Interdependence in Action: The "Advanced Collaborative Support" Program from the HathiTrust Research Center

Wednesday, Oct. 28, 2020

Code of Conduct

HathiTrust events provide an inclusive environment that welcomes inquiry, constructive criticism and debate, and candor. HathiTrust does not tolerate personal attacks, harassment of any kind, verbal or physical violence, or disruptive behavior. All attendees are expected to be respectful of our community's diversity and generous of others' views. A full Code of Conduct and a complete process for handling reports of violations is in development. Until it is available, please bring concerns to us by contacting a member of the HathiTrust staff or by emailing conduct-reports@hathitrust.org.

Technology overview

- Zoom Meeting Features
- Mute and Unmute
- Chat
- Speaker View and Gallery View
- Automated Transcript/Closed Captions
- Support

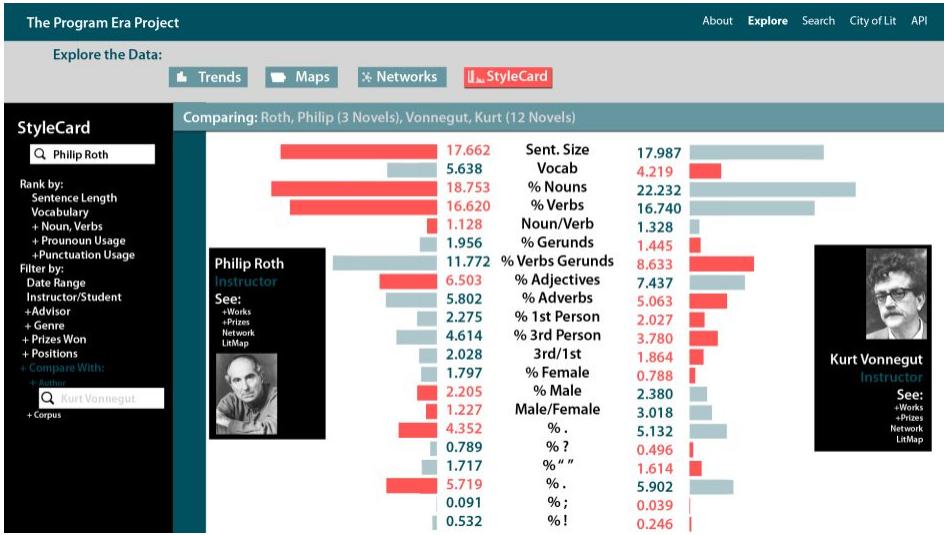
Presenters:

Ryan Dubnicek (HathiTrust Research Center): Introduction
Aduramo Lasode & Cody Hennesy (U. of Minnesota): “Surveying the HathiTrust Collections for Applicability of Energy Recovery Technology for Waste Treatment”
Stephen Krewson (Yale U.): “Deriving Basic Illustration Metadata”
Matthew J. Yoder & Dmitry Mozzherin (U. of Illinois, Urbana-Champaign): “Mapping scientific names to the HathiTrust Digital Library”
Laure Thompson (U. of Massachusetts Amherst): “Building Large-Scale Collections of Genre Fiction”
Glen Worthey (HathiTrust Research Center): Q&A Moderator

Introduction

Advanced Collaborative Support (ACS) Program

- Competitively awarded “grants” of time and resources
- Began in 2015
- 5 rounds of awards, supporting over 20 projects, across many research domains
- Read the project reports: https://wiki.htrc.illinois.edu/x/CADiAQ
- Emphasis on “C” for “Collaborative”
    - ACS has become a model for successful research with HTRC



IWW-affiliated author style cards generated from U. Iowa’s PEP project (read more: https://tinyurl.com/y6euglku)

Advanced Collaborative Support (ACS) Program

Questions about...

- A specific ACS project idea?
- Program logistics?
- The ACS program generally?

Email us!

acs@hathitrust.org

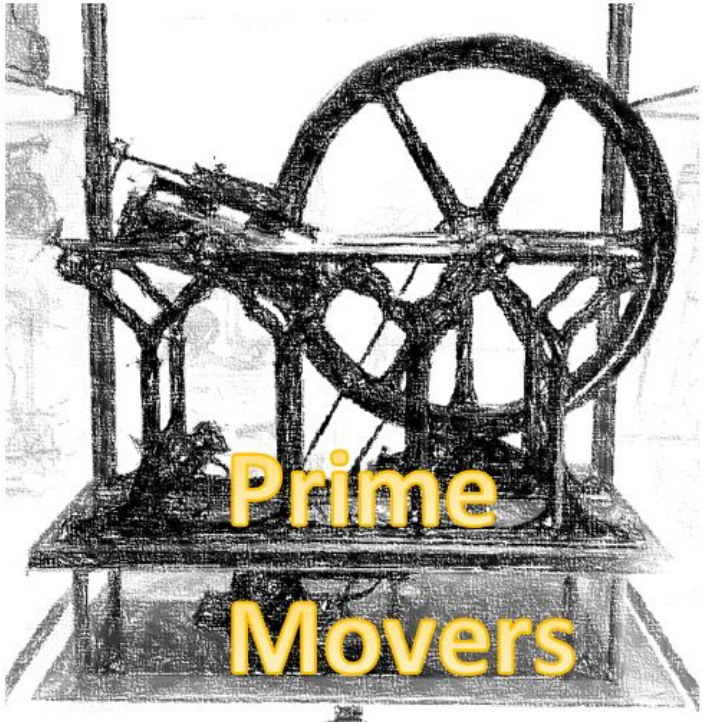
General questions about HTRC? Email us at htrc-help@hathitrust.org

Aduramo Lasode & Cody Hennesy
U. of Minnesota:

Surveying the HathiTrust Collections for Applicability of Energy Recovery Technology for Waste Treatment

Waste and Renewable Energy

- Expanding energy portfolio
- Prime mover technology in centralized treatment
- Future in distributed waste systems
- Technology deciding factors
- Initial study- decoupling factors
- Projected impact



Historic example of prime mover technology.

Text Mining Parameters

- Technology: 6 prime movers
- Year: 1900-2020
- Factors:
  - Efficiency
  - Cost
  - Fuel utilization

Technology	Electric Power (kW)	Reported Efficiency (%)
Gas/Combustion Turbine	2100	16.9
Steam Turbine	64002	49.6
Microturbine	190	28.2
Reciprocating Engine	5000	41
Solid Oxide Fuel Cell	640	51.6

Example of expected data from text mining.

HTRC Preliminary Work

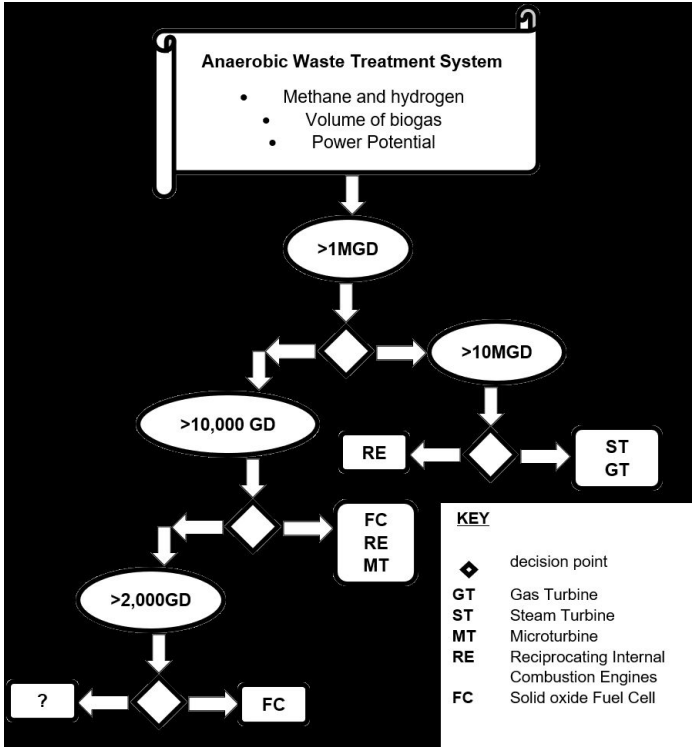
- Identifying initial HTRC corpus
- Evaluating information in HTRC volumes
- Identifying an ‘exclude’ list
- Sampling volumes to test extraction methods
- Measuring confidence of methodology

Highlights

- Technology match- 1 million+ volumes
- Filtering
- ‘Exclude’ lists

Next Steps

- Extraction methods
- Data analysis



Example of final output from data analysis.

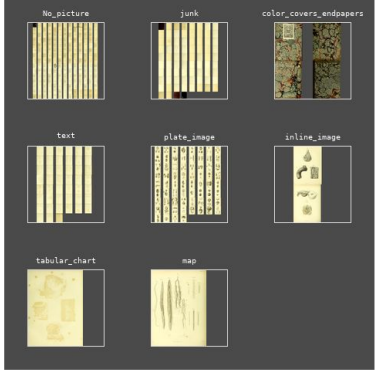
Stephen Krewson
Yale University

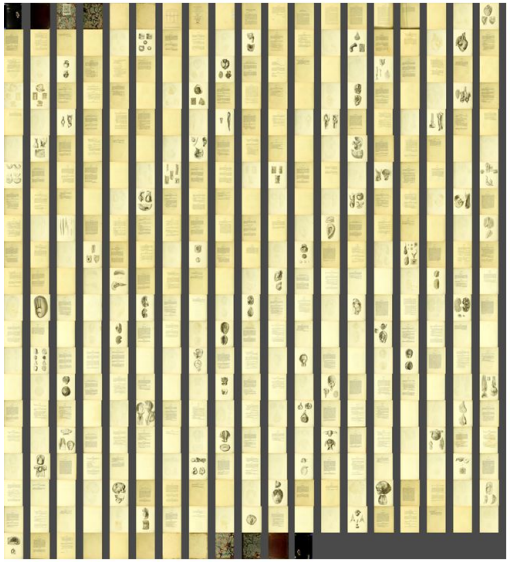
Deriving Basic Illustration Metadata

2,584,888

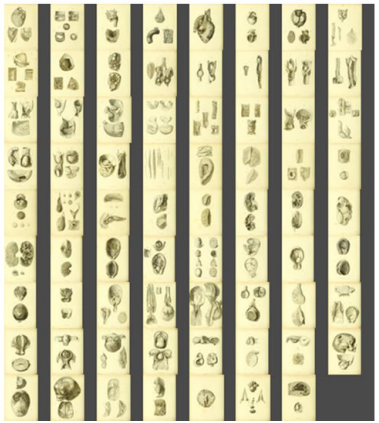
Illustrated regions in Google-digitized volumes, published 1800-50

197 pages with 1+ OCR picture blocks
(grouped by assigned label)



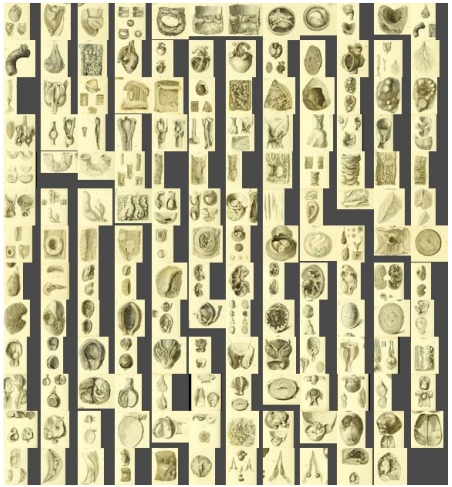


392 total page images in the volume



71 page images from four desired classes:
bookplate, *inline_image*, *map*, *plate_image*

155 visual regions of interest
extracted from 71 retained images



Transfer learning: A two-pass method for illustration extraction

Project outcome

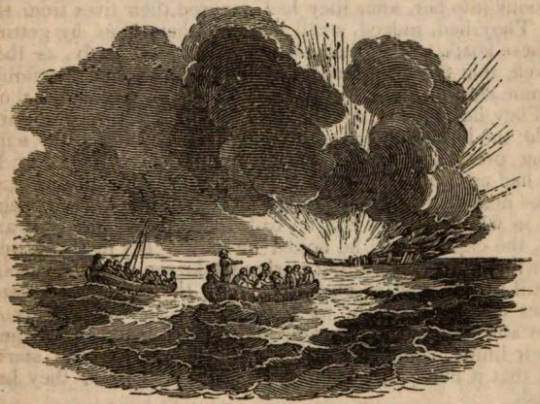
Results

- **183,553** volumes, **1.9M** pages
- Initial **10-class** CNN model for identifying good candidate pages
- Second Mask-RCNN model to localize illustrated regions (ROIs)
- Output: **2.6M** ROIs (**553 GB**)
- All ROI JPEGs vectorized (15 GB), but better to do analysis using dedicated visualization tools

Applications and Agenda

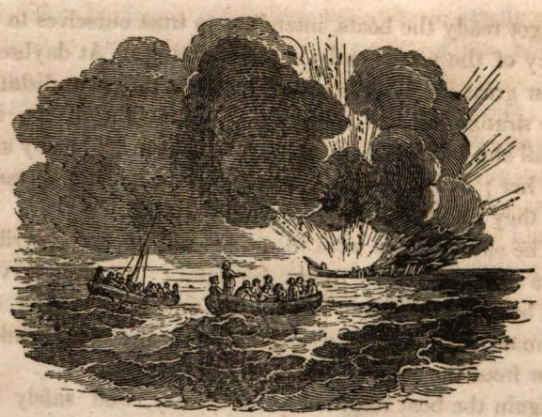
- **Dataset persistence:** JPEGs in private cloud bucket, project archived on Zenodo
- **Literary-historical findings**: Metadata + dataviz leads to discovery of copycat illustrations, stratification of 19C book market
- **Case study**: Munroe & Francis “robinsonades”
- **Interactive Demo**

Identifying similar wood engravings using PixPlot



Ship on Fire at Sea.

Robinson Crusoe, p. 329



hull of the Nancy was seen for some time rolling its burnt sides up and down on each succeeding wave ; at last we lost sight of it. Frank, as well as I, had hastily secured a few things before we stepped into the boat that bore us from the wreck ; my Bible and my father's letter were safe ; but our situation was now sadly altered for the worse, with little hopes of its being made better for a long time to come.

The vessel which had picked us up was bound for Got-

Identical wood engraving in Paul Preston's Voyages, p. 78

“Ship on fire at sea”

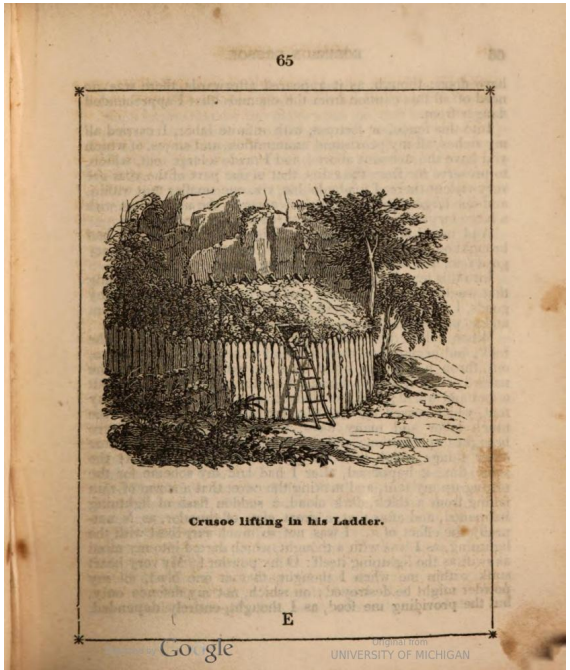
Robinson Crusoe [...] *with new designs on wood by* [Alexander] *Anderson* (1834)

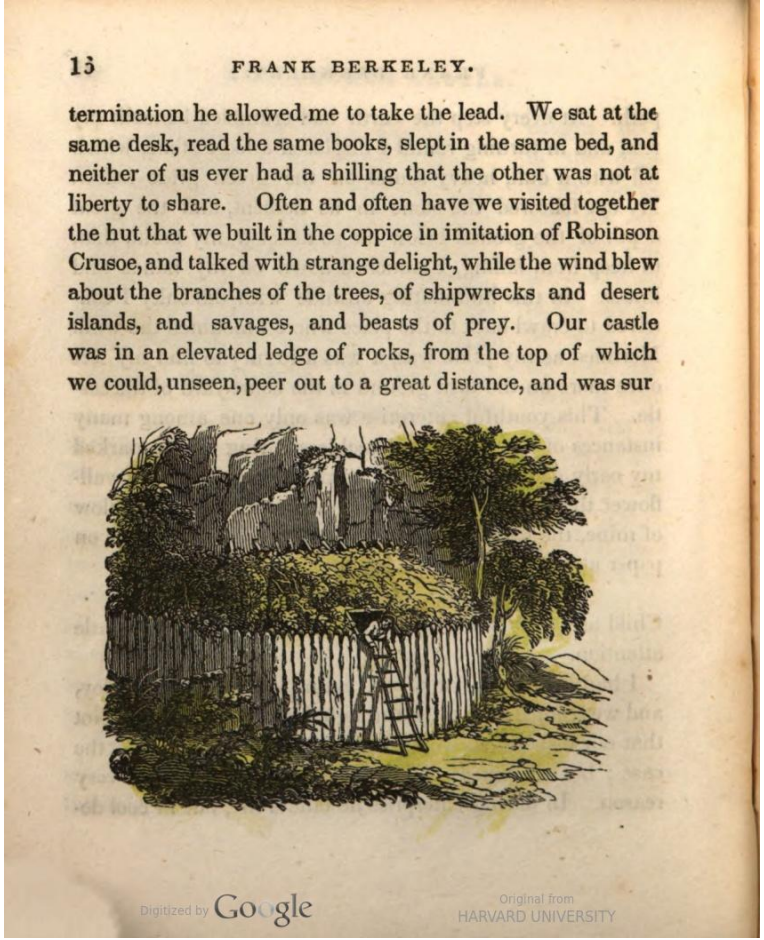
“[O]n a sudden, to our great terror, though we had reason to expect it, the ship blew up in the air; and immediately, that is to say, in a few minutes, all the fire was out, that is to say, the rest of the ship sunk . This was a terrible and indeed an afflicting sight, for the sake of the poor men, who, I concluded, must be either all destroyed in the ship, or be in the utmost distress in their boat, in the middle of the ocean; which, at present, by reason it was dark , I could not see.” (328)

Paul Preston’s Voyages (1847)

“Scarcely had they got clear of the hull, tossing on the waves, when, all in a moment, a billow capsized the boat: neither the captain nor a soul that was in her were ever seen to rise to the water's edge. An explosion of powder almost immediately followed, blowing up part of the deck; yet the hull of the Nancy was seen for sometime rolling its burnt sides up and down on each succeeding wave; at last we lost sight of it. Frank, as well as I, had hastily secured a few things before we stepped onto the boat that bore us from the wreck; my Bible and my father's letters were safe...” (77-78)

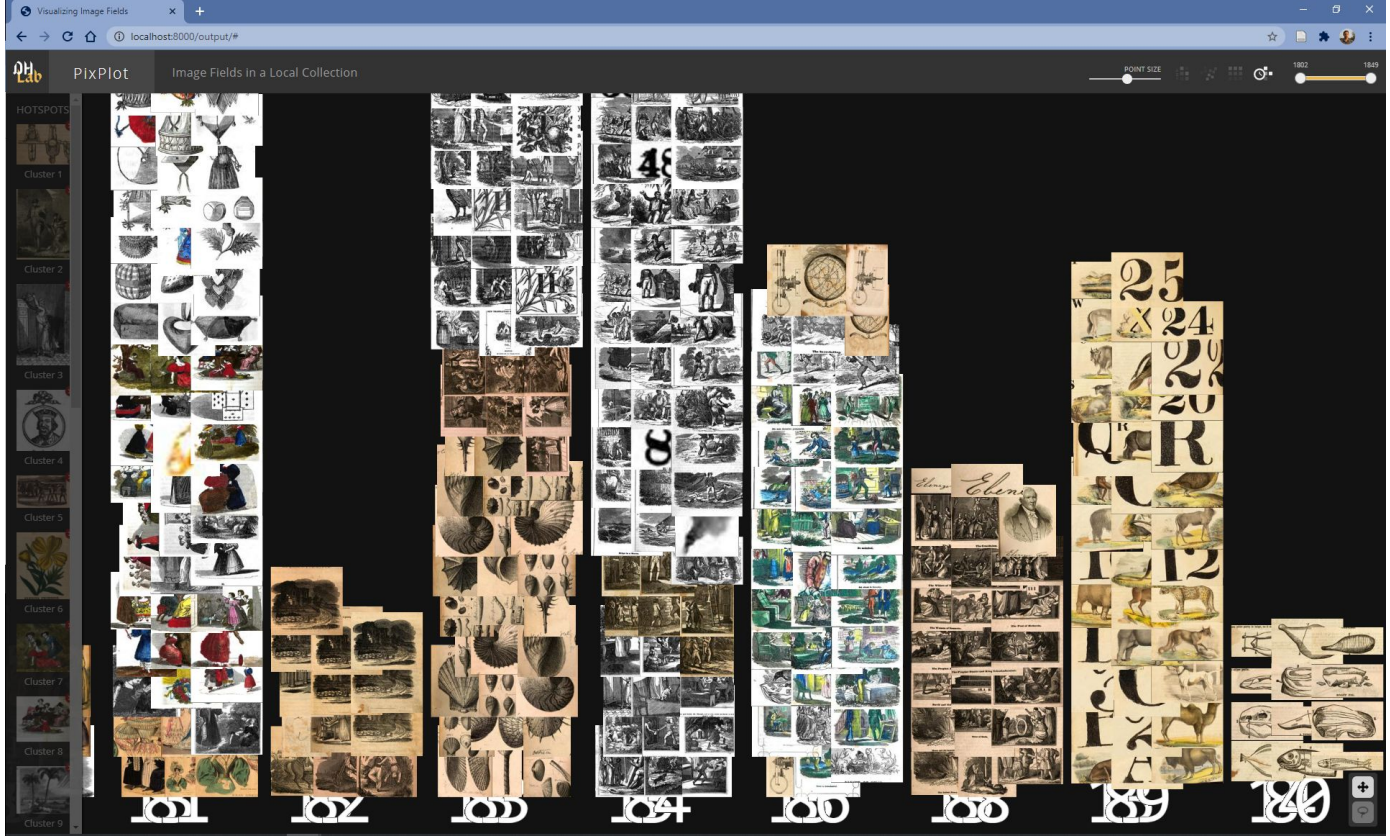
“Often and often have we visited together the hut that we built in the coppice in imitation of Robinson Crusoe, and talked with strange delight, while the wind blew about the branches of the trees, of shipwrecks and desert islands, and savages, and beasts of prey.” (16)





termination he allowed me to take the lead. We sat at the same desk, read the same books, slept in the same bed, and neither of us ever had a shilling that the other was not at liberty to share. Often and often have we visited together the hut that we built in the coppice in imitation of Robinson Crusoe, and talked with strange delight, while the wind blew about the branches of the trees, of shipwrecks and desert islands, and savages, and beasts of prey. Our castle was in an elevated ledge of rocks, from the top of which we could, unseen, peer out to a great distance, and was sur

Another reused Crusoe engraving in Paul Preston’s Voyages



Interactive PixPlot demo!
http://oilpalm2.htrc.illinois.edu/krewson/output/#

Acknowledgements (thank you!)

HTRC — Ryan Dubnicek, Boris Capitanu, Eleanor Dickson Koehl, Glen Worthey

Yale DH Lab — Doug Duhaime, Peter Leonard, Cathy DeRose

Medical Heritage Library — Melissa Grafe, Arthur Belanger

Editors and reviewers at *Programming Historian* and *Cultural Analytics*

. . . and, of course, HathiTrust and University of Indiana HPC

Matthew J. Yoder & Dmitry Mozzherin
U. of Illinois at Urbana-Champaign

Mapping scientific names to the HathiTrust Digital Library

Scientific names and bio-entities (taxon concepts)

Taxonomists

~2.5 M taxon concepts

Nomenclators

~ 7M Names

Literature, Databases, Collections

~ 100M Name Strings

Finding Names in Biodiversity Heritage Library

Scientific names are at the core of biology and are crucial for research. Biological names are anchors for further data-mining.

Volumes	~200,000+
Pages	~ 58M
Scientific Names	~ 200M occurrences ~ 12M Name-Strings ~ 2M “Taxon concepts”

BHL needs a name index. We tried, 10 years ago, and it took us 45 days. Today, an updated algorithms does the same job in 1 day on a laptop. Fast name-finding gives us ability to constantly improve the scientific names index.

https://github.com/gnames/gnfinder

Let’s apply improved algorithms to HathiTrust!

100X bigger than BHL. Name-finding ran on 50 computers for 9 hours

Volumes	~17M
Pages	~ 6B
Scientific Names	~ 3.6B occurrences ~ 30M Name Strings ~ 5.5M Canonical forms

Verification of names took 7 days against ~100 datasets. Further, we removed non-biological volumes to decrease the number of false positives.

https://github.com/gnames/htindex

What is next for scientific names in HathiTrust, BHL, and other sources?

Dramatic quality improvement

25% of names are abbreviated, and need be resolved.
Fine tuning for languages.
Removing false positives (“Habeus corpus”, “Oedipus complex”) with AI, ML.

Global Names Index

When both speed and quality are great, name-finding for the whole web. Creating a truly global name-index.

Laure Thompson
U. of Massachusetts Amherst

Building Large-Scale Collections of Genre Fiction

Curated list of works: author-title pairs

Extensive fan-built database of speculative fiction with rich work-level metadata:

- genre, subgenre
- award nominations & wins
- inclusion in notable book lists

18,809 works by 3,718 authors published from 1900–2010

Book Type	# Works
Novel	13,585
Collection, Anthology, Omnibus	3,567
Novella, Novelette	1,657

Genre	% Works
Science Fiction	53.7%
Fantasy	33.7%
Horror	8.3%

Searching the HathiTrust catalog

Total: 3,241 works & 5,160 volumes

Automated Search

Produced 4,920 work-volume pairs

- 4,382 good matches
- 538 mismatches

Non-Exhaustive Manual Search

Identified 3,155 work-volume pairs

- 778 missed by our automated search

Main causes of mismatches:

- Distinct works with similar titles
- Overlapping content for compilations & contained works

Main causes of misses:

- Series-level records
- Title & author name variation
- Alternate titles & author names

What did we find?

Mostly novels but also compilations

- 2,231 novels, 367 collections, 130 anthologies

Authors with the most work-level matches are fairly prolific

- Top 3: Robert Silverberg, Robert Heinlein, & Isaac Asimov

Out of copyright works are over represented at the volume level

Some prolific authors have little representation

- James Axler, Tanith Lee, C.J. Cherryh, & Mercedes Lackey

...also true for award-winning ones

- Only 2 works by Connie Willis

Other surprise absences:

- Octavia E. Butler's *Parable of the Sower* and its sequel
- Works by S.P. Somtow

Beyond the catalog: content-based methods

We use HathiTrust Extracted Features to produce volume-level signatures which allow us to identify duplicate works and ignore bad catalog-level matches.

Sim.	HTID	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

Sim.	HTID	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

Discussion

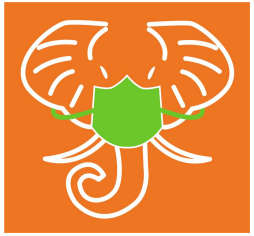
Aduramo Lasode & Cody Hennesy (U. of Minnesota): “Surveying the HathiTrust Collections for Applicability of Energy Recovery Technology for Waste Treatment”
Stephen Krewson (Yale U.): “Deriving Basic Illustration Metadata”
Matthew J. Yoder & Dmitry Mozzherin (U. of Illinois, Urbana-Champaign): “Mapping scientific names to the HathiTrust Digital Library”
Laure Thompson (U. of Massachusetts Amherst): “Building Large-Scale Collections of Genre Fiction”

Futures

Coming Soon: the next call for
Advanced Collaborative Support projects

With support from the **Mellon Foundation**, focusing on **historically under-resourced** and **marginalized textual communities**, and on **identifying gaps** in the HathiTrust collection

Tentative deadline: **November 30, 2020**



Thank you!

Aduramo Lasode
Cody Hennesy
Stephen Krewson
Matthew J. Yoder
Dmitry Mozzherin
Laure Thompson

Ryan Dubnicek & Glen Worthey

- Please complete the survey
- Jump in on the conversation on the HathiTrust Community Slack: #2020-comm-week
- Use @[PRESENTER] when referring to this session