HathiTrust Digital Library

Update On January/February Activities

March 23, 2016

Top News

Christenson, McIntyre, and Payne to Join HathiTrust Team

HathiTrust is pleased to announce the appointment of Heather Christenson, Sandra McIntyre, and Lizanne Payne to lead major services and new programs. Each will begin in May 2016.

Heather Christenson will join HathiTrust from California Digital Library as Program Officer for Federal Documents and Collections. Sandra McIntyre, currently director of the Mountain West Digital Library, will serve as HathiTrust's new Director of Services and Operations. Lizanne Payne, a nationally recognized expert in shared collection management, joins HathiTrust as the Program Officer for Shared Print Initiatives. See the full announcement here: https://www.hathitrust.org/library-leaders-to-join-hathitrust-staff

Heather Christenson

SAVE THE DATE: 2016 HathiTrust Member Meeting

The 2016 HathiTrust Member Meeting will be held on Thursday, November 10, 2016 at the Big Ten Center in Chicago. More details and registration information will be distributed in the coming months.

Sandra McIntyre

HathiTrust Research Center Awarded $1.7 million from the Andrew W. Mellon Foundation

On behalf of the HathiTrust Research Center, the University of Illinois at Urbana-Champaign has been awarded a two-year, $1,170,000 grant from the Andrew W. Mellon Foundation for the "Workset Creation for Scholarly Analysis + Data Capsules: Laying the Foundation for Secure Copyrighted Data in the HathiTrust Research Center, Phase I (WCSA+DC)," project. WCSA+DC will result in an overhaul and rebuild of the HTRC Workset Builder and improvement and scale-up of the HTRC Data Capsule for secure computing. Both systems will be extended to allow computational access to the in-copyright portion of the HathiTrust Digital Library.

HTRC Co-Directors J. Stephen Downie, (Illinois) and Beth Plale, (Indiana University Bloomington), along with Timothy Cole (Illinois) will be leading the efforts as principal investigators. Ted Underwood (Illinois), Kevin Page (Oxford e-Research Centre), James Pustejovsky (Brandeis University) and Annika Hinze (University of Waikato) will also be participating as project partners. In addition to working on overhaul and improvement of Workset Builder and Data Capsule, partners will also be working with the PIs on metadata improvement, construction, vetting and integration of computational tools and pilot use cases.

HTRC is a collaboration between the University of Illinois and Indiana University, with a primary goal of allowing research access to the HathiTrust corpus while still respecting copyright limitations. For more on this project, see: https://news.illinois.edu/blog/view/6367/331211

Lizanne Payne

John Butler to Chair Program Steering Committee

John Butler, Associate University Librarian for Data and Technology at the University of Minnesota, is replacing Bob Wolven as chair of the Program Steering Committee (PSC). John has served on PSC since 2013 and has been deeply involved in HathiTrust governance since the very start. His term begins in March 2016 and will run through February 2018. Wolven has served as chair of the PSC since its inception in 2013, and remains on the Board of Governors.

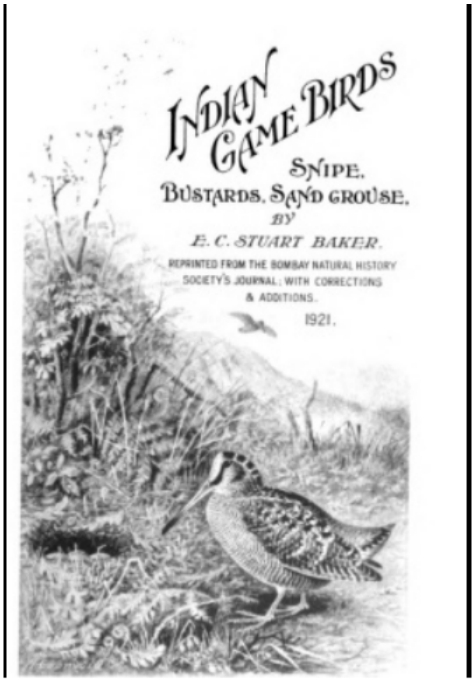
Copyright Reviews Continue as CRMS Grants Conclude

The HathiTrust membership has been making copyright determinations on books in HathiTrust since 2008 with the support of three National Leadership grants awarded to the University of Michigan by the Institute of Museum and Library Services (IMLS). This has been explicitly a collaborative effort, achievable because 20 institutions pledged staff time to engage in review work using the Copyright Review Management System (CRMS) interface. A few of the highlights to date are:

- 331,889 determinations for US books published 1923-1963. Of these 177,398 were found to be in the public domain (53.6%)
- 177,912 determinations for books published in the UK, Canada, and Australia. 144,733 of these were determined as public domain (79.2%).
- Since spring of 2015 61,514 state government documents in HathiTrust and have been identified as potential candidates for review. 9,726 of the 13,387 reviewed so far were found to be in the public domain (73%).
- The project team has produced a CRMS Toolkit that details the CRMS methodology developed after eight years in the context of HathiTrust. The Toolkit will allow the CRMS approach to be replicated and used in a variety of new ways.

The IMLS supported phase of this review work was scheduled to close at the end of February. As the grant period concludes, HathiTrust staff has now taken responsibility from the University of Michigan Library for coordination of these continued reviews. Ten institutions have committed to continuing to review through the end of calendar year 2016, focusing on the fewer than 30,000 currently remaining works from the UK, Canada, and Australia for review.

The CRMS project has had the practical effect of expanding access to public domain materials and improving our understanding of the nature our collective collection. It has helped our community see the impact of cooperative work and expanded our thinking about what is possible with post-1923 publications. HathiTrust continues to collect digitized books in its collections, and during 2016 we will be examining other potential review projects that would leverage the investment reflected in CRMS' accomplishments and sustain this vital work.


Indian Game Birds
Snipe.
Bustards, Sand Grouse.
by
E. C. Stuart Baker.
Reprinted from the Bombay Natural History Society's Journal; with corrections & additions.
1921.

Game Birds of India, 1921
http://hdl.handle.net/2027/mdp.39015019736951 ...

Update On January/February Activities

March 23, 2016

On March 16, Executive Director Mike Furlough and Copyright Review Manager Kristina Eden hosted a webcast explaining the copyright review process. You can view a recording of the webcast here: https://youtu.be/DUJh6d9OPHI

For more on the CRMS project: https://www.hathitrust.org/grants

You can follow HathiTrust on Twitter or Facebook

Subscribe to email updates (via Google Groups)

Eight New Members Joined HathiTrust in 2015

Eight Institutions joined HathiTrust in 2015. Those institutions are:

- Auburn University
- Bryn Mawr College
- Georgia Tech University
- Smith College
- Swarthmore College
- University of Nevada, Las Vegas
- University of Rochester
- University of Wyoming

Total membership has grown to 113 institutions on three continents. HathiTrust staff welcomes our new members and looks forward to several institutions currently in process to join in 2016. Last summer, HathiTrust staff held a webcast for new members. This webcast can be viewed here: https://www.youtube.com/watch?v=0DfMrIt70As&feature=youtu.be

Jock Halliday, or sketches of life in an old city parish. Eninburg, 1883 *http://hdl.handle.net/2027/uc1.ax0002485399 ...*

HathiTrust Research Center

HTRC Ingest of In-Copyright Data

HTRC has been in the process over the last several months of securely mirroring the full HathTrust digital library. This action, begun only after an extensive security review that completed late summer 2015, will provide greater benefit to the HathiTrust membership by enabling non-consumptive computational analytical access to the full HT digital library. These services will be announced as they become available later this year.

Ingest

A new storage installation and upgrade in January has permitted staff to focus more attention on ingest activities. HathiTrust began ingesting Google-digitized content from Michigan State University, adding 27,800+ volumes to the repository (see all content at http://bit.ly/1pHkdf6). In addition, University of Maryland and Northwestern University successfully submitted their first batches of locally digitized materials, respectively adding 286 and 220 volumes to the HathiTrust collection. Staff continued to work with a number of other institutions (namely: Boston College, University of Illinois at Urbana-Champaign, University of Washington, University of Queensland, Cornell University) to answer questions and ingest additional content.

HathiTrust Digital Library

Update On January/February Activities

March 23, 2016

Zephir Update

In January and February 2016, Zephir loaded 270,374 new and 204,440 updated records from HathiTrust content contributors and established metadata submission processes for several new content streams. The Zephir Operations team is invested in making improvements to the metadata submissions process, beginning with refining local processes and coordination with HathiTrust staff at University of Michigan. Technical team members are engaged in cross training across system components. In January, the Zephir Operations team began providing additional, volume–specific data access to the HathiTrust User Support Working Group to support the corrections workflow.

Procedures for Replacing Pages in Google Books

Members of an ad hoc HathiTrust quality working group have developed and shared documentation for Google partners to use in replacement and insertion scanning of pages in Google-scanned volumes. This process (also known as SPIR for Single-Page Insertion and Replacement) was developed a number of years ago for Google partners to use to submit missing or replacement pages for Google-scanned volumes. This documentation is intended to make it easier for Google partners to submit those scanned pages through SPIR and was shared directly with Google Partners Communications Group as well as directly with HathiTrust contributors who are Google partners. Google partners who do not have access to this document should contact feedback@issues.hathitrust.org.

The members of this working group are Kat Hagedorn from the University of Michigan, Michelle Paolillo from Cornell University, Janet Gertz from Columbia University, and Andy Hart from the University of North Carolina, Chapel Hill.

Projects

Copyright Review

	January-February		Overall	
	Public Domain Determinations	All Determinations	Public Domain Determinations	All Determinations
CRMS-US	854	1,230	177,398	331,899
CRMS-World	9,448	16,715	144,733	272,300
Total	10,302	17,945	322,131	604,199

Volumes Added

Ingest numbers and Collection statistics are updated daily and can be found on our website here: https://www.hathitrust.org/visualizations_deposited_volumes_current

Journey to Llasa and Central Tibet by Sarat Chandra Das, 1902. http://hdl.handle.net/2027/yale.39002022247994 ...

US Federal Documents Registry

The Registry's staff have begun receiving daily updates of new and updated records for US federal government documents from the HathiTrust Digital Library. These Staff have also begun enhancing the accessibility of the Registry's interface, bringing it in line with other HathiTrust interfaces.

Work has continued to improve duplicate detection, incorporating simple title matching in the algorithm. In the near future, staff will analyze the roughly 5 million records that do not contain any identifiers. The Registry staff have also begun working to identify records for US federal government documents that are already in HathiTrust and not coded as such in the MARC record.

As of March 1, HathiTrust's collection included 677,593 U.S. Federal Documents.

Development Updates

Improvements to HathiTrust Services for Users with Print Disabilities

In January, we resolved a serious accessibility issue to which they had been alerted in the fall of 2015. When the current method for PDF generation was first implemented, staff used common recommendations at the time for packaging OCR and images into PDFs. This caused a problem in which screen readers read the watermarked copyright notices added to HathiTrust PDFs but ignored the text content of the book.

Staff have been able to change the way text is embedded into PDFs, so that the watermark is ignored and screen readers recognize the OCR text. This also means that users are now able to use the common PDF functionality that allows them to save PDFs in a different format, such as Word documents or text files. This is an initial foray into improving the accessibility of PDFs. Staff are continuing to investigate and work on making PDFs more accessible.

We have been working to provide print disability proxy users (staff at partner institutions who obtain copyrighted content in HathiTrust on behalf of users with print disabilities) with more support in order to increase usage of this service. A user guide has been created and made available at <http://bit.ly/1NOSC2N>. Partners are welcome to borrow and adapt this guide for their own purposes. In addition, a group email address has been created in order to facilitate communication with and among proxies. See https://www.hathitrust.org/accessibility for more information about how this service works and to register a proxy user at your institution.

THE "QUADRANT"

Nº 8

This Machine, though only before the public for the last two years, HAS REVOLUTIONISED THE STYLE OF TRICYCLES. It is much imitated; but as its vital principles are protected by patent rights,' it continues far and away

THE FINEST ROADSTER IN THE WORLD.

Two trips to the emerald isle by "Faed." I. -A racing trip to Dublin, by A. J. Wilson, 1888
http://hdl.handle.net/2027/uc1.c052208372

Repository Availability

Cumulative 12-month availability of repository access: 99.975% (-/+0.000%).

There's an elephant in the library.™

www.hathitrust.org

Full-text Search

Two issues related to relevance ranking were investigated:

- Approaches to relevance ranking when only part of a document is relevant
- Impacts of OCR errors on ranking and methods for improving ranking when such errors are present

Work continued on a testing framework for relevance ranking that would enable principled testing of possible solutions to these and other relevance ranking issues. As part of this work, issues with the 2007 -2010 INEX Book Retrieval Track were analyzed to better understand academic research on how to rank books when only part of the book is relevant (The INEX Book Retrieval Track was an annual event where information retrieval researchers experimented with different methods to improve relevance ranking in book search using a shared test collection of about 50,000 books). The analysis of the INEX Book Track will also inform the design of a HathiTrust test collection for book retrieval.

PDF Downloads

The development team has begun experimenting with an automated PDF workflow that will make it possible to sync built PDFs/EPUBs between Michigan and Indiana sites to support load balancer behavior.

Storage

A complete replacement of primary repository storage was installed at both sites in January and brought online in February. Content is currently being migrated to the new storage, and the old storage will be retired by the end of March. With this recent installation we expect our current storage footprint to meet repository needs through 2019.

Papers and Presentations

Presentations

- Bhattacharyya, Sayan. HTRC Talk to Stanford University Library DH group and subject specialist librarian's group. Green Library, Stanford University. January 20, 2016.
- Bhattacharyya, Sayan. "Why is studying the humanities important?" Illinois Program for Research in the Humanities-sponsored workshop with HT+Bookworm for student teams for 4humanities.org Student Prize Contest. Scholarly Commons, University Library, University of Illinois. February 4, 2016.
- Bhattacharyya, Sayan and Muhammad Saad Shamim "The HathiTrust+Bookworm Project as a Model for Collaborative Research at Large Scale." Presentation in the panel "Developing and Sustaining Collaborative Research in the Humanities." 131st Annual Convention of the Modern Language Association (MLA). January 8, 2016

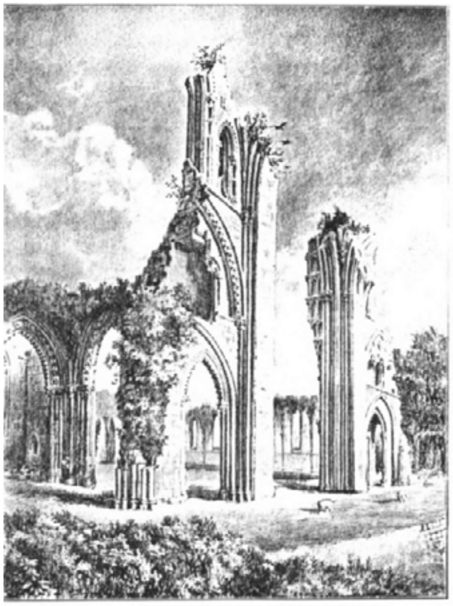
Forecast

Continue work on a unified logging framework for HathiTrust applications.

Support for alternative text formats.

Accessibility audit for HathiTrust apps/interfaces in production.

Implement metadata downloads from the Collection Builder



The story of a psychological experiment which resulted in discovery of Edgar Chapel, 1918
http://hdl.handle.net/2027/hvd.32044014077713 ...

Update On January/February Activities

- Dickson, Eleanor. “Doing Text Analysis with the HathiTrust Research Center’s Tools.” Workshop for librarians at the University of Texas at Austin, January 4, 2016.
- Dickson, Eleanor. University of Illinois Library Savvy Research workshop on Text Analysis. March 8, 2016.
- Dickson, Eleanor. “Text Analysis Methods and Tools.” Brownbag at the Illinois Program for Research in the Humanities. University of Illinois at Urbana-Champaign. February 10, 2016
- Downie, J. Stephen. “DH Panel: Fair Use and the Future of Digital Scholarship.” Scholars Lab, University of Virginia, February 24, 2016
- Downie, J. Stephen. “HathiTrust and the Future of Digital Archive.” Keynote address, International Symposium at University of Tokyo Ito International Academic Research Center. January 25, 2015
- Jett, J. et. al. 2016 The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-Consumptive Research Collections. Journal of Open Humanities Data, X: eX, DOI: http://openhumanitiesdata.metajnl.com/articles/10.5334/johd.3/
- Organisciak, Peter and Sayan Bhattacharyya. “New tools from the HathiTrust Research Center for digitized text analysis at scale: The HathiTrust+Bookworm tool and the Extracted Features dataset.” E-Research Roundtable, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign. Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign. February 10, 2016.

THE BLUE MOSQUE AT TABRIZ.

A history of Persia by Brigadier General Sir Percy Sykes, 1921. http://hdl.handle.net/2027/wu.89009509969 ...

HathiTrust on the Road

HathiTrust staff will be attending the following events in early 2016. Please contact us if you wish to meet us at any of these events:

CRL 2016 Global Resources Collections Forum, Chicago, IL, April 14-15 - Mike Furlough

DPLAfest 2016, Washington, D.C., April 14-15 - Kristina Eden and Angelina Zaytsev

ARL Spring Membership Meeting, Vancouver, BC, Canada, April 26-28 - Mike Furlough

Digital Humanities 2016, Krakow, Poland, July 11-14 - J. Stephen Downie

Update On January/February Activities

March 23, 2016

RESEARCH CENTER

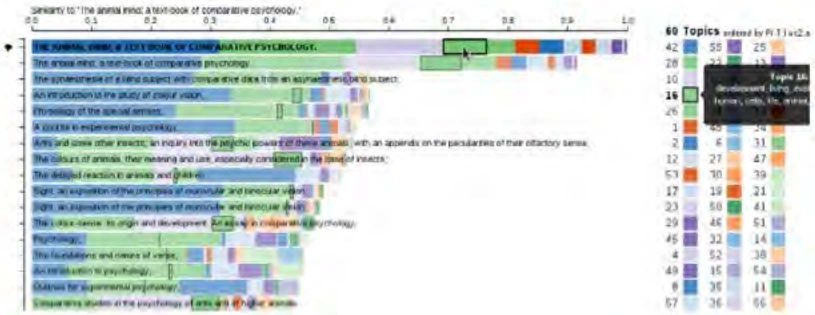
HATHITRUST
RESEARCH
CENTER

LEARN MORE AT HTTP://WWW.HATHITRUST.ORG/HTRC

Taxonomizing the Texts:
Towards Cultural-Scale Models of Full Text

Colin Allen & Jaimie Murdock, Indiana University

This project from Indiana University leverages the HTRC Data Capsule framework to test and visualize topic models. It will show the relationship between a topic model created on a random sample of volumes and the entire category from which it is drawn. Their proof-of-concept models Library of Congress subject headings and visualizes them in an online tool called Topic Explorer.


Similarity to "The animal mind; a text-book of comparative psychology."
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
THE ANIMAL MIND; A TEXT-BOOK OF COMPARATIVE PSYCHOLOGY.
The animal mind; a text-book of comparative psychology.
The synaesthesia of a blind subject with comparative data from an asynaesthesic blind subject.
An introduction to the study of colour vision.
Physiology of the special senses.
A course in experimental psychology.
Ants and some other insects; an inquiry into the psychic powers of these animals, with an appendix on the peculiarities of their olfactory sense.
The colours of animals, their meaning and use, especially considered in the case of insects.
The delayed reaction in animals and children.
Sight, an exposition of the principles of monocular and binocular vision.
Sight, an exposition of the principles of monocular and binocular vision.
The colour-sense: its origin and development. An essay in comparative psychology.
Psychology.
The foundations and nature of verse.
An introduction to psychology.
Outlines for experimental psychology.
Comparative studies in the psychology of ants and of higher animals.
60 Topics
Topic 16
development, living, evolution, human, cells, life, animal

- Extracted Library of Congress Subject Headings for 1,606,302 volumes for building topic models by subject headings.
- Queried to create random sample using a machine with a 1.9 GHz CPU and 64 GB of memory.
- Utilized 6 virtual machines in the HTRC data capsule
- Topic Explorer readily available in the HTRC Data Capsule

HTRC Use Case: Collaborating and Supporting Scholars

Most-accessed volumes Jan/Feb 2016
Solid mensuration, by Willis F. Kern and James R. Bland.
Quicksand, by Nella Larsen.
Advanced accounts; a manual of advanced book-keeping and accountancy for accountants, book-keepers, and business men, edited by R.N. Carter
Annual Report of the Board of Directors of the Lehigh Valley Railroad Company to the Stockholders, 1905-13.
The Marines Magazine, v. 2, (1916-17).
Godey's magazine, v.40-41, 1850.
The human figure, by John H. Vanderpoel.
Families directly descended from all the royal families in Europe (495 to 1932) and Mayflower descendants, by their lineal descendant, Mrs. (Oscar Herbert) Elizabeth M. Leach Rixford
The light of Egypt; or, The science of the soul and the stars, Vol. 1, by Thomas H. Burgoyne.
Roster of the Confederate soldiers of Georgia, 1861-1865, v.1.

The 'submarine menace' in The British Navy at War by W. Macneile Dixon, 1917. http://hdl.handle.net/2027/mdp.39015063971538 ...

User Support Issues	Jan-Feb	Nov-Dec
Content	**272**	**249**
Quality	250	229
Collections	21	19
Cataloging	**227**	**250**
Access and Use	**298**	**253**
Copyright	120	87
Permissions	19	28
Takedown	1	2
Print on Demand	1	0
Inter-library loan	2	6
Full-PDF or e-copy requests	50	64
Datasets	8	5
Data Availability and APIs	7	2
Reuse of content	13	8
Web applications	**66**	**51**
Functionality problems	31	17
Problems with login specifically	10	7
General questions about login	1	1
Partners setting up login	1	1
Usability issues	0	1
Feature requests	4	6
Partner Ingest	**37**	**37**
General	**237**	**190**
Partnership	25	19
Miscellaneous	212	171
Total	**1137**	**1030**

*See User Support Working Group Issue Types for a description of the types of issues included in each category.