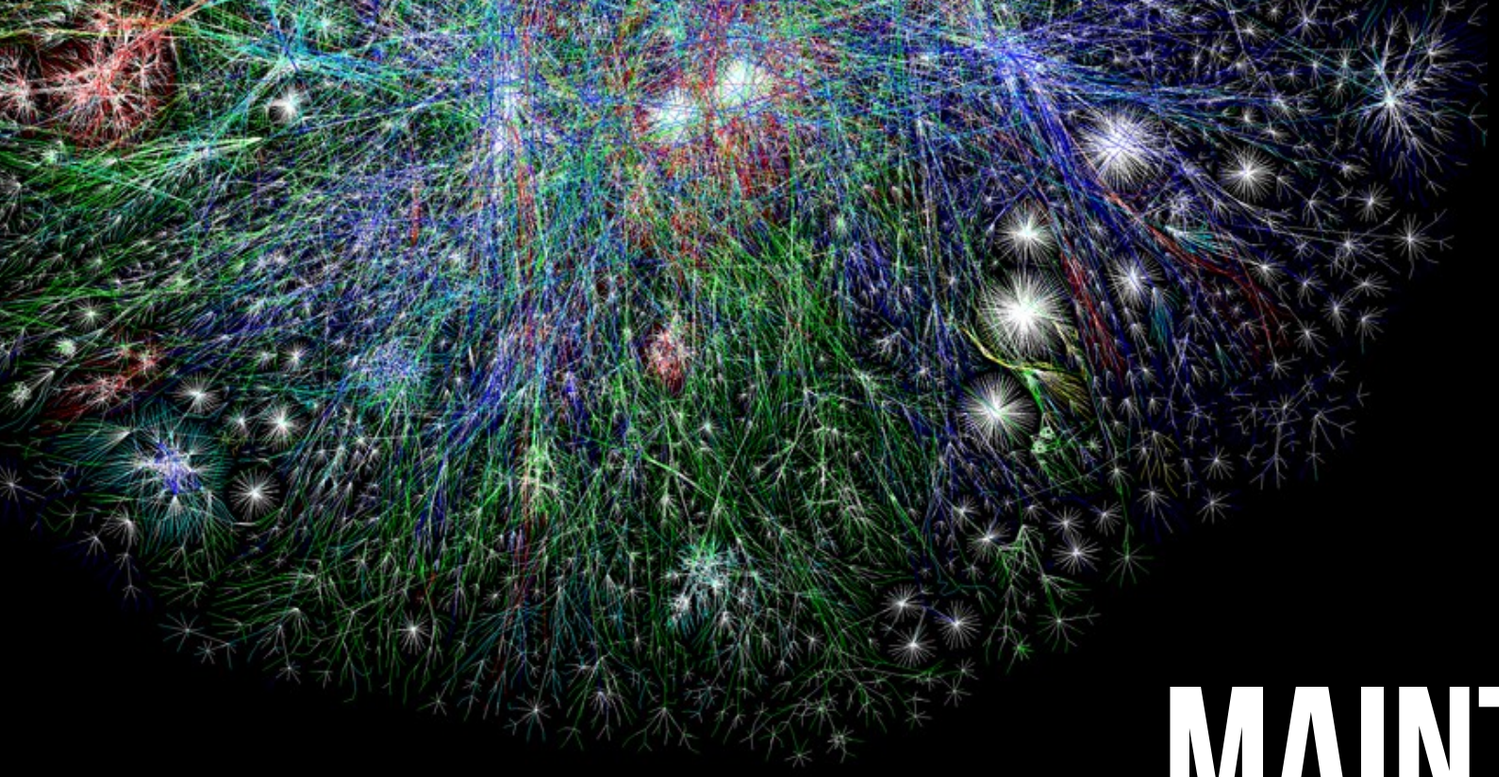


MAINTAINING THE WEB

Web Archiving, Labour and the Internet Archive

UNIVERSITY OF Southampton | Web Science DTC

Jessica Ogden (@jessogden)

Susan Halford

Les Carr

ASSOCIATION OF INTERNET RESEARCHERS (AOIR) CONFERENCE 2017 | TARTU, ESTONIA | 18 - 22 OCTOBER 2017


Library of Congress. Image: https://flic.kr/p/dUt1ir
WEB ARCHIVING


INTERNET ARCHIVE
The Internet Archive


Data Rescue NYC
EXIT
Image: Sam Hodgson,
NYTimes https://nyti.ms/2mxKAGM

- a lot of rhetoric around web archives

- don't know a lot about *how* web archiving operates at scale, for whom and for what purpose

- inherently selective, contingent and ≠ a copy

- increasingly used to make claims about the Web

PROBLEMATISING

https://www.newscientist.com/article/dn4434-internet-mapping-project-weaves-colourful-web/

CONCEPTUALISING MAINTENANCE & REPAIR

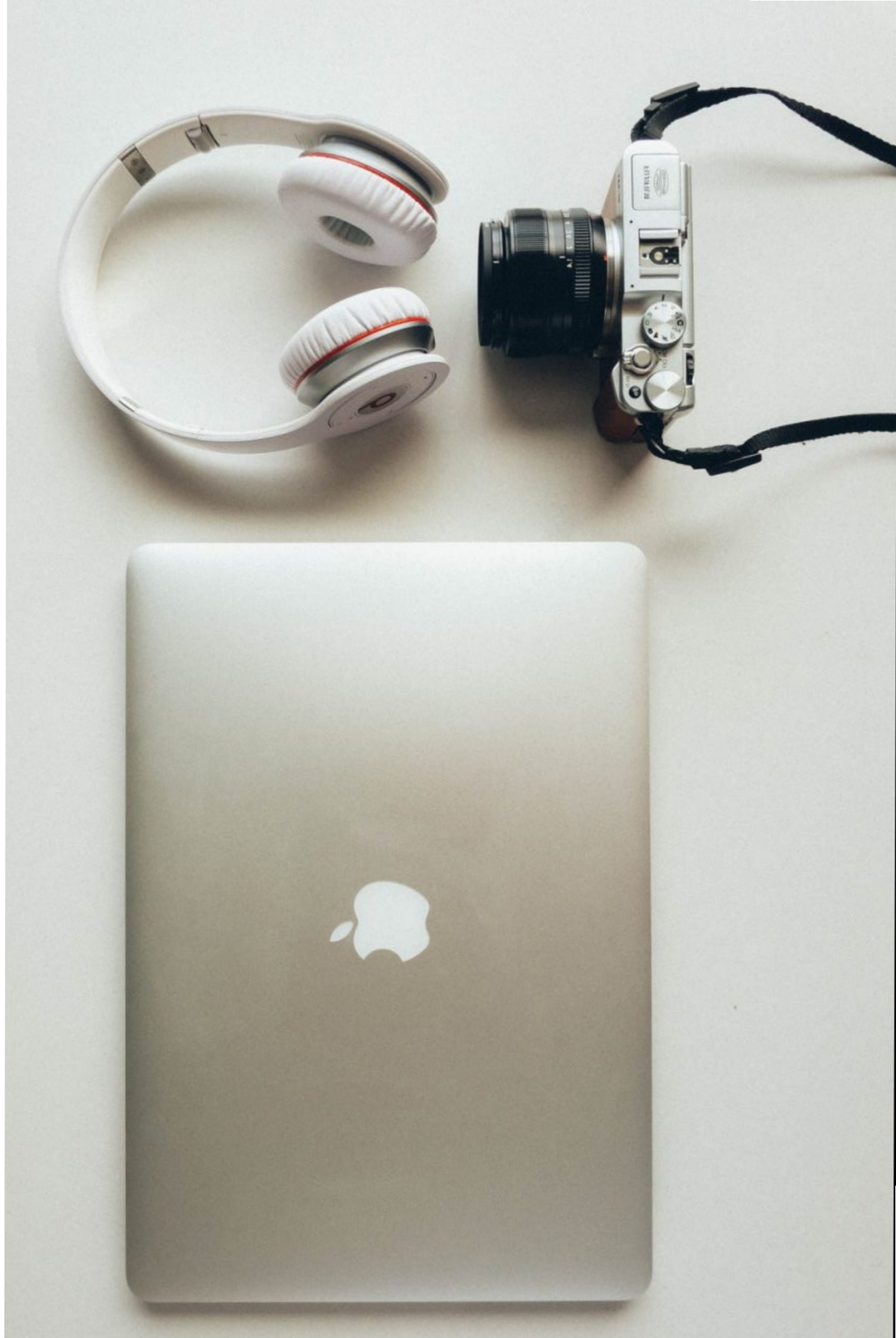
Web archiving as a form of maintenance work for the archived Web.

Web archiving as a form of repair for the Web.

https://archivingtheweb.me/2017/10/15/4s-and-me-part-i/

METHODOLOGY

- 16 un/semi-structured (ethnographic) interviews
- Observation records - what was done, made and used, what was said
- Documentary sources - wiki, policies, reports
- Two-tiered consent, 4 weeks



ARCHIVE

INTERNET

INTERNET ARCHIVE
WayBackMachine

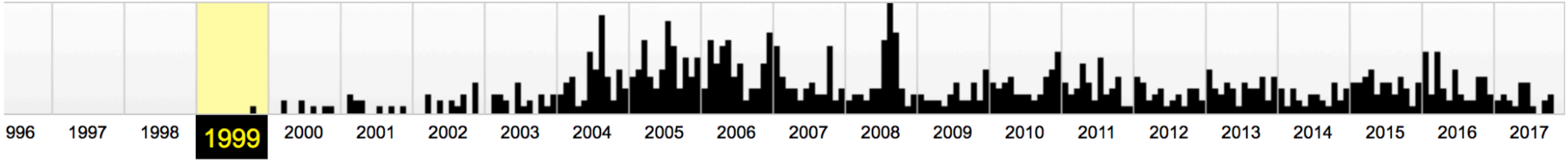
www.aoir.org

Explore more than 306 billion web pages saved over time

Saved **843 times** between October 12, 1999 and October 11, 2017.

Summary of aoir.org

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.



JAN						
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

FEB						
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28						

MAR						
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

APR						
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

a(0).i.r

a(0).i.r

a(0).i.r

a(0).i.r

a(0).i.r

a(0).i.r

a(0).i.r

association(of).internet.researchers

- academic association
- resource and support network
- critical and scholarly Internet research
- independent from traditional disciplines
- existing across academic borders
- international in scope

goals

faq

join

}

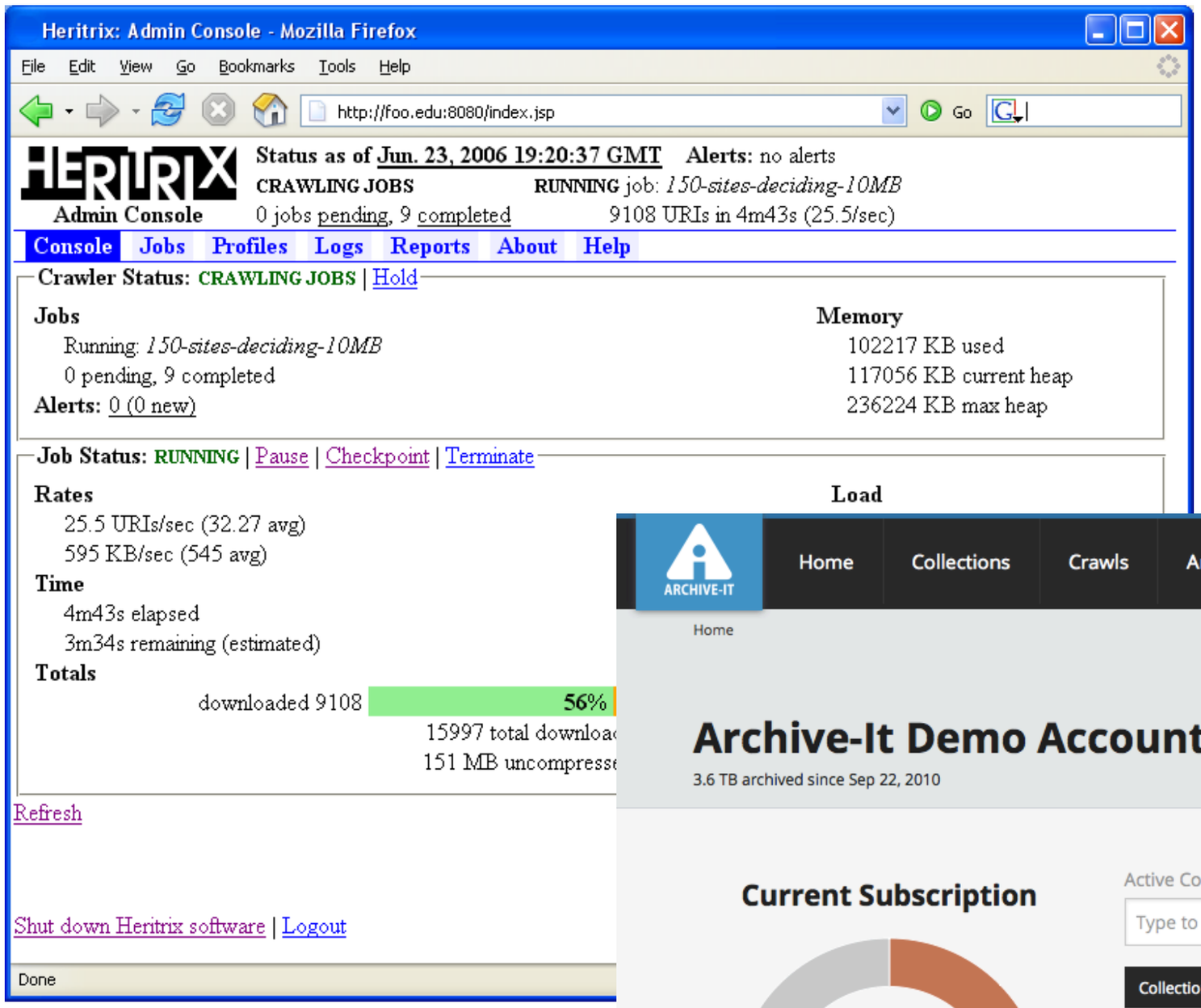
virtual graduate seminar (Spring 2000)

member database

downloadable flier (Acrobat PDF format)

listserve archive

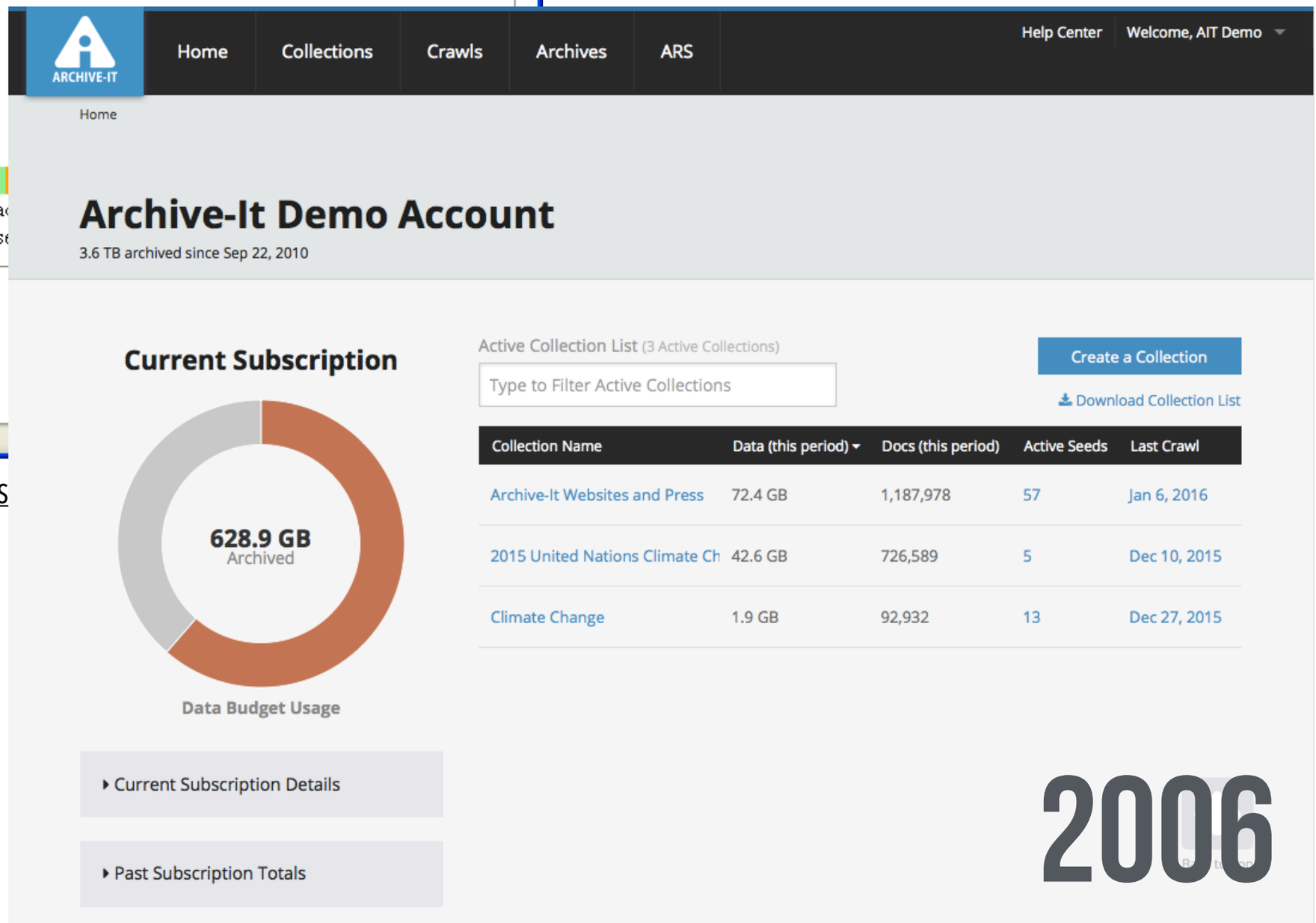
action items - your help needed!



FMCCOWN AT ENGLISH WIKIPEDIA, HTTPS://COMMONS INDEX.PHP?CURID=10700055

INTERNET ARCHIVE

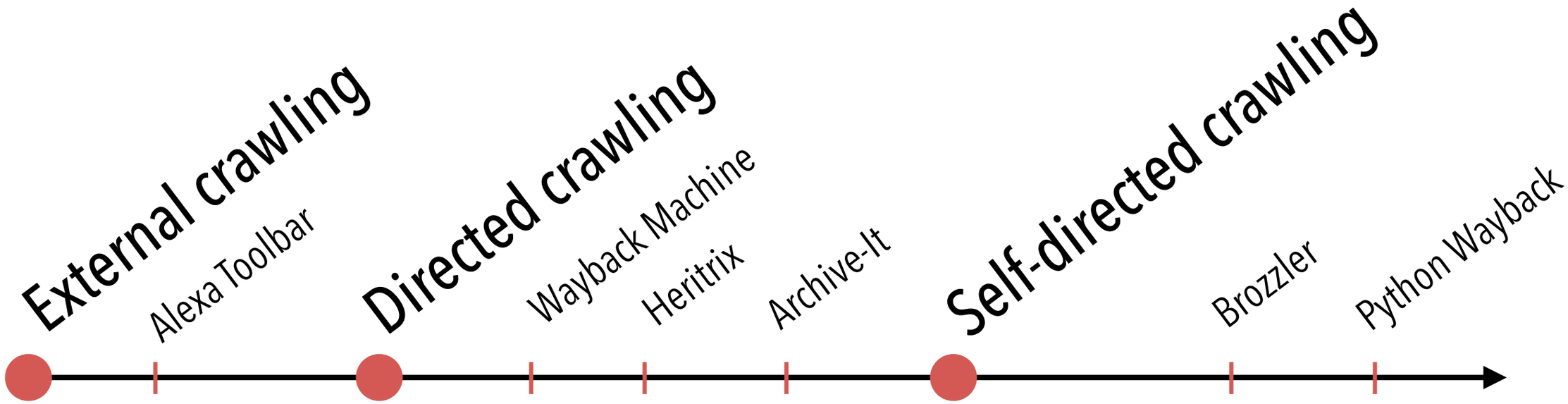
WayBackMachine



2006

MAPPING ROLES & ACTIVITIES

"strategically mishmash"



- User-directed

- Domain Crawls
- Thematic Crawls
- Event-based Crawls

'Librarians'
Engineers
Web Archivists

- Wide Crawls
- Survey Crawls
- Platform/service specific

Engineers
Web Archivists
'The Crowd'

*SCHEMATIC, NOT TO SCALE AND PROBABLY NOT EVEN CHRONOLOGICAL!

WEB ARCHIVAL LABOUR

COLLECTION WORK

SELECTION & MODERATION WORK

REASSEMBLY WORK

REPAIR WORK

*Work in Progress!

COLLECTION WORK

- e.g. Wide Crawls

```
2011-06-23T17:12:08.802Z   200       1299 http://content-5.powells.com/robots.txt LREP http://content-5.powells.com/cgi-bin/imageDB.cgi?isbn=9780385518635 text/plain #014 20110623171208574+225 sha1:YI
UOKDGOLGI5JYHDTXRFFQ5FF4N2EJRV - -
2011-06-23T17:12:09.591Z   200      15829 http://www.identitytheory.com/etexts/poetics.html L http://www.identitytheory.com/ text/html #025 20110623171208546+922 sha1:7AJUMSDTOMT4FN7MBFGGNJU3Z56MLCMW
- -
```

SELECTION & MODERATION WORK

- e.g. Domain Browser Tool

"The domain browser manual tool is for identifying undesirable domains. It's used to establish and prioritise 'shades of gray,' for example only crawl this site if there are no other sites to crawl. It's used as a ranking mechanism for prioritising domains based on time, resources and place in the queue, as certain important URLs can get blocked by many instances of unimportant URLs. […]Each domain/host is assigned a budget and the crawler is paused if it reaches its budget."

SELECTION & MODERATION WORK

- e.g. Domain Browser Tool

"What we did was hired half a dozen people - they would just go through it and get the top 30,000 hosts [...] and they go through 4-5,000, that's what one person can do in a month or so. And then we get actual human interaction to say yes, this is a good website. And then we would delete or modify or prioritise based on that input. So having humans actually spend a little bit of time at the top really helped. We'd love to do it further of course."

SELECTION & MODERATION WORK

- e.g. Studying link graphs

"I do a lot of link analysis where I study the hyperlink structure of our crawls and try to figure out in certain pockets, use some rank methodologies to figure out 'oh these are important resources,' for instance they have a lot of links to them or traffic is really high - let's seed the crawl with those. The most recent wide crawl I took the most linked to pages from every single website, so 230 million websites [...] and instead of crawling the Alexa top million, let's crawl this bit."

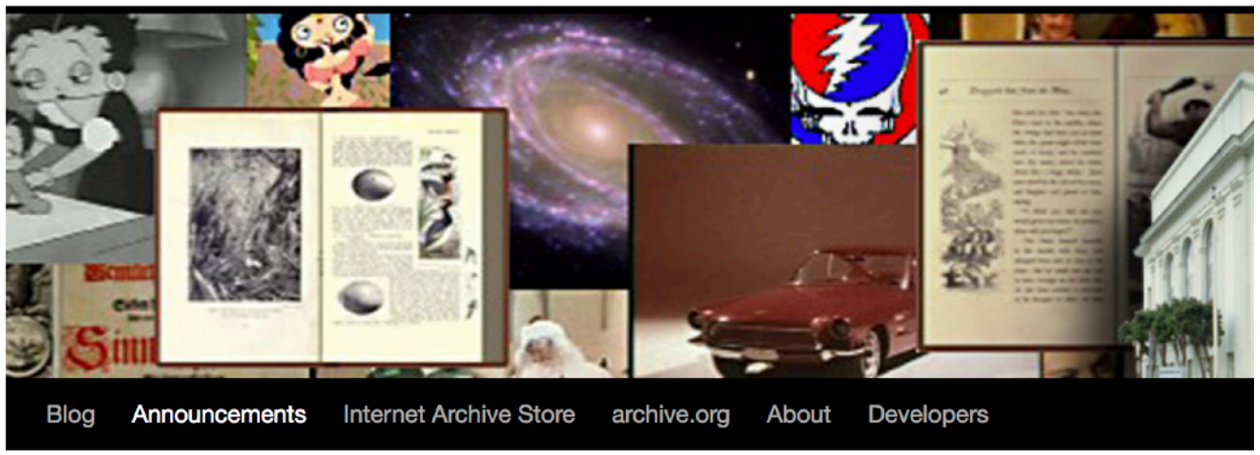
SELECTION & MODERATION WORK

- e.g. Studying link graphs

*"The way I think of it [...] there's three branches, there's **popularity**, there's **novelty** and there's the **risk of going away**. How do you achieve that balance? You want to get stuff that people are using - not just junk that is on the Web that you're just filling up the servers with that won't ever be found useful, like calendar pages, things like that, crap [...] - there's no novelty. It's new? We want to make sure it's preserved because it just came out, it's a new article, it's a new website. And then there's the risk of going away [...] - if you're going to shut down this service - Vine is going away - we jump in and crawl. So as we're crawling the Web can we do a good job of sort of achieving that balance? We don't quite know what the solution is to achieving that balance."*

‘GOOGLE FOR THE ARCHIVED WEB’

Internet Archive Blogs



Save Page Now

http://

SAVE PAGE

Capture a web page as it appears now for use as a trusted citation in the future.

Only available for sites that allow crawlers.

← In the news: Trump Archive, end-of-term preservation, & link rot

If You See Something, Save Son In the Wayback Machine

Posted on January 25, 2017 by Alexis Rossi

In recent days many people have shown inter copies of the web pages they care about most linked to – and they will continue to exist eve removed from the web.

There are several ways to save pages and whc Machine. Here are 6 of them.

[redacted]

Too bad archive.org waxes political; and far left at that.

I really appreciate the fact as a conservative Christian, I can find content here.
I wonder when that material will be removed because it is both Christian and Conservative?

One of the last bastions of public cyber space has gone off in HATE TRUMP land. How disappointing! Can’t we keep politics out?

https://blog.archive.org/2017/01/25/see-something-save-something/

SUMMARISING

- tensions between automation and manual labour reflect complexities of work

- values embedded at inflection points in selection, collection

- underlying role of power in shaping web imaginaries - for whom?

- future work

If you're interested in more:

Ogden, Jessica, Susan Halford, and Leslie Carr. 2017. 'Observing Web Archives: The Case for an Ethnographic Study of Web Archiving'. In Proceedings of WebSci'17, Troy, NY, USA., June 25–28, 2017. ACM. doi: 10.1145/3091478.3091506.

ACKNOWLEDGEMENTS

This work was supported by the **UK Engineering and Physical Sciences Research Council** and the **Web Science Centre for Doctoral Training**, Grant No. EP/G036926/1.

The authors would also like to thank the **Internet Archive and staff** for opening their doors and being so generous with their time and feedback.