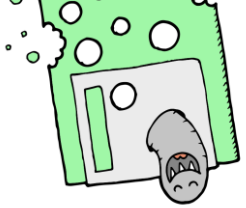
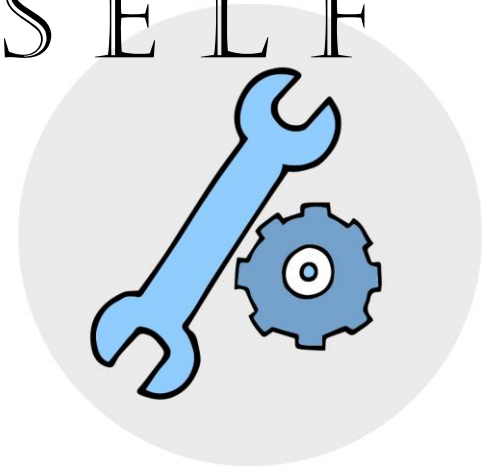
THE UNIVERSITY of EDINBURGH



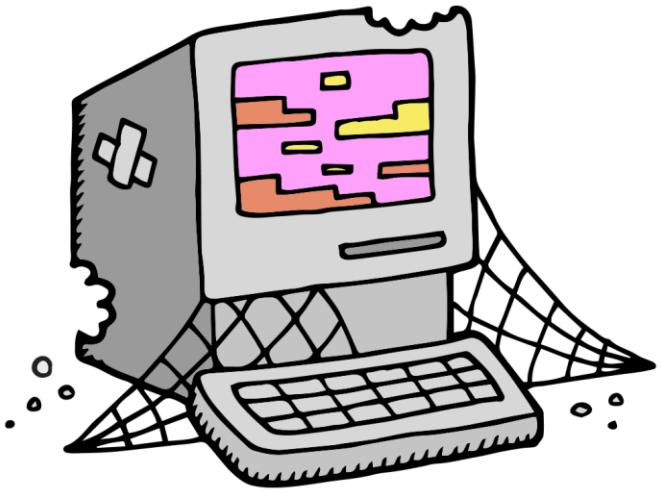
Web Archiving:

DO IT YOURSELF



Sara Day Thomson
Digital Archivist
Centre for Research Collections

Why archive the web?



Political movements & activism

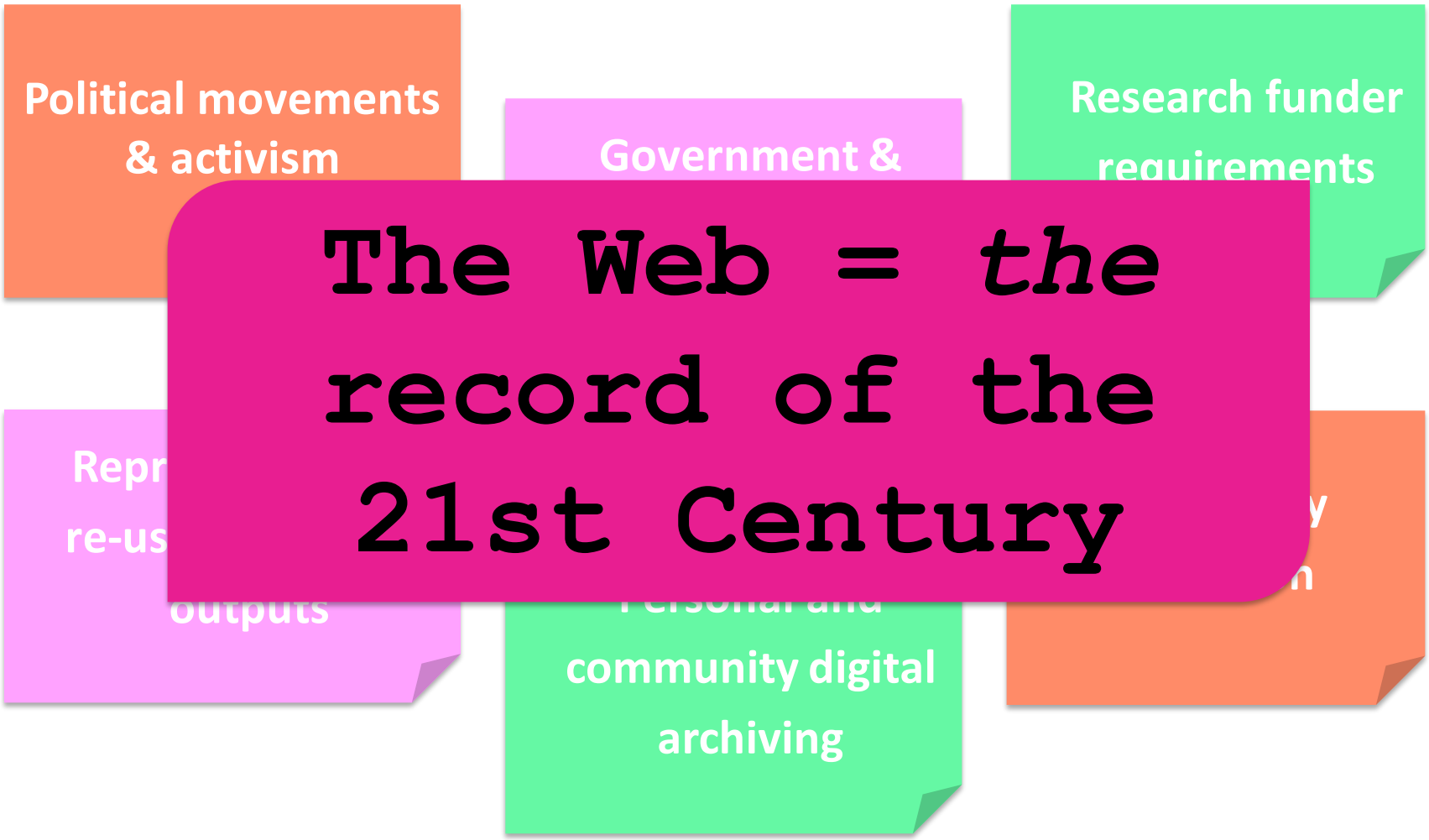
Government & corporate accountability

Research funder requirements

Reproducibility & re-use of research outputs

Personal and community digital archiving

Scholarly Citation


Political movements & activism
Government &
Research funder requirements
Repr
re-us
outputs
Personal and community digital archiving
The Web = the record of the 21st Century

Why archive web pages myself?

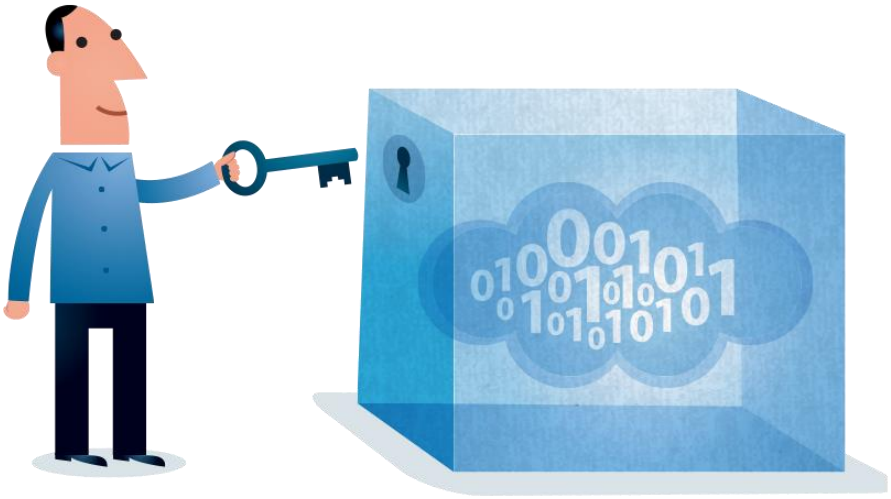


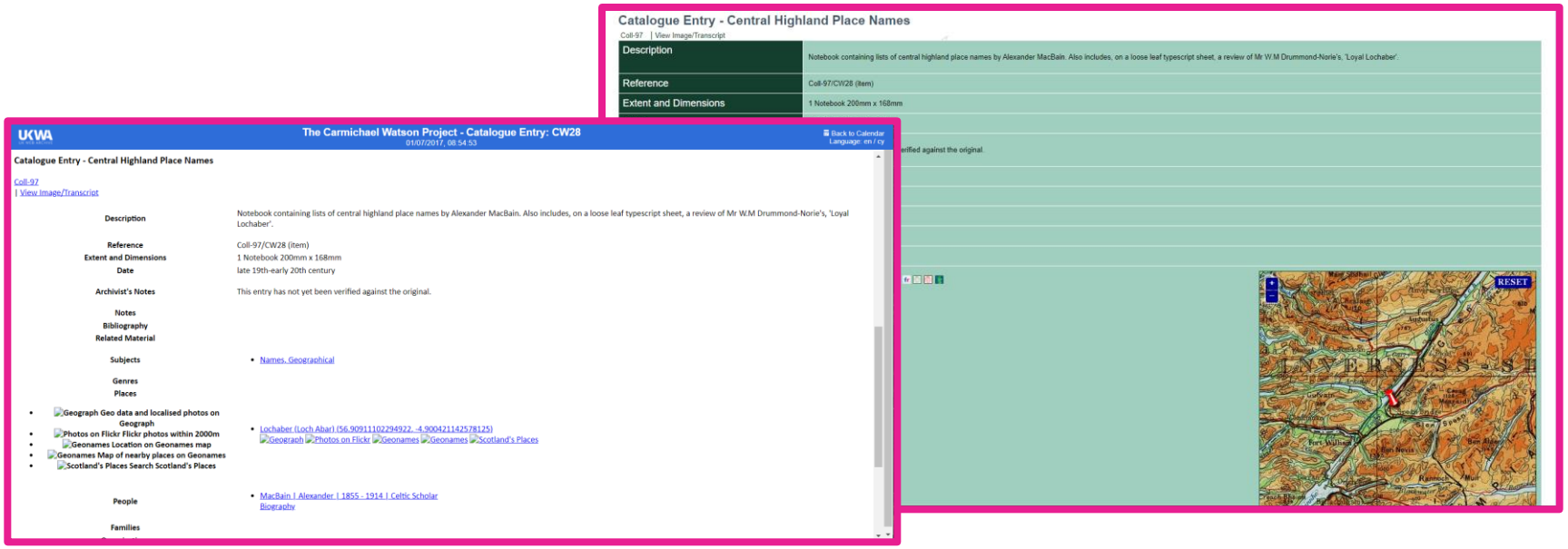
Image: https://digitalbevaring.dk/

Why DIY…

Web page(s) not archived by UKWA or other web archiving programme

Why DIY…

Web page(s) is archived, but is broken or incomplete



Why DIY…

Requirement for an offline copy

Image: No Wifi by Sergey Demushkin from NounProject.com

Why DIY…

Requirement for a different format, e.g. social media data


```
}, {
  "retweeted" : false,
  "source" : "<a href=\"http://twitter.com/download/android\"
  rel=\"nofollow\">Twitter for Android</a>",
  "entities" : {
    "hashtags" : [ {
      "text" : "DPC",
      "indices" : [ "67", "71" ]
    } ],
    "symbols" : [ ],
    "user_mentions" : [ {
      "name" : "Sharon McMeekin",
      "screen_name" : "SharonMcMeekin",
      "indices" : [ "3", "18" ],
      "id_str" : "124154729",
      "id" : "124154729"
    } ],
    "urls" : [ ]
  },
  "display_text_range" : [ "0", "139" ],
  "favorite_count" : "0",
  "id_str" : "1147130277296246785",
  "truncated" : false,
  "retweet_count" : "0",
```

Why DIY…

Files required for project documentation, research data management, or personal digital archiving

Image: https://digitalbevaring.dk/

When to DIY…

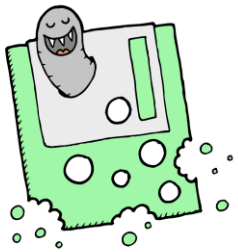
- When citing a web resource in a publication
- As soon as a new web page goes online
- Before big changes (e.g. upgrade) or deletion of out-of-date content
- Before migrating to a new service or platform

Image: https://digitalbevaring.dk/

Web pages evolve – if you don't begin collecting at the point of publication, the record of changes over time will be lost

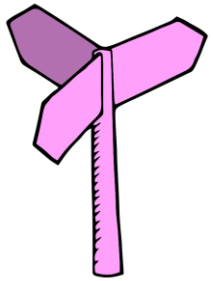
Link Rot

Web page(s) disappears or returns an error (404)



Link Drift

Content on web page or entire domain changes



Study conducted by Harvard Law School:
>70% of links in a sample of law journals don't work
>50% of links in Supreme Court opinions don't work

Harvard Law study: https://perma.cc/48QZ-DT6L

DIY Approaches

- Check other web archives – it might already be archived
- Ask a web archive to do it for you (nominate a web page(s))
- Manually capture a web page(s) with a Webrecorder tool
- Download code and files for cheap & cheerful local copy

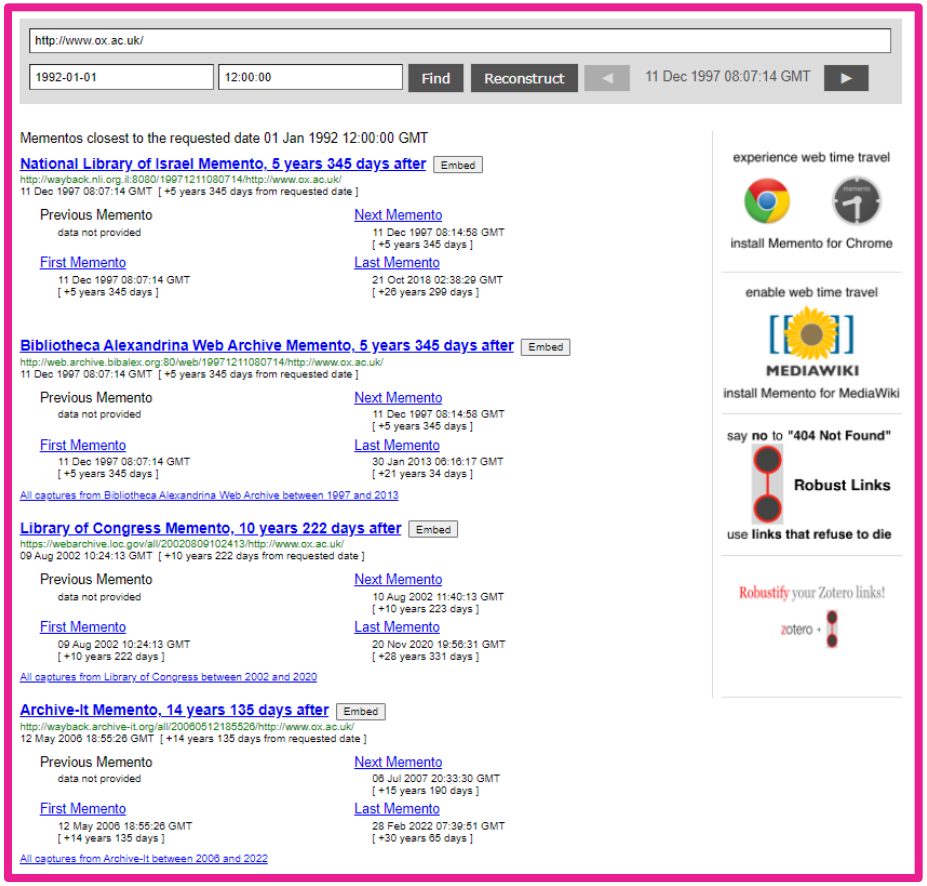
- Use a citation service
- Harvest machine-readable data (social media)
- Download your data directly from a platform

Check other web archives…

https://timetravel.mementoweb.org/

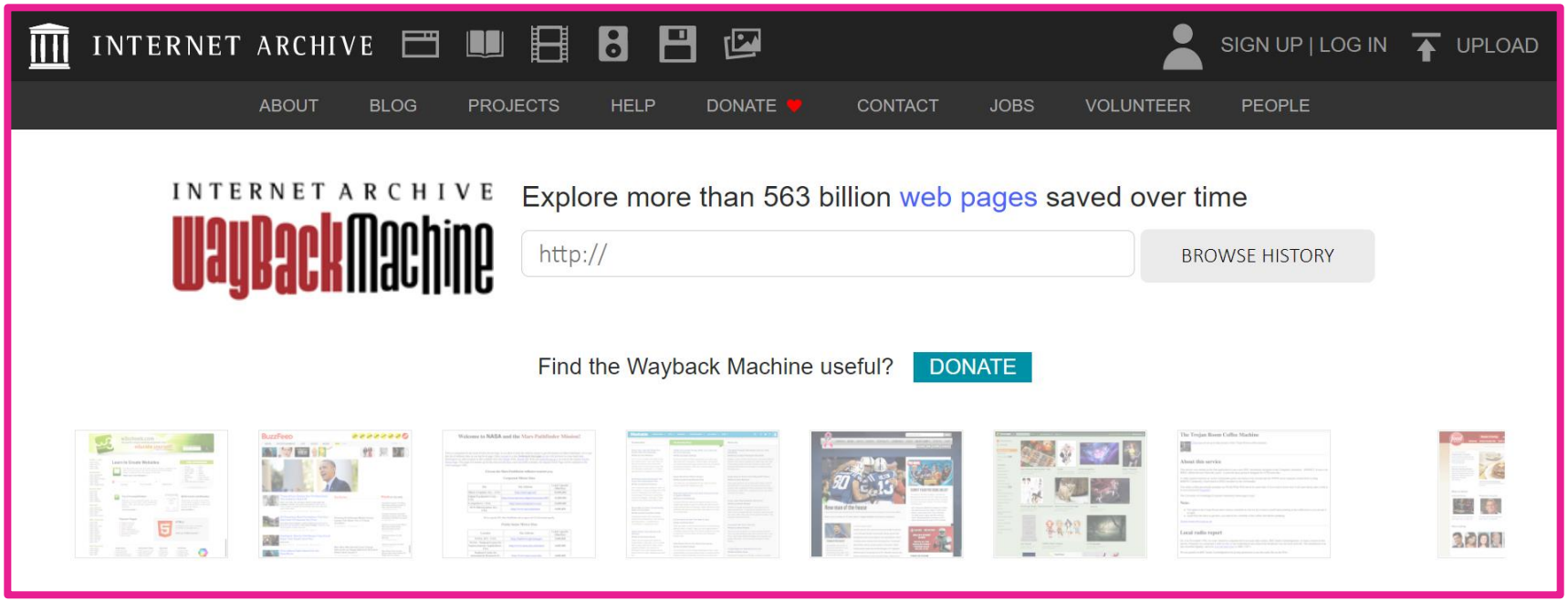
Check other web archives…

https://ox.ac.uk



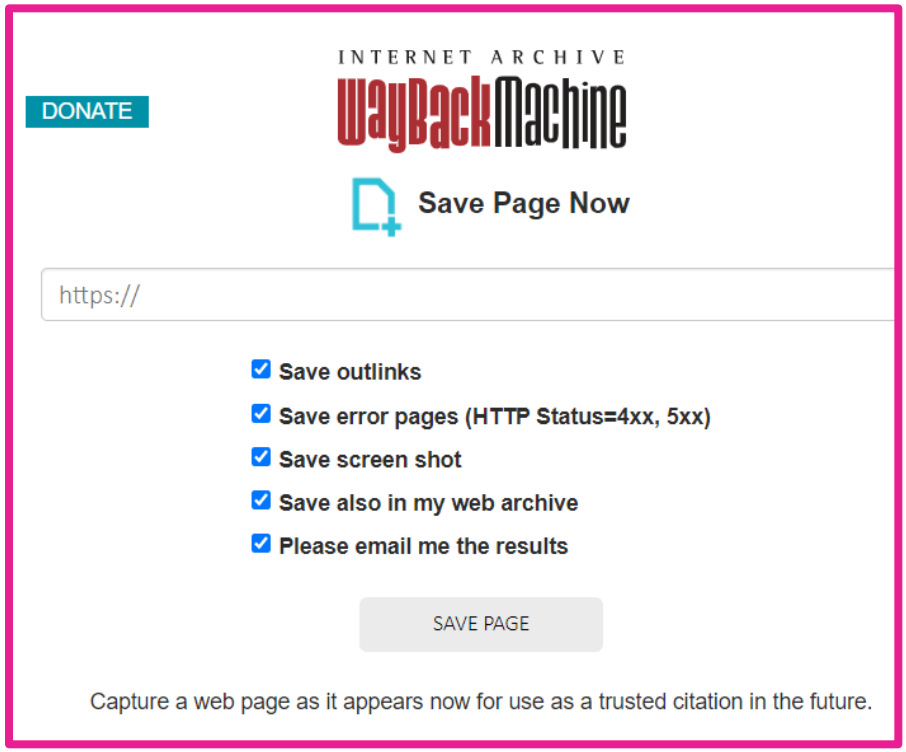
etc etc

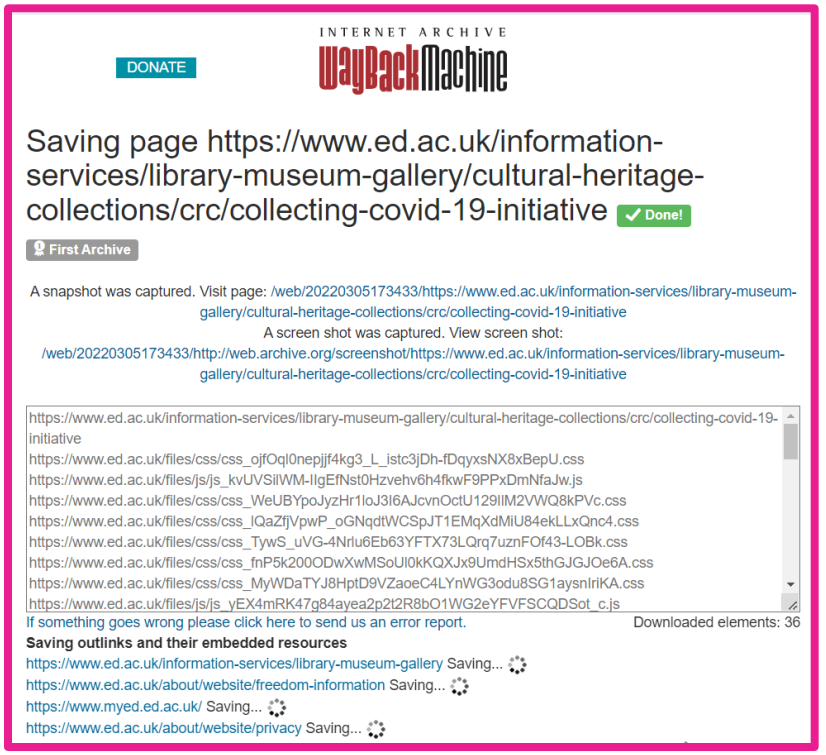
Check other web archives…



https://web.archive.org/

Ask a web archive to do it for you …





https://web.archive.org/save/

THE UNIVERSITY of EDINBURGH

http://www.webarchive.org.uk/en/ukwa/info/nominate

Save a UK Website

We archive websites:

1. That are on a .uk or other UK geographic top-level domain such as .scot or .cymru
2. That are published in the UK.

We do *not* archive:

1. Online Sound or Video platforms, in which audio-visual material is the predominant content
2. Private Intranets and Emails
3. Personal data in social networking sites or websites only available to restricted groups.

Save a UK website

Fields marked with * are mandatory

Personal details you provide on this form are protected by UK Data Protection Law. Please view our Privacy Statement.

* Name

Please enter your name

Email address

Please enter your email address

We need your email address if you require a response from us. It would also be helpful if you are the owner of the nominated website and would like us to make archival copies of the website available through the UK Web Archive.

* Title of website

Please enter the title of website

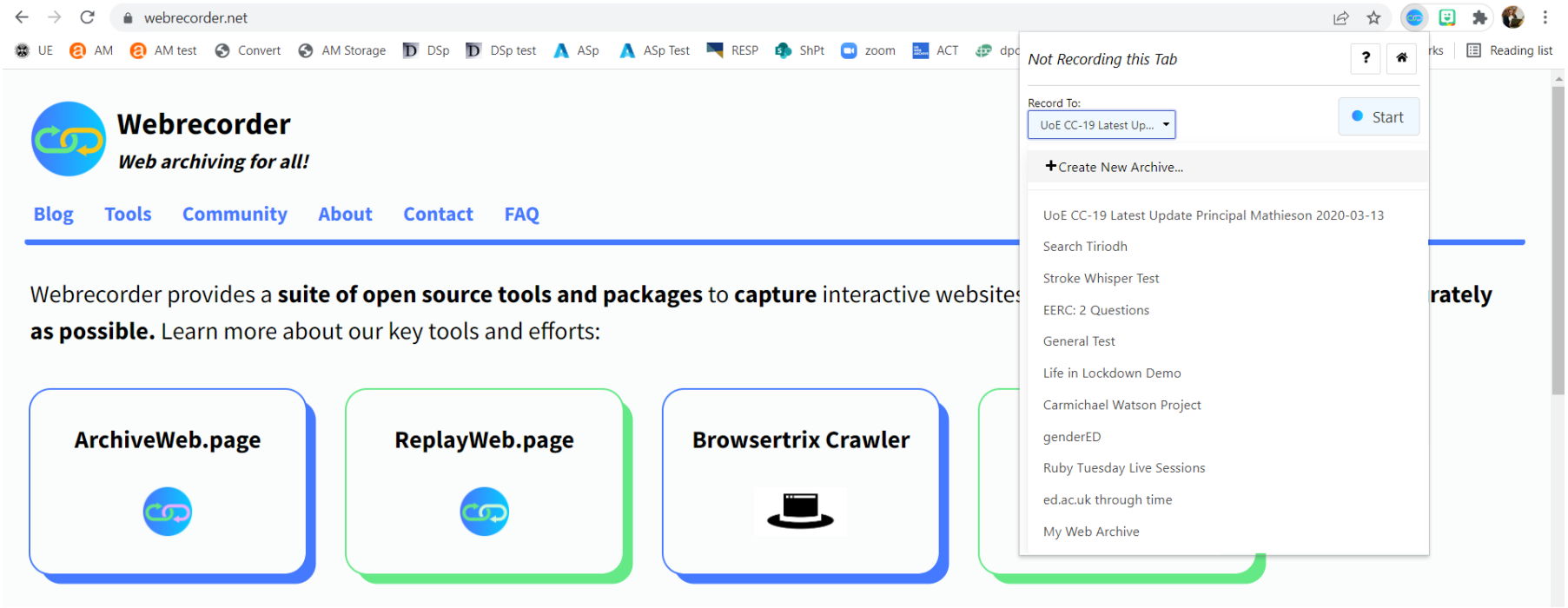
* URL of the website

Please enter URL of website

Any other information

Use this space to tell us for example, if you own this website, or if it is about to go offline

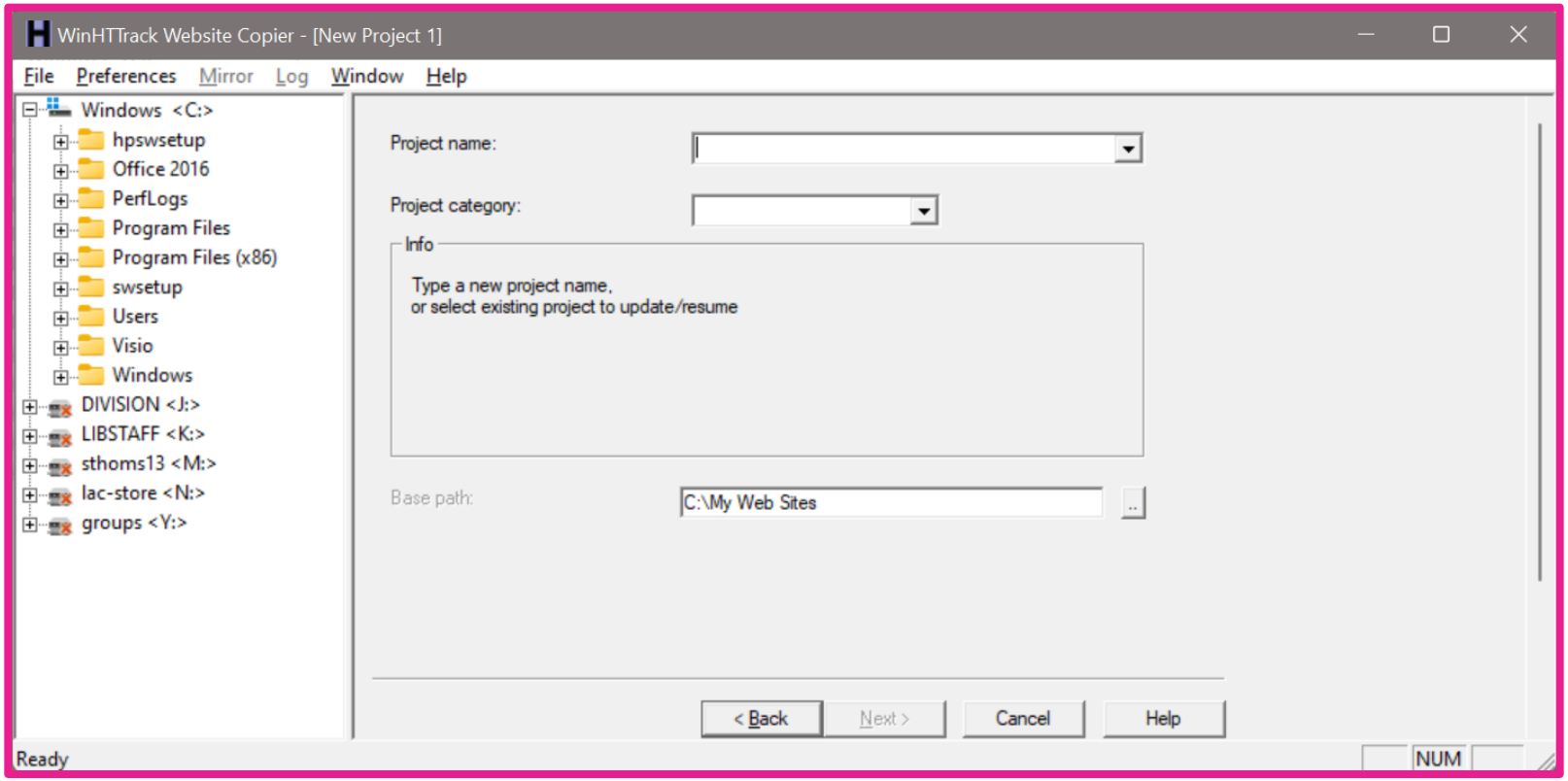
Manually capture a web page(s) with Webrecorder...



https://webrecorder.net/

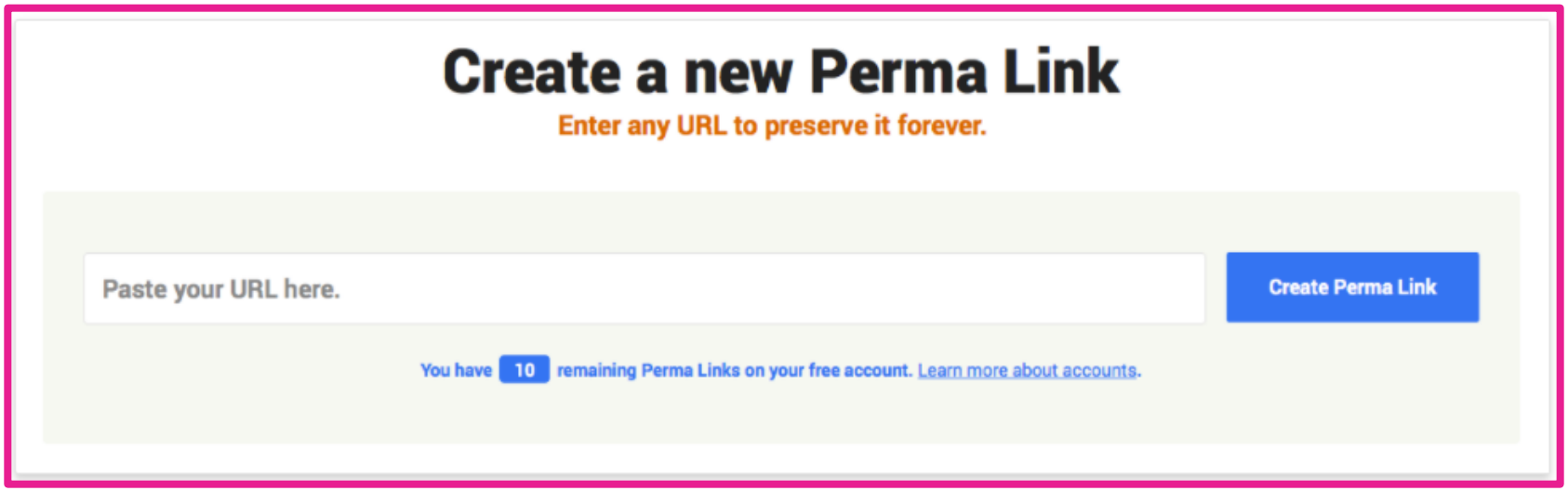
HTTrack WEBSITE COPIER
Free software offline browser

Download code and files for cheap & cheerful local copy…



https://www.httrack.com/

Use a citation service…



https://perma.cc/

Disclaimer: Paid Service (legal publications)

twarc.py

Harvest machine-readable data (social media) …

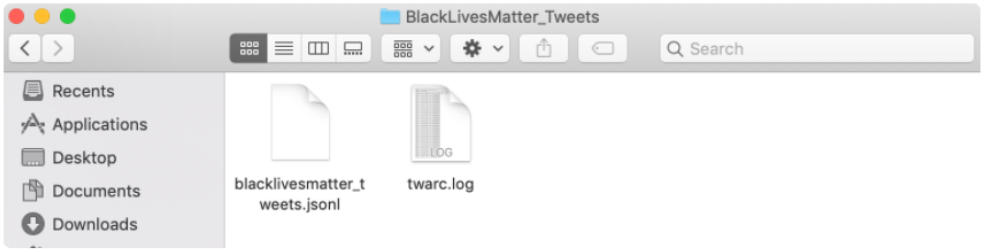
Type the following into your command prompt and press enter:

```
twarc search blacklivesmatter > tweets.jsonl
```

- Your collection may take some time to return all tweets. You can tell when the process has completed if your terminal has returned to the shell ($) prompt:

```
eam5hc$ twarc search blacklivesmatter > blacklivesmatter_tweets.jsonl
eam5hc$
```

- While the prompt is running, check to make sure your command was successful by looking inside the twarc folder. Inside you should see your tweets.jsonl file and a twarc.log file (You don't need to open these files, just make sure they've been created.):

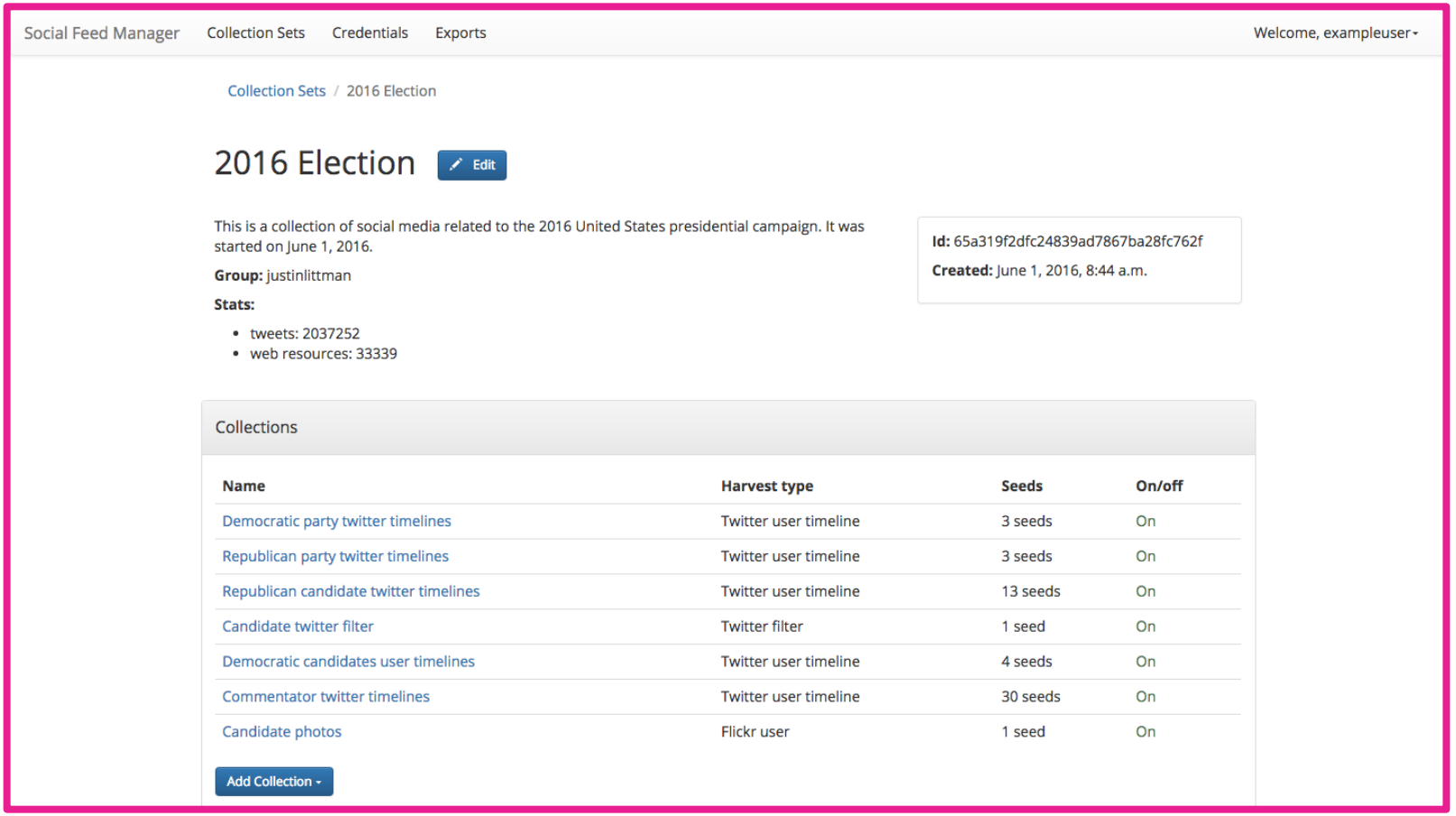


https://scholarslab.github.io/learn-twarc/

THE UNIVERSITY of EDINBURGH

Social Feed Manager

Harvest machine-readable data (social media) …



https://social-feed-manager.readthedocs.io/

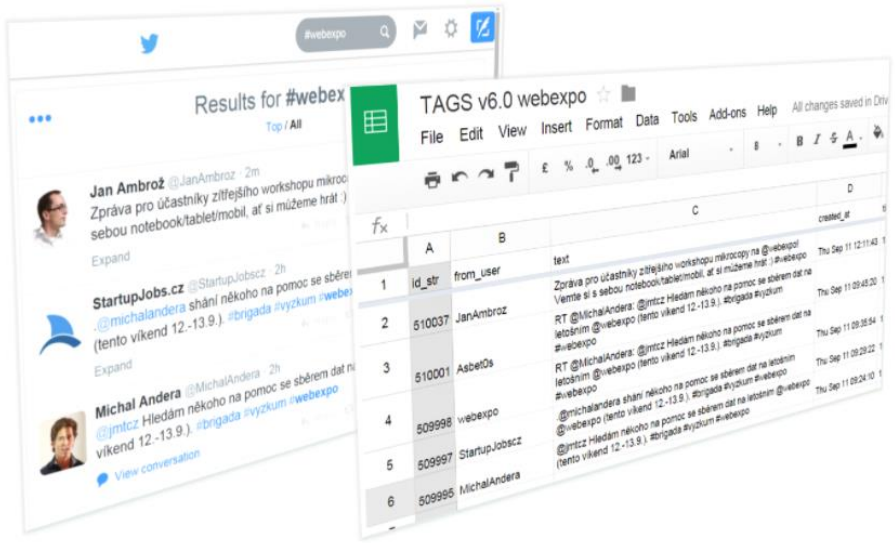
Harvest machine-readable data (social media) …

#TAGS

TAGS – Twitter Archiving Google Sheet

https://tags.hawksey.info/

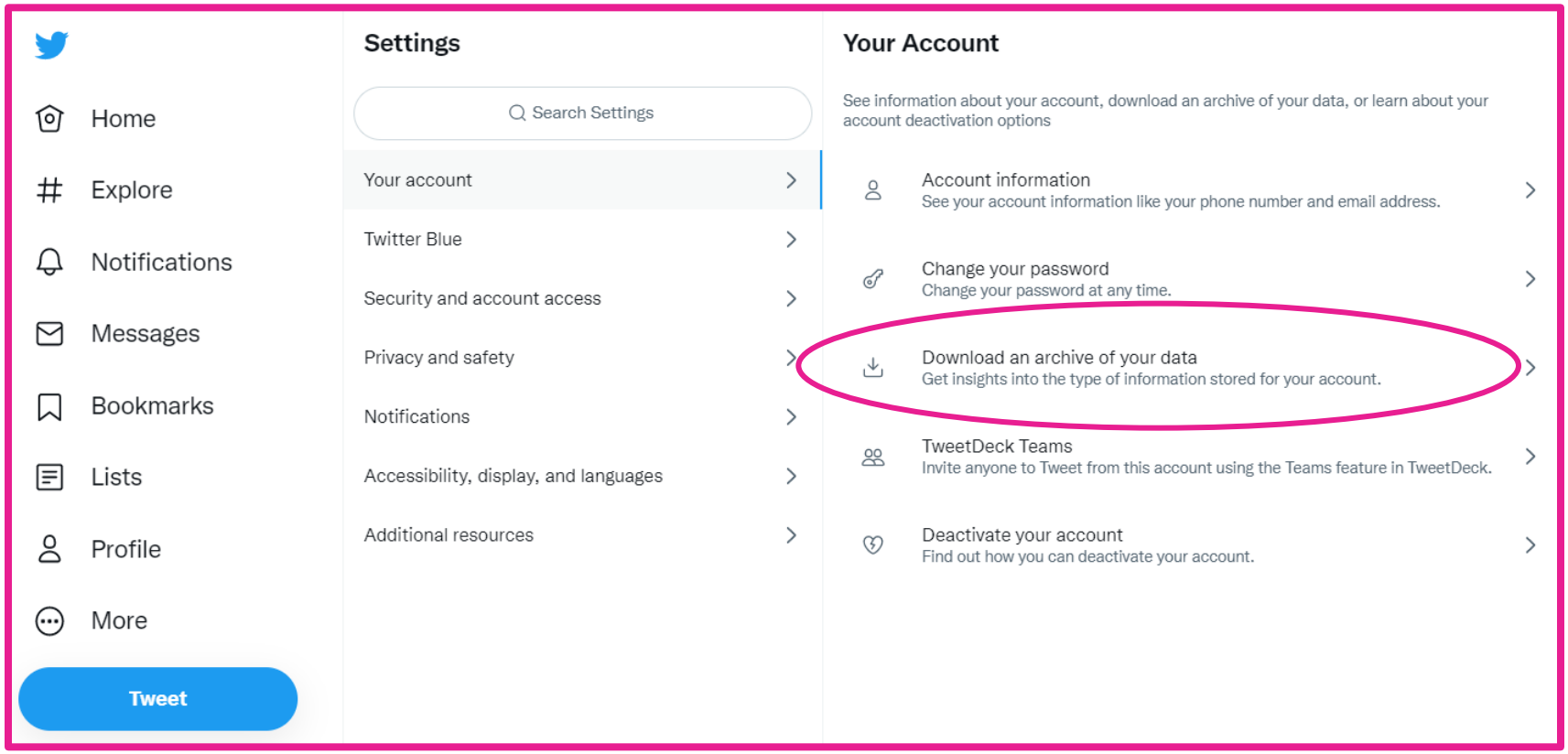
Get TAGS



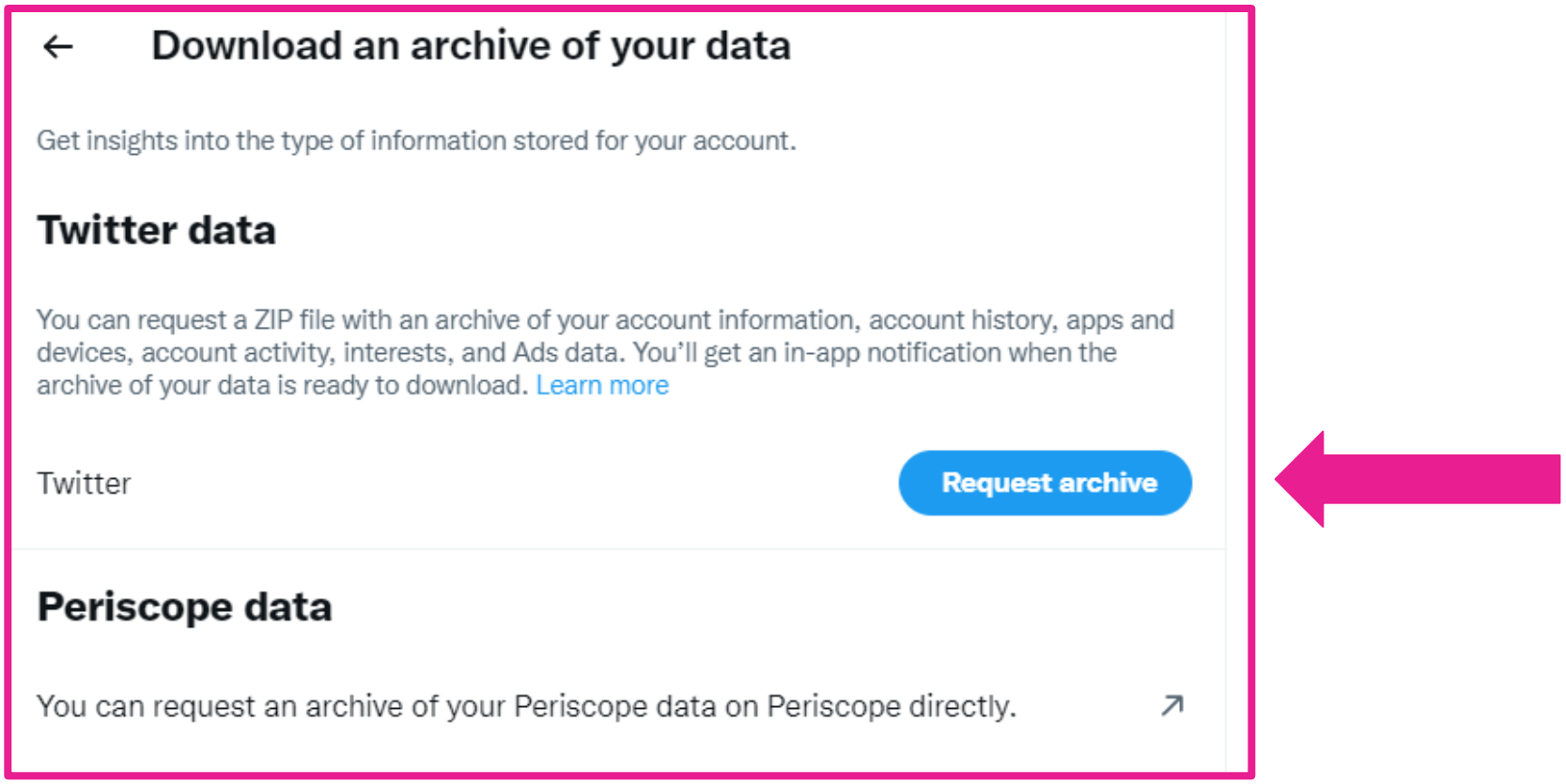
To start using TAGS select one of the versions below then follow the steps below:

TAGS v6.0
TAGS v6.1

Download your data directly from a platform…

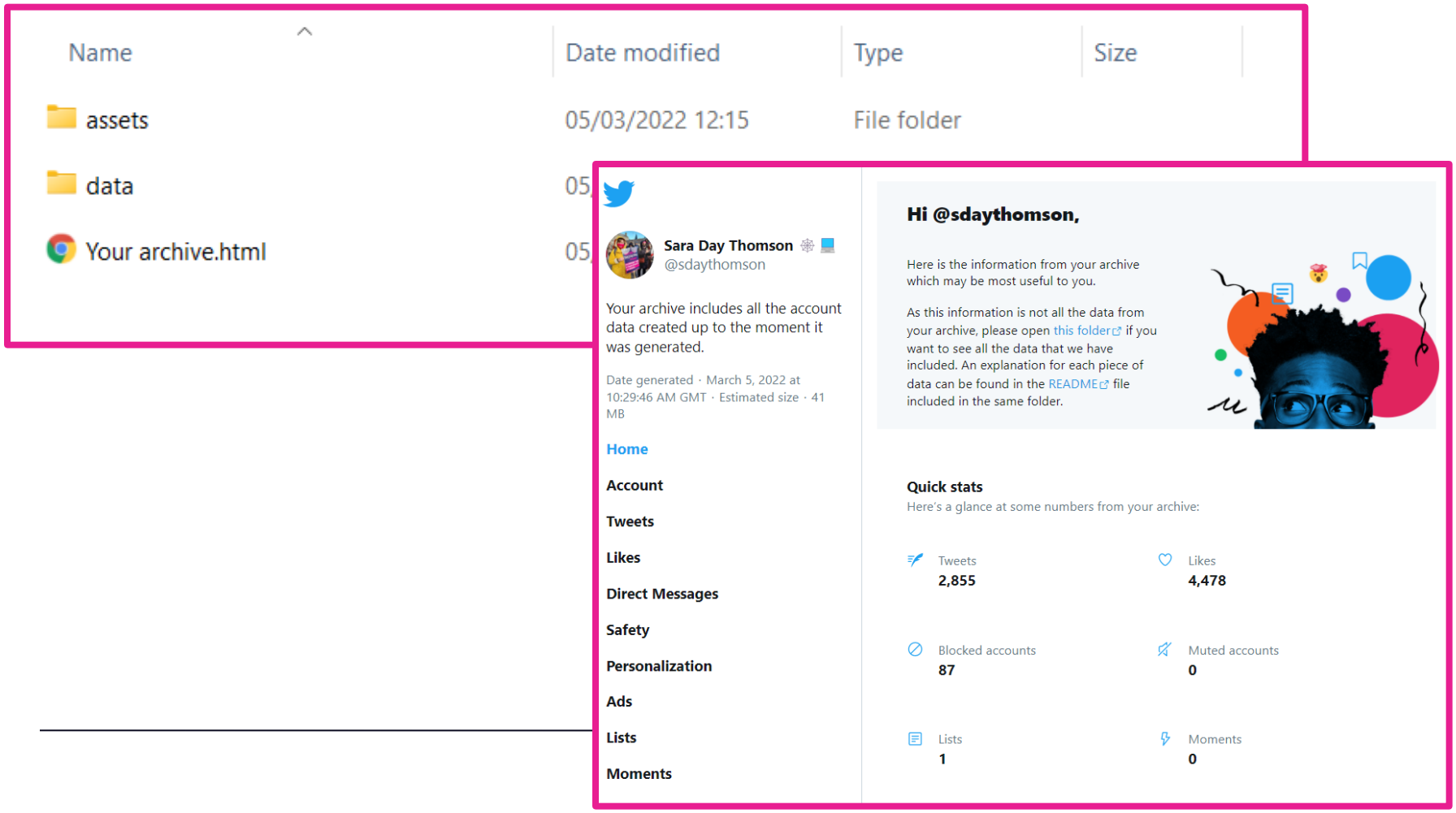


Download your data directly from a platform…



THE UNIVERSITY of EDINBURGH

Download your data directly from a platform…



Now for a Demo

Webrecorder

Web archiving for all!

Webrecorder

Web archiving for all!

About | Tools | Blog | FAQ | Contact | Jobs

http://webrecorder.net/

Latest News

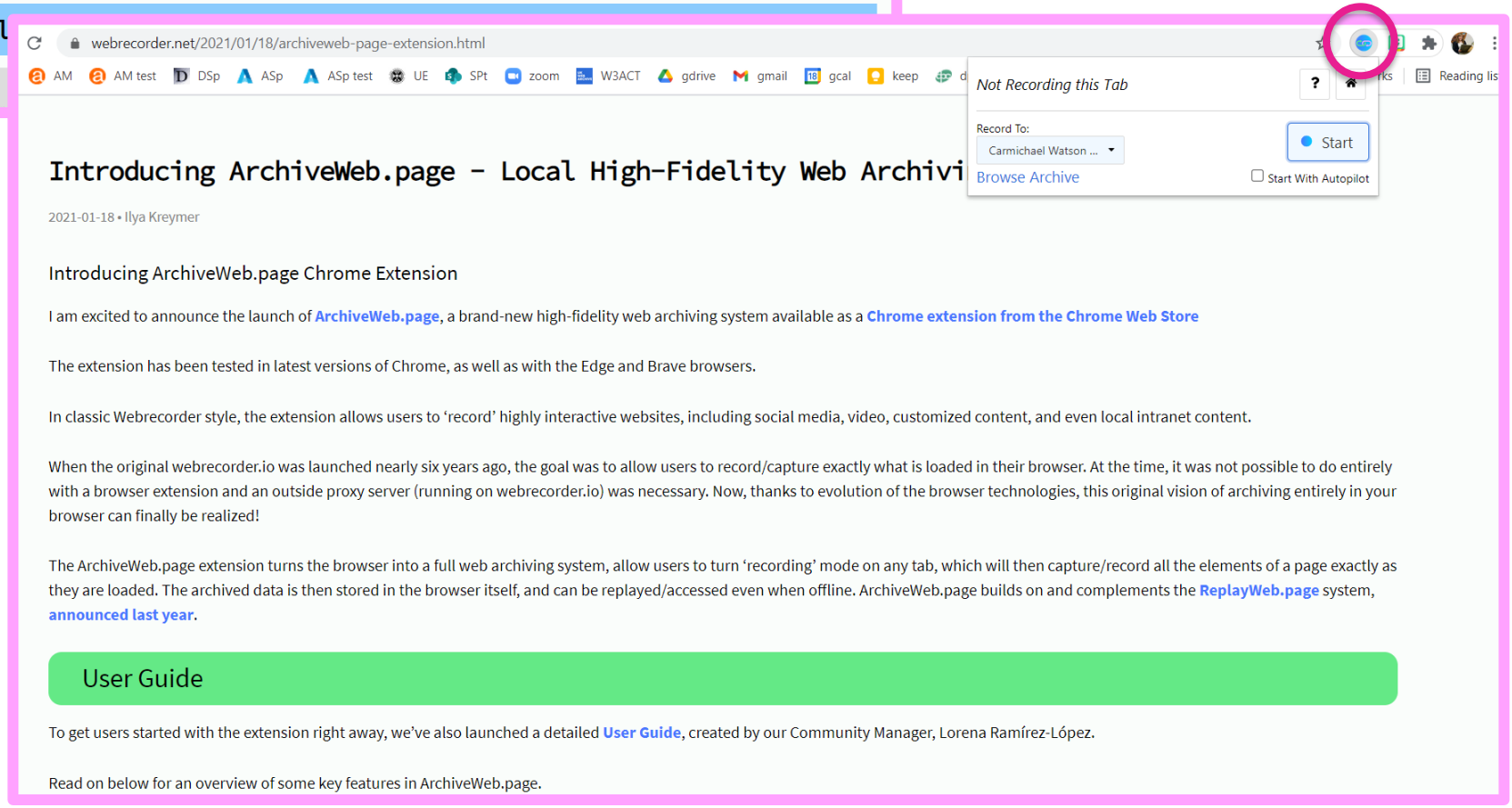
Read about the **ArchiveWeb.page Chrome Extension for high-fidelity web archiving**.

Read about the **release of the new OldWeb.today browser emulation system**.

Featured Tool

ARCHIVEWEB.PAGE

Webrecorder Toolset



Crawler

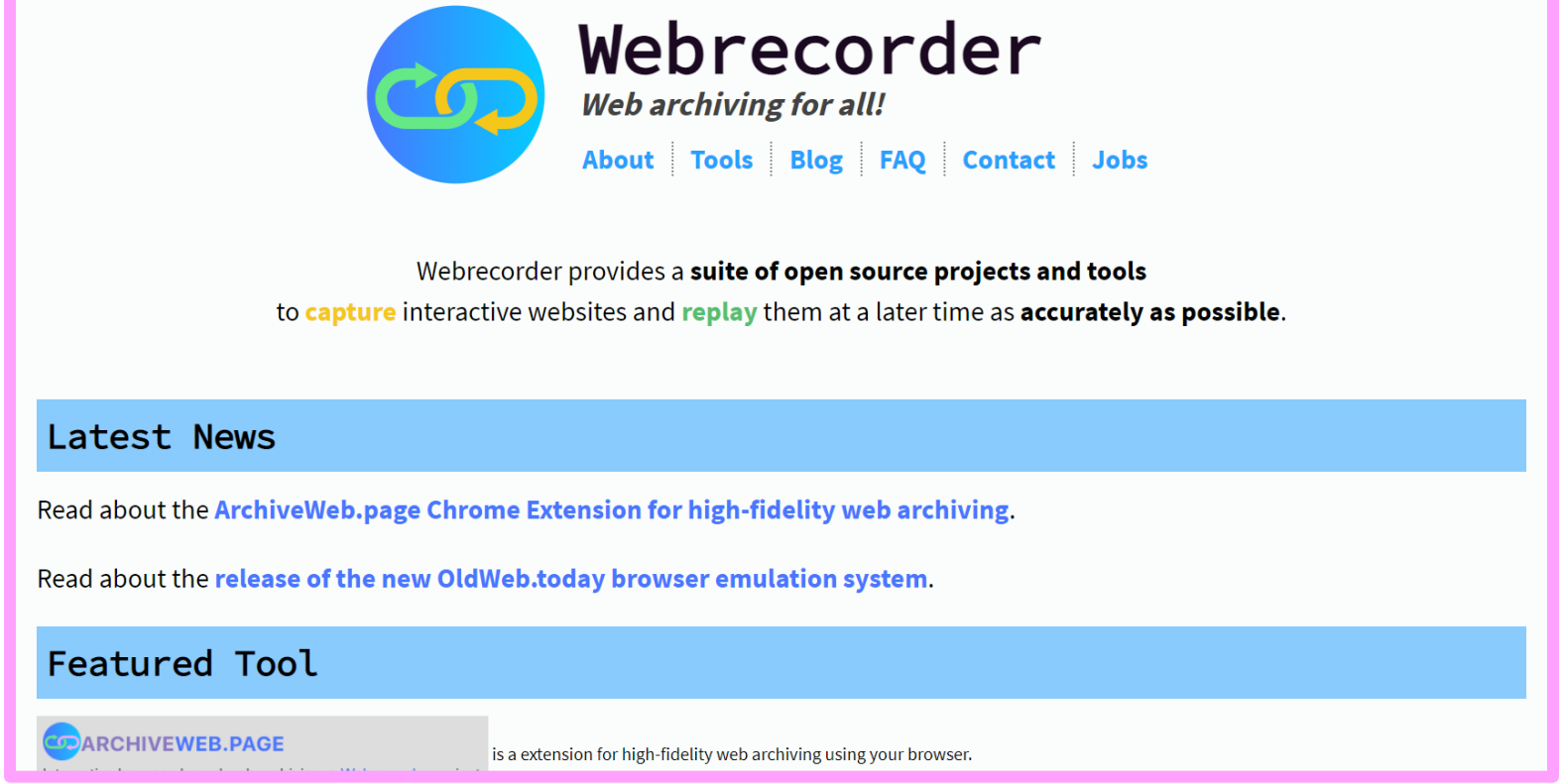
- Automated
- Capable of capturing high volume of URLs
- Typically captures minimal metadata
- Doesn't always capture AV or dynamic content
- Requires IT configuration and institutional infrastructure

vs

Webrecorder

- Manual
- More effective for small number of URLs
- Supports close qualitative analysis and more detailed metadata
- Excels at capturing high fidelity copies
- Easy to use for non-experts

https://webrecorder.net/

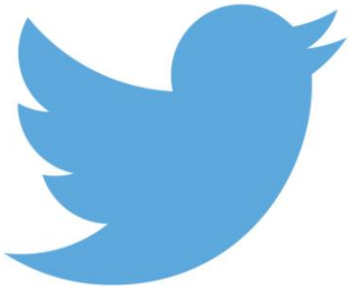


#TAGS

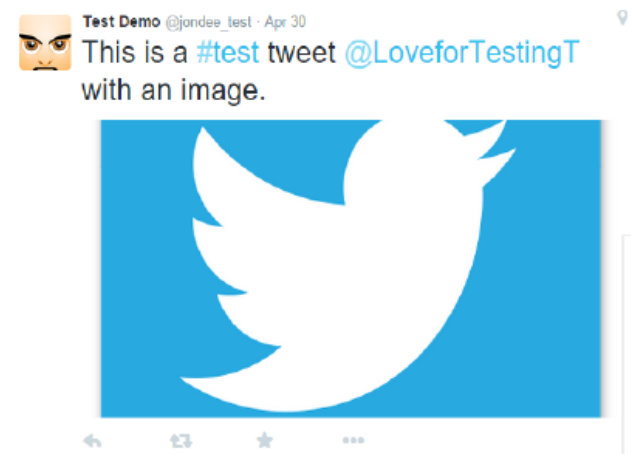
API Harvesting

- Application Programming Interface – backdoor to platform
- Harvest the underlying data
  - Machine-readable (JSON, XML)
  - Rich metadata
- Only available on certain platforms
- Licenced under Terms of Service
- Supports linking datasets from different sources
- More complex personal data issues

APIs



Twitter Data



Twitter Interface

JSON


```
{
    "created_at": "Thu Apr 30 21:53:11 +0000 2015",
    "id": 593895901623496700,
    "id_str": "593895901623496704",
    "text": "This is a #test tweet @LoveforTestingT with an image. http://t.co/ZvgHovKZq4",
    "source": "<a href=\"http://twitter.com\" rel=\"nofollow\">Twitter Web Client</a>",
    "truncated": false,
    "in_reply_to_status_id": null,
    "in_reply_to_status_id_str": null,
    "in_reply_to_user_id": null,
    "in_reply_to_user_id_str": null,
    "in_reply_to_screen_name": null,
    "user": {
        "id": 2993982541,
        "id_str": "2993982541",
        "name": "Test Demo",
        "screen_name": "jondee_test",
        "location": "Denver, CO",
        "url": null,
        "description": "this is a test account.",
        "protected": false,
        "verified": false,
        "followers_count": 2,
        "friends_count": 43,
        "listed_count": 0,
        "favourites_count": 0,
```

Tweet Elements


```
{ "user": {
    "id": 6253282,
    "id_str": "6253282",
    "name": "Twitter API",
    "screen_name": "TwitterAPI",
    "location": "San Francisco, CA",
    "url": "https://developer.twitter.com",
    "description": "The Real Twitter API. Tweets
about API changes, service issues and our Developer
Platform. Don't get an answer? It's on my website.",
    "verified": true,
    "followers_count": 6129794,
    "friends_count": 12,
    "listed_count": 12899,
    "favourites_count": 31,
    "statuses_count": 3658,
    "created_at": "Wed May 23 06:01:13 +0000 2007",
```

Twitter Terms & Conditions: *Re-sharing*

> *I.F.2.b.i. You may not distribute more than 1,500,000 Tweet IDs to any entity (inclusive of multiple individual users associated with a single entity) within any given 30 day period, unless you are doing so on behalf of an academic institution and for the sole purpose of non-commercial research or you have received the express written permission of Twitter.*

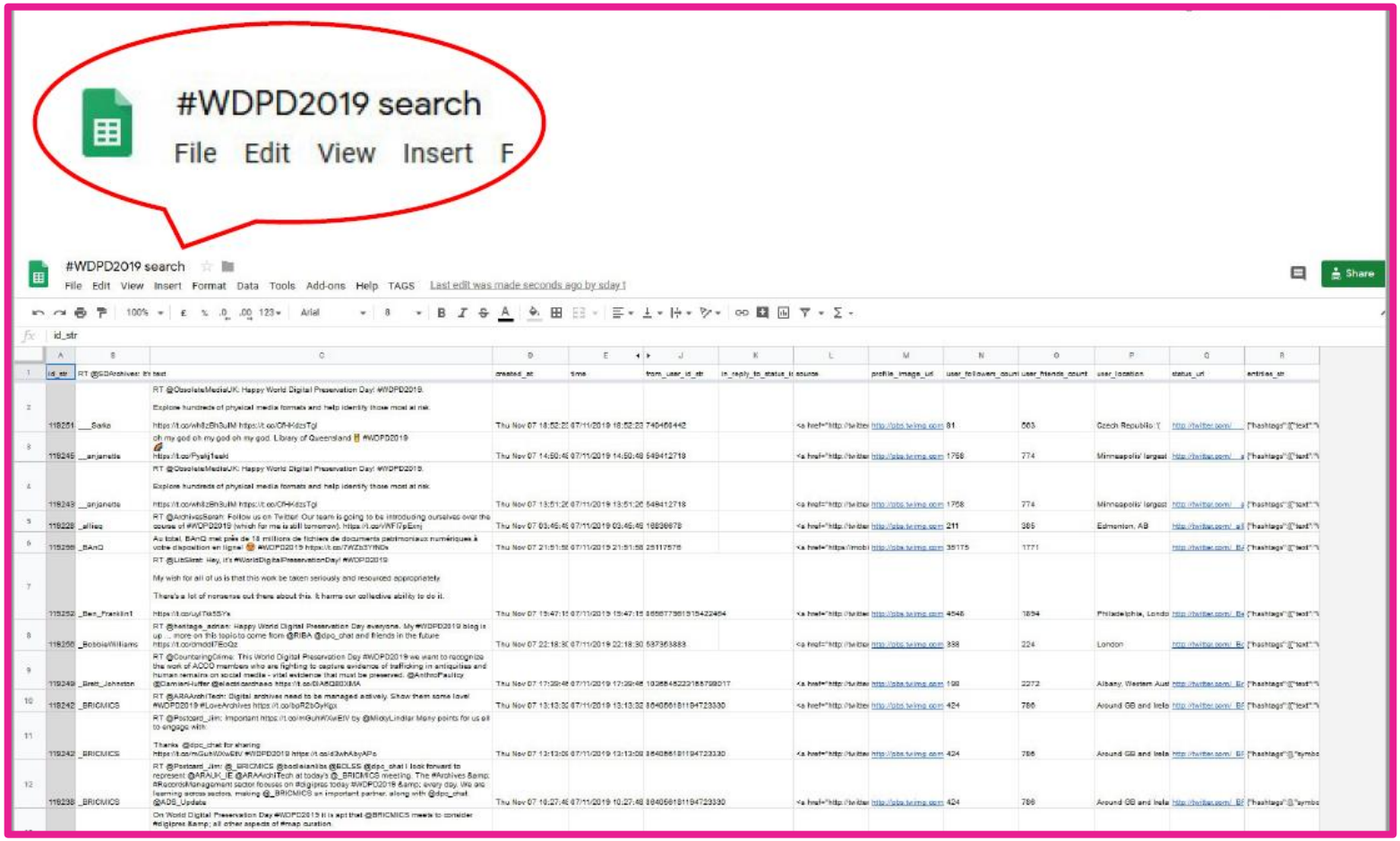
Twitter's Developer Policies for Researchers, Archivists, and Librarians

Justin Littman Follow
Jan 7, 2019 · 16 min read

https://medium.com/on-archivy/twitters-developer-policies-for-researchers-archivists-and-librarians-63e9ba0433b2

THE UNIVERSITY of EDINBURGH

https://tags.hawksey.info/



Questions?

Sara.Thomson@ed.ac.uk @sdaythomson