

Participatory Web Archiving

Opening the Black Box of Save Page Now

JESSICA OGDEN

University of Southampton

jessica.ogden@soton.ac.uk

@jessogden

ED SUMMERS

University of Maryland

edsu@umd.edu

@edsu

SHAWN WALKER

Arizona State University

shawn.w@asu.edu

@walkeoh

THE WEB THAT WAS - RESAW 2019 | AMSTERDAM | JUNE 22, 2019

Web Archive
built by Robots
and 1,000 librarians
Save Page Now



Image: David Rinehart (2016)

https://archive.org/details/ia20thanniversaryevent_images/page/n29



Brewster Kahle

@brewster_kahle

Following



651,621,510,000 web URL's now in the Wayback Machine by [@internetarchive](#) . Billions and Billions of web pages! users hitting "save page now" at 100 per second: web.archive.org

7:56 PM - 9 May 2018

68 Retweets 159 Likes



If You See Something, Save Something – 6 Ways to Save Pages In the Wayback Machine

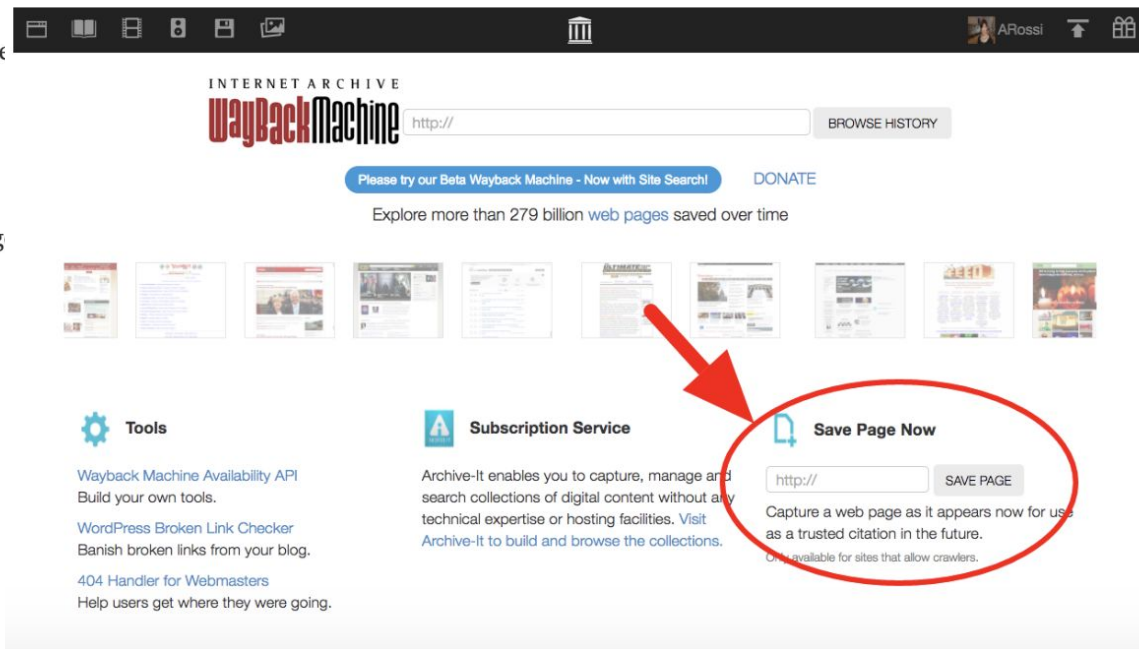
Posted on [January 25, 2017](#) by [Alexis Rossi](#)

In recent days many people have shown interest in making sure the [Wayback Machine](#) has copies of the web pages they care about most. These saved pages can be cited, shared, linked to – and they will continue to exist even after the original page changes or is removed from the web.

There are several ways to save pages and whole sites so that the Machine. Here are 6 of them.

1. Save Page Now

Put a URL into [the form](#), press the button, and we save the page permanent URL for your page.



The screenshot shows the top navigation bar of the Wayback Machine website. Below the navigation bar, the "Wayback Machine" logo is displayed next to a search input field containing "http://". To the right of the search field is a "BROWSE HISTORY" button. Below the search field, there is a blue button that says "Please try our Beta Wayback Machine - Now with Site Search!" and a "DONATE" button. Underneath these buttons, a line of text reads "Explore more than 279 billion web pages saved over time". A row of ten small thumbnail images of various web pages is shown below the text. At the bottom of the page, there are three columns of content. The first column is titled "Tools" and lists several utilities. The second column is titled "Subscription Service" and describes the Archive-It program. The third column is titled "Save Page Now" and features a form with a "SAVE PAGE" button. This "Save Page Now" section is circled in red, and a red arrow points from the text "the form" in the article to the "SAVE PAGE" button.

Tools

- Wayback Machine Availability API
Build your own tools.
- WordPress Broken Link Checker
Banish broken links from your blog.
- 404 Handler for Webmasters
Help users get where they were going.

Subscription Service

Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit [Archive-It](#) to build and browse the collections.

Save Page Now

http://

Capture a web page as it appears now for use as a trusted citation in the future.
Only available for sites that allow crawlers.

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



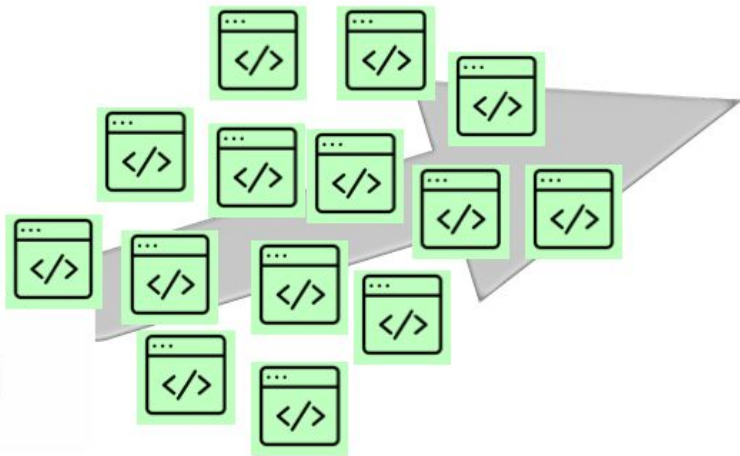
Browser



INTERNET ARCHIVE



Bot



SPN has changed over time

SPN v1

- Heritrix
- *No difference between web & API submissions*

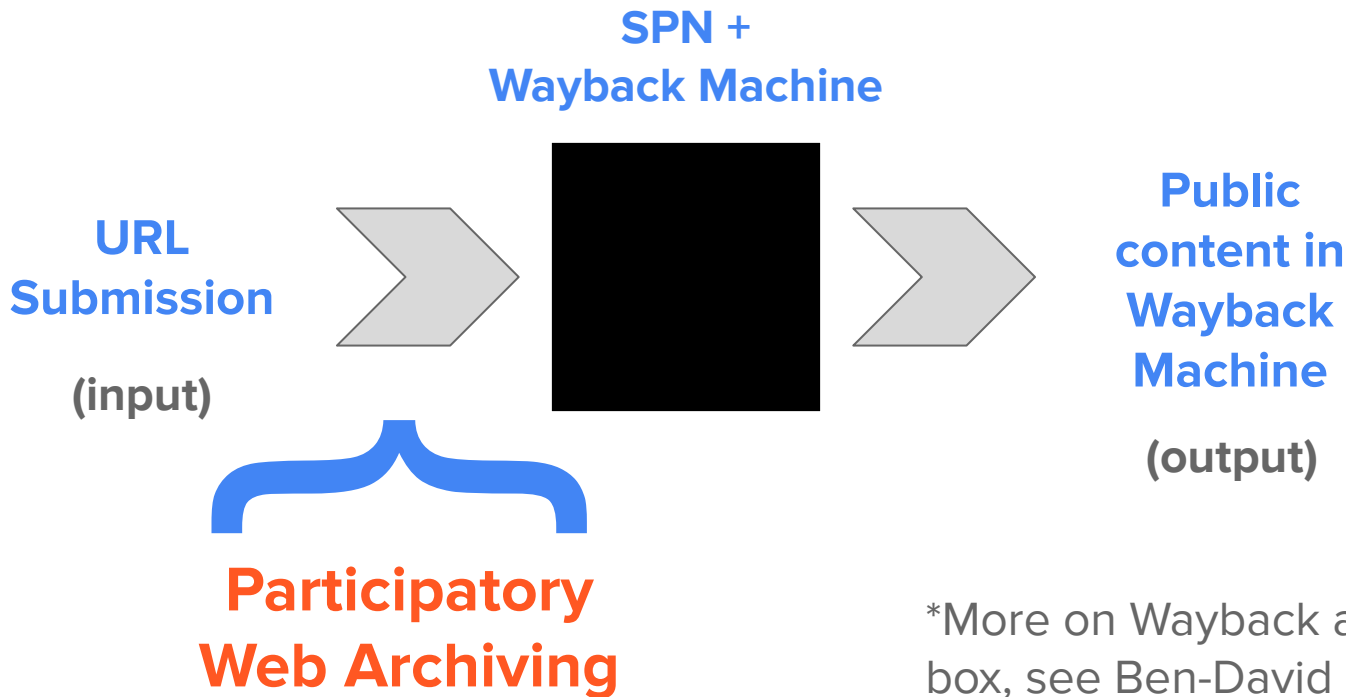
SPN v2

- Heritrix for API submissions
- Browser-based archiving for web submissions

SPN v3

- Server-based headless browser for API submissions
- Browser-based archiving for web submissions

SPN as black box



*More on Wayback as black box, see Ben-David (2018)

Motivating Research Questions

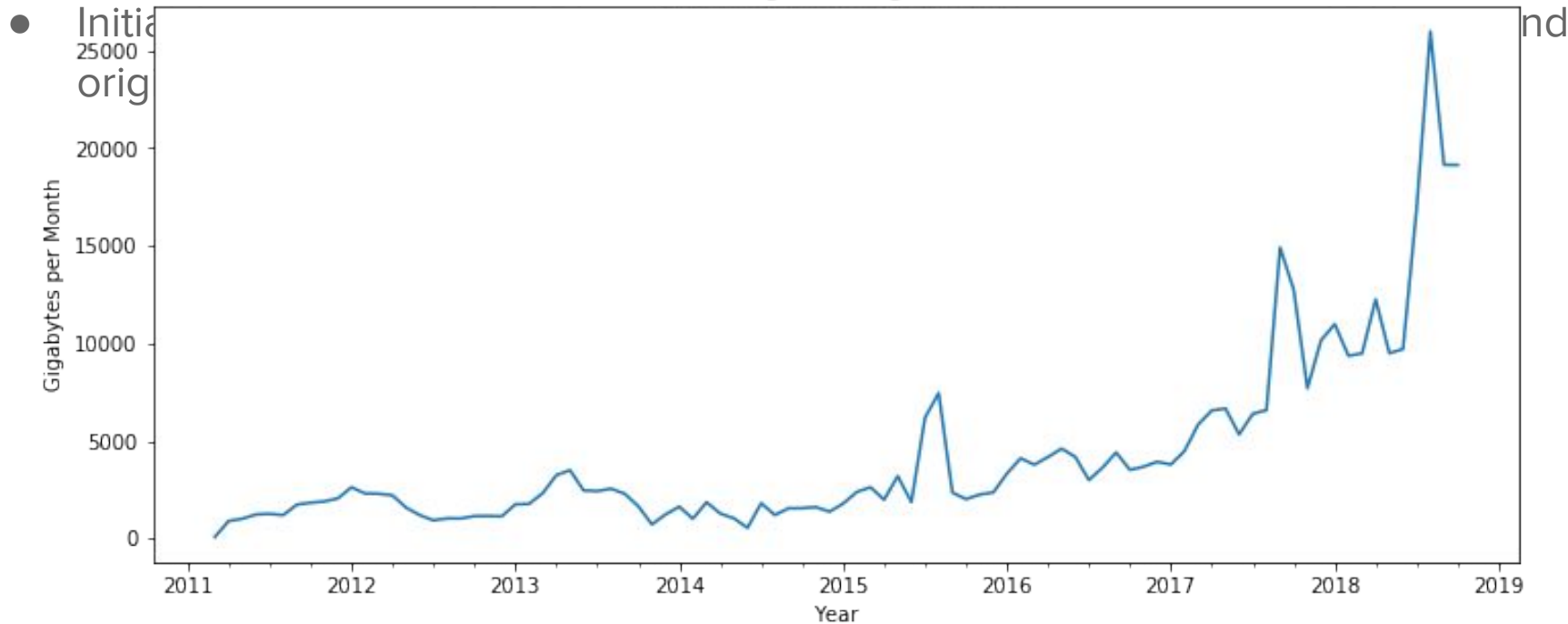
Aim: to understand SPN as form of participatory WA infrastructure

- **RQ1:** What is saved via SPN and how has the ‘collection’ changed over time?
- **RQ2:** To what extent are SPN resources available on the live Web and in other web archives?
- **RQ3:** In what ways is automation a factor in archival production and what purposes does it serve?

Methodology

Methodological Frame (High Level)

Save-Page-Now Ingest Rate

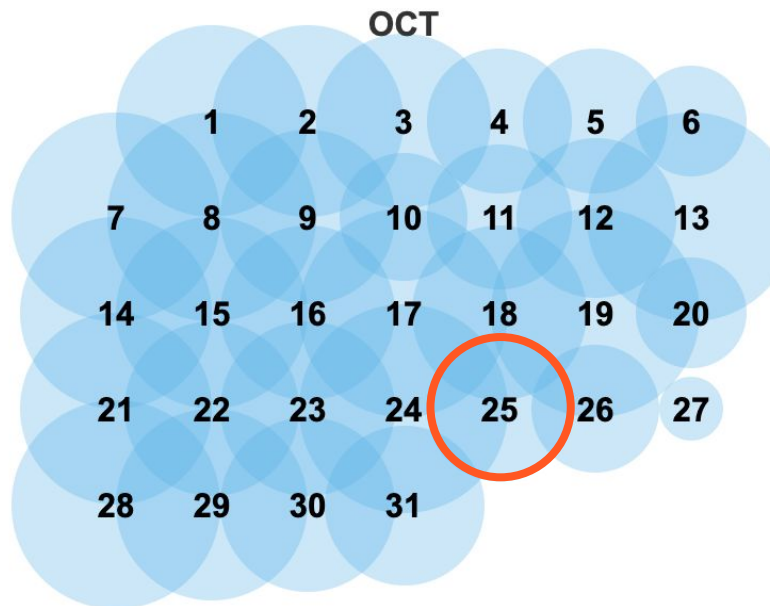




Sample:

- One day per year
- 5 years of data
- Oct 25, 2013 - 2018

SPN first made public on
homepage on October 25, 2013





WARCs
downloaded via
API



+



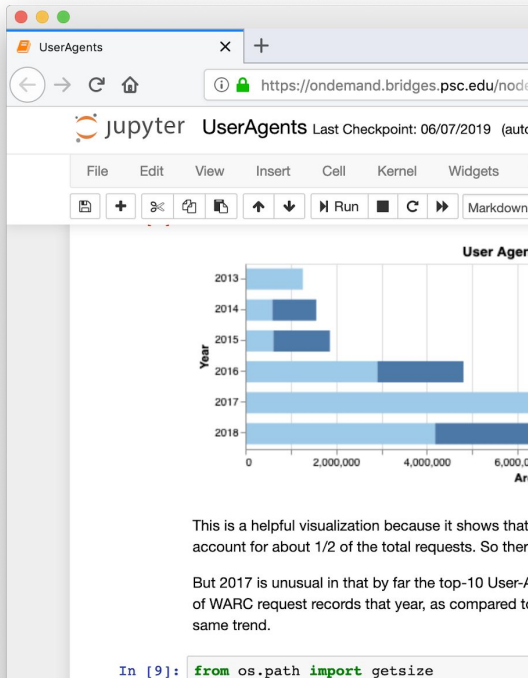
warcio

+



Methodological Frame (High Level)

- Importance of collaborative working



Zoom

[mosh] swalke28@informbat2: /etc/mysql/mysql.conf.d

swalke28@br005: pylon5(ec5fp4p/edsu/spn/liveweb-20181025233542

```
rchive.org/details/archive.org_bot)
User-Agent: mediawords bot (http://cyber.law.harvard.edu)
User-Agent: Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/69.0.3497.100 Safari/537.36
User-Agent: mediawords bot (http://cyber.law.harvard.edu)
User-Agent: Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/69.0.3497.100 Safari/537.36
User-Agent: Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/69.0.3497.100 Safari/537.36
User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/69.0.3497.100 Safari/537.36
User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrom
e/60.0.0.1508 Safari/537.36
User-Agent: Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) SamsungBr
owser/7.0 Chrome/59.0.3071.125 Safari/537.36
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_6) AppleWebKit/537.36 (KHTML, like
Gecko) Chrome/69.0.3497.100 Safari/537.36
User-Agent: python-requests/2.18.1
User-Agent: Mozilla/5.0 (compatible; archive.org bot; Wayback Machine Live Record; +http://a
```

RQ1: SPN Collection Development



Hey guys!



Check out my imageboard

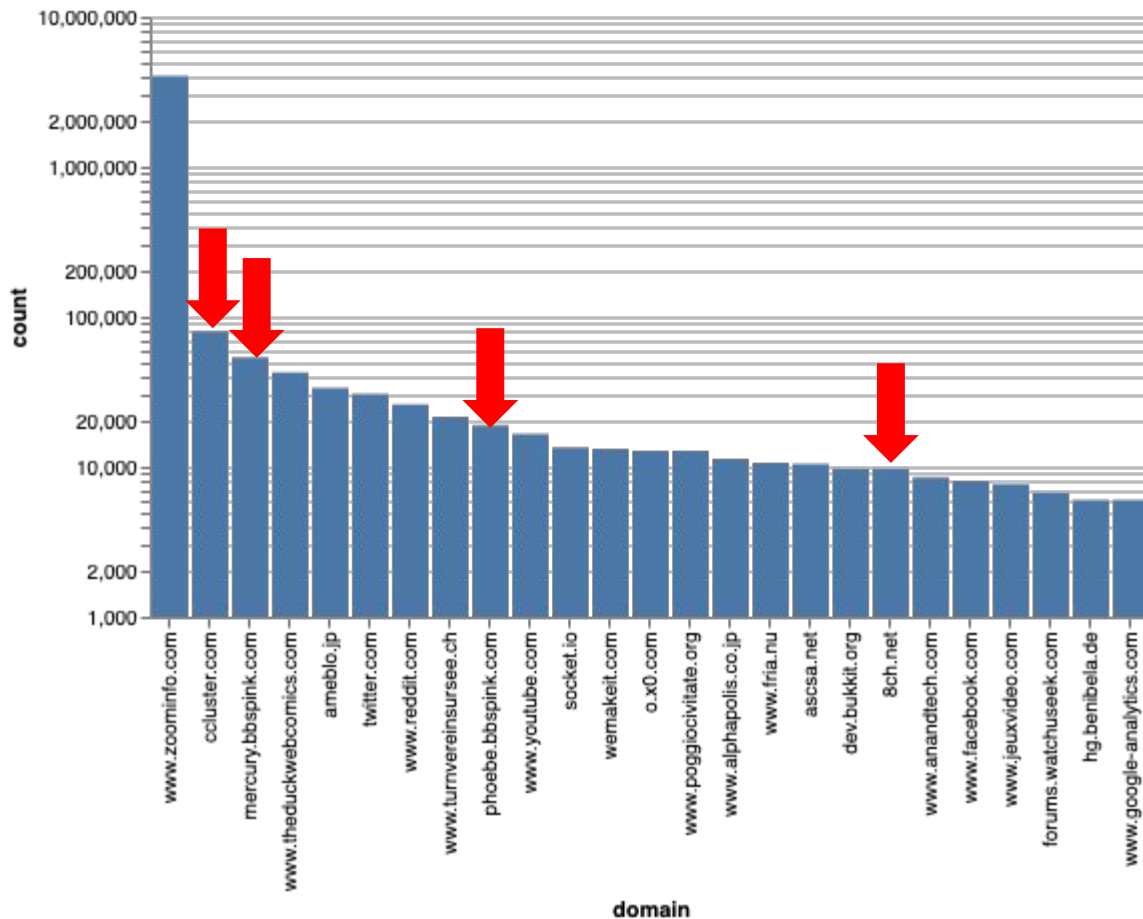
It's called 600chan

It has a /b/ board!!!!

It's also better than 4chan and 8chan combined

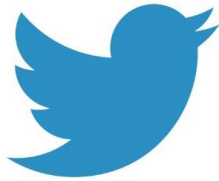
Plz visit my website, no one goes there...

Popular Domains 2017

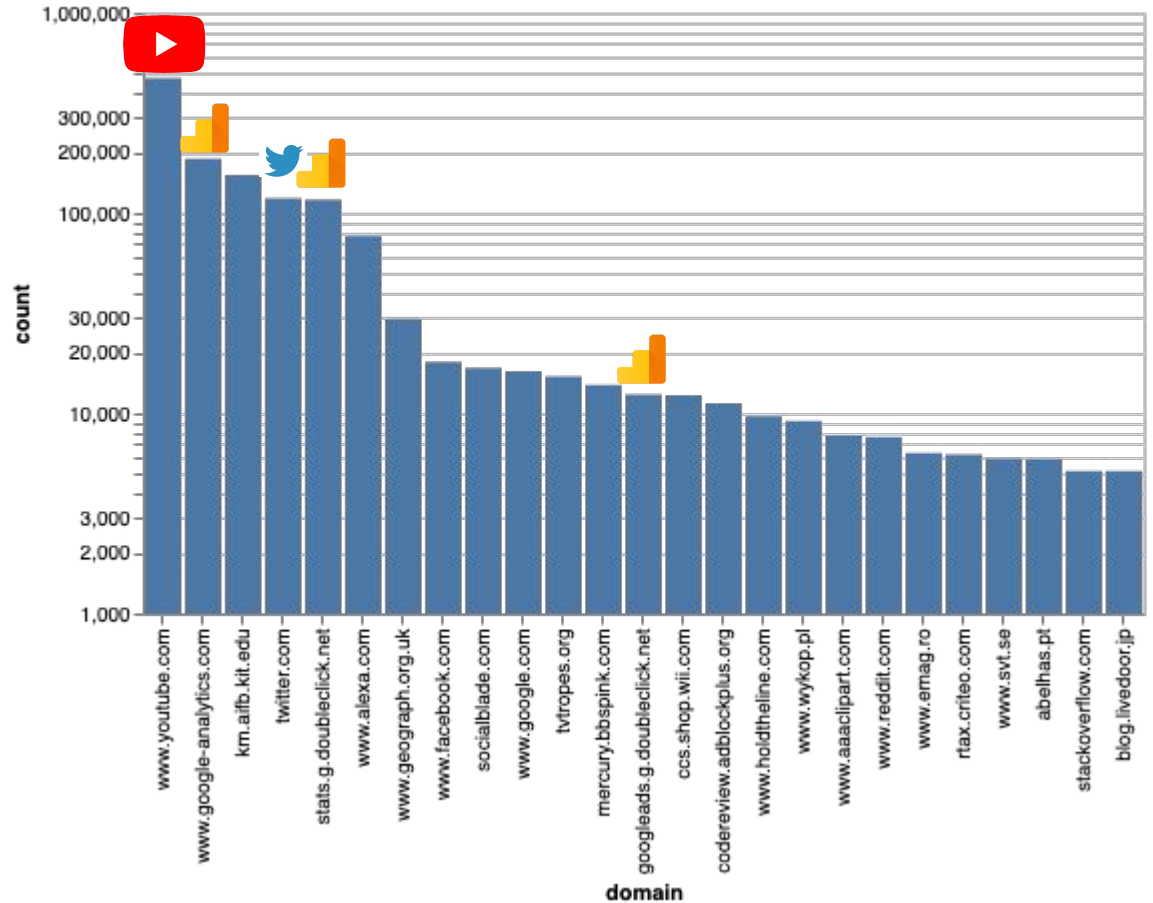




Google Analytics

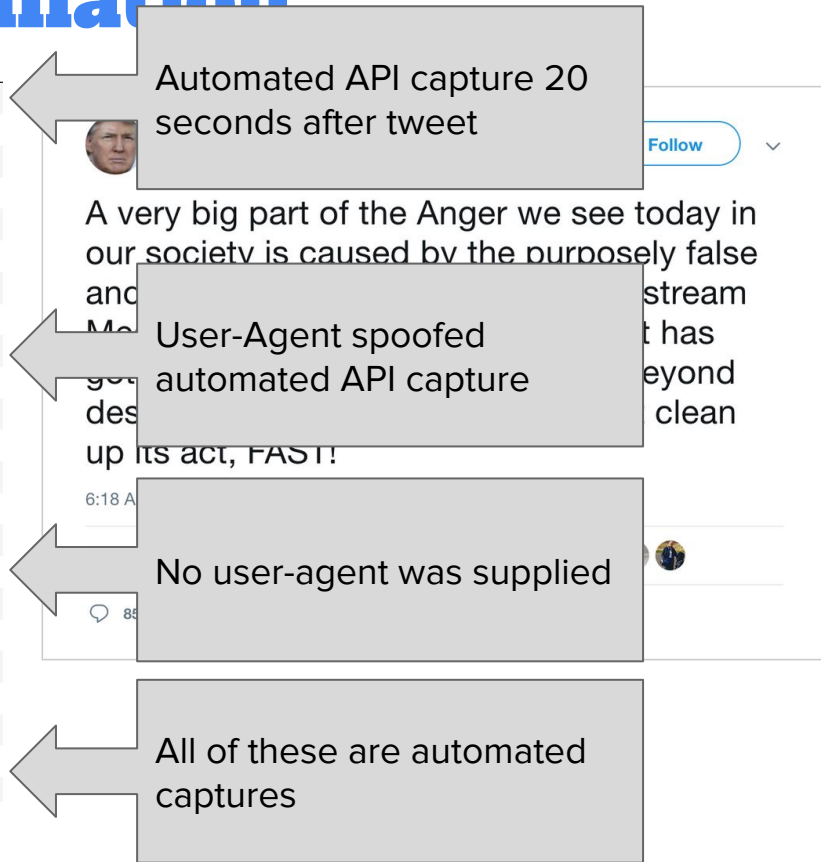


Popular Domains 2018



User-agents and Automation

	date	user_agent
0	2018-10-25T11:18:39Z	python-requests/2.13.0
1	2018-10-25T11:37:07Z	Mozilla/5.0 (compatible; archive.org_bot; Wayback Machine Live Record; +http://archive.org/details/archive.org_bot)
2	2018-10-25T12:00:22Z	python-requests/2.18.1
3	2018-10-25T12:01:45Z	Mozilla/5.0 (compatible; archive.org_bot; Wayback Machine Live Record; +http://archive.org/details/archive.org_bot)
4	2018-10-25T12:11:19Z	python-requests/2.18.1
5	2018-10-25T12:21:36Z	python-requests/2.18.1
6	2018-10-25T12:31:56Z	python-requests/2.18.1
7	2018-10-25T12:43:07Z	python-requests/2.18.1
8	2018-10-25T13:19:18Z	Chrome 41.0.2227.0
9	2018-10-25T13:54:52Z	Firefox 40.1
10	2018-10-25T13:58:42Z	Mozilla/5.0 (compatible; archive.org_bot; Wayback Machine Live Record; +http://archive.org/details/archive.org_bot)
11	2018-10-25T13:58:43Z	Mozilla/5.0 (compatible; archive.org_bot; Wayback Machine Live Record; +http://archive.org/details/archive.org_bot)
12	2018-10-25T14:32:02Z	Safari 5.1.7
13	2018-10-25T15:13:25Z	Chrome 41.0.2227.0
14	2018-10-25T15:37:12Z	Mozilla/5.0 (compatible; archive.org_bot; Wayback Machine Live Record; +http://archive.org/details/archive.org_bot)
15	2018-10-25T15:37:12Z	Mozilla/5.0 (compatible; archive.org_bot; Wayback Machine Live Record; +http://archive.org/details/archive.org_bot)
16	2018-10-25T15:53:29Z	Firefox 33.0
17	2018-10-25T16:32:46Z	Chrome 41.0.2228.0
18	2018-10-25T17:08:11Z	Chrome 41.0.2228.0
19	2018-10-25T17:44:50Z	Firefox 36.0
20	2018-10-25T18:21:49Z	Chrome 41.0.2228.0
21	2018-10-25T18:58:06Z	Safari 6.0
22	2018-10-25T19:34:05Z	Chrome 41.0.2227.1



Automated API capture 20 seconds after tweet

A very big part of the Anger we see today in our society is caused by the purposely false and stream

User-Agent spoofed automated API capture

No user-agent was supplied

All of these are automated captures

Follow

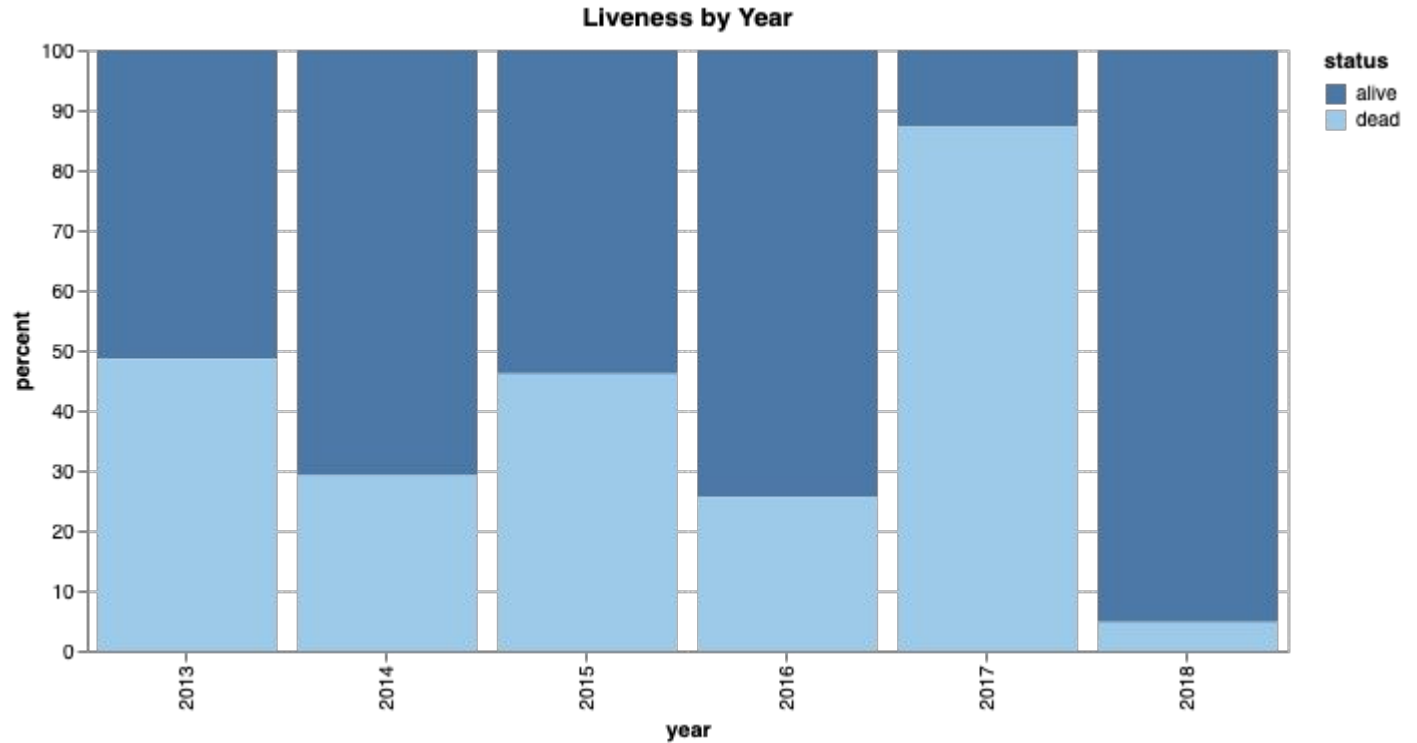
6:18 A

85

The image shows a screenshot of a tweet with several annotations. A grey box at the top points to the tweet's header, stating 'Automated API capture 20 seconds after tweet'. A second grey box points to the tweet's text, which reads 'A very big part of the Anger we see today in our society is caused by the purposely false and stream'. A third grey box points to the tweet's user-agent string, stating 'User-Agent spoofed automated API capture'. A fourth grey box points to the tweet's user profile picture, stating 'No user-agent was supplied'. A fifth grey box points to the tweet's content, stating 'All of these are automated captures'. The tweet also features a 'Follow' button and a timestamp of '6:18 A'. The number of replies is shown as '85'.

RQ2: Measuring Live(li)ness

- Sample of the URLs in dataset
- 2017 - ZoomInfo URLs die



RQ3: Assessing the Role of Automation

Assessing the Role of Automation

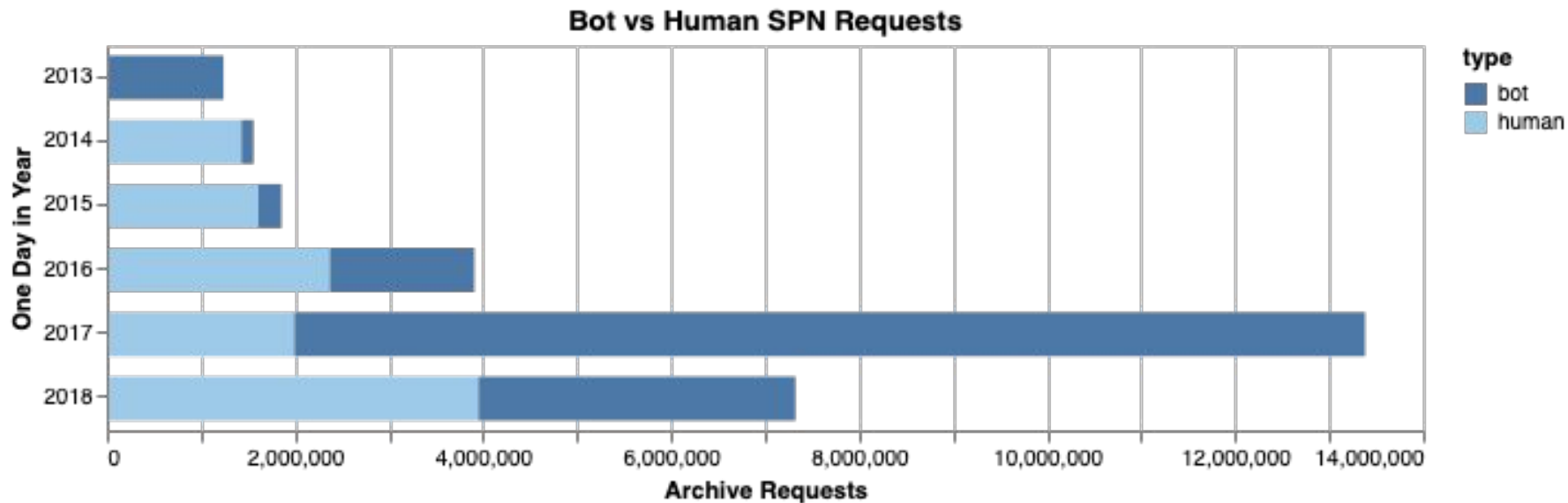
1) Hypothesised two types of SPN interactions -

- Software as **intermediary** (home page tool, browser extension) - SPN *mediates* intent of user and transports to archive
- Software as **actant** (API, cron) - SPN *transforms* selection and transports to archive

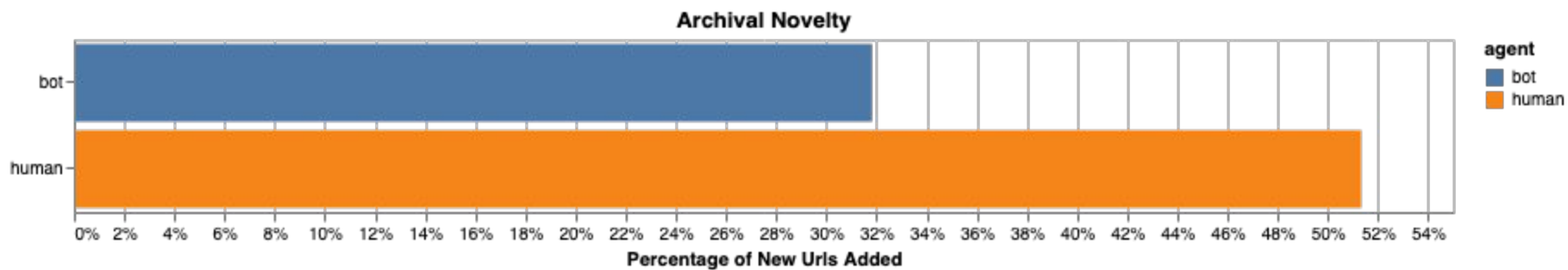
2) User-Agents as proxy for detecting automation

- User-Agents and grouping User-Agent Families
- Percentage of traffic from different types of Users

Bots vs Browsers?



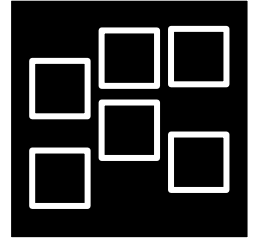
Novelty



Limitations

- Sample strategy - it's **one day a year**.
- Detecting automation
- Conceptualizing web archival attributes (e.g. 'novelty')
- URLs - delineating targets → inferring intentionality

Reflexivity as Strategy



- Complex layering of data and findings (boxes within boxes), ways that proxies/data abstraction often creates more questions + uncertainty
- Problematizing *situatedness* (Haraway 1988) of so-called big data (+ extra complexities of SPN) - recognizing that data views are always partial, '*cooking data with care*' (Bowker 2005; Geiger and Halfaker 2018)
- Value of triangulation, epistemological flexibility - WA research as *boundary work* (Star and Griesemer 2015; Gieryn 1983) - requires a lot of translation

Future work...

Acknowledgements



XSEDE

Extreme Science and Engineering
Discovery Environment

This work used the Extreme Science and Engineering Discovery Environment (XSEDE) Bridges and Bridges Storage at the Pittsburgh Supercomputing Center through allocation TG-ECS180012. Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562.