliminated by examining additional metadata or computational comparison of the volume contents. Note that some of these automatic (mis)matches could be of interest since they represent compilational works that might not be included in more novel-centric bibliographies. In our case, WWE tends to be more focused on novels than the various compilation reprints.

In addition to this automatic search, we conducted a separate, non-exhaustive manual search using the HathiTrust catalogue search interface. This manual search identified 3,155 work-volume matches of which 786 were missed by the automated search. These missed matches stem from variation in catalogue records. Some of these misses were caused by author and name variation, such as "and" versus "&", "Eight" vs "8", and "Color" vs "Colour". Others were caused by the inclusion or exclusion of subtitles such as *The dispossessed* versus *The Dispossessed: An Ambiguous Utopia*. In a few cases the variation was the fault of our initial list since WWE work title fields can contain multiple titles (e.g. "Daybreak - 2250 A.D. (Star Man's Son, 2250 A.D.)") as well as other metadata (e.g. "Driftglass (collection)").

Unsurprisingly, the manual search was able to successfully find volumes within alternate titles (e.g. Alfred Bester's *The Stars My Destination* and *Tiger! Tiger!*) as well as alternate author names (e.g. Alice Mary Norton instead of Andre Norton). As touched on earlier, many volumed series records were identified by hand. In addition to J. R. R. Tolkien's *Lord of the Rings* series, this also affects many series of anthologies. These anthology series can be ordered by volume number or by year, but this information is variably documented within catalogue records.

Part II: Analyzing the matched HathiTrust volumes

As expected, we find mostly novels: 2,231 novels, 367 collections, 130 anthologies, 24 omnibuses, 21 novellas, and 2 novelettes. Despite the uneven breakdown of genres for our WWE working list, our matched works have similar proportions for all genres. We find 15% of Science-Fiction works (1507 works / 2218 volumes), 14% of Fantasy works (901 works / 1623 volumes), and 17% of Horror works (390 works / 261 volumes).

Examining the breakdown of matches across decades, we find—perhaps unsurprisingly—that the works found within HathiTrust provide proportionately more cover for early works. However, these proportions are misleading since there is disproportionately more coverage for

more recent decades. In term of raw numbers, the top three decades with the most matched works are the 1980s, the 1970s, and the 2000s* (including 2010).

Decade	Total Works	# Matched Works	% Matched Works	# Matched Volumes
1900s	77	59	77%	371
1910s	71	51	72%	239
1920s	85	42	49%	128
1930s	110	47	43%	111
1940s	153	76	50%	175
1950s	559	226	40%	430
1960s	1095	317	29%	530
1970s	2025	493	24%	687
1980s	3296	516	16%	647
1990s	4569	464	10%	531
2000s*	6769	484	7%	533

We find that the out of copyright volumes have much higher average representation than the other decades. The top ten most represented works, with between 14 and 31 copies, are all out of copyright with the most recent being Franz Kafka's *The Trial*.

Title	Author	Year	# Matched Volumes
Just So Stories	Rudyard Kipling	1902	31
The Wind in the Willows	Kenneth Grahame	1908	25
Zuleika Dobson	Max Beerbohm	1911	25
Before Adam	Jack London	1906	21
Puck of Pook's Hill	Rudyard Kipling	1906	21
The Iron Heel	Jack London	1908	18
Penguin Island	Anatole France	1908	17
The Wizard of Oz	L. Frank Baum	1900	16

The Trial	Franz Kafka	1925	15
The Inheritors	Joseph Conrad & Ford Madox Ford	1901	14

The most-represented authors we found in HathiTrust tend to be fairly prolific, but some of the most prolific authors in our WWE working list have little to no representation. We found none of the 94 novels written by James Axler, the shared pen name for authors writing the *Deathlands* series, only seven of 82 works by Tanith Lee, only eleven of 73 works by C. J. Cherryh, and only five of the 55 works by Mercedes Lackey. While these authors are active in the later decades (i.e. 1970s–2010s) the volumes found within HathiTrust are spread across these decades.

Author	**# Works (% Found)**	**# Volumes**
Robert A. Heinlein	37 (76%)	75
Michael Moorcock	34 (36%)	42
Robert Silverberg	34 (25%)	40
Andre Norton	33 (29%)	36
John Brunner	31 (48%)	48
Isaac Asimov	30 (45%)	48
Stephen King	30 (45%)	42
Philip K. Dick	28 (57%)	44
Philip José Farmer	28 (43%)	38
Roger Zelazny	25 (43%)	40

For the book lists in WWE, we have the best coverage for Locus Best SF Novels of All-Time (90%), Easton Press Masterpieces of Science Fiction (82%), David Pringle's Science Fiction: The 100 Best Novels (77%), The Classics of Science Fiction (77%). For awards lists, the Hugo and Nebula award winners have the best coverage at 64% and 62% respectively. While the very recent awards (e.g. Shirley Jackson, Red Tentacle, David Gemmel) and Australian-specific awards (Aurealis) have very poor coverage, with less than 10% of winners and nominees found. Surprisingly, we have very little by Connie Willis (2 works and 2 volumes) despite being the most decorated science fiction writer (11 hugos, 7 nebulas).

Other surprising absences we appeared to find in HathiTrust include that there are no copies of Octavia E. Butler's *Parable of the Sower* or its sequel nor any works by S. P. Somtow. We could also find any copies of these highly listed novels: *Hyperion* by Dan Simmons (1989), *China*

Mountain Zhang by Maureen F. McHugh (1992), *Cyteen* by C. J. Cherryh (1988), *Ammonite* by Nicola Griffith (1993), *I am Legend* by Richard Matheson (1954), and *The Day of the Triffids* by John Wyndham (1951).

Part III: Content-based approaches using HathiTrust Extracted Features

We convert extracted-features JSON files to something more like text using a Python script. The fact that features are listed by part of speech makes a simple stopword filtering process simple: we ignore any token with POS in the set {"DT", "PRP", "PRP$", "IN", "CC", "MD", "CD", "WRB", "WDT"}. This removes determiners, pronouns, prepositions, numbers, and other words that are less useful for semantic analysis. Removing these words from the start has the additional benefit of reducing dataset size, as they tend to account for about half of the running tokens in a corpus.

Deduplication methods are difficult based on only metadata, but using extracted features it can be made relatively easy and reliable. We use random projection (Johnson and Lindenstrauss) to create a *signature* for each volume. For each distinct word in the vocabulary we generate a random 100-dimensional Gaussian vector. For each instance of a word in a volume we add the appropriate vector, and then normalize the resulting 100-dimensional vector to have length 1.0. The inner product between normalized vectors is equivalent to cosine similarity. Empirically, volumes whose signature vector similarity is greater than 0.97 are very likely to contain the same work. To provide intuition for these results, here are the 10 most similar volumes to a copy of *The Chessmen of Mars* by Burroughs. As expected, the volume is closest to itself (cosine 1.0), but the second most similar (0.988) is also a copy of the same work. After that, the cosine similarity drops to 0.8, for a different work by the same author.

Cosine	HT ID	Title / Author	Year
1	nyp.33433112045251	The chessmen of Mars / by Edgar Rice Burroughs ... ; illustrated by J. Allen St. John.	1922
0.988	osu.32435017883182	The chessmen of Mars / by Edgar Rice Burroughs ... ; illustrated by J. Allen St. John.	1922
0.802	osu.32435017174004	Thuvia, maid of Mars / by Edgar Rice Burroughs, illustrated by J. Allen St. John.	1920
0.798	uc1.32106006727876	Ware hawk / Andre Norton.	1983
0.797	mdp.39015053531805	Pirates of Venus / by Edgar Rice Burroughs ; introduction to the Bison Books ed. by F. Paul Wilson ; afterword by Phillip R. Burger ; glossary by Scott Tracy Griffin ; frontispiece by J. Allen St. John ; illustrations by Thomas Floyd.	2001
0.793	osu.32435018600130	The gods of Mars / by Edgar Rice Burroughs ; frontispiece by Frank E. Schoonover.	1918

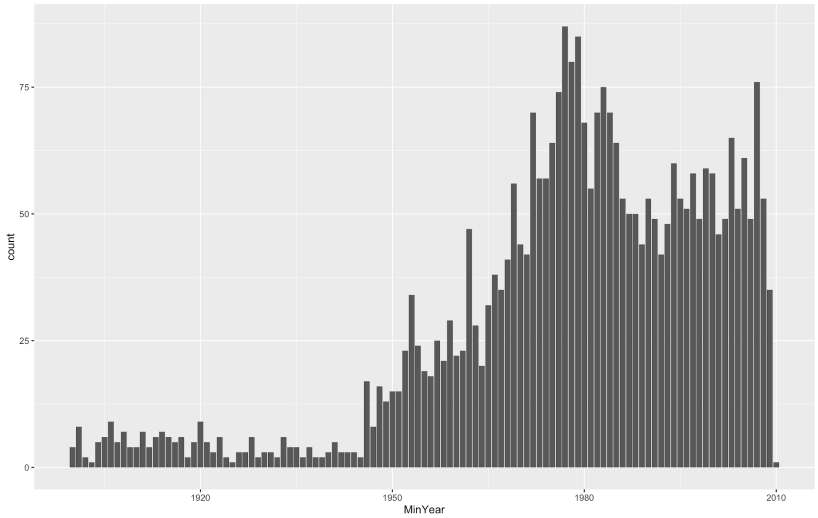
0.792	mdp.39015054164291	The moon maid / Edgard Rice Burroughs ; introduction to the Bison Books edition by Terry Bisson ; afterword by Richard J. Golsan ; Red Blood vs. the red flag by Phillip R. Burger ; illustrations by Thomas Floyd.	2002
0.79	msu.31293103176974	Ayesha, the return of She; by H. Rider Haggard.	1905
0.786	mdp.39015066062350	A haunted house, and other short stories.	1943
0.786	pst.000014058820	Ayesha : the return of She / Illustrated by Hookway Cowles.	1972

The lowest scoring volumes by similarity to this query volume include several copies of Borjes' *Ficciones* and *Figures of Earth; A Comedy of Appearances / James Branch Cabell*, which appears to be a 17-page excerpt from the beginning of a volume.

As a counter example, we identified four volumes with the exact same title (*The Hugo winners, edited by Isaac Asimov*.) that do *not* match this similarity criterion. These volumes represent a series containing Hugo-winning short fiction for each year. While they are identical from the perspective of metadata, they are not with respect to contents. As an example, the 10 most similar volumes for one of these four Hugo Winner volumes are listed below. All are short story anthologies, which may contain all or some of the same stories.

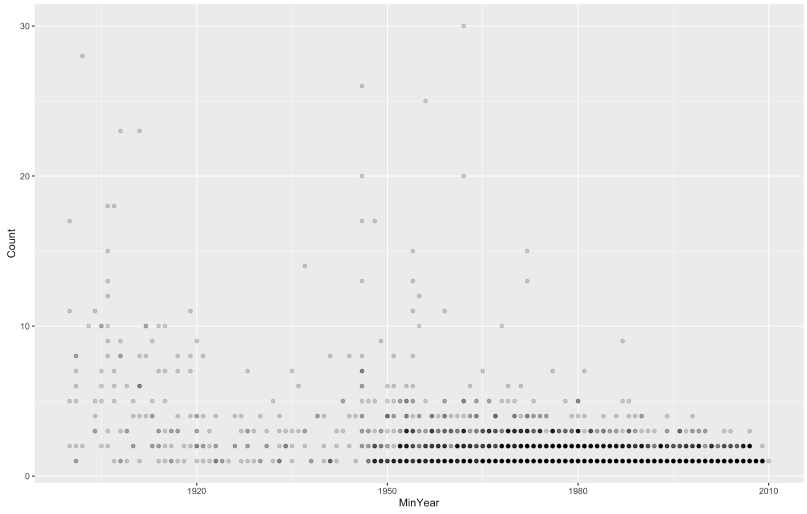
Cosine	HT ID	Title / Author	Year
1	mdp.39015013315810	The Hugo winners, edited by Isaac Asimov.	9999
0.963	pst.000012384754	The Science fiction hall of fame.	9999
0.962	mdp.39015000656127	Dangerous visions; 33 original stories. Illus. by Leo and Diane Dillon.	1967
0.96	inu.30000011372871	Dangerous visions : 33 original stories / edited by Harlan Ellison ; illustrations by Leo and Diane Dillon.	1967
0.96	pst.000012691142	Dangerous visions : 33 original stories / edited by Harlan Ellison ; illus. by Leo and Diane Dillon.	1972
0.958	pst.000012769308	The Science fiction hall of fame.	9999
0.958	mdp.49015002156918	The year's best science fiction.	1991
0.956	mdp.39015008543244	World's best science fiction.	1969
0.955	uc1.32106007301978	The Nebula awards	1999
0.955	uc1.$b348189	The best of science fiction, edited with an introduction by Groff Conklin. Preface by John W. Campbell, jr.	1946

The same works (Borjes and Cabell) appear as the most dissimilar volumes to this volume as well.

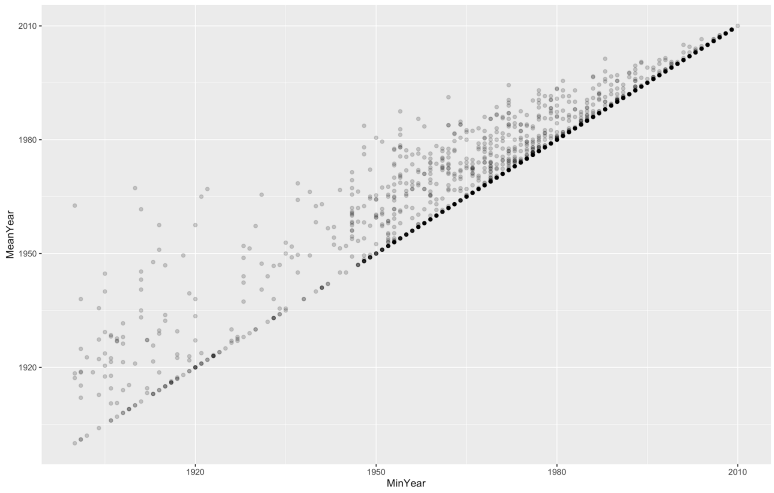


The cosine-based de-duplication process reduced the number of volumes from 5135 to 3182. The first figure shows the distribution of the minimum year for de-duplicated volumes. Listed publication years are, as is well known, highly unreliable as an indicator of the original publication date of a work. Duplication, however, can be useful in estimating a more reliable date. For example, we map 10 volumes to the title *Ayesha* by H. Rider Haggard. Six of these list the correct date, 1905, but the others list 1949, 1965, 1972, and 1977. Taking the minimum of these dates is usually a reliable way to find the correct date. There are exceptions, including *The Hitchhiker's Guide to the Galaxy* (1979), which maps to three volumes from 1989, 1996, and 2005. Most works have a minimum date after 1950, with a peak in the late 1970s.

Earlier works tend to have more copies. The next figure shows one point for each de-duplicated work. The *x*-axis shows the same value as the previous figure, the earliest year within a cluster of similar volumes. The *y*-axis shows the size of the cluster. Although there are relatively few distinct works from before 1950, the majority of the works that are well-represented in the collection as multiple copies appear in the earlier half of the time period. There is a noticeable drop in the number of copies per work around 1920, presumably because of libraries that did not digitize in-copyright works. For works with a minimum date after 1980 almost all works are attested in only one or two copies. The most common works are anthologies of stories by Harlan Ellison (1962) and *The Jungle Book* by Kipling.



This distribution is somewhat troubling, as the core of the genre focus is expected to be mid- to late-20th century. Yet the volumes most likely to be retrieved are from the early 20th century. Certainly some of these works are squarely Science Fiction, such as *The Scarlet Plague* (1912) by Jack London, which contains a character in 2072 describing his early life before the apocalyptic plague of 2013. But others, such as *The Jungle Book* or *Peter Pan* are more marginal. **While we find that 20th century genre fiction does exist and is findable, the collection shows strong bias towards out-of-copyright works.**



The next figure also shows the minimum year on the *x*-axis, and the mean of all attested years on the *y*-axis. There is a distinct "line" along the diagonal, indicating works whose minimum and average years are the same. For works with a minimum year before 1920, there is a large variation in the value of the *pubDate* field. There is a "dead zone" from 1920 to 1950. After that point there is considerable reprinting, but rarely more than about 20 years. In the most recent time period there is relatively little difference, but this is likely to be due to the reduced quantity of duplicate volumes.

We are currently preparing a public interface for a topic model trained on the resulting de-duplicated volumes. We used the *Authorless Topic Models*[1] pre-processing method that we previously developed for this application. This process selectively removes words that show high correlation with specific sections of a collection (in this case volumes). These are often the names of characters and settings, which would otherwise dominate any cluster-based analysis. For example, from *Peter Pan* the most frequently removed words are *Wendy, Hook, Nana, Tink, Tootles,* and *Smee*. The more common name *Peter*, though presumably frequent in the work, is downsampled less frequently because it is less surprising relative to other works. The resulting topics are sometimes still closely related to prolific authors (one for Stephen King, another for E.E. "Doc" Smith's *Lensman* series) or words with *fl* and *fi* ligatures, but most topics relate to cross-cutting themes such as religion, ships and water, eating food outdoors, understanding languages, driving cars, controlling spaceships, writing letters, and planets. Interestingly, our two Connie Willis books (*Passage* and *Doomsday Book*) both score high in a topic about telephone calls: indeed, many of her works have as a central plot device the need to send and receive messages. A topic about cigarettes and tobacco links Stephen King, Michael Chabon, Thomas Pynchon, and Haruki Murakami.

[1] See https://www.cs.cornell.edu/~laurejt/papers/authorless-tms-2018.pdf & https://github.com/laurejt/authorless-tms

Part IV: A less-successful attempt at another genre

Due to our initial success with speculative fiction, we were interested to try an additional category of genre fiction. We chose *Romance* due to its combination of massive and enduring popularity, and complicated dynamics of prestige. We were encouraged by the fact that the speculative fiction collection includes a number of works that cross over into Romance, such as Gabaldon's *Outlander*. One issue with Romance we ran into was that its fan-based bibliographies tended to run extremely new: predominantly from the 2000s or later. While this was somewhat true for WWE, it was far less true, especially for award winning novels and authors. In the end we constructed our list from novels tagged as "Romance" by users within LibraryThing.

We were not able to make significant progress in handling Romance fiction. Although we were able to find a set of title/author pairs, and HTRC was able to return a list of matching IDs, results were less successful than for speculative fiction. Many of these were similar to problems experienced in speculative fiction, but to a greater degree. As with speculative fiction, we were most successful in finding popular or prestige literature, such as Austen's *Pride and Prejudice*, and literature on the boundaries of the genre (our source lists *Harry Potter* due to its themes of romantic relationships). Another problem was publication styles, especially in the earlier half of the 20th century. The important publisher Harlequin, for example, shifted from reprinting other works to publishing in its own right during this time period, which made metadata based on publisher information less reliable. It is worth noting we only found a handful of novels published by Harlequin within the HathiTrust catalog.

Part V: Things we learned about the catalog

In addition to the above process, we also invested time into better understanding how volumes are added and organized into HathiTrust. The following is an overview of the Zephir system for work identification:

- HathiTrust Record Number is cluster id
- Records are clustered by the following elements:
    - OCLC numbers
    - existing cluster id = bibliographic id # (assigned by ILS)
    - previous ILS number
- Records are **not** clustered by titles, authors, or ISBNs
- Each cluster is represented by a single catalog record, whichever scores highest
    - This is the only record that is sent to HathiTrust

We were also able to submit corrections for a number of small errors that we encountered during the process of searching for works. Through the process of checking records, we updated a handful of records as we came across them. A particularly exciting example was updating the catalog record for Terry Brooks's *Antrax*. It was mistitled as "Anthrax".

- https://catalog.hathitrust.org/Record/004204708

    - Issue: The record 004204708 has the wrong title; it's listed as "Anthrax" but should be "Antrax".
    - Status: This has been corrected!
- https://catalog.hathitrust.org/Record/001025869
    - Issue: Volume mdp.39015005607919 should be marked as "pt.2". It is not the full trilogy but rather the second volume The Two Towers.
    - Status: Forwarded to Bibliographic corrections
- https://catalog.hathitrust.org/Record/007039835
    - Issue: The title of this record should end with (1960) not (1959).
    - Status: Discrepancy between cover and title page. Contacting contributing institution to add alternate title to record
- https://catalog.hathitrust.org/Record/006738647
    - Issue: Volume uc1.b3869277 should be titled Orbit 9.
    - Status: Contributing institution contacted, but this institution does not typically submit corrections.
- https://catalog.hathitrust.org/Record/007119250
    - Issue: The volume uc1.32106006364613 is of The Fifth Galaxy Reader (1961) rather than the first Galaxy Reader of Science Fiction published in 1952.
    - Status: Contributing institution contacted, but this institution does not typically submit corrections.
- https://catalog.hathitrust.org/Record/011406301
    - Issue: 008 field was improperly coded, but has been corrected. (Thanks to Michelle)

Part VI: Conclusion / Next Steps

We find that HathiTrust is useful for at least one popular genre. Works are available and findable in sufficient numbers to lead to a usable collection, but not all known works are findable. This result appears to be largely a question of availability and digitization, though there are significant challenges in matching based on title/author pairs. Moreover, we observe patterns in the lack of coverage, particularly in terms of publication year. This may reflect potentially troubling shifts in library collection practices away from genre fiction.

Our plans for further work include advocacy for corpus building, improved metadata-based searching, and expansion of content-based searches. First, our ability to add new works to the HathiTrust collection directly is limited, but we hope that this study will provide some guidance for further digitization work. Second, we have identified a need for more flexible metadata-based search capabilities. In particular, searching based on series titles in addition to individual work titles has proven to be critical for SF, and is likely to be similarly important for other popular genres such as Mystery, Thriller, and Romance. Finally, we are particularly excited about the potential for content-based search and filtering. Metadata may be the most useful and effective information possible within a limited storage range, but having access to hundreds of kilobytes to megabytes of data from full text representations is much more powerful. One promising direction could be to use hashing methods such as random projection to index large numbers of works (such as the 300,000 works identified as fiction by Underwood et al.) and

then search for near neighbors to known matches in our dataset. Based on our current results, we will almost certainly be able to find additional copies of known works, but also to find additional works by represented authors or works within series. The potential for discovering completely new authors and works is also possible, though more challenging as we reach the boundaries of our genre set (for example, *Clan of the Cave Bear* might return similar non-SF works about living in the wild). On the other hand, such "false positives" could challenge genre boundaries in productive ways. In any case, content-based analysis shows great promise.