OBSERVING ARCHIVES

Web Archival Labour as Socio-technical Practice

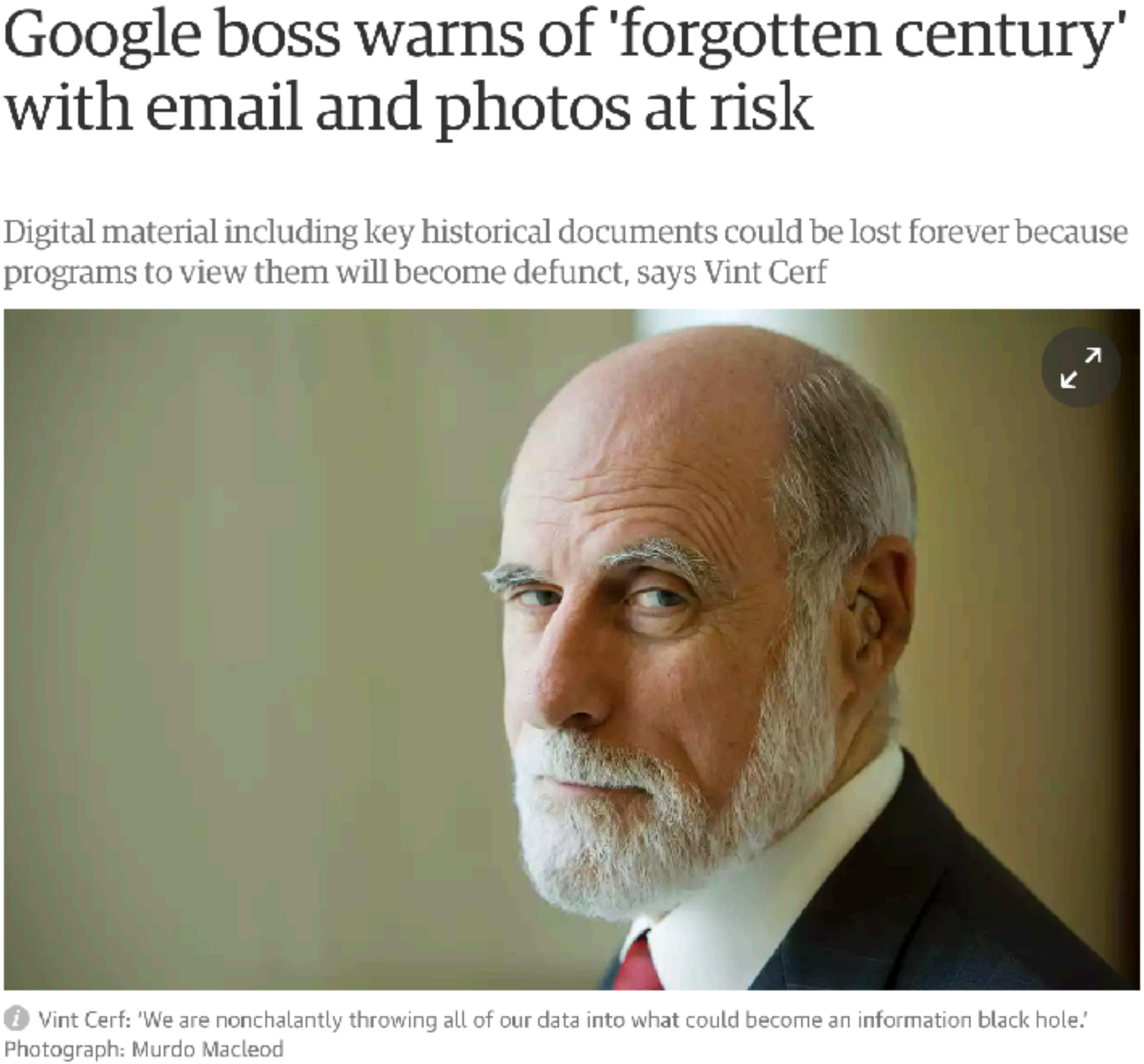
JESSICA OGDEN, SUSAN HALFORD, & LES CARR

UNIVERSITY OF Southampton | Web Science DTC

EPSRC

Research Councils UK Digital Economy Transforming Business and Society

SOCIETY FOR SOCIAL STUDIES OF SCIENCE (4S) CONFERENCE 2017 | BOSTON | 1 SEPTEMBER 2017



Google boss warns of 'forgotten century' with email and photos at risk

Digital material including key historical documents could be lost forever because programs to view them will become defunct, says Vint Cerf

Vint Cerf: 'We are nonchalantly throwing all of our data into what could become an information black hole.' Photograph: Murdo Macleod

Guardian: https://www.theguardian.com/technology/2015/feb/13/google-boss-warns-forgotten-century-email-photos-vint-cerf

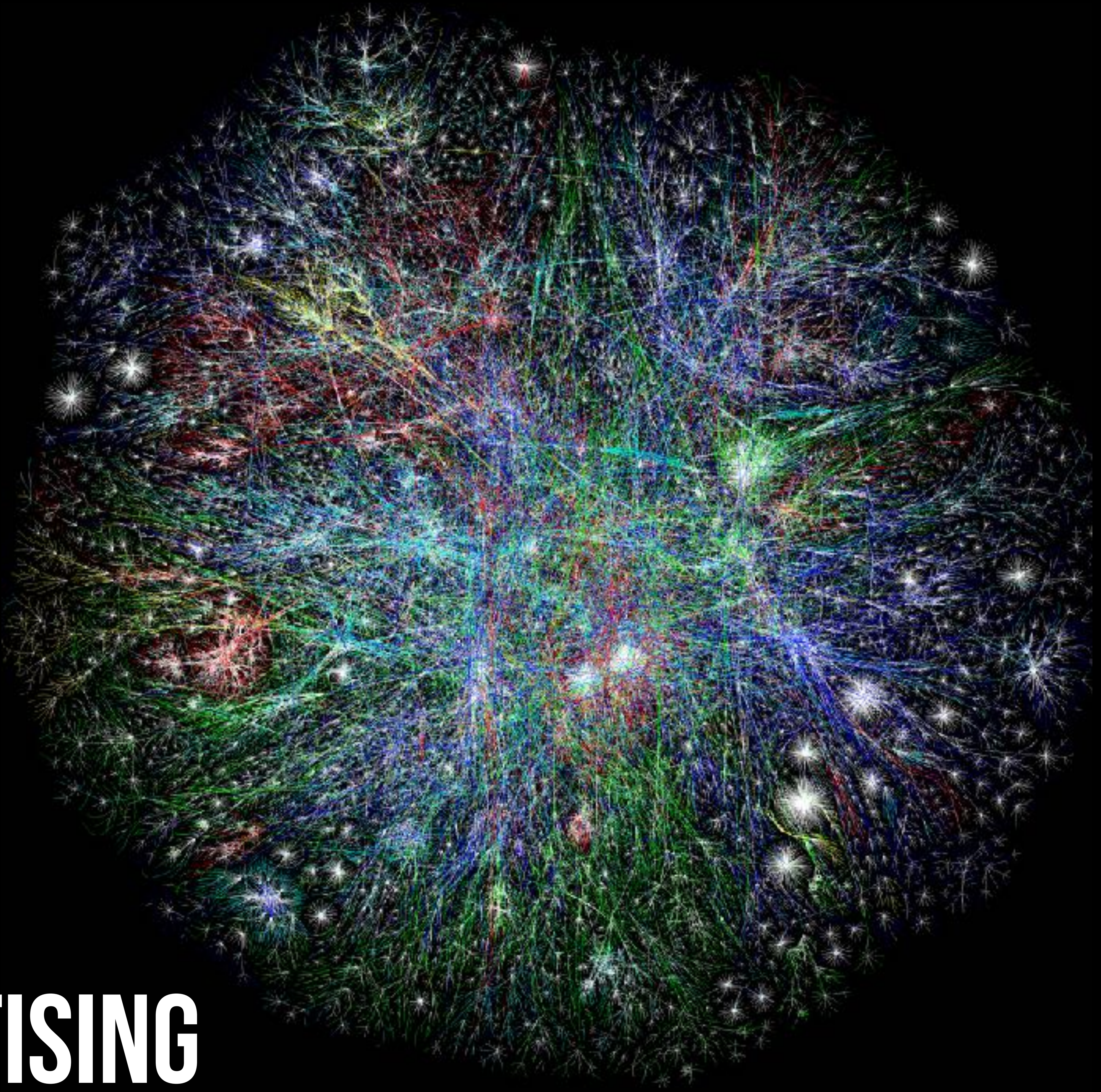
WEB ARCHIVING

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

The Internet Archive

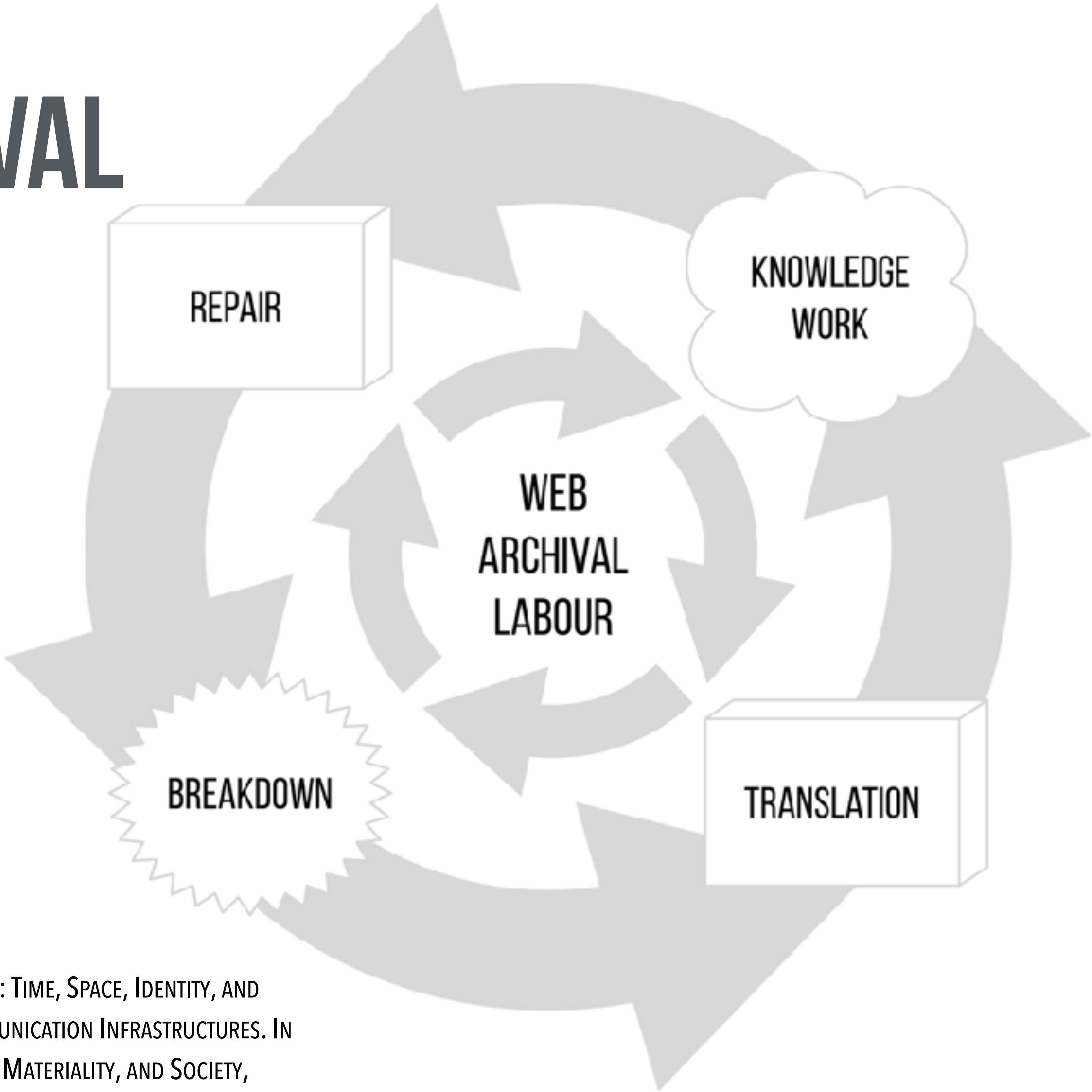
Data Rescue NYC

Image: Sam Hodgson, NYTimes https://nyti.ms/2mxKAGM

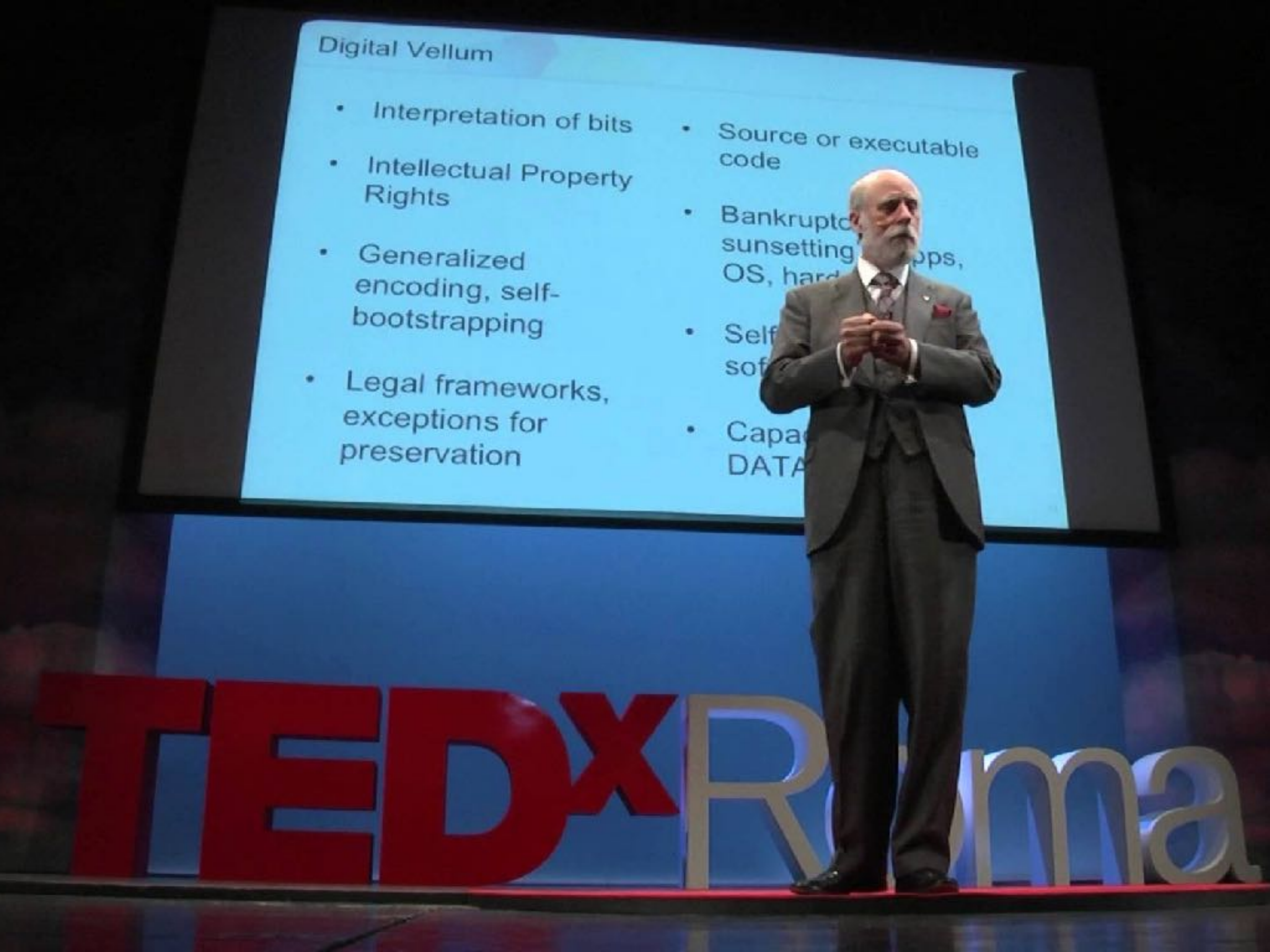


PROBLEMATISING

WEB ARCHIVAL LABOUR



Gregory J. Downey. 2014. Making Media Work: Time, Space, Identity, and Labor in the Analysis of Information and Communication Infrastructures. In Media Technologies: Essays on Communication, Materiality, and Society, Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot (Eds.). MIT Press, Cambridge, Massachusetts; London, England, 141–165.



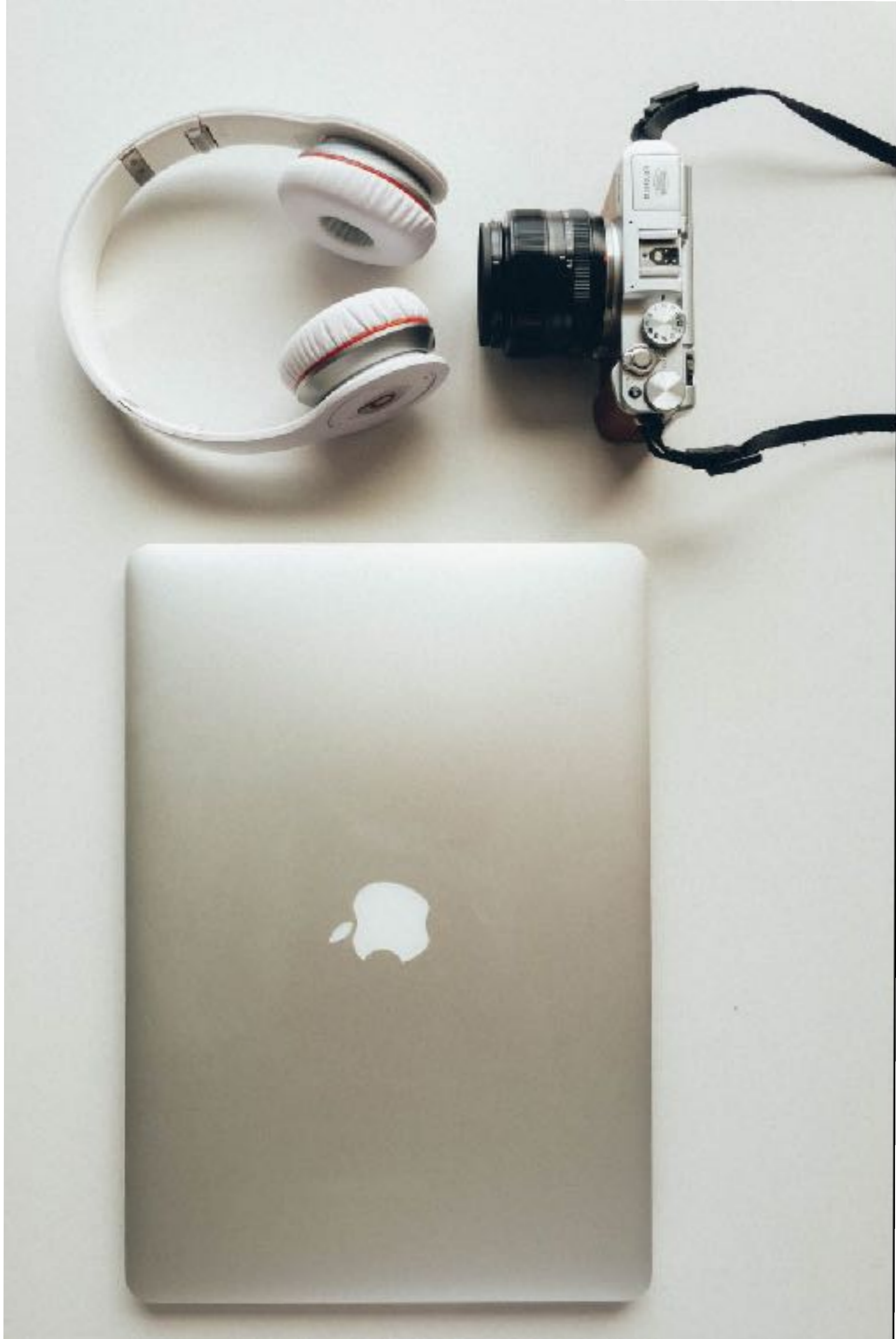
Digital Vellum

- Interpretation of bits
- Intellectual Property Rights
- Generalized encoding, self-bootstrapping
- Legal frameworks, exceptions for preservation

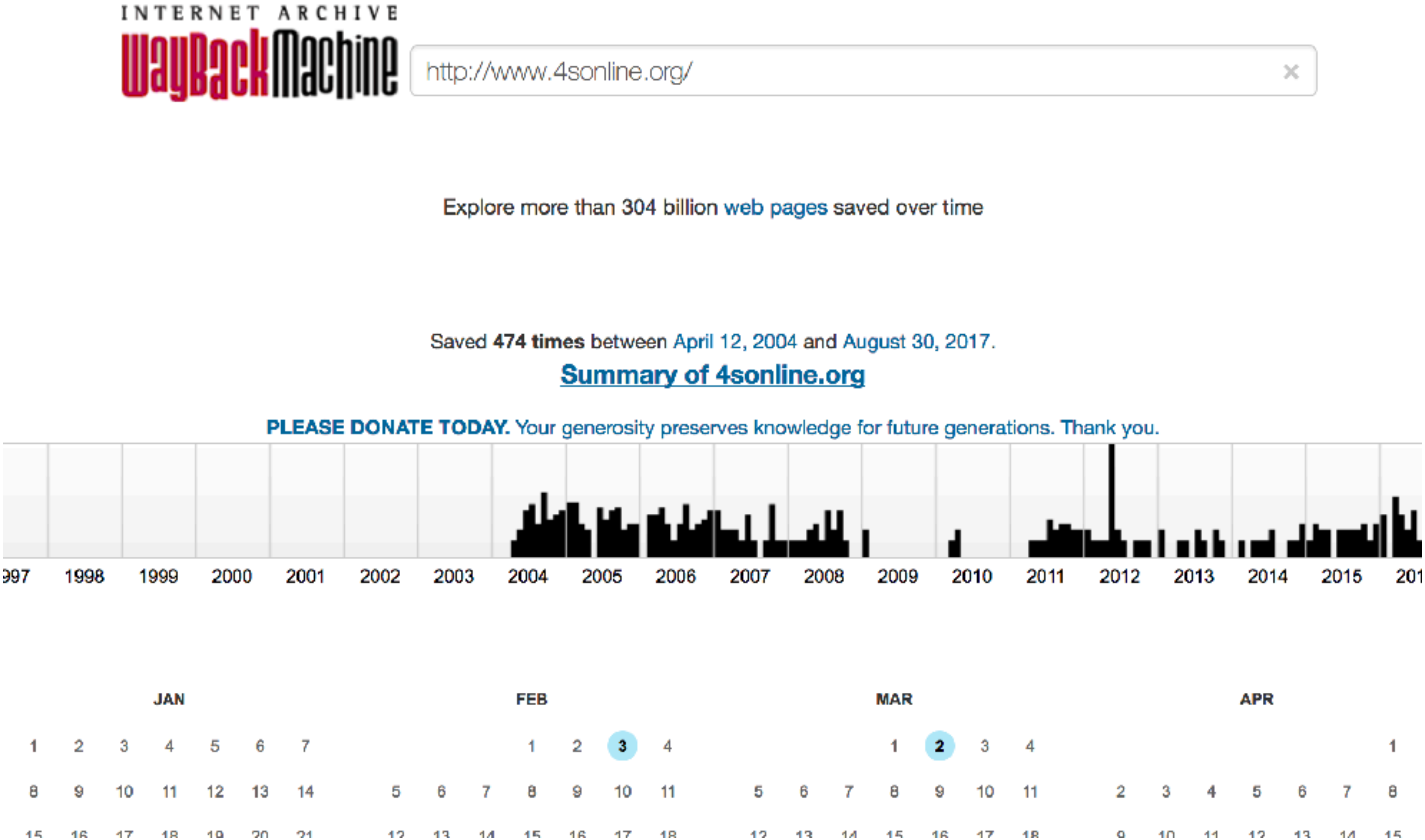
- Source or executable code
- Bankrupto sunsetting pps, OS, har
- Self sof
- Capa DATA

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



INTERNET ARCHIVE



Explore more than 304 billion web pages saved over time

Saved **474 times** between April 12, 2004 and August 30, 2017.

Summary of 4sonline.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

FEB

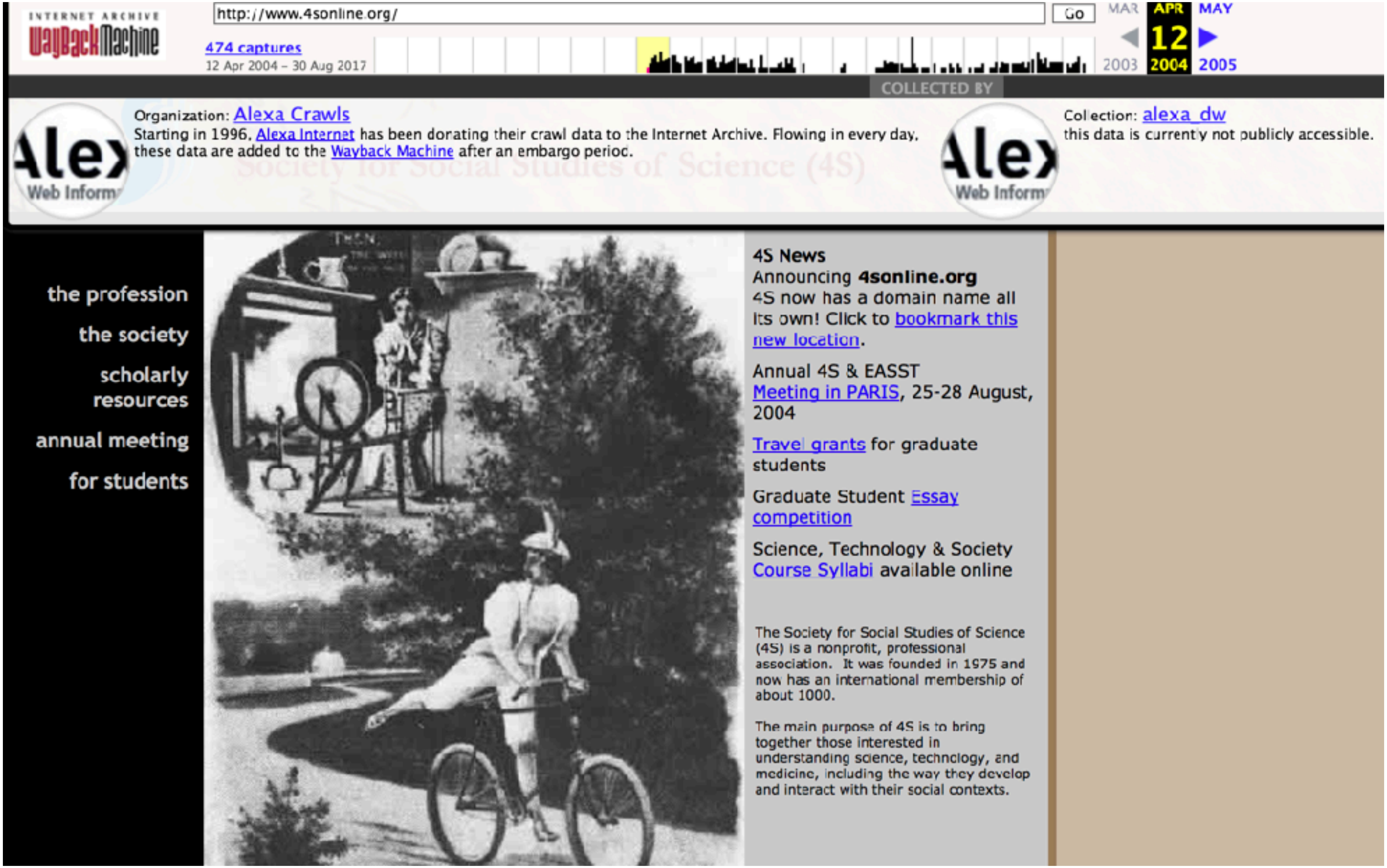
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18

MAR

			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18

APR

						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15



Organization: Alexa Crawls

Starting in 1996, Alexa Internet has been donating their crawl data to the Internet Archive. Flowing in every day, these data are added to the Wayback Machine after an embargo period.

Collection: alexa_dw

this data is currently not publicly accessible.

Society for Social Studies of Science (4S)

4S News

Announcing **4sonline.org**

4S now has a domain name all its own! Click to bookmark this new location.

Annual 4S & EASST Meeting in PARIS, 25-28 August, 2004

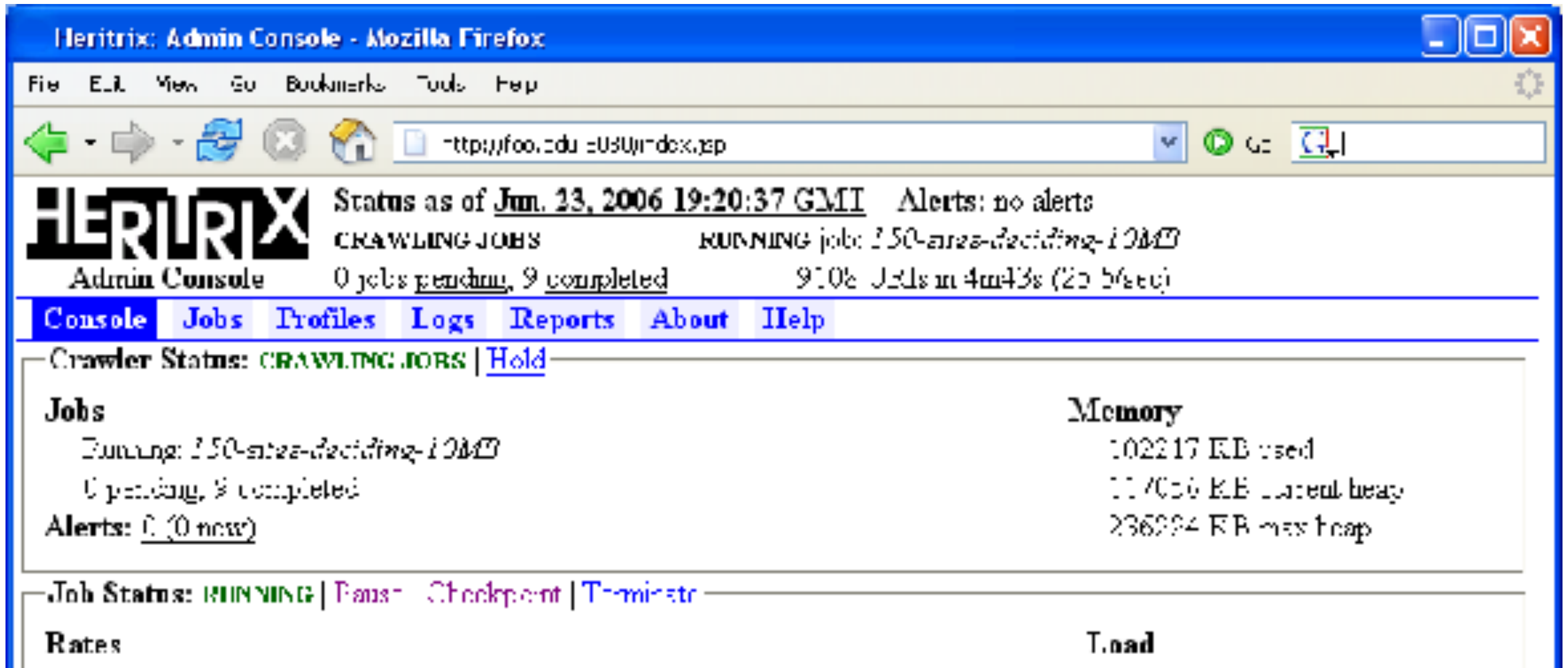
Travel grants for graduate students

Graduate Student Essay competition

Science, Technology & Society Course Syllabi available online

The Society for Social Studies of Science (4S) is a nonprofit, professional association. It was founded in 1975 and now has an international membership of about 1000.

The main purpose of 4S is to bring together those interested in understanding science, technology, and medicine, including the way they develop and interact with their social contexts.


Heritrix: Admin Console - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://foo.edu:8080/index.jsp
HERITRIX
Console
Jobs
Profiles
Logs
Reports
About
Help

Status as of Jun. 23, 2006 19:20:37 GMT Alerts: no alerts

CRAWLING JOBS RUNNING job: 150-sites-deciding-10MB

Admin Console 0 jobs pending, 9 completed 9108 URIs in 4m43s (25.5/sec)

Crawler Status: CRAWLING JOBS | Hold

Jobs
Running: 150-sites-deciding-10MB
0 pending, 9 completed
Alerts: 0 (0 new)

Memory
102217 KB used
117056 KB current heap
236224 KB max heap

Job Status: RUNNING | Pause | Checkpoint | Terminate

Rates
25.5 URIs/sec (32.27 avg)
395 KB/sec (515 avg)

Load

Time
4m43s elapsed
5m54s remaining (estimated)

Totals
downloaded 9108 56%
15997 total downloa
151 MB uncompresse

Refresh

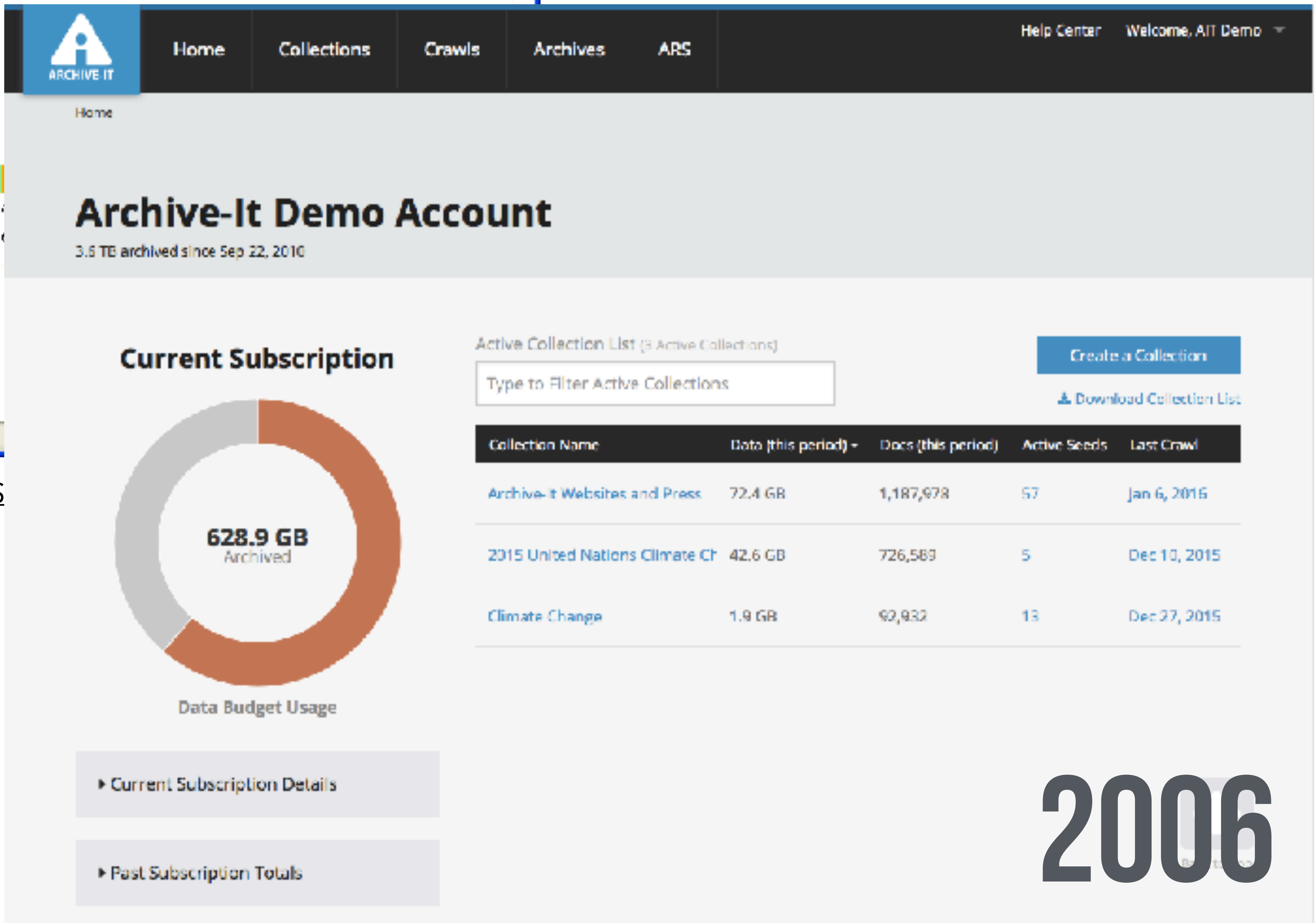
Shut down Heritrix software | Logout

Done

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

INTERNET ARCHIVE

WayBackMachine



Archive-It Demo Account

3.5 TB archived since Sep 22, 2010

Current Subscription

628.9 GB Archived

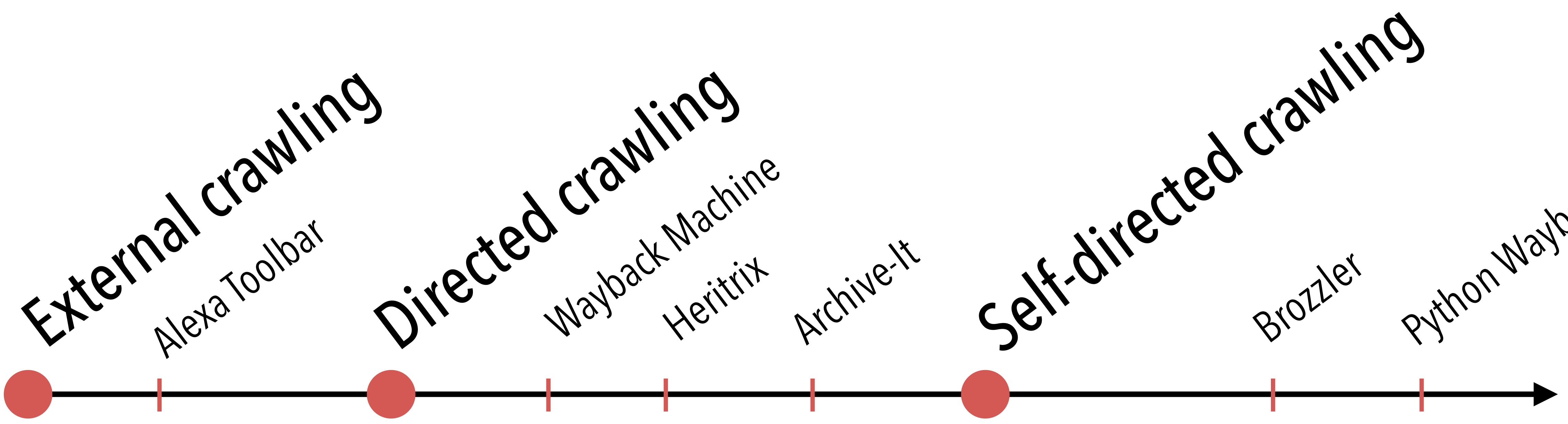
Data Budget Usage

Active Collection List (3 Active Collections)

Collection Name	Data (this period)	Docs (this period)	Active Seeds	Last Crawl
Archive-It Websites and Press	72.4 GB	1,187,978	57	Jan 6, 2016
2015 United Nations Climate Ch	42.6 GB	726,583	5	Dec 10, 2015
Climate Change	1.9 GB	92,932	13	Dec 27, 2015

2006

MAPPING ROLES & ACTIVITIES



External crawling

Alexa Toolbar

Directed crawling

Wayback Machine

Heritrix

Archive-It

Self-directed crawling

Brozzler

Python Wayback

- 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!

BREAKDOWN AND REPAIR

- Breakdown as moments when contingencies of assemblage are revealed (Star & Ruhleder 1996)
- Repair and maintenance reveals 'ethics of care' afforded to technologies over time (Jackson 2014)

Examples

- crawler traps
- missing capture elements
- debugging capture vs replay
- patch crawling

STAR, SUSAN LEIGH, AND KAREN RUHLEDER. 1996. 'STEPS TOWARD AN ECOLOGY OF INFRASTRUCTURE: DESIGN AND ACCESS FOR LARGE INFORMATION SPACES'. INFORMATION SYSTEMS RESEARCH 7 (1).

STEPHEN J. JACKSON. 2014.RETHINKING REPAIR. IN MEDIA TECHNOLOGIES: ESSAYS ON COMMUNICATION, MATERIALITY, AND SOCIETY, TARLETON GILLESPIE, PABLO J. BOCZKOWSKI, AND KIRSTEN A. FOOT (EDS.). MIT PRESS, CAMBRIDGE, MASSACHUSETTS; LONDON, ENGLAND, 221–239.

VALUES & 'ETHICS OF CARE'

- Popularity, novelty and levels of 'precariousness' all inform collection tasks
- Organisation/archivists value a Web with temporal depth - a **'Google for the historical Web'**

Reflect (expanding?) public insecurities around web-based truth claims

Highlight controversies over power dynamics, sustainability questions

ECONOMIES OF WEB ARCHIVING

- Emergent economy of external (repair) services around the Wayback Machine
- Sharing economy and reliance on volunteer labour; knowledge production
- Politics of knowledge production beyond selection to consider repair

Examples

- download and revive ageing websites
- capturing links shared on social media; Wikipedia
- broken link repair; browser extensions

SUMMARISING

- web archiving as a form of maintenance
- ability to repair and maintain WAs intimately tied to innovation on the Web
- changing the landscape of the Web; creating stakeholders
- increasing reliance on parallel project of WAs - power dynamics

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.

EPSRC

Research Councils UK Digital Economy Transforming Business and Society

INTERNET ARCHIVE