Columbia Law School
Scholarship Archive

Sabin Center for Climate Change Law | Research Centers & Programs

2016

Designing and Maintaining an EIS Database: Lessons Learned in Developing Library-based Digital Repositories

Robert T. Cartolano

Follow this and additional works at: https://scholarship.law.columbia.edu/sabin_climate_change

Part of the Environmental Law Commons

COLUMBIA UNIVERSITY
LIBRARIES

Designing & Maintaining an EIS Database:

Lessons Learned in Developing Library-based Digital Repositories

Robert Cartolano
Associate VP, Digital Program & Technology Services
June 16, 2016

Topics

- Overview of Libraries
- Six Lessons Learned
  - Find Partners and Collaborate
  - Listen to Your Users
  - Preservation and Access
  - Decouple Front-end from Back-end
  - Metadata, metadata, metadata
  - Innovate at the Front-end
- Q & A

Overview of Libraries

- 21 Libraries
- 12 million volumes
- 160,000 journals and serials
- 450 professional and support staff
- 4,000+ study seats
- 1,100 workstations
- Significant technology infrastructure
- Thousands of external vendors
- Over $10M annual spend on E-resources

http://library.columbia.edu/about.html

New Type of Research Library

"While the core features traditionally associated with research libraries continue to play a central role, our services embrace a broader range, extending the conception of the research library."

- James Neal, Former VP for Information Services and University Librarian

1. Find Partners, Collaborate

- Internal and external partners
- Partners share development efforts
- Leverage open source efforts
- Communities support each other
- Communities can help obtain external funding for targeted projects

Academic Preservation Trust (APTrust)

- University of Virginia, 17 partner institutions
- APTrust is working toward TRAC certification
- Repository to be built from the ground up taking TRAC certification into account.
- Certification Working Group established and will be advising and consulting with the APTrust staff and partners on TRAC objectives.
- Initial development work is proceeding at the level of Digital Object Management and Infrastructure.

http://aptrust.org/

External Partners

- **Blacklight**
  - discovery software
  - http://projectblacklight.org/
- **Fedora**
  - digital preservation repository software
  - http://www.fedora-commons.org/
- **Hydra**
  - digital content management
  - http://projecthydra.org/
- **DuraSpace**
  - 138 member institutions
  - http://duraspace.org/

2. Listen to Your Users

- Gather data - Assessments, surveys, focus groups, needs analysis
- Real-world user testing - capture activity, video session
- Engage users in any way you can

3. Preservation and Access

Preservation - prolong the **existence** and **accessibility** of the collections for current and future students and scholars

http://library.columbia.edu/services/preservation.html

Backup ≠ Preservation

Digital Library Collections

We must preserve and provide access to …

- Local Digitization Projects
- Preservation-Related Digitization
- Institutional Repository / Data Sets
- Born Digital Archival Content
- Archived Web Sites
- Super Dark Archives – *highly secure*

Census Data

http://www.census.gov/data.html

ICE Pod - Antarctic Research

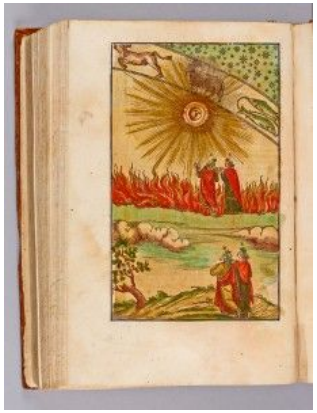
http://www.ldeo.columbia.edu/res/pi/icepod/

Ancient Ink

http://cise.columbia.edu/AncientInk

SEDAC

http://sedac.ciesin.org/



16th Century Dante Books

http://digitaldante.columbia.edu/

IEDA
INTERDISCIPLINARY
EARTH DATA ALLIANCE

http://www.iedadata.org/

duPont-Columbia Awards Videos

http://www.journalism.columbia.edu/dupont

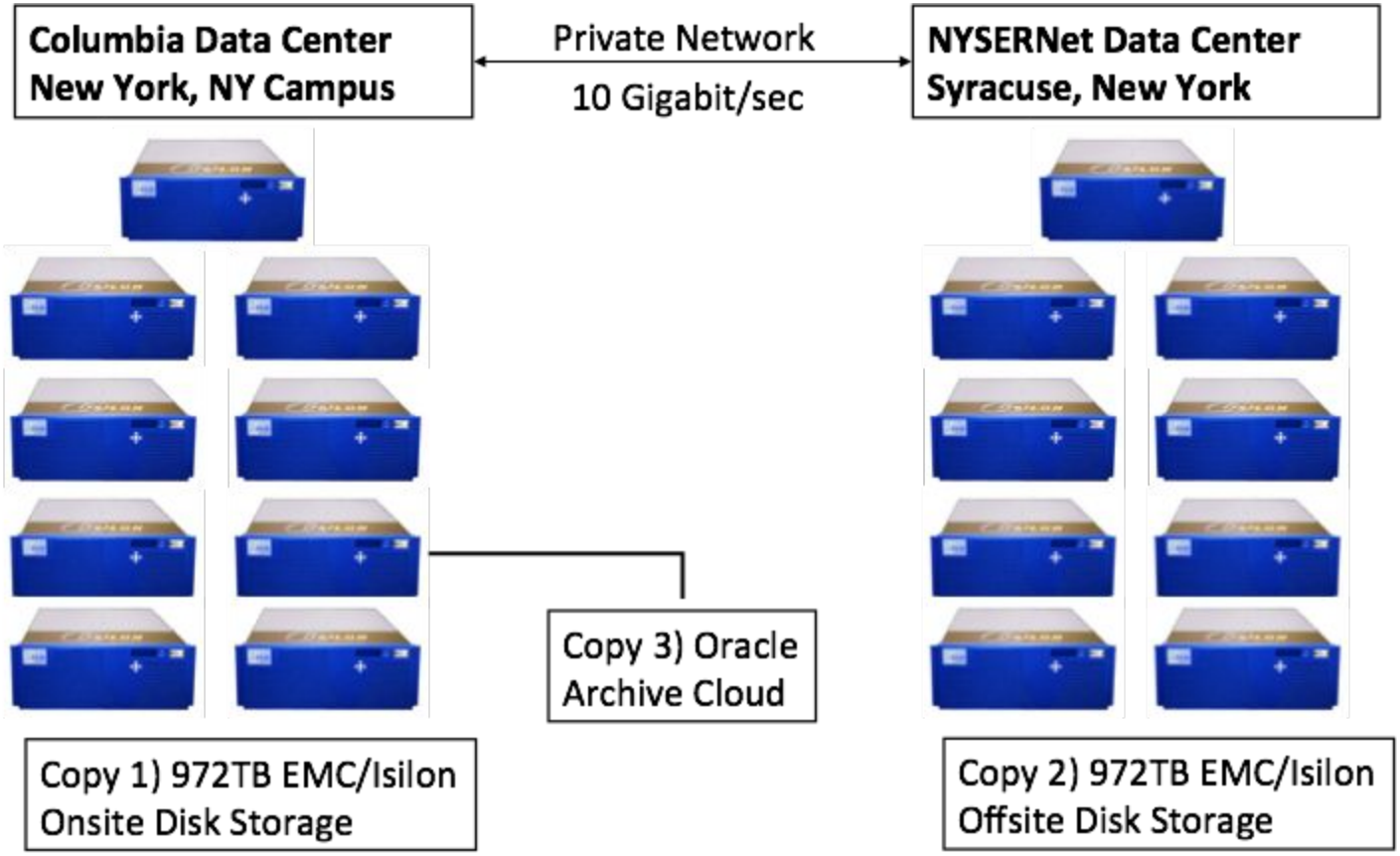
COLUMBIA UNIVERSITY LIBRARIES / INFORMATION SERVICES

4. Decouple Front-end & Back-End

- Separation provides flexibility
  - Risk-averse storage approach
  - Different storage for preservation, access
  - Accommodate multiple front-end systems
  - Leverage national/cloud archive solutions
- Swap out back-end storage
  - 2009 - 70 Terabytes, 4 copies
  - 2013 - 300 Terabytes, 3 copies
  - 2016 - 800 Terabytes, 3 copies

2016 Library Digital Preservation Storage

- **Two copies on disk, third copy archived in cloud**
- **800 Terabytes effective storage (~2PB total)**



Fedora Software Platform

http://fedorarepository.org/

Flexible **E**xtensible **D**igital **O**bject **R**epository **A**rchitecture

- Open source digital repository software
- Internationally developed, supported, adopted
- Platform for Columbia's digital preservation system

Fedora Overview: http://dx.doi.org/10.7916/D8FF3QDR

Columbia Profile: http://www.fedora-commons.org/profile/101

Key Features of Fedora 4

http://fedorarepository.org/

- Authentication/Authorization
- Backup
- Clustering
- Content modeling
- Fixity services
- Linked data (native RDF support)
- Search (admin and external)
- SPARQL endpoints (semantic protocol, RDF query language)
- Advanced storage capabilities
  - Policy-driven, self-healing, very large files
  - External data source “projection” (aka “virtual ingest”)
- Transactions
- Versioning

5. Metadata, Metadata, Metadata

- Citable metadata in standard formats with persistent identifiers fosters long-term access and data management
- Leverage human and computer-generated descriptive metadata for search and discovery
- Assign rights, administrative, provenance, technical, structural and preservation metadata for long-term curation

Leverage Standards

- Portland Content Data Model (PCDM)
- RDF/Linked Data for Libraries
- Text mark-up standards (XML, TEI)
- Preservation Standards (ISO 14721 - OASIS; ISO 16363 - Trustworthy Repositories)
- New and emerging web, content and technology standards

6. Innovate at the Front-End

http://projectblacklight.org/

Blacklight Achieves...

- Search capabilities that meet user expectations
- Speed
- Flexibility to customize
- Unified Discovery
- Vibrant community - continuous, iterative development

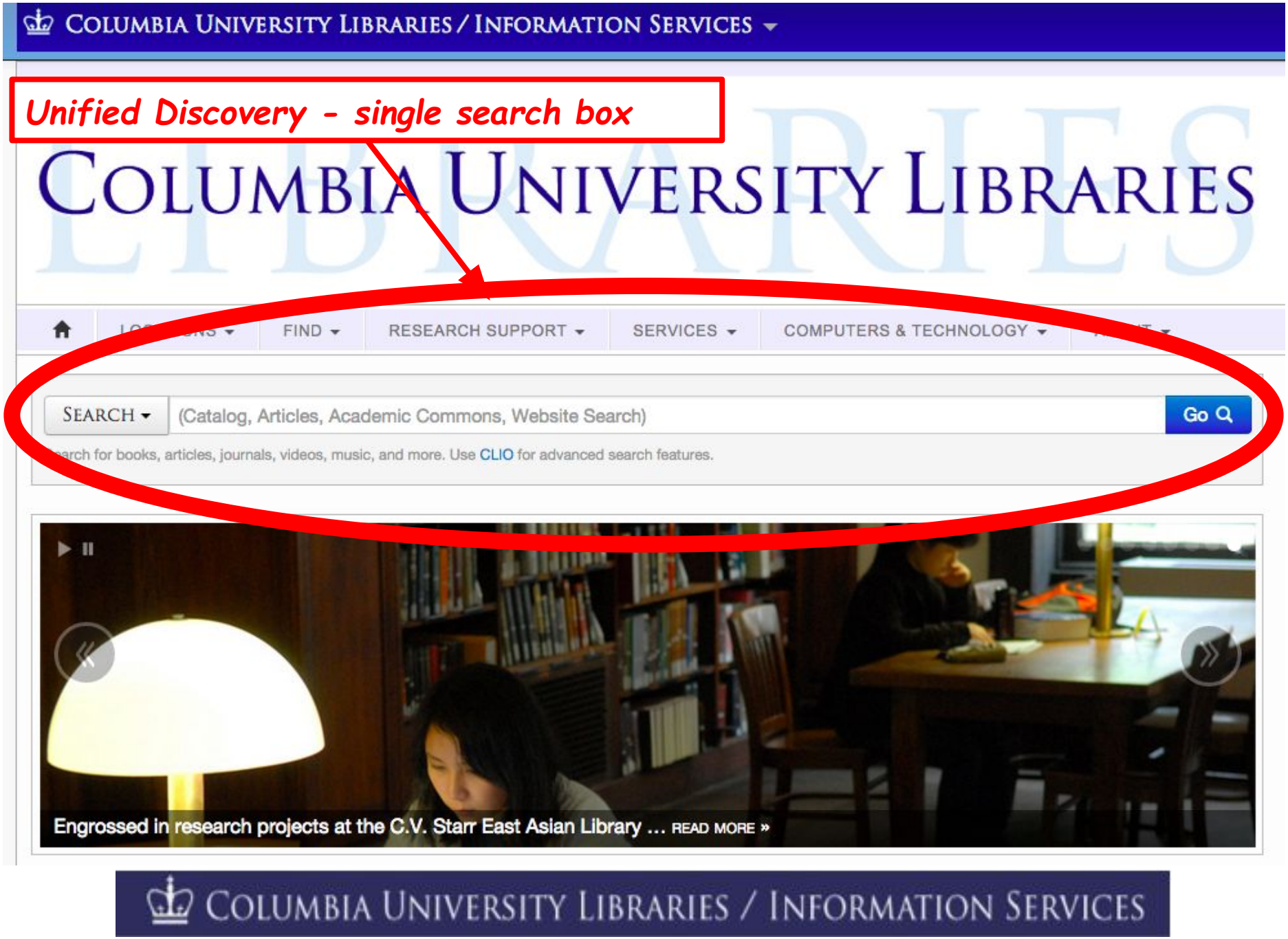
Blacklight Does Much More...

- Relevance ranking ✔
- Stemming ✔
- Field weighting ✔
- Spell-checking ✔
- Refining original queries ✔
- Support for popular query operators ✔
- Boolean bag o' goods ✔
- Flexible default query processing
- In-line query limiters
- Duplicate detection
- Sort Flexibility ✔
- Character Sets (Unicode UTF-8) ✔
- Faceting ✔
- Advanced search ✔
- Easily customized search results
- Best bets ✔
- Search logging and reporting ✔
- Well-rounded administrative interface

Blacklight Community

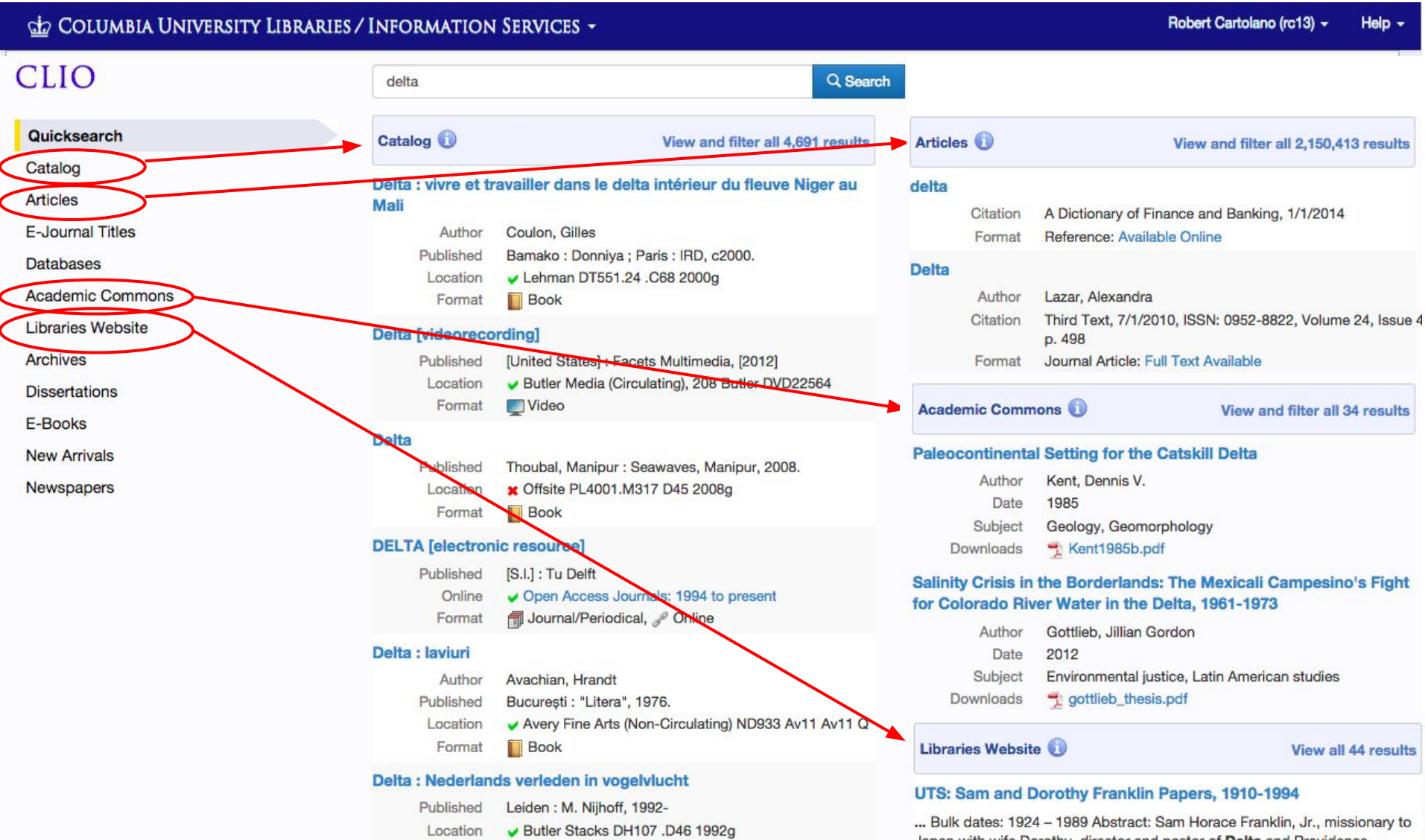
- 469 Members (328 in April 2014)
  - blacklight-development@googlegroups.com
- Over 100 Organizations (estimated)
- Part of the Hydra Project

200 HydraNauts from 60 institutions at Hydra Connect 2015
https://wiki.duraspace.org/display/hydra/Hydra+Connect+2015


COLUMBIA UNIVERSITY LIBRARIES / INFORMATION SERVICES
Unified Discovery - single search box
COLUMBIA UNIVERSITY LIBRARIES
FIND
RESEARCH SUPPORT
SERVICES
COMPUTERS & TECHNOLOGY
SEARCH
(Catalog, Articles, Academic Commons, Website Search)
Go
arch for books, articles, journals, videos, music, and more. Use CLIO for advanced search features.
Engrossed in research projects at the C.V. Starr East Asian Library ... READ MORE »
COLUMBIA UNIVERSITY LIBRARIES / INFORMATION SERVICES

CLIO

http://clio.columbia.edu/



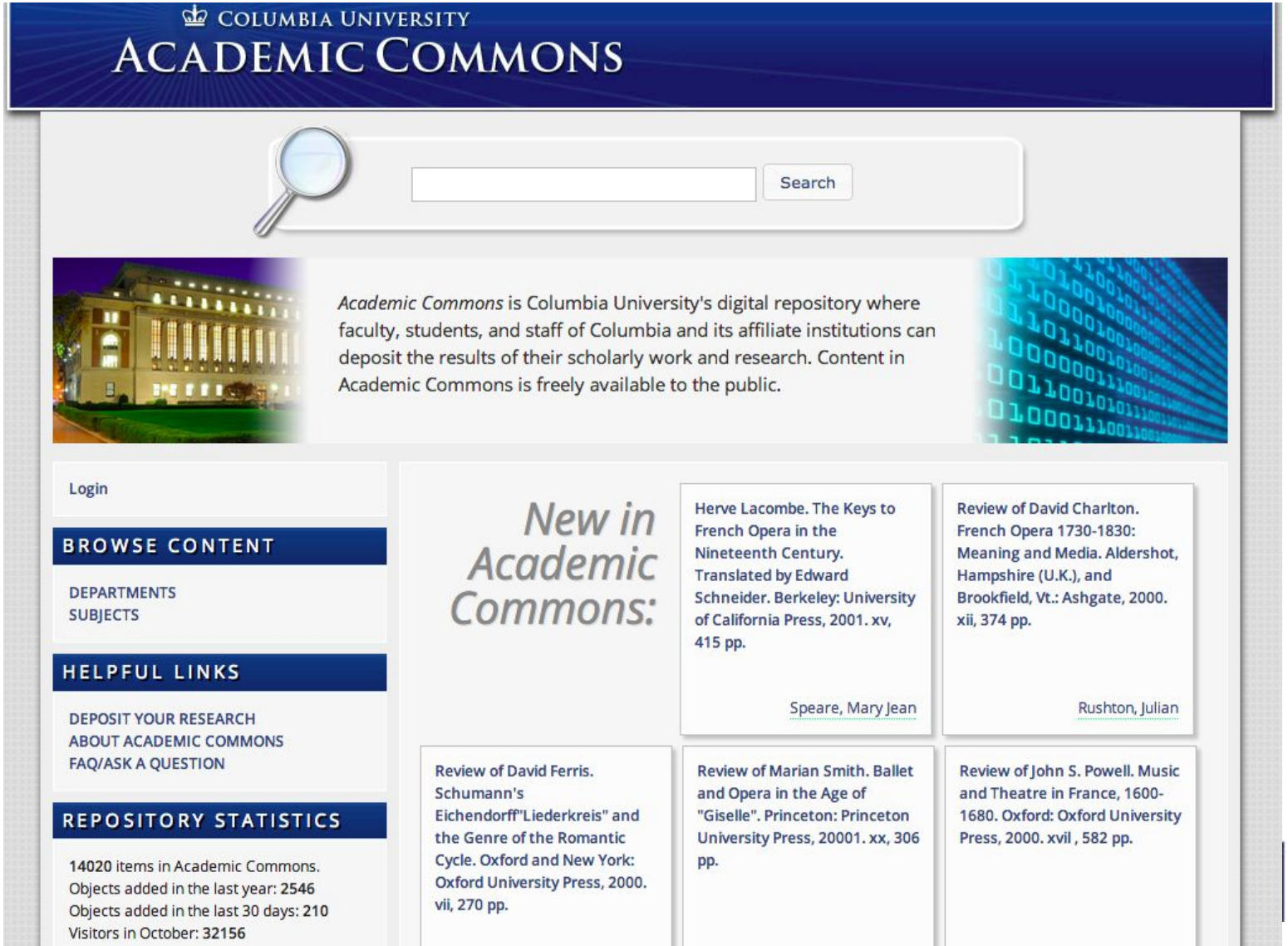
CLIO Technologies

https://blogs.cul.columbia.edu/clio/

- Blacklight/SOLR front-end (Library infra)
- Ex Libris Voyager back-end (CUIT infra)
- EZProxy proxy service for remote, authenticated access to e-resources (Library infra)
- Summon Article Search - 1.4 billion articles
- Link Resolver - 2M+ permanent URLs
- CY2015: 1.9M sessions, 9.9M page views
- Includes:
  - Integration w/Law Library collection (400,000+ items)
  - HathiTrust ebook titles

Academic Commons

http://academiccommons.columbia.edu/



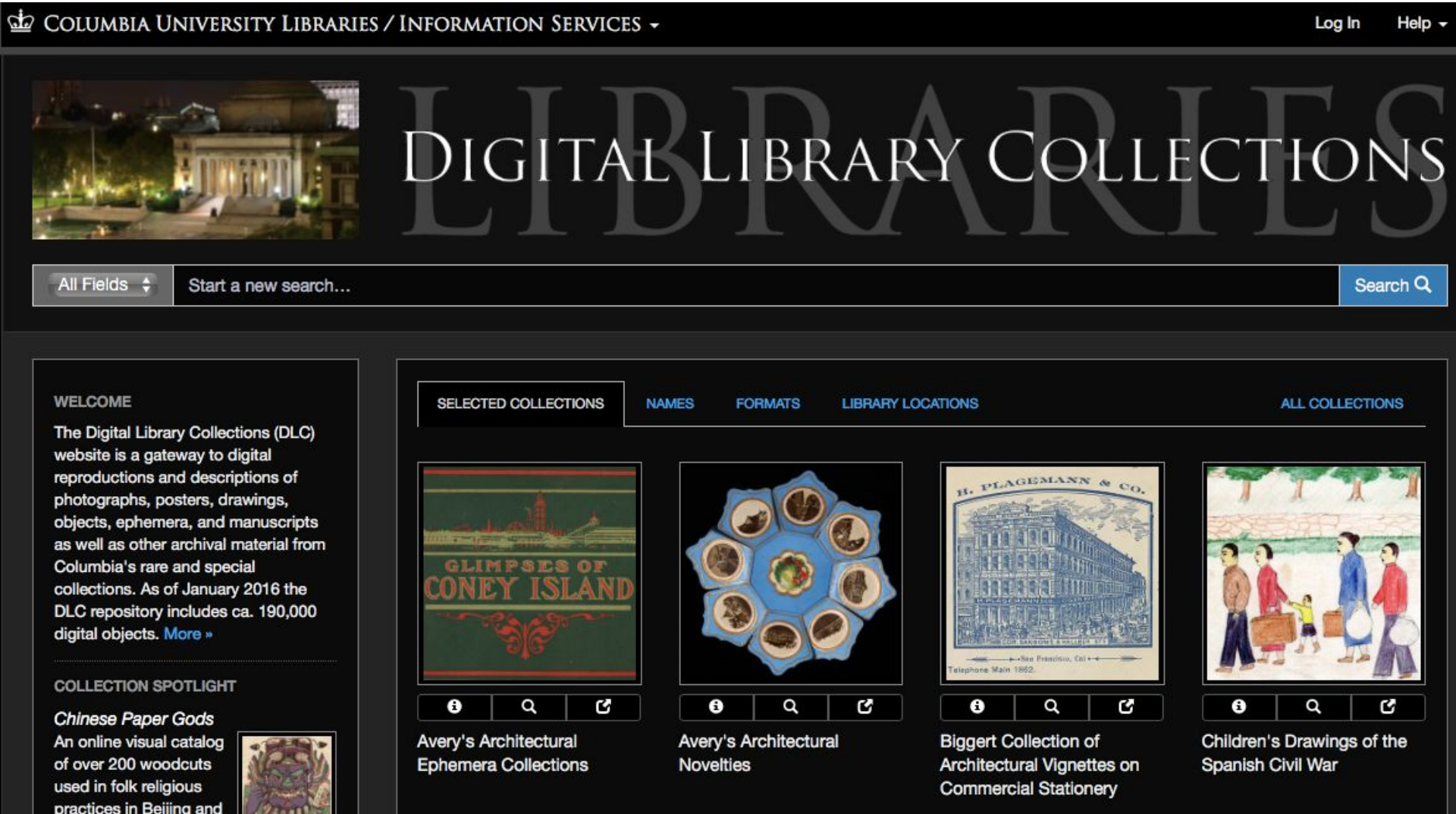
Academic Commons

http://academiccommons.columbia.edu/

- Columbia's repository for faculty, students
- 19,000 items
- Dissertations, journal articles, more
- Blacklight/SOLR front-end
- Fedora repository back-end

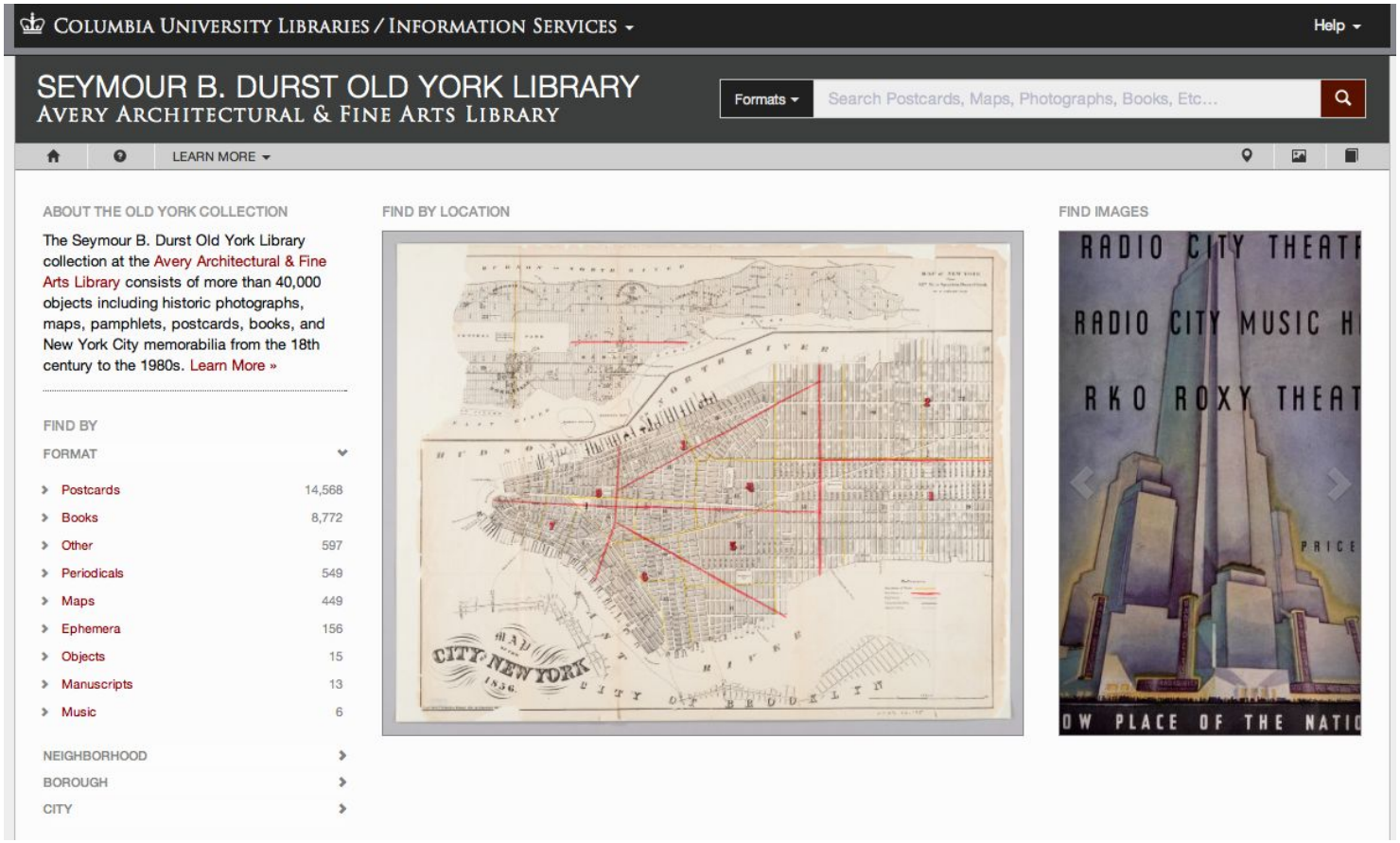
Digital Library Collections

http://dlc.library.columbia.edu/



Durst Old York Library

http://library.columbia.edu/locations/avery/seymour-b-durst-old-york-library.html

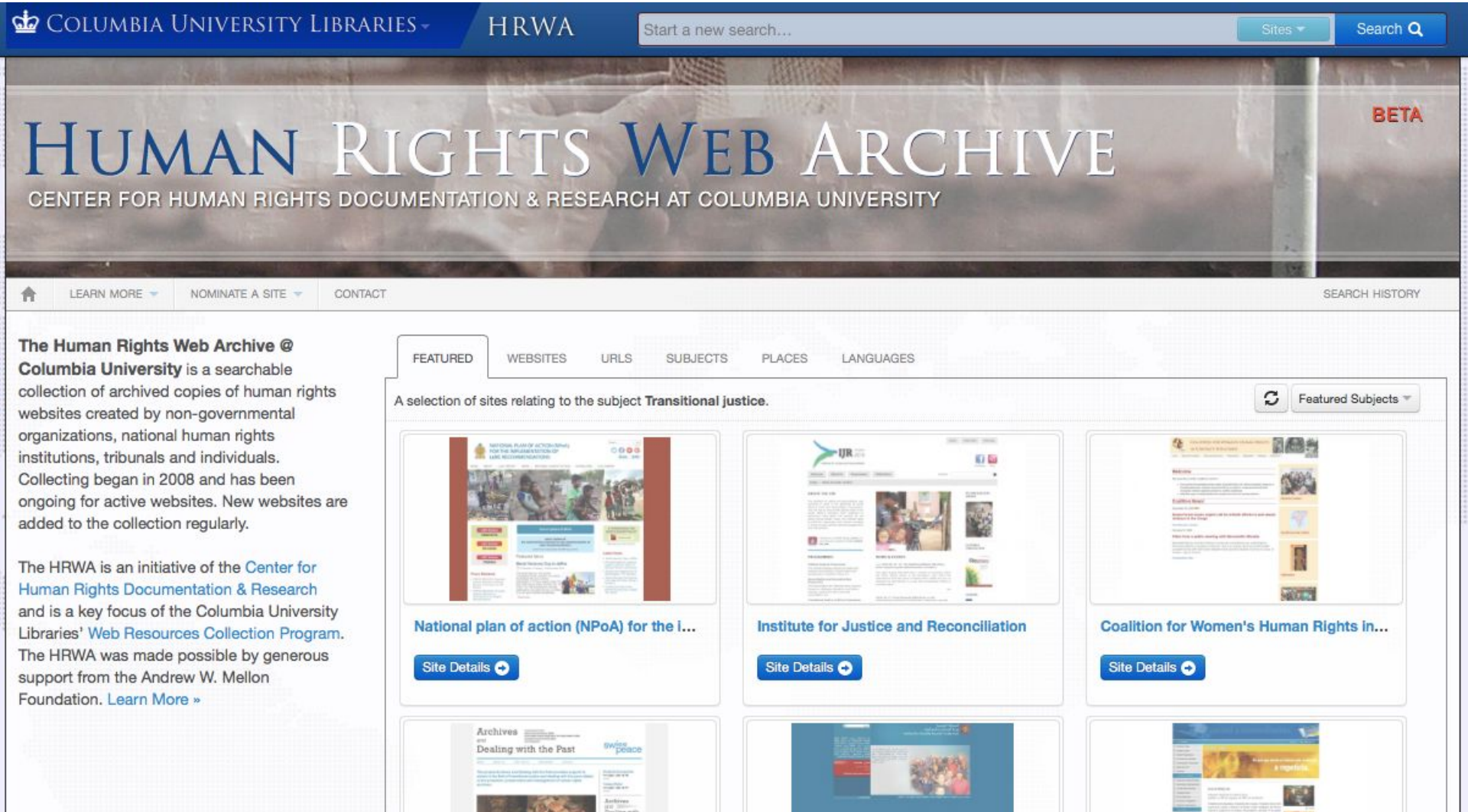


COLUMBIA UNIVERSITY LIBRARIES / INFORMATION SERVICES

Web Archiving

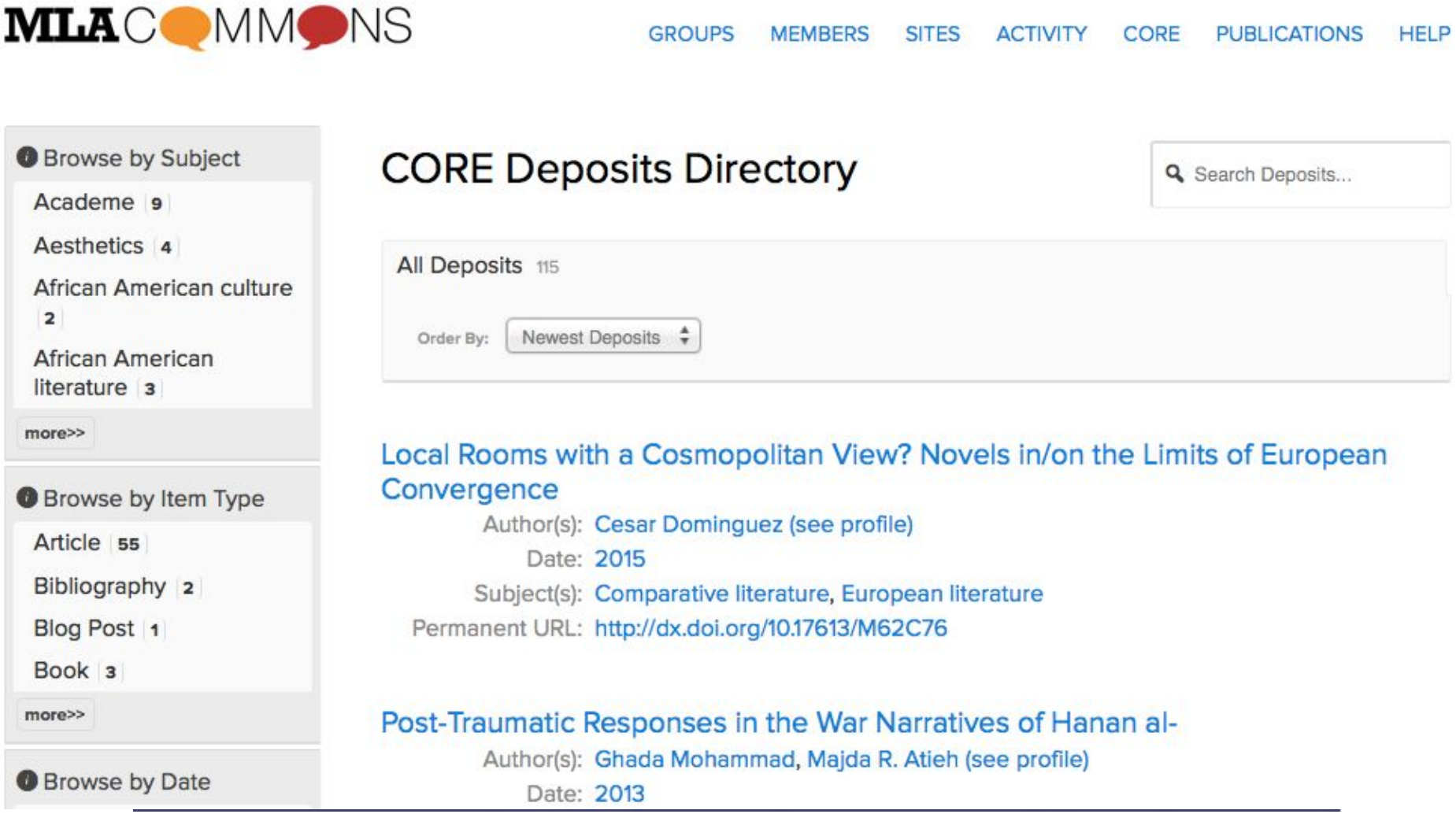
600+ sites, 50M indexed pages, 9TB, Internet Archive

http://hrwa.cul.columbia.edu/



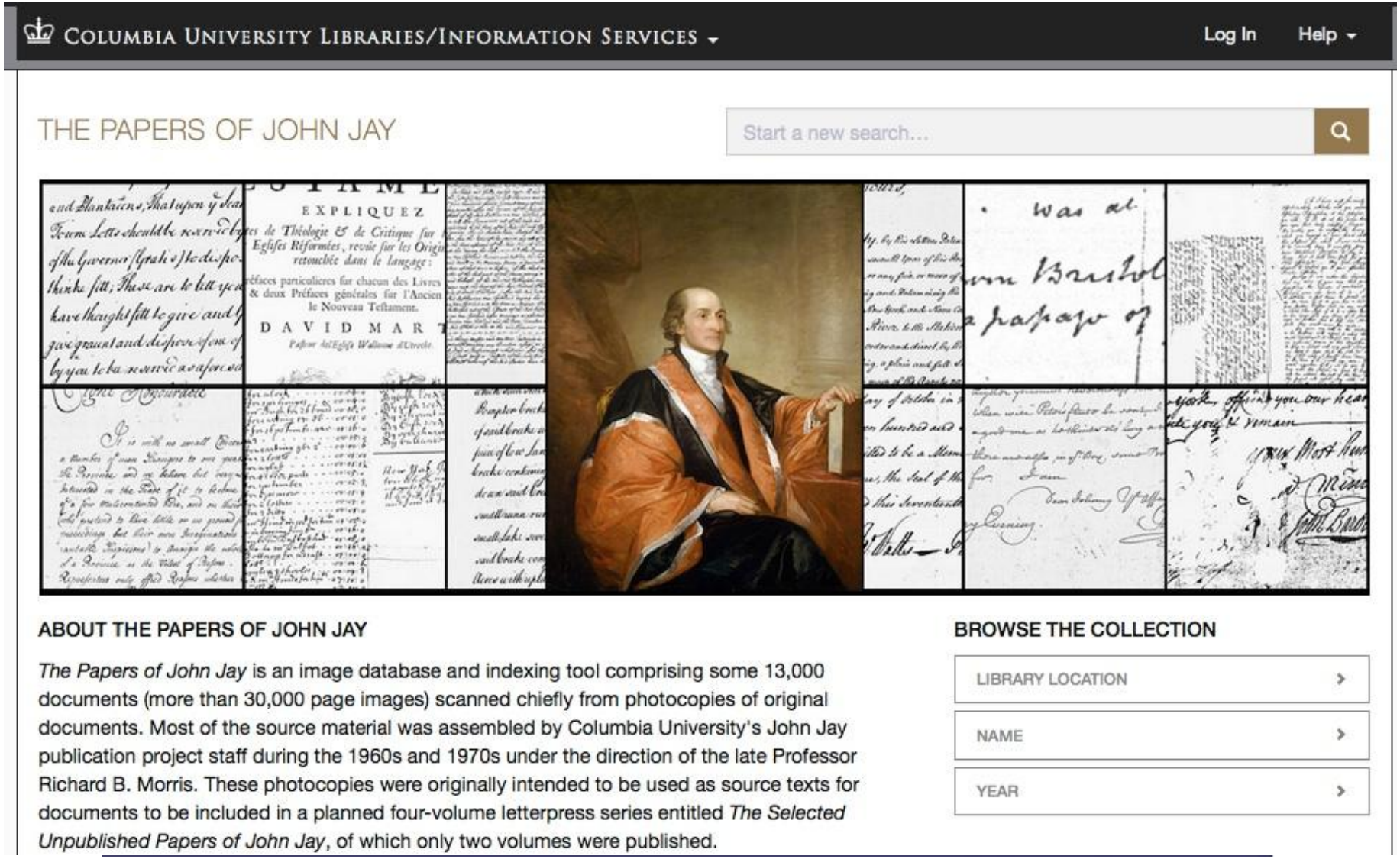
MLA Core Beta

Beta Site at: https://commons.mla.org/core/
Press Release: https://commons.mla.org/core/press-release/



Jay Papers Digital Collection

Sneak Preview at: https://dlc.library.columbia.edu/jay



Ford IFP Archive

Sneak Preview at: https://dlc.library.columbia.edu/ifp

Additional Partnerships & Services

- ICPSR - http://clio.columbia.edu/catalog/1360887
- arXiv.org - http://arxiv.org/
- Social Sciences Data: http://library.columbia.edu/locations/dssc/data.html
- HathiTrust: http://www.hathitrust.org/
- Internet Archive: https://archive.org/details/ColumbiaUniversityLibraries

The Five Opens

- **Open Code** – sustainable platforms
- **Open Data Standards** – manage over time
- **Open API/Protocols** – foster interoperability
- **Open Communities** – go further together
- **Open License** – share knowledge broadly

Based on Mara Hancock's Three Opens

Questions