

BOTS, SEEDS AND PEOPLE

WEB ARCHIVES AS INFRASTRUCTURE

Ed Summers
University of Maryland
[@edsu](#) / ehs@pobox.com

Ricardo Punzalan
University of Maryland
[@archivalflip](#) / punzalan@umd.edu

slides: <http://bit.ly/bots-seeds-people>
paper: <https://arxiv.org/abs/1611.02493v1>

OVERVIEW

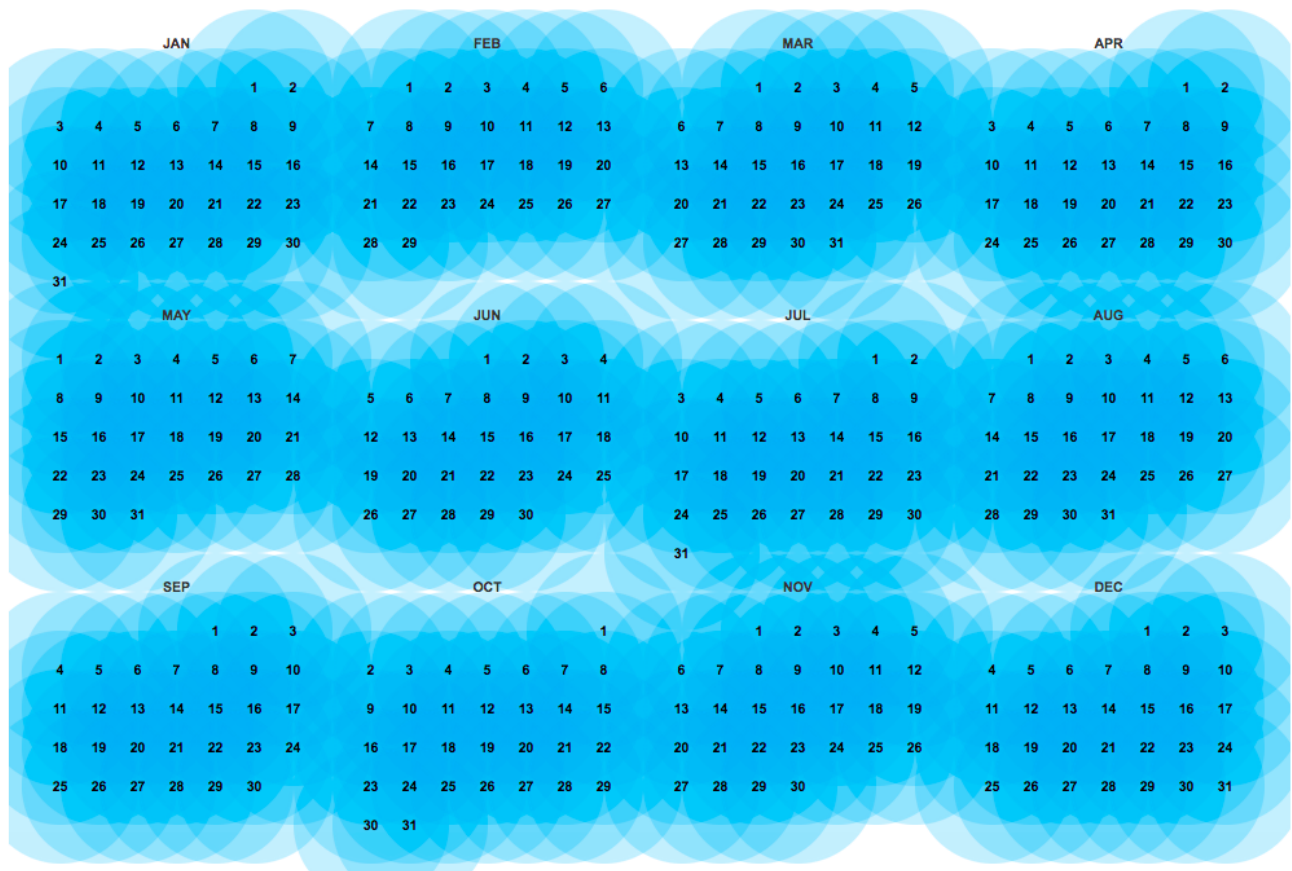
- Appraisal in Web Archives
- Research Question
- Methodology
- Findings
- Future Work

How much of the web is in the Internet Archive?

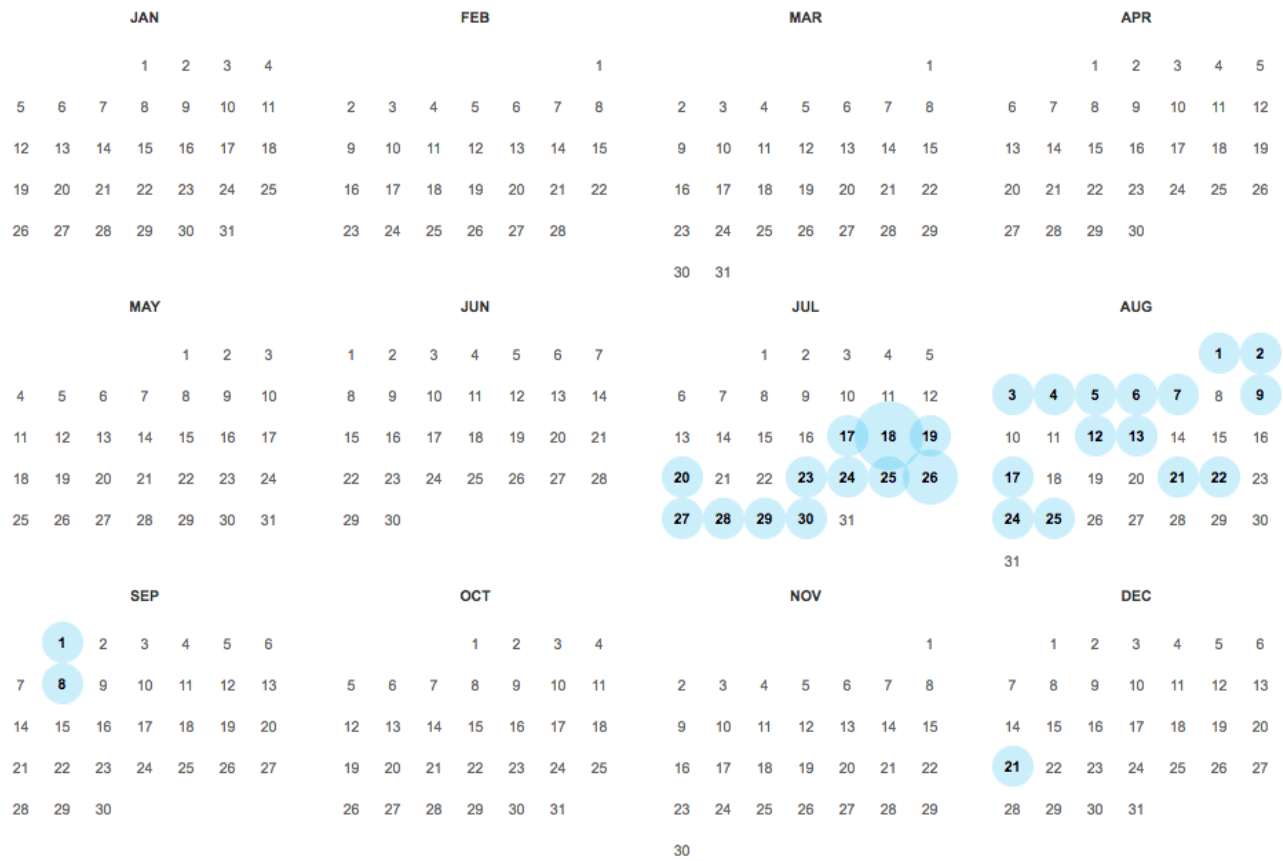
$$273,000,000,000^1 / 1,000,000,000,000^2 = .273 ???$$

1. Alpert, J. and Hajaj, N. (2008). [We knew the web was big...](#) Google.

2. Goel, V. (2016). [Defining Web pages, Web sites and Web captures](#). Internet Archive.



Archival coverage of the NYTimes homepage in 2016.



Archival coverage of Igor Strelkov's VKontakte profile in 2014.

APPRAISAL

The process of identifying materials offered to an archives that have sufficient *value* to be accessioned.

[Appraisal](#) in *A Glossary of Archival and Records Terminology*. Society of American Archivists.

RQ: HOW IS APPRAISAL BEING ENACTED IN *WEB* ARCHIVES?

- Selection strategies in web archiving
- Socio-technical factors that influence selection practices

Kitchin, R. (2016). Thinking critically about and researching algorithms. *Information, Communication & Society*, 1–16.

1. Source Code
2. Reflexively producing code
3. Reverse engineering
4. Design & designers
5. Socio-technical assemblage
6. The world

METHODOLOGY

39 contacted (email)

33 responded

28 interviewed

F (13) / M (15)

university, non-profit, library/museum

archivists, developers, researchers

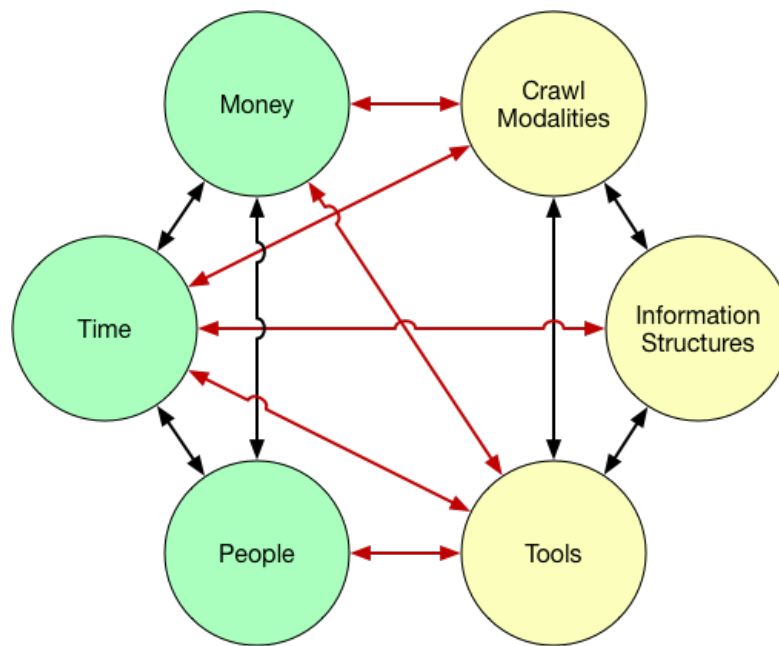
semi-structured interviews



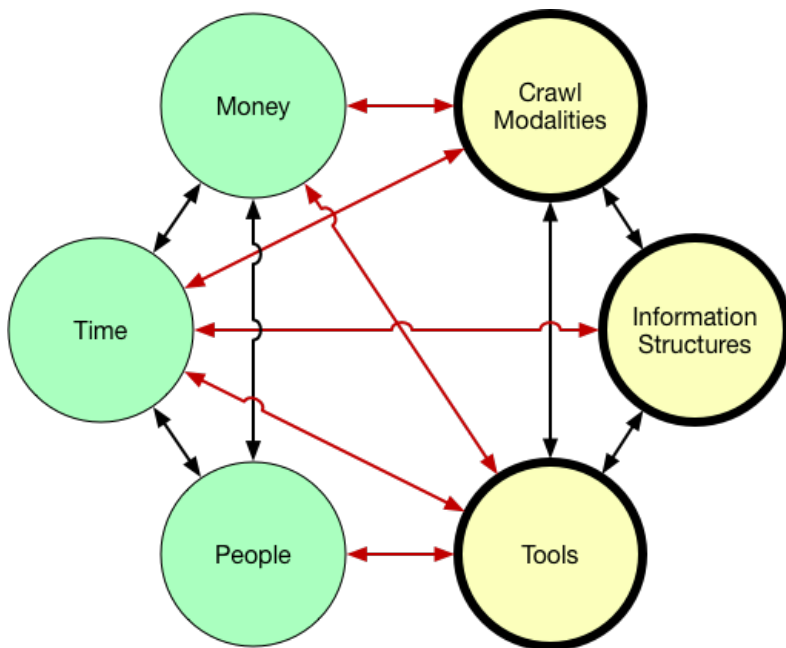
memoing + field notes

coding / thematic analysis

FINDINGS



TECHNICAL



Crawl Modalities

domains, websites, topics,
events, documents

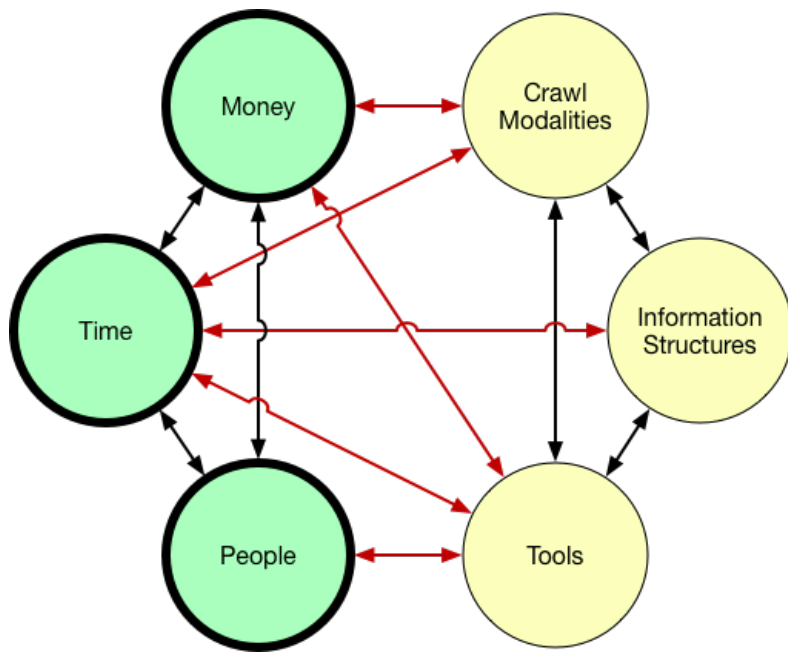
Information Structures

hierarchies, networks, streams

Tools

services, storage systems, open
source utilities, spreadsheets,
forms, email, issue trackers

SOCIAL



People

teams, lone-arrangers,
developers, collaborations

Time

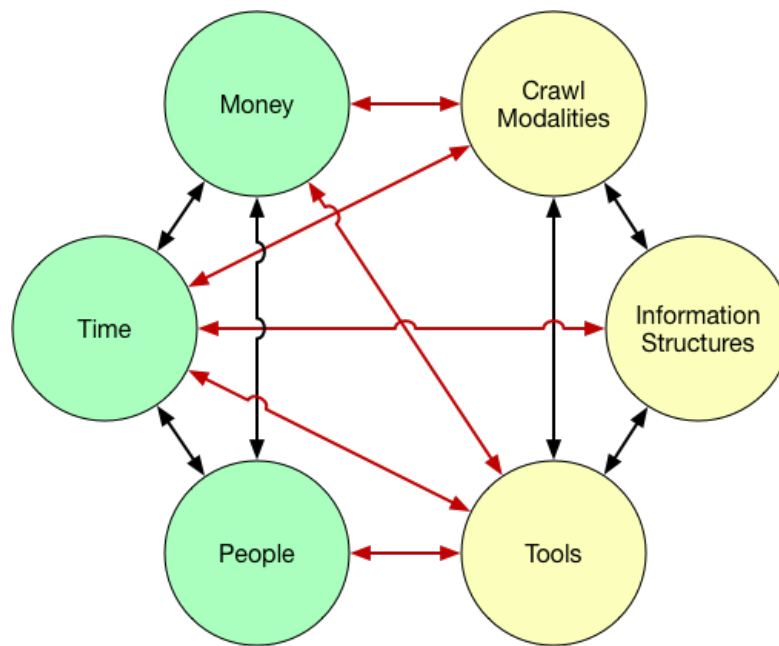
limits, scheduling, always-on,
reading, reviewing

Money

grants, subscriptions,

infrastructure

BREAKDOWN & REPAIR



FUTURE WORK

Kitchin, R. (2016). Thinking critically about and researching algorithms. *Information, Communication & Society*, 1–16.

1. Source Code
2. Reflexively producing code
3. Reverse engineering
4. Design & designers
5. Socio-technical assemblage
6. The world


THANKS!



fix-it by [Derek Bridges](#)

EXTRAS

- [Codebook](#)
- [Interview Questions](#)



A screenshot of a terminal window titled "2. less". The window displays a list of URLs, each followed by a comma-separated list of metadata fields: URL, Group, Status, Frequency, Type, and Access. The URLs include various sources like Twitter, Facebook, and academic institutions. The metadata fields contain values like "True", "NONE", "normal", "DAILY", "rss", and "True". The list ends with a file named "seed-list.csv".

```
Seed URL,Group,Status,Frequency,Type,Access
https://twitter.com/search?q=%23datarescue&src=typd,,True,DAILY,rss,True
http://datarescuesfbay.org/,True,NONE,normal,True
https://www.facebook.com/datarefugeaustin/,True,NONE,normal,True
https://datarescuenyc.com/,True,NONE,normal,True
https://www.facebook.com/events/189177784386985/,True,NONE,normal,True
http://calagator.org/events/1250471401/,True,NONE,normal,True
http://guides.lib.ucdavis.edu/aiap_events/,True,NONE,normal,True
https://www.a2datarescue.com/,True,NONE,normal,True
http://www.climatedataprotection.net/,True,NONE,normal,True
https://www.facebook.com/events/358268607865735/,True,NONE,normal,True
https://docs.google.com/document/d/14d58n2C6R4CqxYiRCPjs4YrjqbnPsdHS19LaGT9gC0I/,True,NONE,normal,True
https://www.facebook.com/events/1814020465528122/,True,NONE,normal,True
http://datarescue.web.unc.edu/,True,NONE,normal,True
https://datarescuehouston.wordpress.com/,True,NONE,normal,True
http://researchdata.yale.edu/datarescuenhv/,True,NONE,normal,True
https://www.eventbrite.com/e/seattle-data-rescue-event-tickets-32105338933/,True,NONE,normal,True
https://www.lib.umn.edu/about/datarescue/,True,NONE,normal,True
seed-list.csv
```

An example of a seed list from [Archive-It](#)

quotes