OBSERVING WEB ARCHIVES

The Case for an Ethnographic Study of Web Archiving

JESSICA OGDEN | SUSAN HALFORD | LES CARR

UNIVERSITY OF Southampton | Web Science DTC

Corresponding Author: jessica.ogden@soton.ac.uk

WEB SCIENCE CONFERENCE 2017 | RENSSELAER WEB SCIENCE RESEARCH CENTER, RPI | 25-28 JUNE 2017

OUTLINE

- Web archiving as a field of practice
- Problematising practice and the field; a theoretical framework
- Methodology and empirical work
- Preliminary results of a study at the Internet Archive

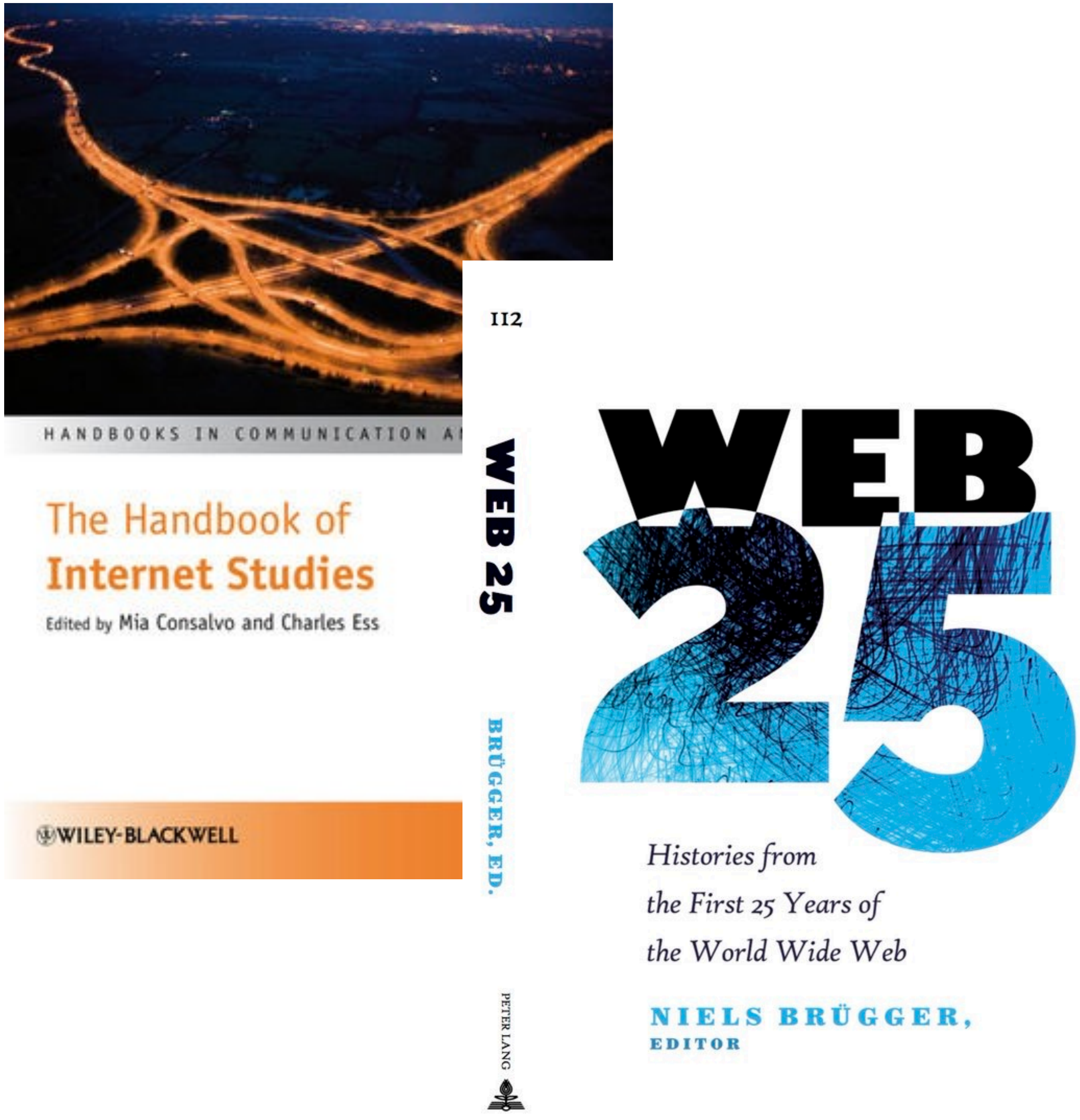
WEB ARCHIVING

https://flic.kr/p/dUt1ir

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

PRACTICE

Julien Masanès (Ed.)
Web Archiving
archiving websites
a practical guide for information management professionals
adrian brown


HANDBOOKS IN COMMUNICATION A
The Handbook of Internet Studies
Edited by Mia Consalvo and Charles Ess
WILEY-BLACKWELL
112
WEB 25
BRÜGGER, ED.
PETER LANG
WEB 25
Histories from the First 25 Years of the World Wide Web
NIELS BRÜGGER, EDITOR

PRACTICE

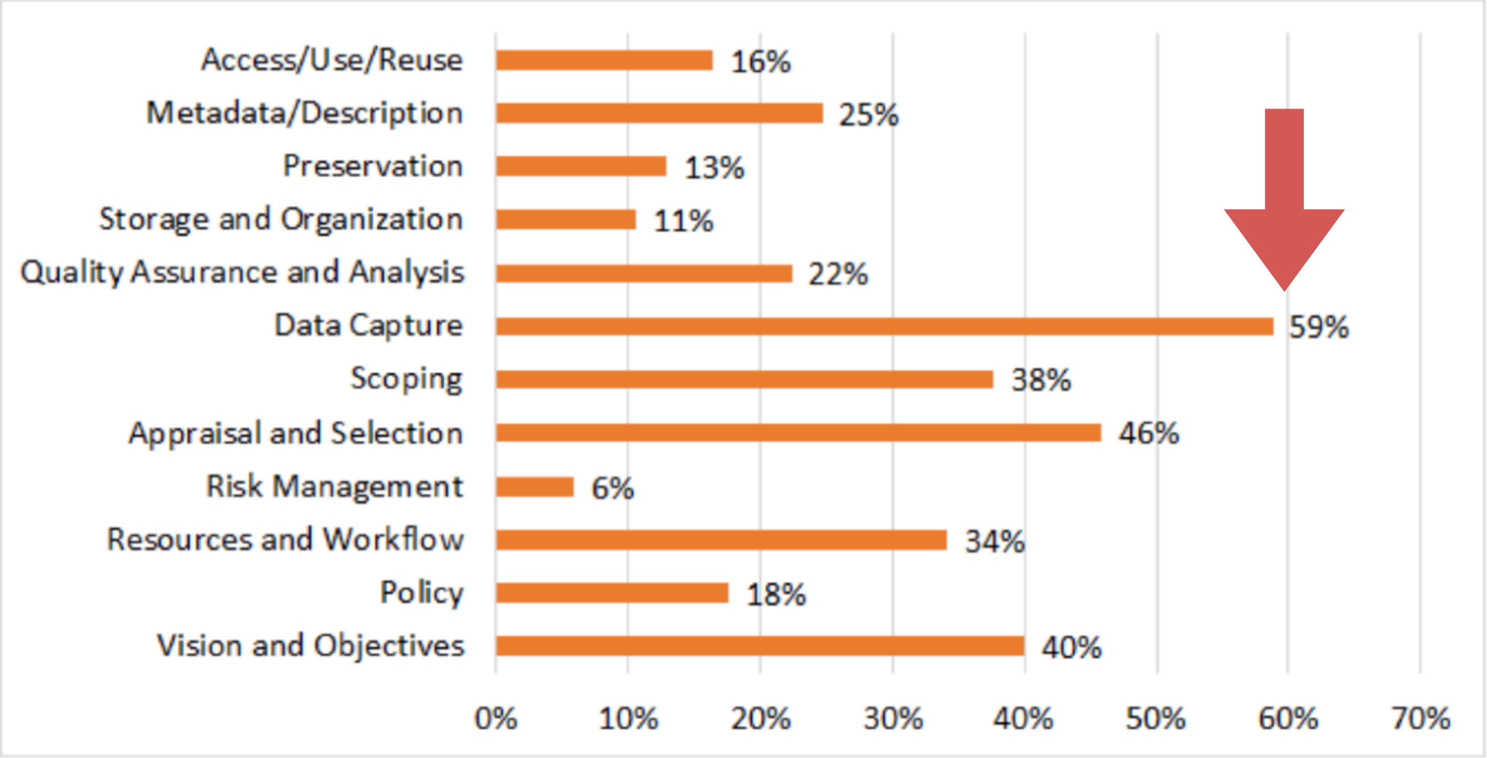


FIGURE 3: PERCEPTIONS OF MOST PROGRESS IN LAST TWO YEARS

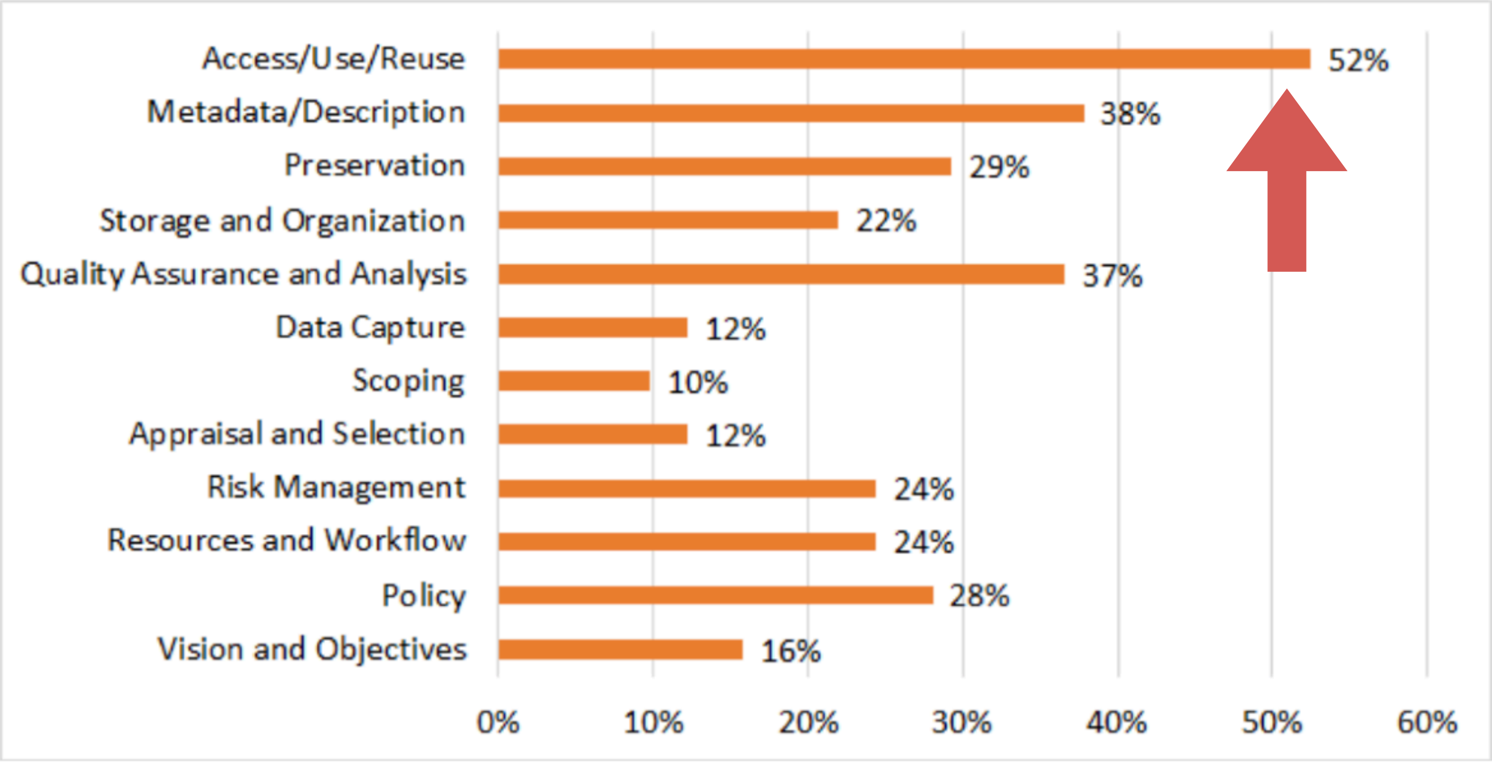


FIGURE 4: PERCEPTIONS OF LEAST PROGRESS IN LAST TWO YEARS

BAILEY ET AL. 2017

Daniel Gomes et al. (2011) A Survey on Web Archiving Initiatives.

National Digital Stewardship Alliance (2012) Web Archiving Survey Report.

Bailey et al. (2014) Web Archiving in the United States: A 2013 Survey

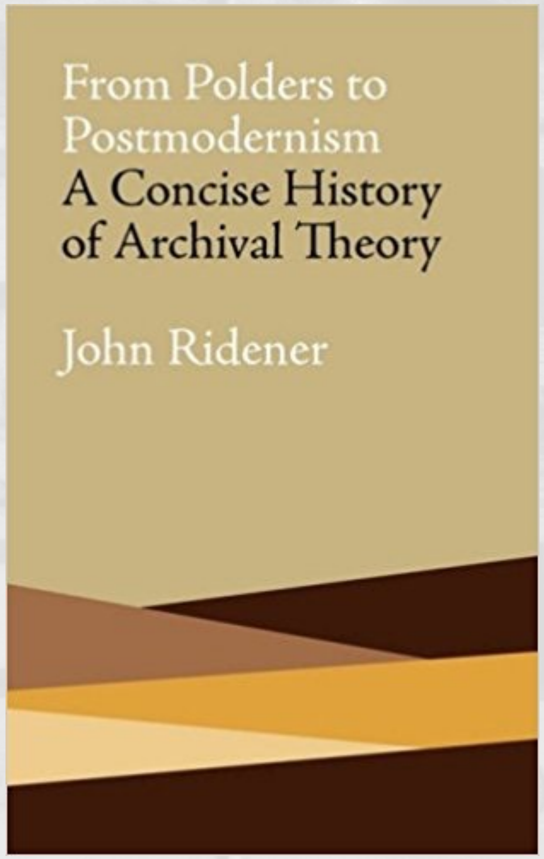
Bailey et al. (2017) Web Archiving in the United States: A 2016 Survey

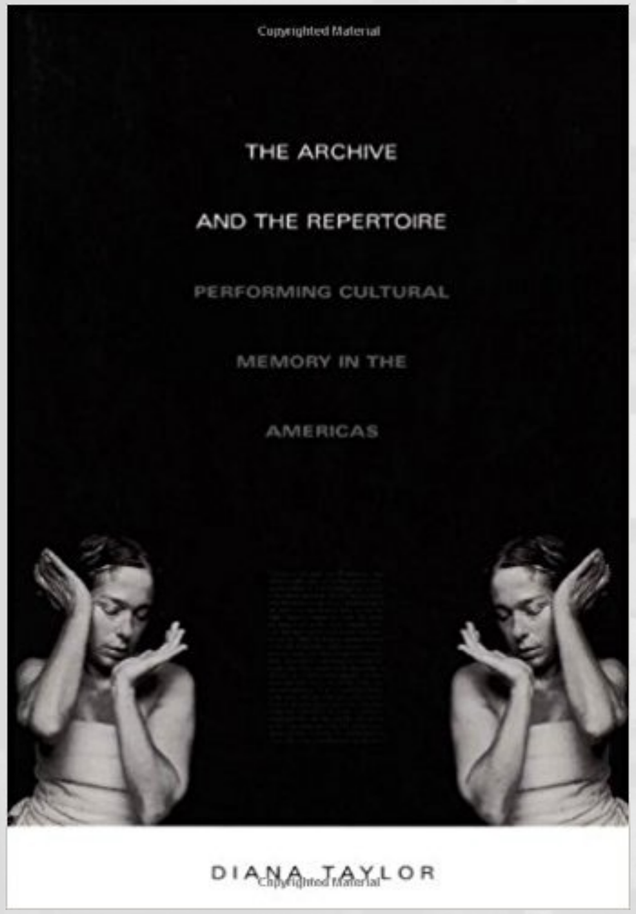
PROBLEMATISING

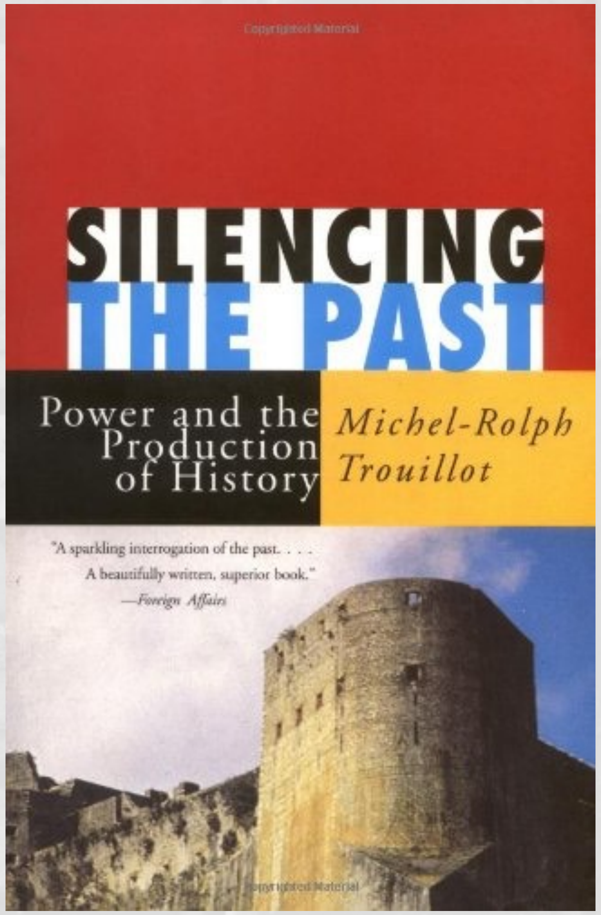
- Existing overviews, best practice documents, surveys are partial
- Web archiving itself is inherently selective, contingent and ≠ a copy of the Web(s)
- Need to understand the role and agency of non/human actors in process
- Understanding the why, when + how is important for claims

FRAMING ENGAGEMENT

> *We will move from databases to knowledge bases. We will move, in the language of the post-modernists, to re-contextualize our activities: we will reorient ourselves from the content to the context, and from the end result to the original empowering intent, that is, from the artifact (the actual record) to the creating processes behind it, and thus to the actions, programmes, and functions behind those processes.*
>
> -Terry Cook


From Polders to Postmodernism
A Concise History of Archival Theory
John Ridener

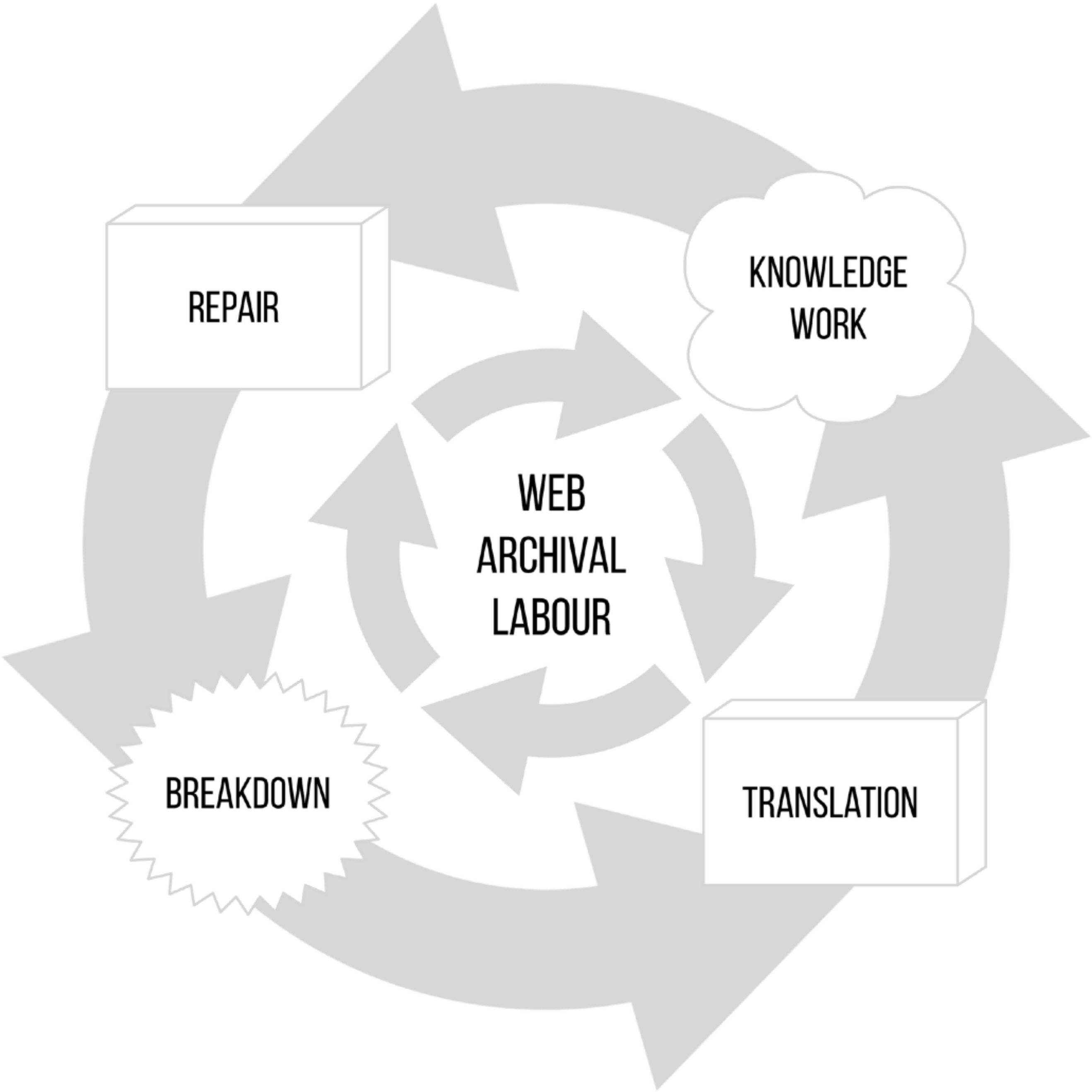



Copyrighted Material
SILENCING THE PAST
Power and the Production of History
Michel-Rolph Trouillot
"A sparkling interrogation of the past. . . . A beautifully written, superior book." —Foreign Affairs

ROLE OF THEORY

- From archival science to the archival turn
- The archive as generative, subjective not a 'view from nowhere'
- Understanding non/professional practice, role of politics
- Materiality of knowledge - practice as labour, tangible implications of intervention

LABOUR

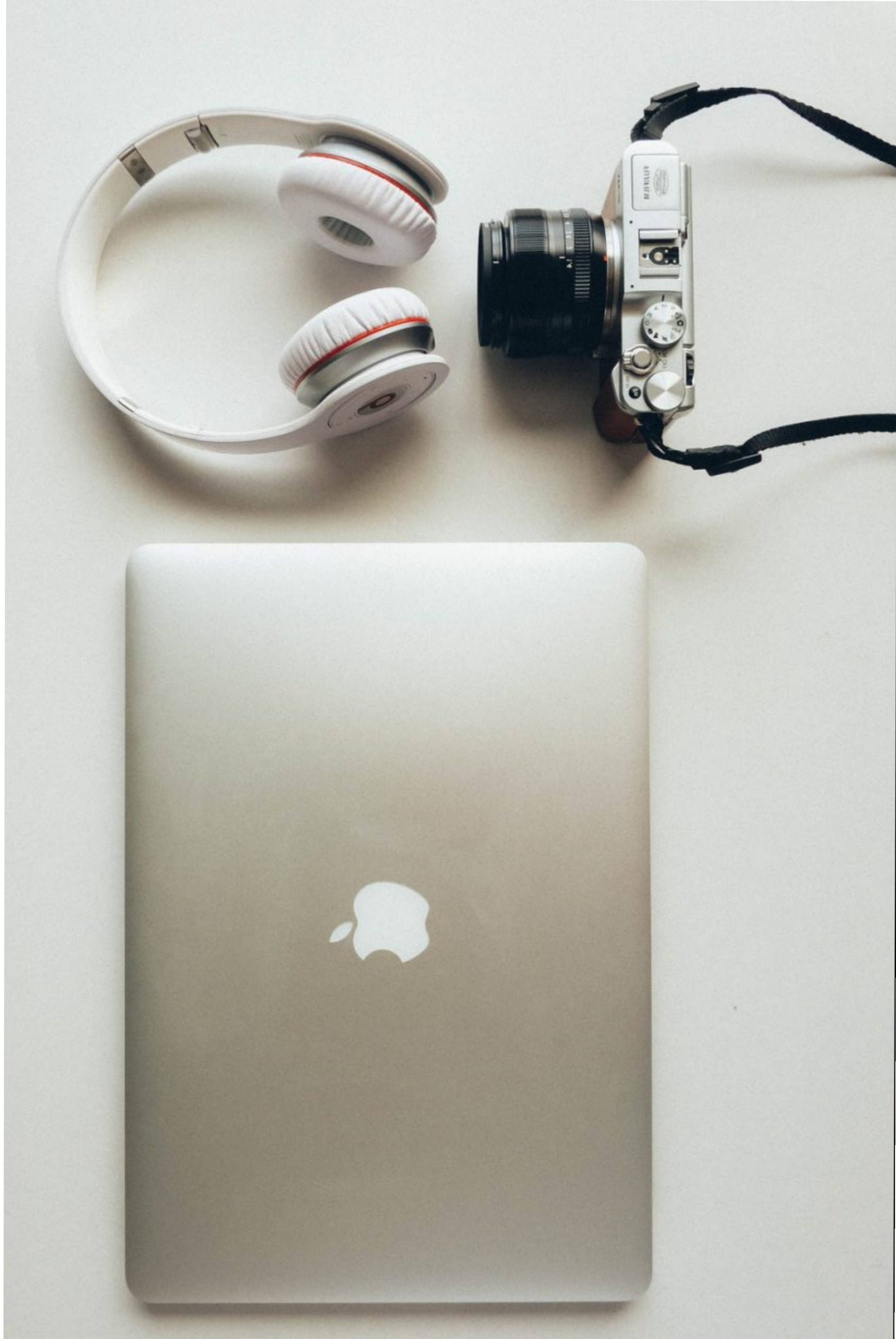


GREGORY J. DOWNEY. 2014. MAKING MEDIA WORK: TIME, SPACE, IDENTITY, AND LABOR IN THE ANALYSIS OF INFORMATION AND COMMUNICATION INFRASTRUCTURES. IN MEDIA TECHNOLOGIES: ESSAYS ON COMMUNICATION, MATERIALITY, AND SOCIETY, TARLETON GILLESPIE, PABLO J. BOCZKOWSKI, AND KIRSTEN A. FOOT (EDS.). MIT PRESS, CAMBRIDGE, MASSACHUSETTS; LONDON, ENGLAND, 141–165.

- Drawing on ‘information labor’ as outlined by Downey (2014)
- Labour = the work it takes to produce and transform web archives into information sources
- Process of making information ‘useful’; putting it into ‘circulation’
- Human and algorithmic; often obscure, hidden

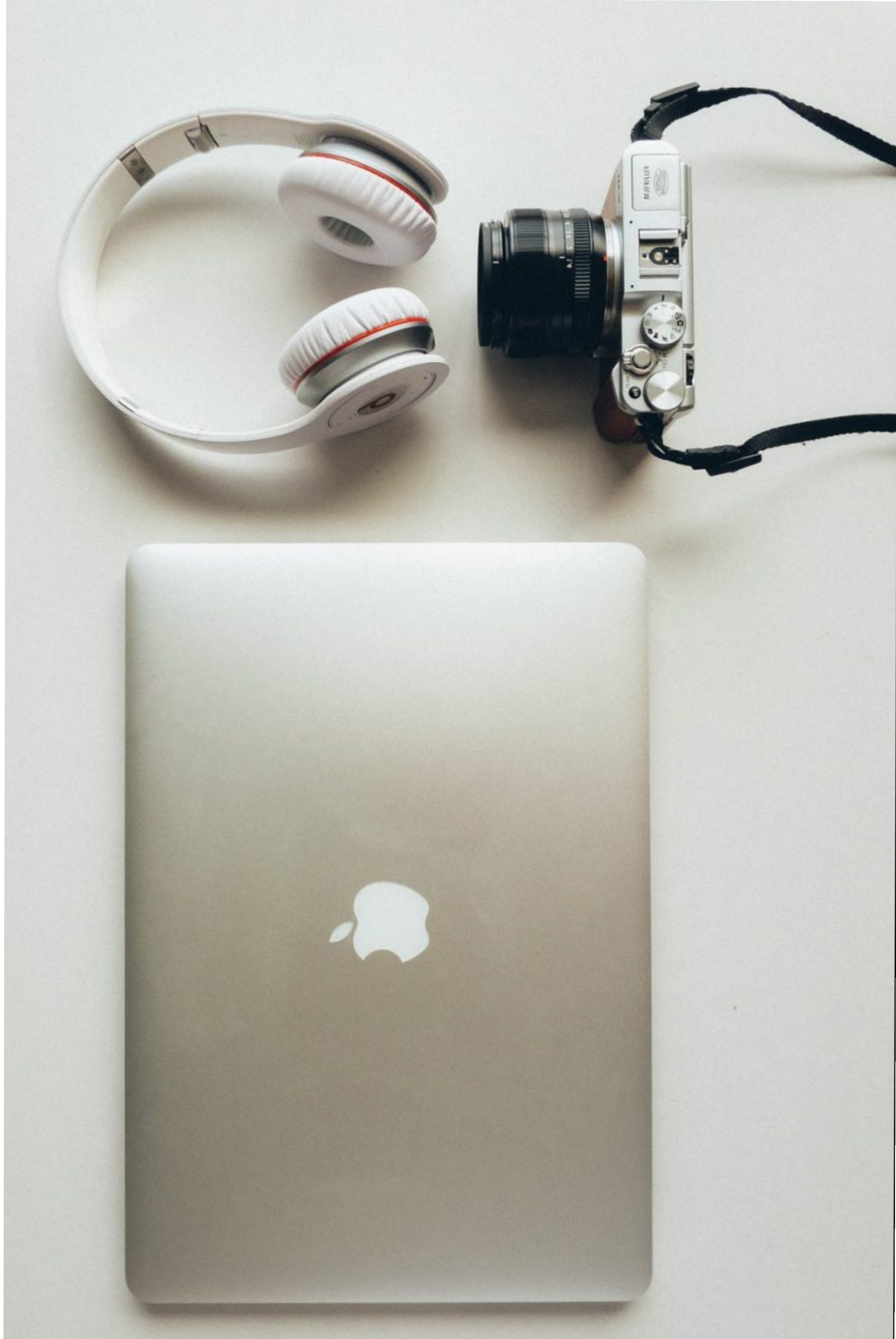
METHODOLOGY

- Ethnographic methods were chosen to document patterns of work
- 'Archival ethnography' (Gracy 2004)
- Understanding socio-cultural meaning, not a behavioural study
- Draws on extensive body of STS literature around record creation, scientific labs



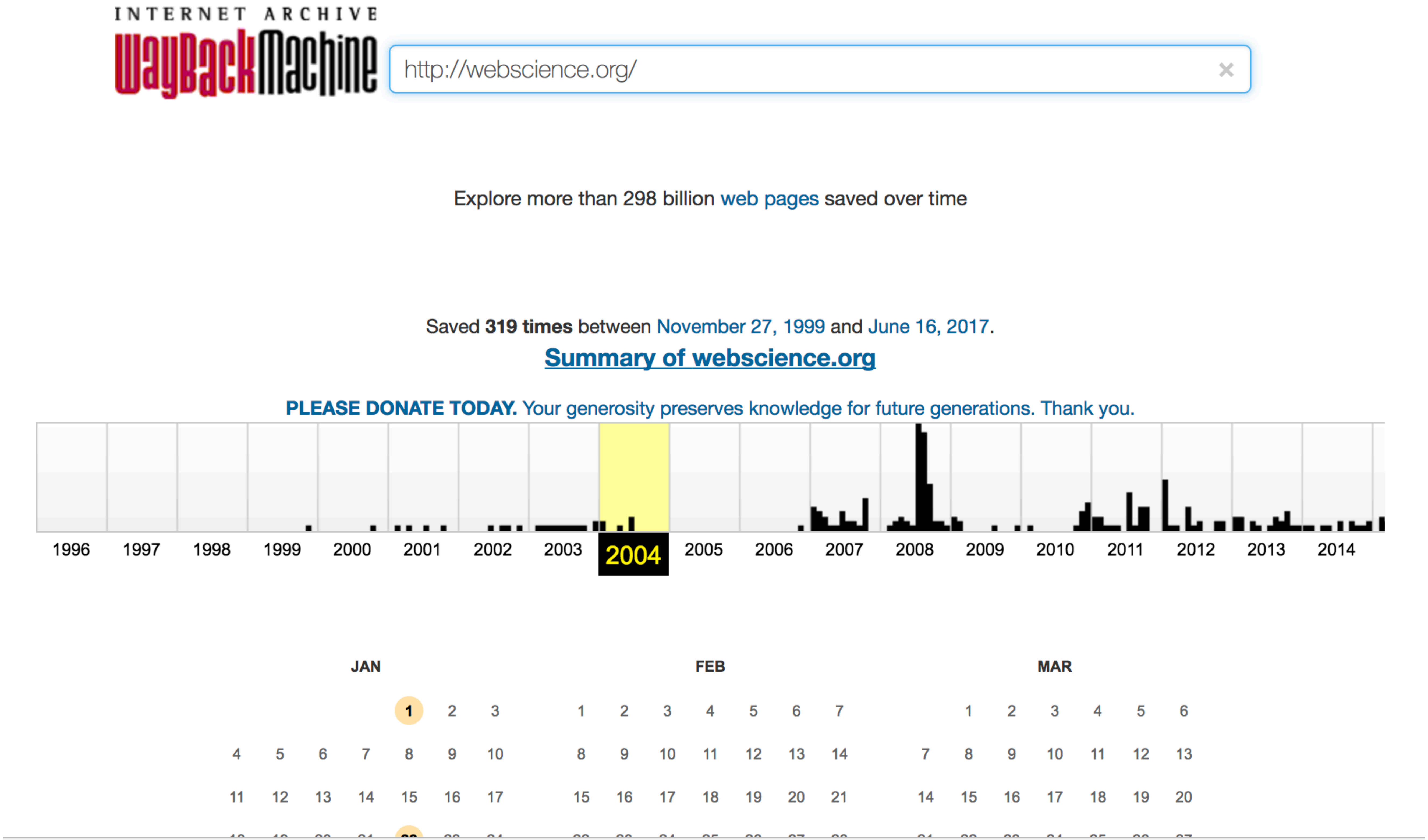
METHODOLOGY

- 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
- Thematic analysis (Spradley 1979) of 'things informants know'



THE SITE

- Non-profit digital library founded in 1996 by Brewster Kahle
- Headquarters in San Francisco
- Began as mechanism to capture web pages as by-product of indexing
- Expanded remit to include books, audio, film/video, images, docs, video games and software



http://webscience.org/

Explore more than 298 billion web pages saved over time

Saved **319 times** between November 27, 1999 and June 16, 2017.

Summary of webscience.org

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

JAN

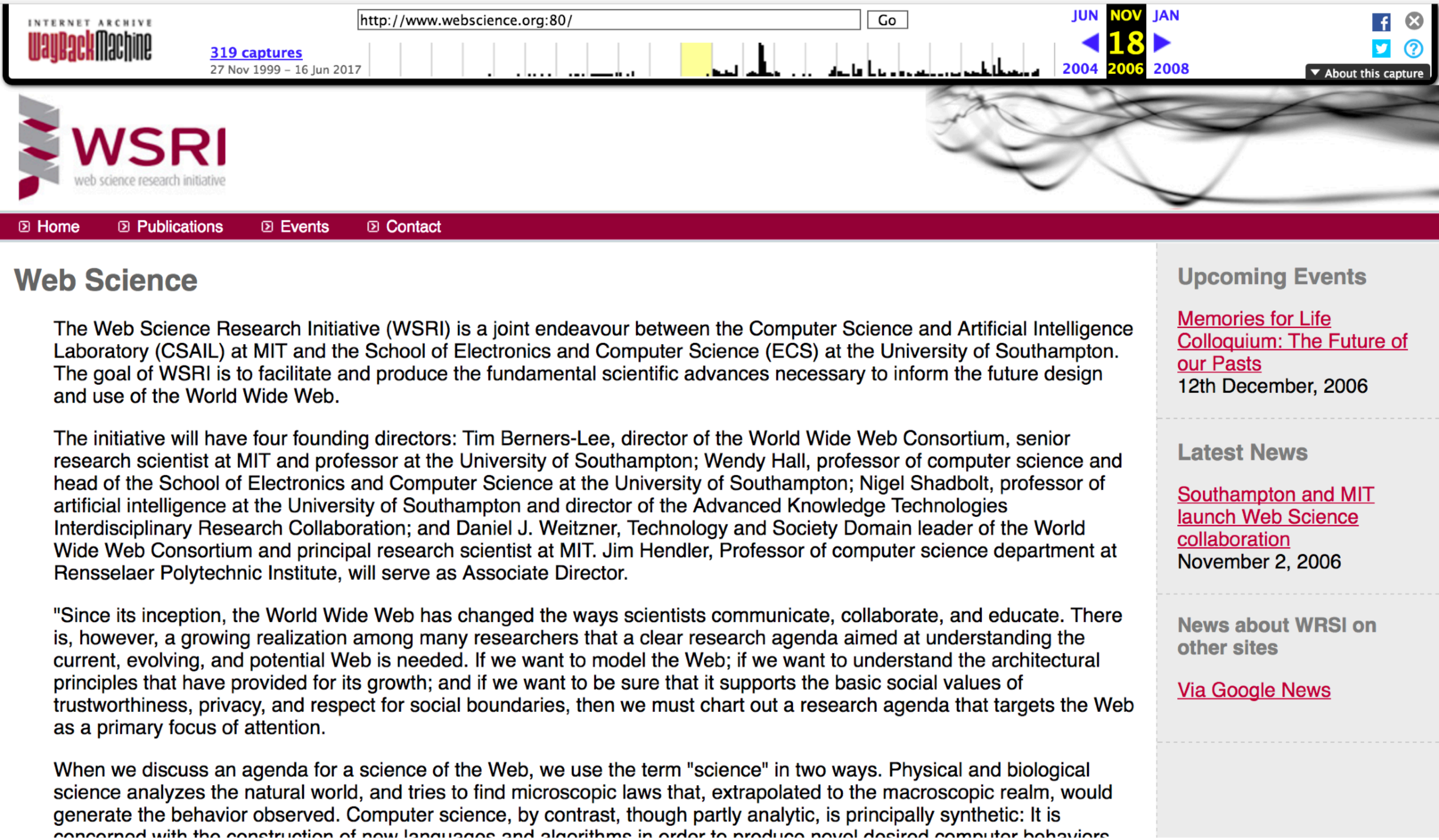
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	**22**	23	24

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



Web Science

The Web Science Research Initiative (WSRI) is a joint endeavour between the Computer Science and Artificial Intelligence Laboratory (CSAIL) at MIT and the School of Electronics and Computer Science (ECS) at the University of Southampton. The goal of WSRI is to facilitate and produce the fundamental scientific advances necessary to inform the future design and use of the World Wide Web.

The initiative will have four founding directors: Tim Berners-Lee, director of the World Wide Web Consortium, senior research scientist at MIT and professor at the University of Southampton; Wendy Hall, professor of computer science and head of the School of Electronics and Computer Science at the University of Southampton; Nigel Shadbolt, professor of artificial intelligence at the University of Southampton and director of the Advanced Knowledge Technologies Interdisciplinary Research Collaboration; and Daniel J. Weitzner, Technology and Society Domain leader of the World Wide Web Consortium and principal research scientist at MIT. Jim Hendler, Professor of computer science department at Rensselaer Polytechnic Institute, will serve as Associate Director.

"Since its inception, the World Wide Web has changed the ways scientists communicate, collaborate, and educate. There is, however, a growing realization among many researchers that a clear research agenda aimed at understanding the current, evolving, and potential Web is needed. If we want to model the Web; if we want to understand the architectural principles that have provided for its growth; and if we want to be sure that it supports the basic social values of trustworthiness, privacy, and respect for social boundaries, then we must chart out a research agenda that targets the Web as a primary focus of attention.

When we discuss an agenda for a science of the Web, we use the term "science" in two ways. Physical and biological science analyzes the natural world, and tries to find microscopic laws that, extrapolated to the macroscopic realm, would generate the behavior observed. Computer science, by contrast, though partly analytic, is principally synthetic: It is concerned with the construction of new languages and algorithms in order to produce novel desired computer behaviors

Upcoming Events

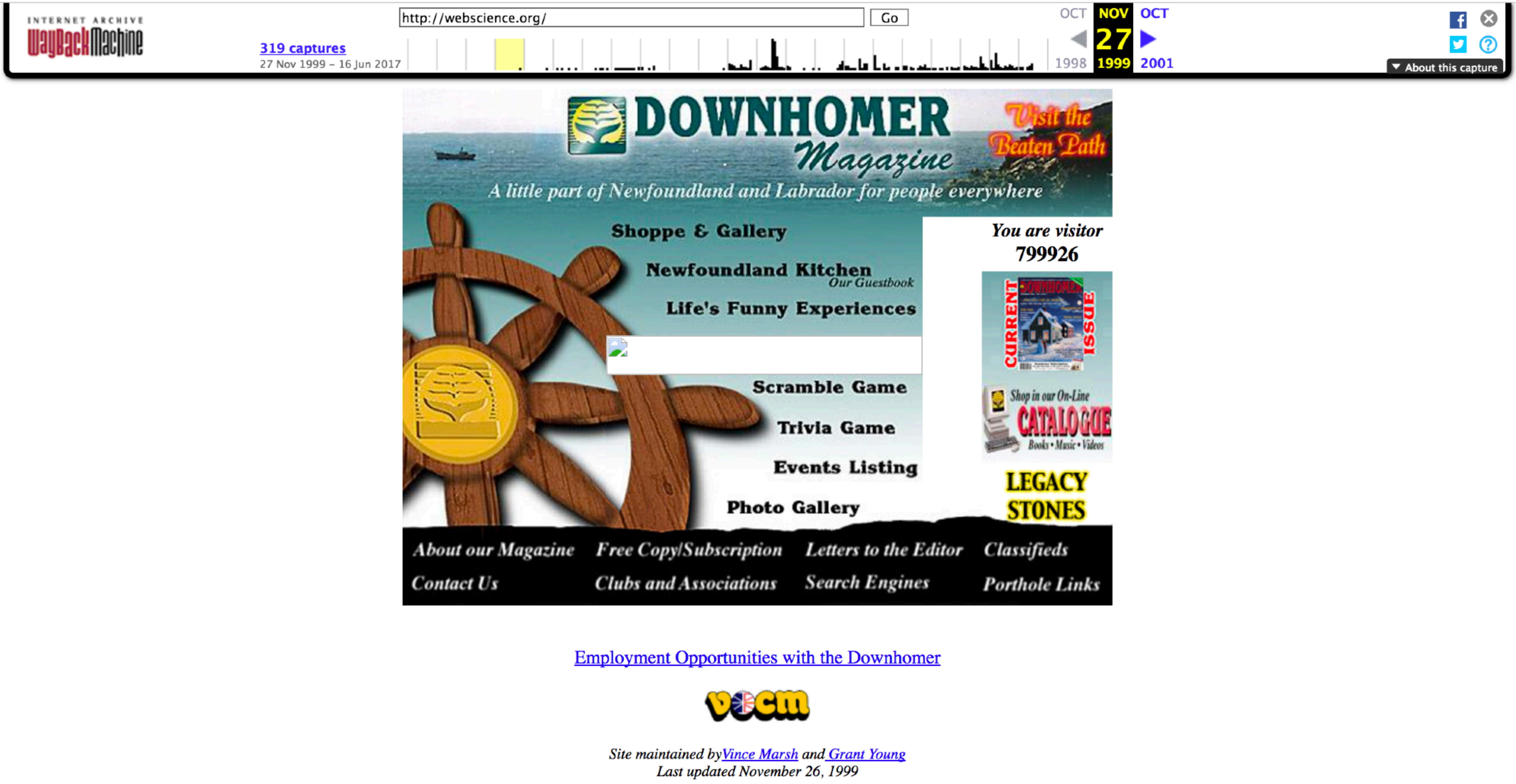
Memories for Life Colloquium: The Future of our Pasts
12th December, 2006

Latest News

Southampton and MIT launch Web Science collaboration
November 2, 2006

News about WRSI on other sites

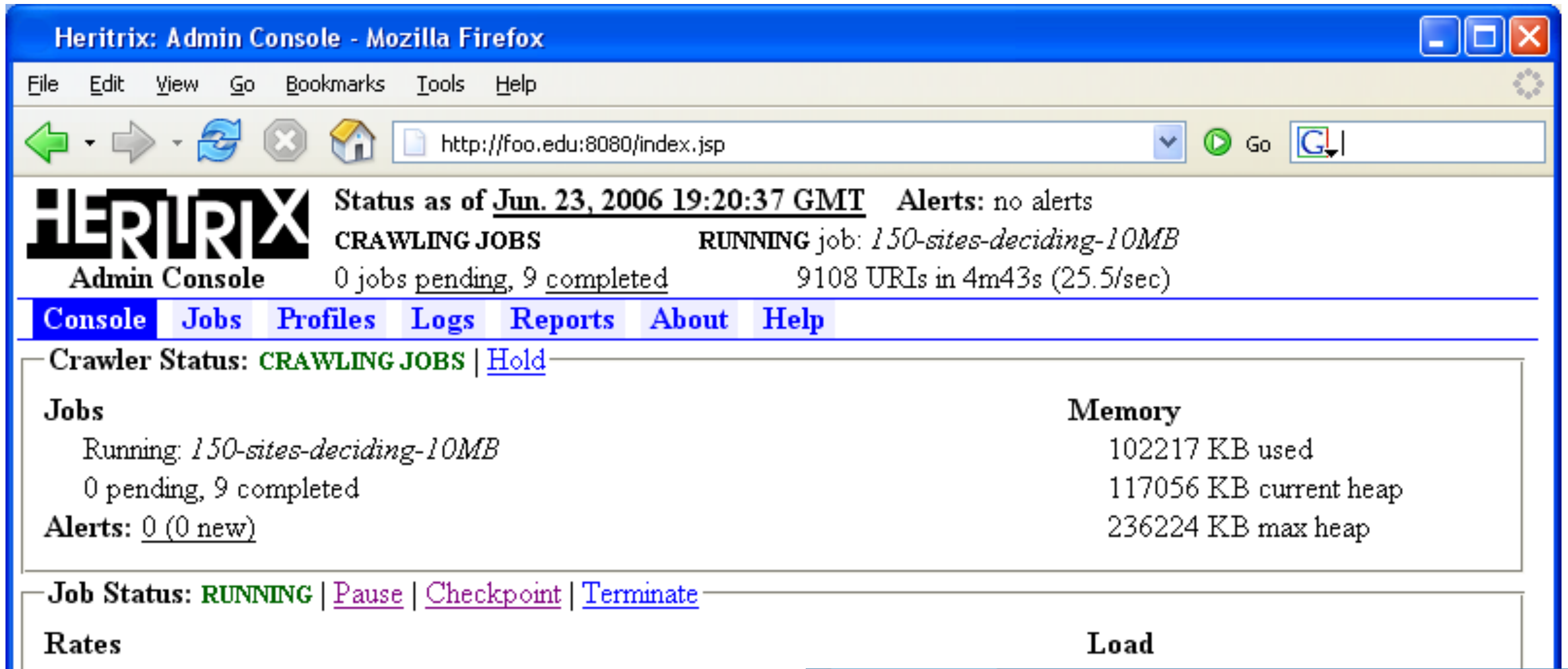
Via Google News



You are visitor
799926

Employment Opportunities with the Downhomer

Site maintained byVince Marsh and Grant Young
Last updated November 26, 1999


Heritrix: Admin Console - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://foo.edu:8080/index.jsp
Go
HERITRIX
Admin Console
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*
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)
595 KB/sec (545 avg)
Time
4m43s elapsed
3m34s remaining (estimated)
Totals
downloaded 9108 56%
15997 total downloa
151 MB uncompresse

Load

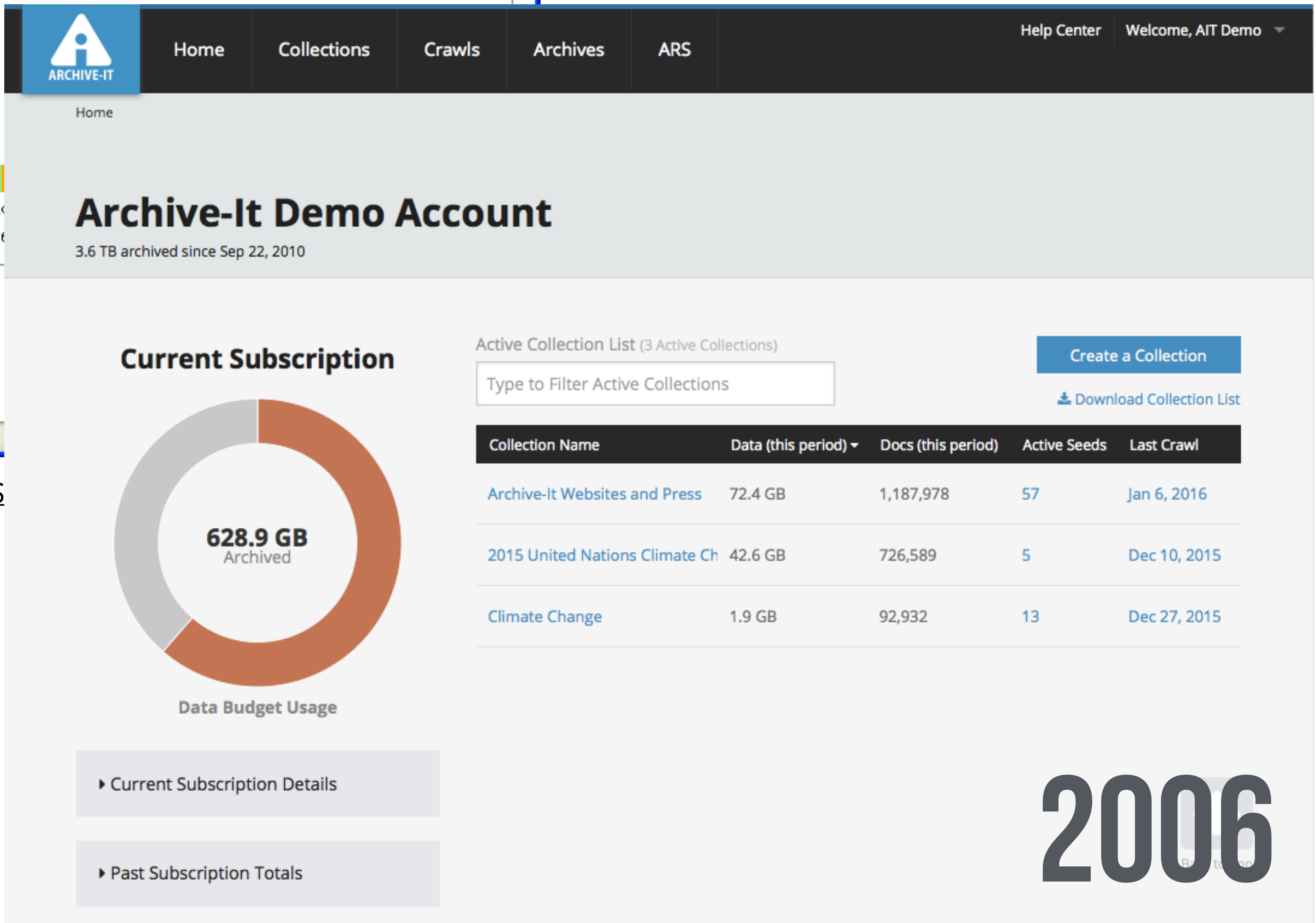
Refresh

Shut down Heritrix software | Logout

Done

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

INTERNET ARCHIVE
WaybackMachine



Archive-It Demo Account

3.6 TB archived since Sep 22, 2010

Current Subscription

628.9 GB Archived

Data Budget Usage

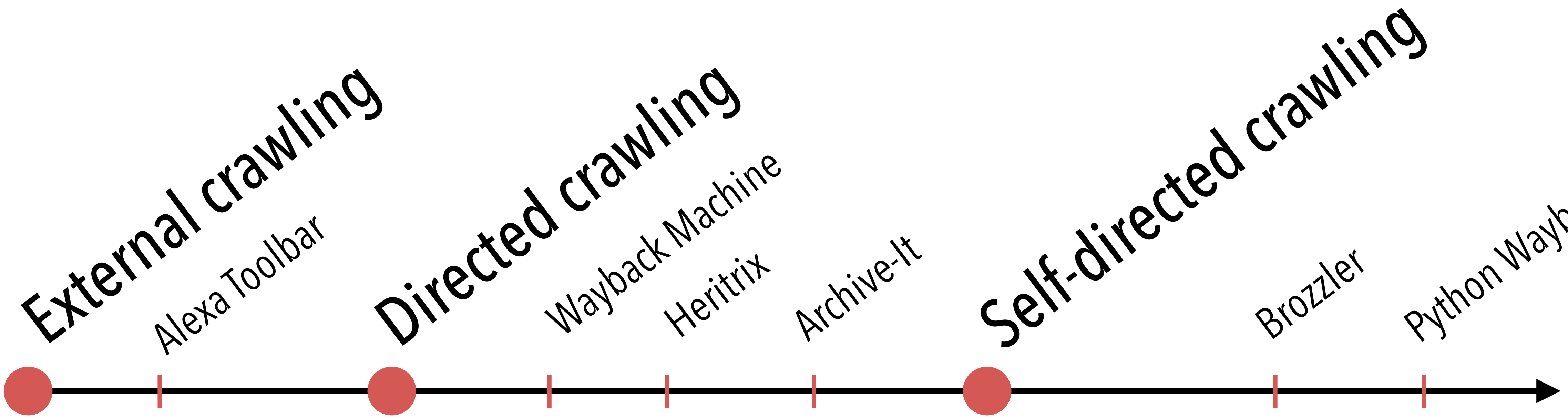
Active Collection List (3 Active Collections)

Type to Filter 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	726,589	5	Dec 10, 2015
Climate Change	1.9 GB	92,932	13	Dec 27, 2015

2006

MAPPING PRACTICE 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!

WEB ARCHIVAL LABOUR

Knowledge Work	Examples
Defining selection priorities	popularity, novelty, ‘precarity’
Allocating resources	staff, domain/host storage limits
Analysing corpus; curating	link-based analysis, tagging
Maintaining crawls	monitoring activities

WEB ARCHIVAL LABOUR

Breakdown & 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
- patch crawling
- other quality assurance tasks

SUMMARY & FUTURE WORK

- Complex system of knowledge/ maintenance work for prioritising web archiving
- Archive is leveraging corpus and crowd for identifying domains
- Developed multiple tagging and analysis tools for curation

Further examining algorithmic (non-human) labour

Expand to include other communities, case studies

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