HATHI TRUST

Library record matching and aggregation

GDD Network

Natalie Fulkerson, Collection Services Librarian, HathiTrust

December 18, 2019

Today's Plan

HathiTrust overview

Analysis

- Defining the problem
- Developing a methodology
- Insights/future directions

Aggregation

- Background
- Guiding principles
- Developing a proto-registry

Q&A

What is HathiTrust?

Founded in 2008, HathiTrust is a **not-for-profit collaborative of academic and research libraries preserving 17+ million digitized items**. HathiTrust offers reading access to the fullest extent allowable by U.S. copyright law, computational access to the entire corpus for scholarly research, and other emerging services based on the combined collection. HathiTrust members steward the collection — the largest set of digitized books managed by academic and research libraries — under the aims of **scholarly, not corporate, interests**.

Research goals

1. Evaluate various methods for determining overlap between the 4 collections
2. Gain experience working with unfamiliar records
3. Think about how we could do this work at (enormous) scale

Can we identify sufficient match points across records to understand overlap?

Our Process

Print Holdings overlap analysis

- Supports collection development
- Informs fee calculation
- Relies on OCLC number for matching

Bibliographic records stored in Zephir

- MARC format
- Contributed by our member libraries as part of ingest
- Multiple copies of the same work means multiple records that describe the same instantiation
- Clustered on OCLC number

Library Records Received

	# of digitized records	# of print records
British Library	516,212	-
National Library of Scotland	10,919	9,640,360
National Library of Wales	2,290	3,224,243
HathiTrust	16, 987, 842	-

Approach #1

⇒ Match library holding records to the HathiTrust collection using OCLC number (OCN)

OCNs present in library records (in the MARC 035 field; digitized items only):

	# records	# OCNs	# matching	% matching
British Library	516,212	611	130	0.025
National Library of Scotland	10,919	561	243	2.22
National Library of Wales	2,290	744	101	4.41

Print holdings:

	# print records	# OCNs	# matching	% matching
National Library of Scotland	9,640,360	466,302	~80,000	~0.99
National Library of Wales	3,224,243	221,382	31,861	1.03

Approach #2:

⇒ Look for other useable identifiers

For example - ISBN

	# digitized records	# print records	# ISBNs - digital	# ISBNs - print	% ISBNs - print
British Library	516,212	-	34	-	-
National Library of Scotland	10,919	9,640,360	55	2,709,837	28
National Library of Wales	2,290	3,224,243	17	3,128,171	97*
HathiTrust		-	2,645,141	-	16

Exploratory Methods

General approach:

- Identify various methods for string matching on title fields

- Pilot each method on a small set of records
  a. Does the method produce verifiable results?
  b. What are the limitations?
  c. Will the method scale to a larger recordset?

- Apply the method to the full dataset

Exploratory method #1

⇒ Literal string match of raw title fields in library datasets (MARC 245|abc) against HathiTrust records (MARC 245|abc)

	# digitized records	# matches	# matches to multiple records	% overlap
British Library	516,212	4,559	2,255	0.88
National Library of Scotland	10,919	343	131	3.14
National Library of Wales	2,290	51	9	2.23

Exploratory method #2

⇒ Literal string match of normalized title fields in library records against HathiTrust

- Processing consisted of:
  - Downcasing
  - Removing non-alpha characters

	# digitized records	# matches	# matches to multiple clusters	% overlap
British Library	516,212	39,815	18,298	7.71
National Library of Scotland	10,919	1,746	837	16
National Library of Wales	2,290	253	83	11.04

Exploratory method #3

⇒ Word-by-word match of BL titles against HathiTrust

1. For each BL title, downcase, eliminate stopwords, and produce a “bag of words”
2. Search HathiTrust for each of the words in the bag (not a literal string search, word order is not important)
3. Determine precision and recall, calculate an average confidence score, rank by score

Output is a list of candidate OCNs for each record, with a corresponding confidence score.

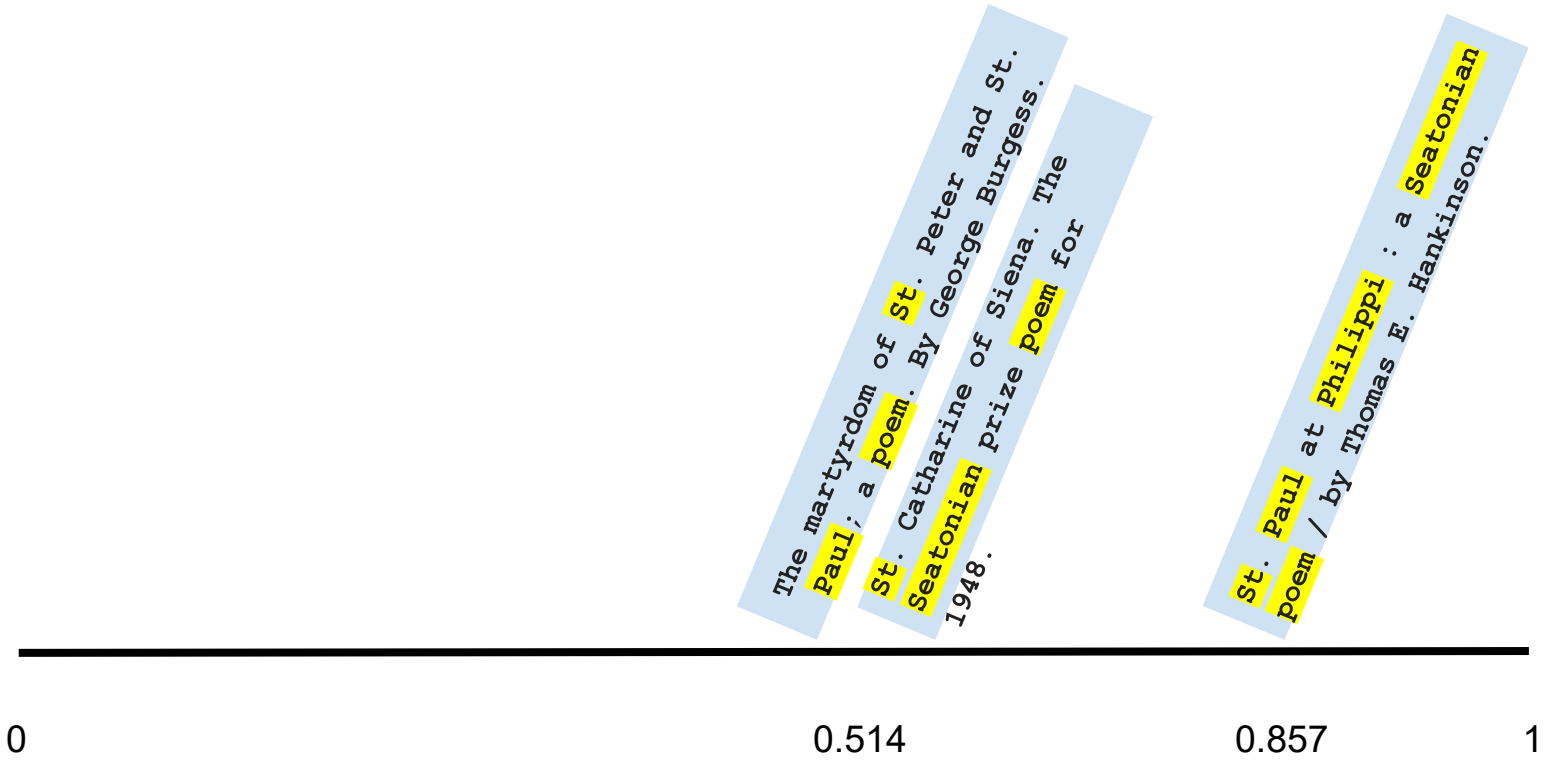
Example

BL title: "St. Paul at Philippi. A Seatonian poem."

Bag of words: st,paul,philippi,seatonian,poem

score	p	r	match_title
0.514	0.600	0.429	The martyrdom of St. Peter and St. Paul; a poem. By George Burgess.
0.514	0.600	0.429	St. Catharine of Siena. The Seatonian prize poem for 1948.
0.857	1.000	0.714	St. Paul at Philippi : a Seatonian poem / by Thomas E. Hankinson.

BL title: "St. Paul at Philippi. A Seatonian poem."
Bag of words: st,paul,philippi,seatonian,poem



Exploratory method #4

⇒ Machine learning

Query: Can you train a support vector (binary) classifier to distinguish between title matches and non-matches?

Machine Learning process:

- Setup
- Training/Iteration phase
- Implementation phase

Two approaches

- Extract specified fields from **43K Zephir records**
- Calculate **Damerau–Levenshtein distance**; create pairwise comparison vectors
- **10M** of resulting **955M vectors** selected at random
- **1M** of these selected for training
- Run classifiers against training data
- Run “trained” classifier against the remaining 9M recordsset

And

- Extract specified fields from **80K Zephir and 80K NLS records***
- Calculate **cosine similarity**; create pairwise comparison vectors
- **19M** of resulting **6.4B** vectors selected
- **260K** of these selected for training
- Run classifiers against training data
- Run “trained” classifier against the remaining 18.7M records

Preliminary Results

Zephir-to-Zephir; D-L distance:

	# of clusters in test set	# of correctly predicted clusters	Precision	Recall
Polynomial	5989	5558	.982	.911
RBF SVC	5989	5948	.975	.968

NLS-to-Zephir; cosine similarity:

	# of clusters in test set	# of correctly predicted clusters	Precision	Recall
Regression	83894	74029	.937	.882
Stochastic Gradient Descent	83894	71538	.928	.853
Linear SVC	83894	74025	.928	.882
RBF SVC	83894	69731	.940	.831

Insights

As expected, few records provided by UK partners contain OCNs

Presence of other identifiers (ISBN etc.) is likely variable.

Other methods show varying levels of promise:

- Literal string matching is simple but very rigid; word-by-word matching is more complicated but presents a more nuanced view
- Machine learning methods are more accurate, but require training time and are resource-intensive to run

Limitations

- Zephir clustering relies OCN, so is only as good (or bad) as the OCN assigned to each record.
- We suspect that Author and Date fields are not well-standardized, so don't work well as string components for literal matching.
- Stop word management is complicated by language differences.
- ML approaches used minimal training data; more work is needed

Conclusions

Duplicate detection is hard...

- Short titles, long titles, common titles
- Different manifestations of the same work

...involves tradeoffs

- Resource-intensive methods yield better results

Implications for aggregation:

⇒ Duplicate detection (overlap) vs. Clustering - how to express relationships to registry users?

Aggregation

Background

“Proto-registry” datafile was specified in grant proposal, but no functionality promised.

Scoped to ensure that records were reasonably complete and comparable between institutions.

Using the HathiFile as a model:

- Identify common fields
- Assess prevalence across project partner records

Proto-Registry fields

HathiFile Data Element	Description - Brief	MARC field/s
Volume Identifier	Permanent item identifier	
Title		245|a
Imprint	Publisher + Date of publication	260|bc
Publication Place		008 (bytes 15-17)
Author		100|abcd; 110|abcd
(URI)	Link to digital object	856
(Publication Place)		260|a

org	local_reco	35	ocns	245abc	260abc	008_15_1	100abcd	110abcd	item_id	856	260a
htdl	5023300	(MiU)005(	61073654	HÄ“ morp	AthÄ“na :	gr	Puchner, Walter, 194		mdp.3901	https://hd	AthÄ“na :
htdl	10754554	sdr-nrlf.b1	10575634	Bulletin de	Varsovie, I	pl			uc1.b4936	https://hd	Varsovie,
htdl	600227	sdr-ucla37	1537777	Bulletin of	[Brussels?	be			mdp.3901	https://hd	[Brussels?] :
htdl	1867750	sdr-nrlf.b1	21657885	The case f	Los Angele	cau	Besant, Annie, 1847-		uc1.$b291	https://hd	Los Angeles, Calif. :
htdl	1.03E+08	sdr-osu.b(	95634719	Shicheng \	Taibei Shi	ch			osu.32435	https://hd	Taibei Shi :|å°åŒ—å„ :
htdl	1.02E+08	sdr-uva.u1	7688573	History of	Cebu City,	ph	Klassen, Winand W.		uva.x0005	https://hd	Cebu City, Philippines :
htdl	526137	sdr-ucsd.b	1765929	Transactic	New York	nyu		Society of	mdp.3901	https://hd	New York :
htdl	5416845	sdr-nrlf.b1	8178129	Shinku tar	TÅkyÅ : SÅ	ja			uc1.31822	https://hd	TÅkyÅ :|æ±äº¬ :
htdl	5729491	sdr-wu473	2576091	Personnel	New York,	nyu	Yoder, Dale, 1901-		uc1.$b389	https://hd	New York,
htdl	997373	(MiU)0009	20831820	Uthalaput	Mujaphph	xx	Siá¹ƒha, RÄmasaÃ±jÄ		mdp.3901	https://hd	Mujaphpharapura,
htdl	1.01E+08	sdr-chi122	8918694	Memorab	New York	nyu	Xenophon		chi.38590	https://hd	New York :
htdl	8929825	sdr-njp253	29619705	Histoire ge	A Paris : C	fr			nyp.33433	https://hd	A Paris :
htdl	522520	(MiU)0005	1755752	Wissensch	Leipzig : D	gw			uva.x0017	https://hd	Leipzig :
htdl	1.02E+08	sdr-uva.u3	49312658	The ten de	[New York	nyu	Mooney, Kelly.		uva.x0044	https://hd	[New York] :
htdl	6038360	sdr-inu109	829968	Lexique d'	Paris, A. C	xx	Pelletier, AndreÌ.		inu.30000	https://hd	Paris,
htdl	1164993	sdr-nrlf.b1	6295767	The Librar	[Washingt	dcu		Library of	uc1.b3921	https://hd	[Washington,
htdl	8234924	sdr-nrlf.b1	47833519	Latin pron	[Ithaca, N.	xx	Peck, Tracy, 1838-192		uc1.$b734	https://hd	[Ithaca, N.Y.,
htdl	503731	sdr-uiuc41	2239457	Catalog of	Washingtc	dcu			osu.32435	https://hd	Washington, D.C. :
htdl	6674310	sdr-hvd00	15561819	Das Gift in	Leipzig, F.(	gw	Harnack, Erich, 1852-		uc1.b3558	https://hd	Leipzig,
bl	18619584			A True Co	London, 1	||||		England a	DRT Digital Store 812		London,
htdl	8683738	sdr-osu.b5	25769476	The Encyc	Philadelpl	pau			nyp.33433	https://hd	Philadelphia :
htdl	6567063	sdr-nrlfGL	13514054	Die volksw	Leipzig, W	gw	Trier, Julius		uc1.$b393	https://hd	Leipzig,

Future Directions

Additional matching work:

- Cleanup, fine tuning
- Test hybrid approach

UI/UX work:

- Whether/how to express duplication

New lines of inquiry:

- Can statistical methods play a role in connecting different instantiations of the same work (different editions, formats, translations)

THANK YOU