

SAVE OUR SITE

Web Archiving at the IU Libraries University Archives

Mary Mellon, Digital Archivist

November 2, 2022

marmello@iu.edu

Wednesday Noon Digital Scholarship Series



#WNDSS @iubarchives

OUTLINE

Why we collect

What we collect

How we collect – Archive-It

Challenges

Current Projects

What's next?

Indiana University *School of Public and Environmental Affairs*

BLOOMINGTON

Undergraduate Degrees

Environmental Science
Public Affairs
Public Health

Graduate Degrees

Environmental Science
Public Affairs

Doctoral Degrees

Environmental Science
Public Affairs
Public Policy

GARY

Undergraduate Degrees

Criminal Justice
Health Services Management
Public Affairs

Graduate Degree

Public Affairs

SOUTH BEND

Undergraduate Degrees

Criminal Justice
Public Affairs
Health Services Management

Graduate Degree

Public Affairs



INDIANAPOLIS

Undergraduate Degrees

Criminal Justice
Health Services Management
Public Affairs
Public Health

Graduate Degrees

Health Administration
Planning
Public Affairs

FORT WAYNE

Undergraduate Degrees

Criminal Justice
Health Services Management
Public Affairs

Graduate Degree

Public Affairs

KOKOMO

Undergraduate Degree

Criminal Justice

Making a world of difference

- [Indiana University Bloomington](#)
- [Indiana University Kokomo](#)
- [Indiana University Northwest Gary](#)
- [Indiana University South Bend](#)
- [Indiana University Purdue University Indianapolis](#)
- [Indiana Purdue Fort Wayne](#)

- [Statewide Faculty and Staff Directory](#)
- [Faculty Openings](#)



[Message from the President](#)
[IU Strategic Directions](#)
[About IU: Events, Facts & Student Information](#)
[The Campuses: IUB, IUPUI, IUE, IPFW, IUK, IUN, IUSB, IUS](#)
[Units & Services: Administrative Offices & Multi-campus Departments](#)
[Finding IU People](#)
[Searching IU by Topic](#)

WHY WE COLLECT

Screenshot of December 19,
1996 IU homepage



UNIVERSITY ARCHIVES MISSION

“As an archives, our primary mission is to collect, organize, preserve and make accessible records documenting Indiana University's origins and development and the activities and achievements of its officers, faculty, students, alumni and benefactors.”

RESEARCH VALUE

Curricula and course descriptions

News releases

Newsletters, reports, student publications

Faculty/staff listings and organizational charts

Meeting agendas and minutes

Speeches, announcements, and other communications

Some records are solely available in web formats

WEBSITES ARE EPHEMERAL



404

Page not found

The Page you are looking for doesn't exist or an other error occurred.
Go back, or head over to weebly.com to choose a new direction.

Everything to know about the IU graduate worker strike

The IDS has organized its coverage to catch you up and keep you up to date on this multi-year dispute.



WHAT WE CRAWL

Screenshot of IDS feature on IU graduate student worker strike, April 29, 2022

WHAT THE UA CRAWLS

Official websites and social media accounts of IU and IU Bloomington departments, schools, administrative units, student organizations

Websites for IU/IUB events, programs, courses, faculty, etc. that are part of the iu.edu and indiana.edu domains

Indiana Daily Student

Websites for events, programs, etc., that are cosponsored by IU but hosted on third-party sites

Websites, blogs, etc., of faculty and alumni whose papers we collect (*not university records)

Event-based, curated collection – COVID-19

WHAT THE UA DOESN'T CRAWL

IUPUI and regional campus websites

Transactional websites and service portals

Password-protected pages

Digital repositories and databases

External websites that discuss IU

Welcome to Indiana University Bloomington Libraries

[Libraries at Indiana University](#)

Information, Hours, IU Library WWW/Gopher sites

[Online Catalogs, Indexes, Electronic Journals & Texts](#)

[IUCAT](#) - IU libraries' Online Catalog

[Other Catalogs](#) - OCLC's WorldCat, worldwide library catalogs...

[Indexes](#) - Expanded Academic Index, ERIC, MLA, PsycINFO...

[Electronic Journals Collection](#) - Cataloged Titles, CIC...

[Electronic Texts](#) - Britannica Online, Poetry Databases...

[Local Databases](#) - Lilly Library Manuscripts Collection...

[Library Services](#)

Interlibrary loan, renewals, distance education, ask a librarian...

[Internet Favorites](#)

Internet resources and search tools.

■ [IUCAT](#) ■ [Ask a Librarian](#) ■ [Search this WEB](#) ■ [IUB Libraries Home](#) ■ [IUB Home](#) ■ [Staff Home](#)

Last updated: September 13, 1996

URL: <http://www.indiana.edu/~libweb/index.html>

Questions and Comments: libweb@www.indiana.edu

[Copyright](#) 1995, The Trustees of [Indiana University](#)

WEB ARCHIVING WITH ARCHIVE-IT

Screenshot of IU Libraries
homepage, December 19, 1996

ARCHIVE-IT

Browser-based, subscription service from Internet Archive

Includes data hosting and technical support

Non-profit, widely used by academic libraries

Great training resources

- Archive-It video curriculum: <https://support.archive-it.org/hc/en-us/articles/216489103-Archive-It-Video-Curriculum->
- Guide for new Archive-It users: <https://support.archive-it.org/hc/en-us/articles/360041250172-Guide-for-new-Archive-It-users->

IU LIBRARIES ARCHIVE-IT ACCOUNT

Websites -2006 – present

Social media: 2017-present

1.5 TB/year data budget

Currently used by University Archives and Government Information, Maps, and Microform Services (GIMMS)

Other campus archives have separate Archive-It accounts

Public interface: <https://www.archive-it.org/organizations/72>

UA ARCHIVE-IT COLLECTIONS

- Web sites

<https://www.archive-it.org/collections/219>

- Social Media accounts:

<https://www.archive-it.org/collections/8920>

- Coronavirus Days Web Archive:

<https://www.archive-it.org/collections/14261>

COLLECTION BASICS

Organized by “seed”, or URL e.g.

Can set crawls for individual seeds or groups of seeds

Test crawl option to prevent unintended data collection

Can schedule automatic crawls at various frequencies – ranges from multiple times per day to annual

Manual quality assurance (QA) workflow to troubleshoot capture issues e.g. missing style sheets, images, or videos

CRAWL FREQUENCY CONSIDERATIONS

Data budget

How often content on site changes

Time/resources for regular QA

Current events

Data de-duplication

Recommended crawl frequencies for university websites:

<https://uisapp2.iu.edu/confluence-prd/display/IULAO/Running+an+Archive-It+crawl>



COLLECTION MAINTENANCE

Content suppression (failed captures)

Updating URLs

Adjusting crawl frequency

Deactivating seeds

Tracking data budget

This page has not been archived here

Why might this happen?

- This page might not have been included in this organization's collecting plan.
- This page might have prevented Archive-It from collecting it.
- This page may have been collected but needs more time to appear in Wayback. If you just collected this page, please wait 24 hours for storage and indexing.

What can you do?

- Click to see other pages from <http://www.indiana.edu/>.
- You can also try searching for <http://www.indiana.edu/~libraru> on the [live web](#), in [other Archive-It collections](#), and the [Global Wayback Service](#).

CHALLENGES

(Standard Archive-It
error message)

DYNAMIC CONTENT



<https://www.nytimes.com/2020/04/07/style/internet-archive-library-congress.html>

SOCIAL MEDIA

Rapidly changing code prevents consistent capture of social media pages

Can be data hogs

Difficult to track creation, deactivation of accounts



SOCIAL MEDIA: IU BICENTENNIAL

Instagram

- Used Conifer to capture Instagram page
- Account data also exported through Instagram


Twitter

- Captured through Archive-It
- Account data also exported through Twitter

Facebook

- Partially captured through Archive-It
- Unable to export account data due to departure of account owner

FLASH RETIREMENT



The Indiana University Art Museum was unable to detect the latest version of the Flash Player on your computer.

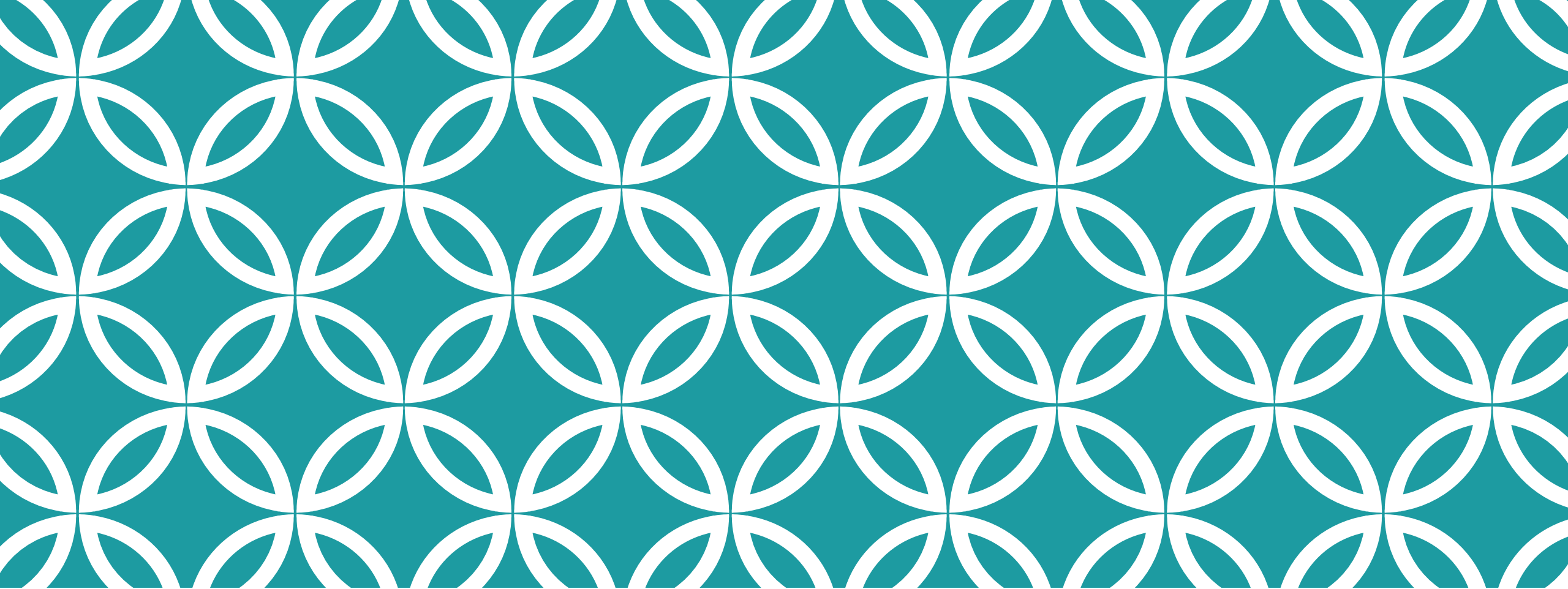
In order to take advantage of the new enhancements to the IUAM website, you should upgrade your computer with the latest version of the Flash Player. This free, one-time upgrade to your computer will install in less than 1 minute and is available from [Macromedia](#).

If you think you have received this message in error, and you have the latest version of Flash installed, [click here](#).

The Indiana University Art Museum
1133 E. 7th Street | Bloomington, IN 47405-7509

Telephone: 812.855.5445
Email: iuam@indiana.edu

Gallery Hours:
Tuesday - Saturday, 10am to 5pm
Sunday, 12pm to 5pm



CURRENT PROJECTS





WAYBACKFILL

Archive-It service that imports selected crawls from Internet Archive to individual collections

Recently added IU websites from 1996-2005

Currently identifying legacy URLs that need seeds and metadata

So much metadata

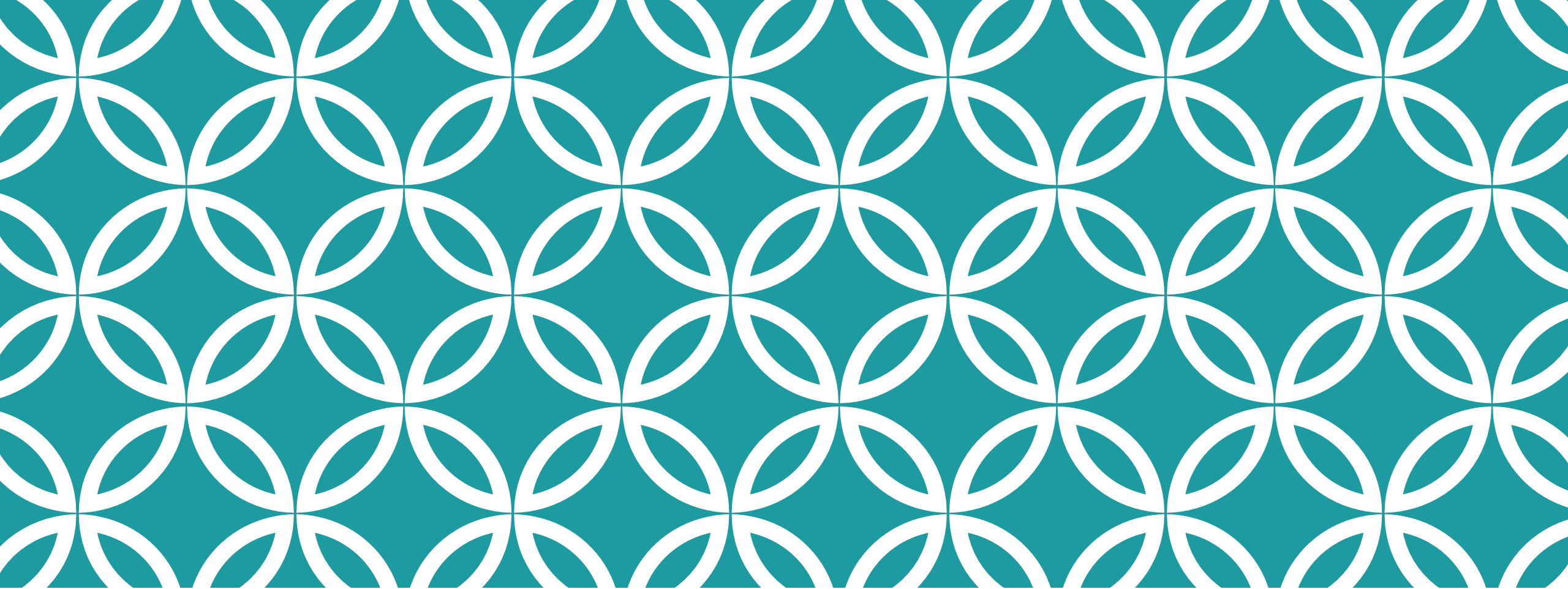


COLLECTION CLEANUP

Eliminating duplicate seeds (http vs. https)

“Continues” / “Continued by” linkages

Failed crawl suppression requests



WHAT'S NEXT? |

EDUCATION AND OUTREACH

Raise awareness of website and social media collections

- Preserve legacy content
- Historical research source
- Reference assistance
- Administrative uses

Build partnerships with IU Studios and other units that manage websites

Web archiving as records management

- University Records Retention and Disposition Policy (UA-18)

Archiving to support text and data mining, other research methods?



IS THERE TIME FOR A DEMO?



QUESTIONS?