Unsustainability and Retrenchment in American University Web Archives Programs

Gregory Wiedeman, University at Albany, SUNY
Amanda Greenwood, Union College

UAlbany Web Archives Program, 2012-2020

- 2012 - capturing albany.edu
  - Small part of University Archivist position
- 2016 - expanded to capture outside collecting areas
  - New York State political organizations
  - Some collecting for 2018 state elections
- Our progress was creating a maintenance backlog
  - Crawls would timeout and need re-scoping
  - Did not have time allocated for this

UAlbany's web archives program is typical

- Many college & university archives
  - 61% of respondents to NDSA Web Archiving survey
  - ~60% of Archive-It organizations in 2018
    - Next is public libraries at ~7%, subsidized
  - IA uses us as curators via the Archive-It program
- Many collect because of state records laws
  - Require the preservation of state records on the web

Web archives are “pathetically underfunded”

- David Rosenthal, “Losing the Battle to Archive the Web,” Digital Preservation Handbook (2017).
- 67% of respondents reported ~0.25 FTE in 2017 NDSA survey
  - Increase from 2016 and 2013

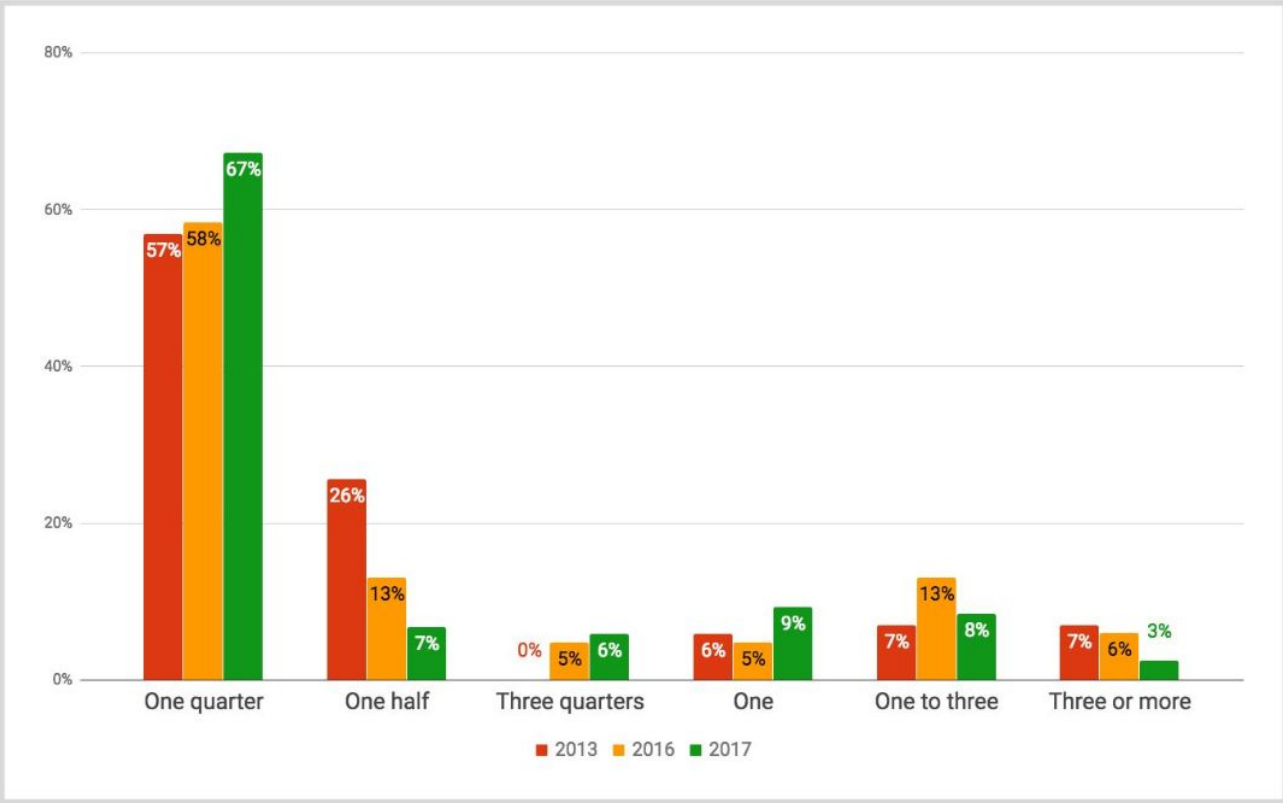


FIGURE 8: Full time staff (FTE) dedicated to Web archiving

2020

- FY 2020-2021 Anna Radkowski-Lee Web Archives Graduate Assistantship
  - Management of over 30 web archives collections
  - Better document NY state politics and death penalty
- 2020 February - Archive-It 7.0
  - Youtube-dl, data increases
- March-April leave
- Pandemic

2020-2021 Program Reset

- Let’s be more thoughtful about collecting
- Little known use outside of Wayback Machine
- Rescope active collections while staying under Archive-It budget
- Allocate for maintenance

Appraisal for Web Archives

- In traditional archives, appraisal saves time and labor
- Appraisal for web archives is different
  - It’s very easy to crawl a lot
  - It takes more labor to crawl *efficiently*
- Lack of staff doesn’t block collecting
  - But it creates waste that is hard to account for

Waste in web archives

- Web archives can be a wasteful collecting method
- Crawlers in traps, forever
- 2018 state candidates crawl with YouTube link
  - Archive-It Sites with automated scoping rules
- Cost externalities
  - Compute and storage are “cheap”
  - Cost of servers responding to crawler
  - This is physical infrastructure that creates emissions

Waste in web archives

- Access waste is hard to measure
- The less efficient web archives are, the harder they are to use
- The difficulties and lack of use for web archive is a direct reflection of the limited staff we devote to collecting
- Costs of waste and inefficiency in web archives deserves further study

The Web Archiving Experience

- Time-consuming job
  - Graduate student position = 20 hours/week
- Multiple test crawls usually needed
  - Sometimes lasted up to 4 weeks
  - Often no idea when they will finish
- Difficult to plan work around
- No control over website updates
- Your time commitment decided by website changes and Archive-It technology updates

The Web Archiving Experience

- No prior web development experience
- Required a lot of University Archivist's help with rescoping at first
- Needed to learn HTTP codes and parameters
- Difficult to understand host reports
  - Little to no documentation on why
  - Archive-It Help Center is helpful generally but not on any specifics

Maintenance inflation

- The web is more complicated, dynamic
- Arms race between web technology and archiving technology
- More tools, more options, more time
- The same websites take more time to scope even 3-5 years later

	2016-2017		2020-2021	
Collection	**Test crawls**	**Scoping time**	**Test crawls**	**Scoping time**
98 Acres in Albany	1	3 days	1	<1 day
Atlantic States Legal Foundation	1	<1 day	1	<1 day
CSEA	1	4 days	15	223 days
Correctional Association of New York	2	<1 day	1	<1 day
Empire Center	1	1 day	8	219 days
Environmental Advocates of New York	4	183 days	4	62 days
IUE/CWA Local 81359	2	72 days	6	107 days
National Association of Social Workers	1	1 day	10	204 days
NYCLU	2	87 days	7	234 days
NYS Council of School Superintendents	1	<1 day	2	82 days
New York State Republican Committee	1	<1 day	4	103 days
NYS Right to Life Committee	1	<1 day	4	189 days
NYS School Administrators Association	2	85 days	2	2 days
Parks & Trails NY	1	6 days	3	58 days
United University Professions	1	<1 day	6	207 days
Mean	1.5	29.9 days	4.9	112.9
Median	1	1	4	103

Maintenance inflation

- CSEA (https://cseany.org/)
  - 2016 February 9-13 (5 days)
    - 2 test crawls, ~2 ½ days each
    - 0 scoping rules
  - 2020 October-2021 April (6 months)
    - Initial test crawl in August 2020
    - 12 Test crawls, about ~3 ½ days each
    - 10 scoping rules

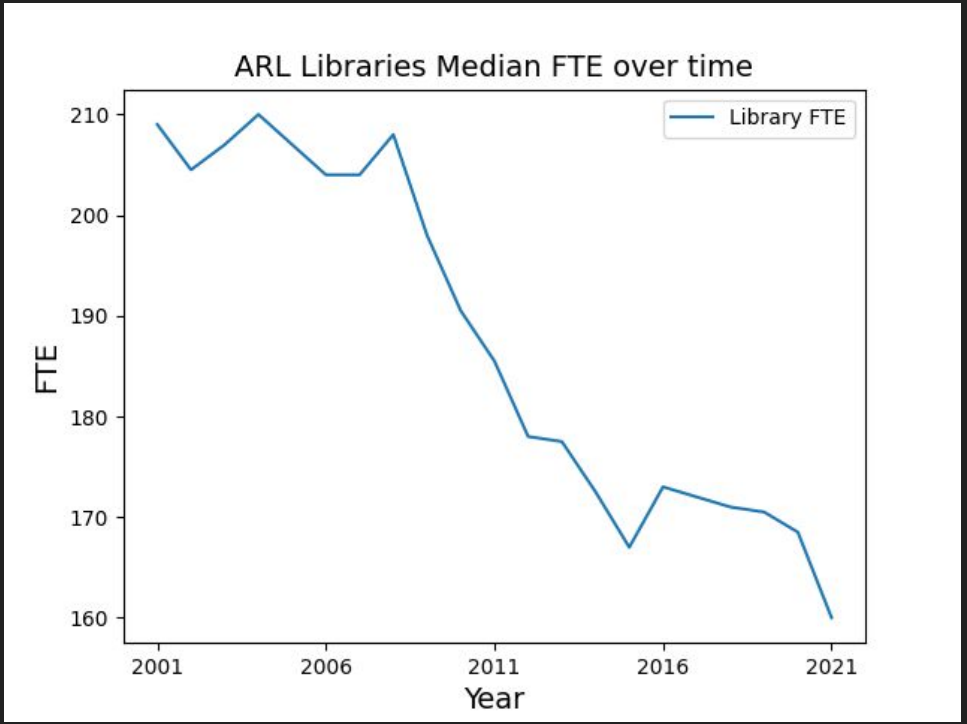
Rational case for increased staffing

- Maintenance for web archives grows over time
- Traditionally, many archives are collected years after they are created
  - Much of New York State politics and death penalty activism now happens on the web
  - Need to both collect 1970-1980s and today
  - Delay results in unacceptable loss

Staff loss in Association of Research Libraries (ARL)

- ARL libraries have major staff decreases since 2008
- Eira Tansey: Cycle of poverty

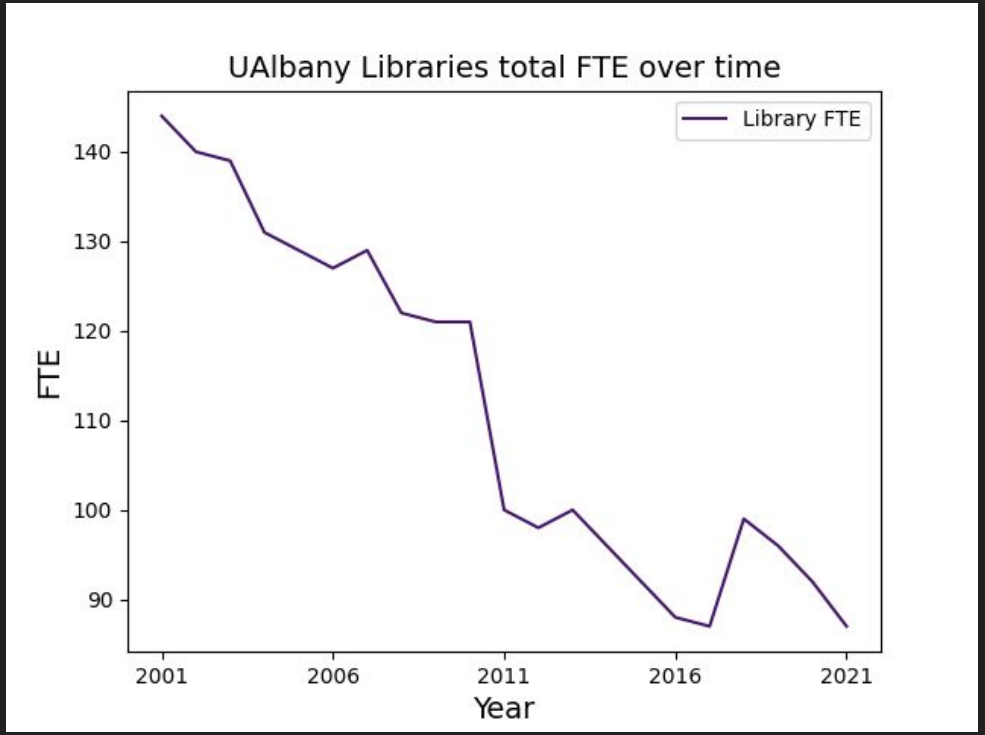
Eira Tansey, “Archives without archivists” in Reconstruction: Studies in Contemporary Culture, Vol. 16, No. 1 (Spring 2016).



Staff loss at UAlbany Libraries

150 FTE in 2000

75 FTE in 2022



The Broader Resource Environment

- UAlbany as a useful case study
- Administrators are re-allocating staff from Libraries to units that demonstrate impact on student recruitment & retention
- Potential 2026 enrollment drops
- Archives have a broader mission

Unit	FY 2014	FY 2023	Change	Pct Change
University Libraries	98.5	71.8	-26.7	-27.1%
Academic Affairs Support	210.0	231.8	21.8	10.4%
Academic Affairs	943.2	1018.3	75.1	8.0%
University	2137.8	2158.5	20.7	1.0%

More than just austerity

- It starts with post-2008 austerity
- Crisis in how many American colleges & universities are resourced
  - A lack of value for the mission of research libraries
- 2022 New York State budget surplus
  - Albany AI Supercomputing Initiative
    - $75 Million in permanent state funding
    - 49 tenure track positions in every college & school
    - 0 hires for University Libraries

Conclusions

- Lack of staff forces us to collect very wastefully
- Maintenance for web archives increases over time
- Over the past year UAlbany stopped collecting the web in our major collecting areas
- UAlbany’s experience is typical
- This is an existential crisis for research libraries
- Improvements in practice or technology won't address these structural forces
- It’s hard to see how this gets better

References

David Rosenthal, “Losing the Battle to Archive the Web,” Digital Preservation Handbook (2017).

David Rosenthal, “The Amnesiac Civilization: Part 2,” DSHR’s Blog (2017).

Matthew Farrell, Edward McCain, Maria Praetzellis, Grace Thomas, Paige Walker, “Web Archiving in the United States: A 2017 Survey,” NDSA (2017).

Jefferson Bailey, “Let’s put our money where our ethics are,” National Forum on Ethics and Archiving the Web (March 23, 2018).

Ben Goldman, “It's Not Easy Being Green(e): Digital Preservation in the Age of Climate Change,” in Archival Values: Essays in Honor of Mark Greene, eds. Christine Weideman and Mary A. Caldera (Chicago: Society of American Archivists, 2019): 174-187.

Keith L. Pendergrass, Walker Sampson, Tim Walsh, Laura Alagna, “Toward Environmentally Sustainable Digital Preservation,” *American Archivist* 82 (1) (2019).

Eira Tansey, “Archives without archivists” in Reconstruction: Studies in Contemporary Culture, Vol. 16, No. 1 (Spring 2016).

Eira Tansey, “(Un)common Knowledge,” Digital Library Federation (DLF) Forum (November 2, 2021).

Anam Mian and Holly Gross. ARL Statistics 2021. Washington, DC: Association of Research Libraries, 2023.