Approaches to Selection for Web Archives

DigitalPreservationCoalition

This training session was developed in partnership by the International Internet Preservation Consortium (IIPC) and the Digital Preservation Coalition (DPC)

IIPC
netpreserve.org

Common Approaches

- Broad or “Global”
- Domain Crawling
- Event or Theme Driven
- Selective or Representative
- Records Management
- Hybrid

Broad or “Global”

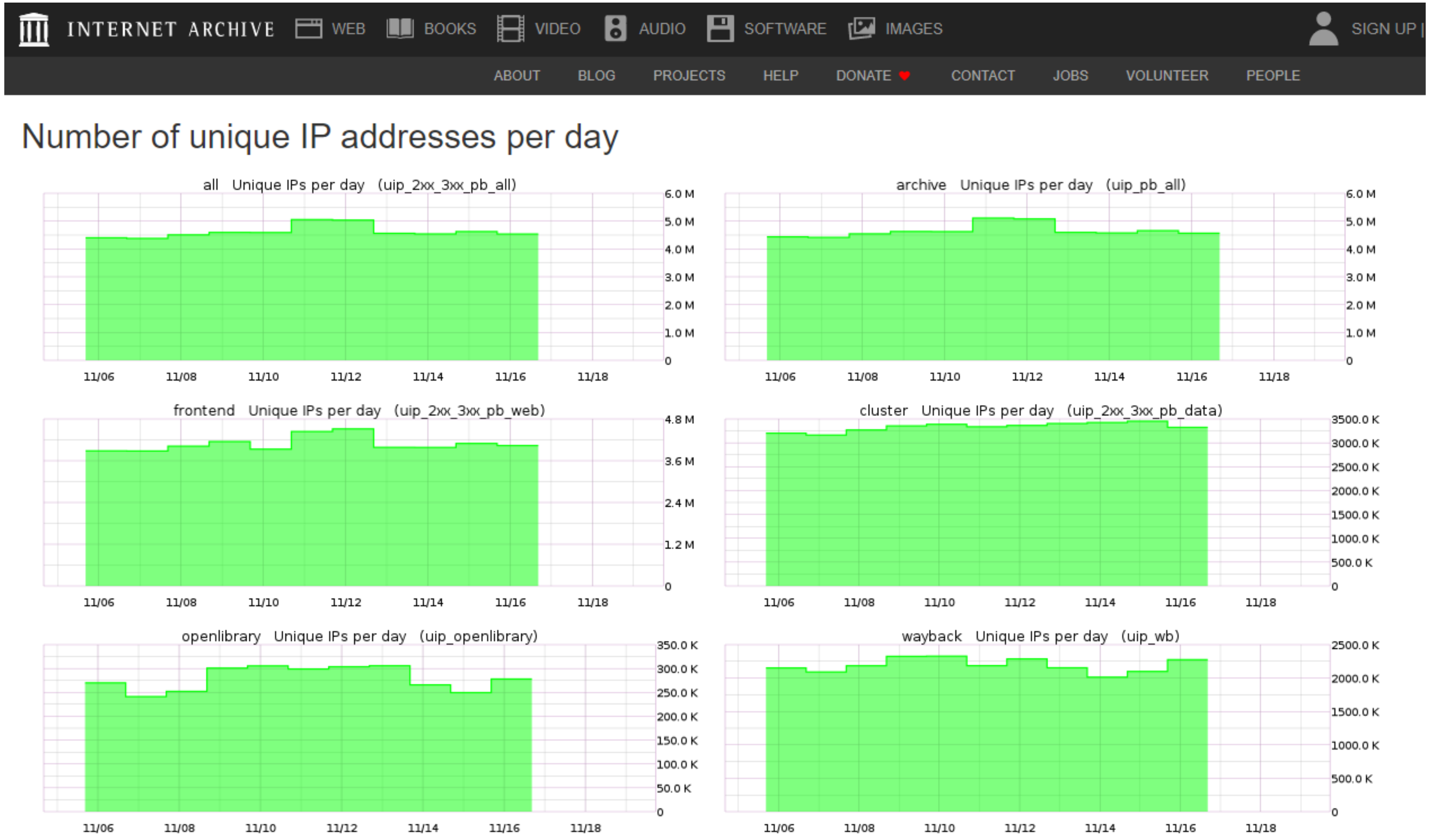
- Harvesting as much of the public web as possible with automated crawlers
- Aims for volume of content rather than perfect individual web pages
- Can provide context for other selective or curated collections
- Low fidelity in some instances due to lack of manual quality assurance and technical limitations of capture tools
- Might include a ‘Take Down’ policy for requests to remove content from the archive

Example: Internet Archive

IIPC netpreserve.org INTERNATIONAL INTERNET PRESERVATION CONSORTIUM

dpc DigitalPreservationCoalition

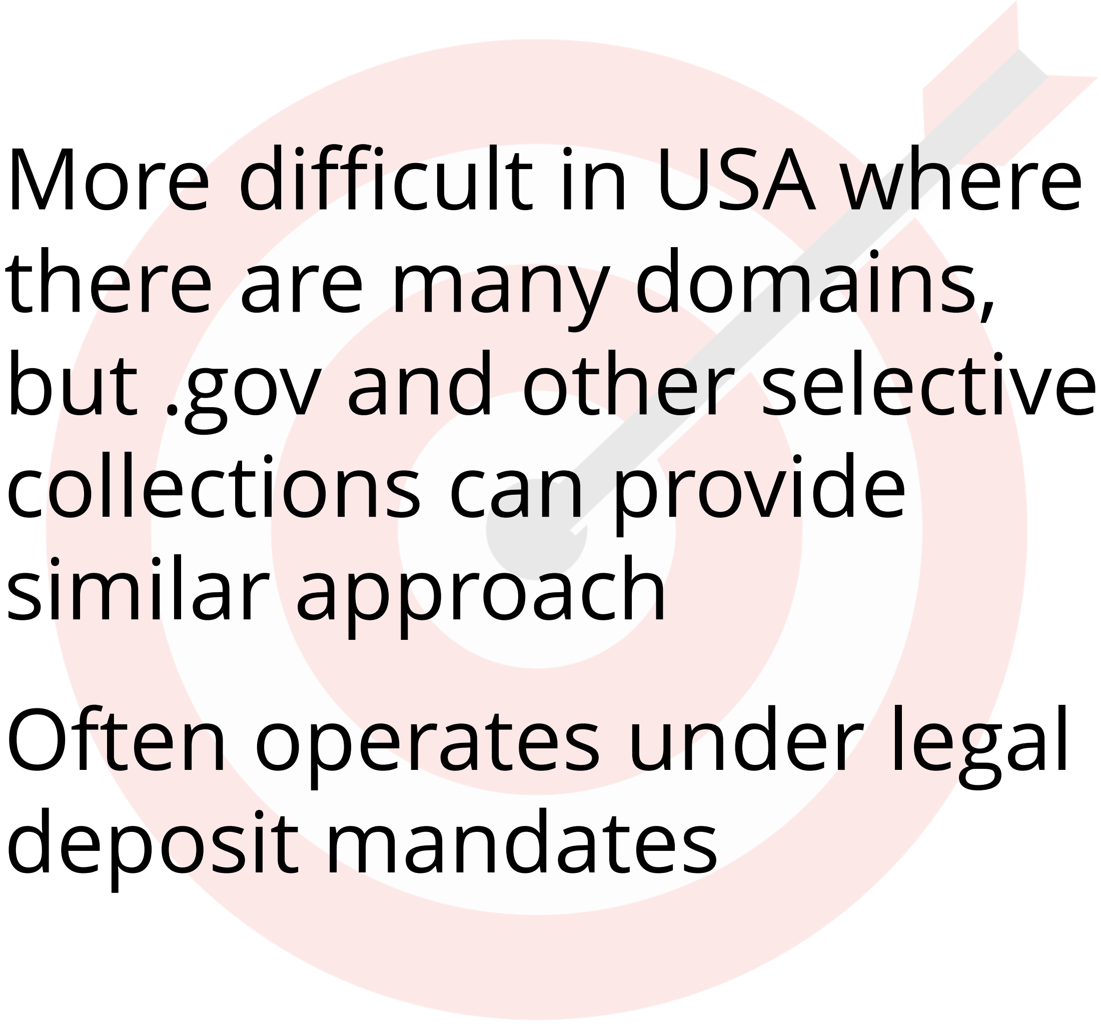
- Attempts to crawl entire public web
- Available through the Wayback Machine
- Currently preserves 475+ billion web pages*
- Collection occupies 45+ Petabytes of server space



Number of unique IP addresses per day

https://archive.org/stats/

Domain Crawling

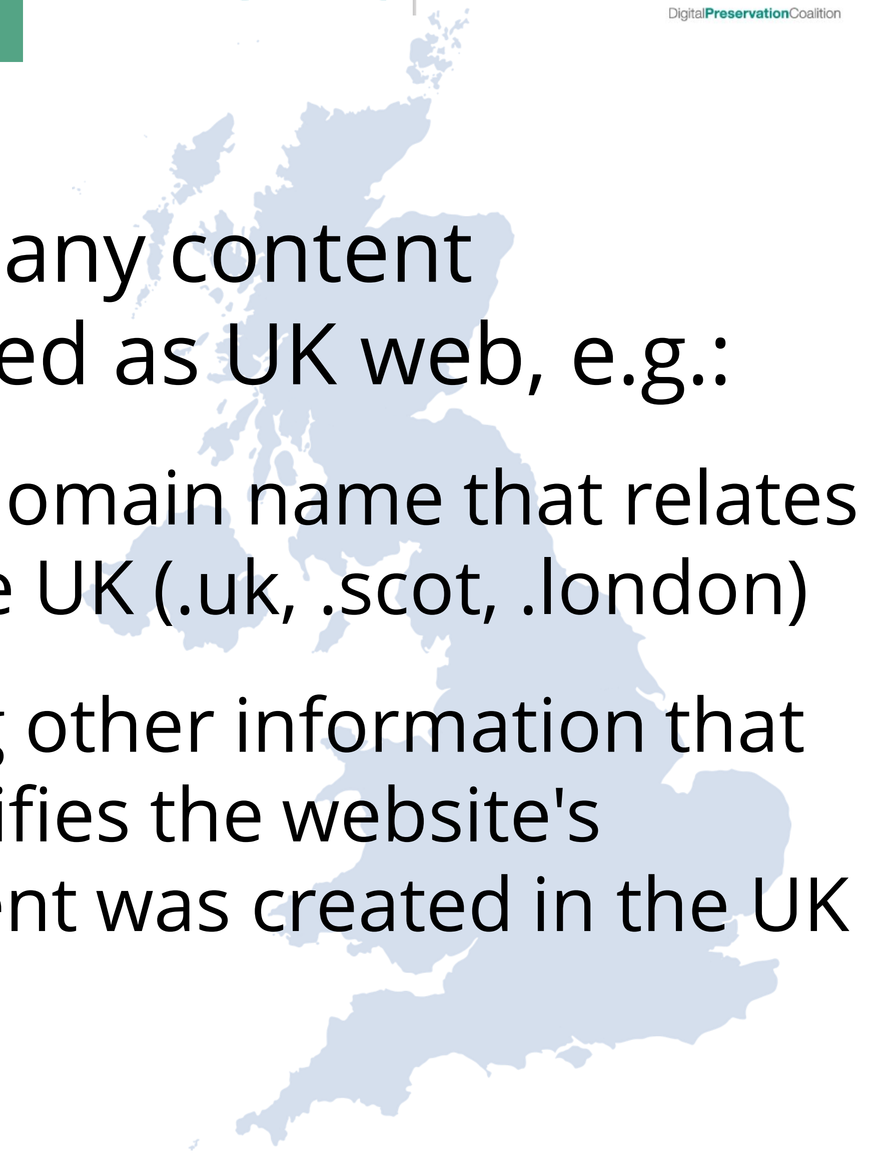


- Generally carried out by national libraries or other large collecting institutions
- Targeting URLs hosted in a particular country (e.g. .fr in France, .co.uk in the United Kingdom, .jp in Japan)
- More difficult in USA where there are many domains, but .gov and other selective collections can provide similar approach
- Often operates under legal deposit mandates

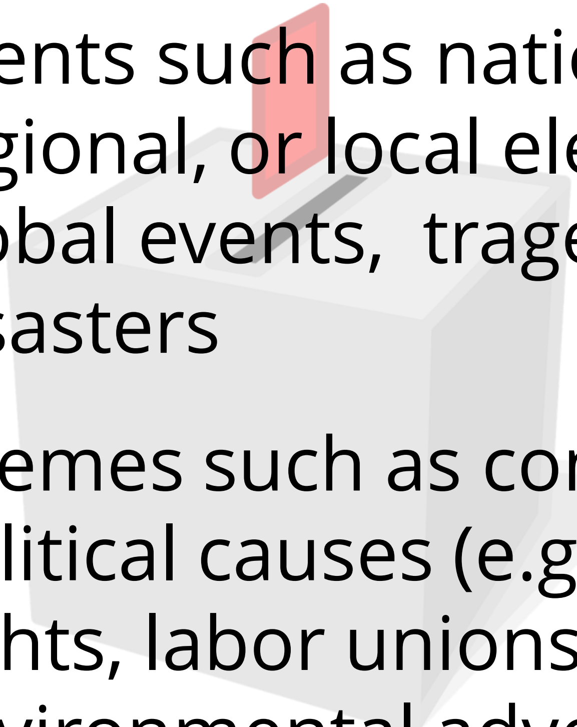
Example: UK Web Archive Domain Crawl

- Based on Non-print Legal Deposit 2013 mandate
- Automated crawl of entire UK web once per year
- As of 2017, comprised of 500 Terabytes of data, increasing by 60-70+ TB per year

- Collect any content identified as UK web, e.g.:
  - Any domain name that relates to the UK (.uk, .scot, .london)
  - Using other information that identifies the website's content was created in the UK

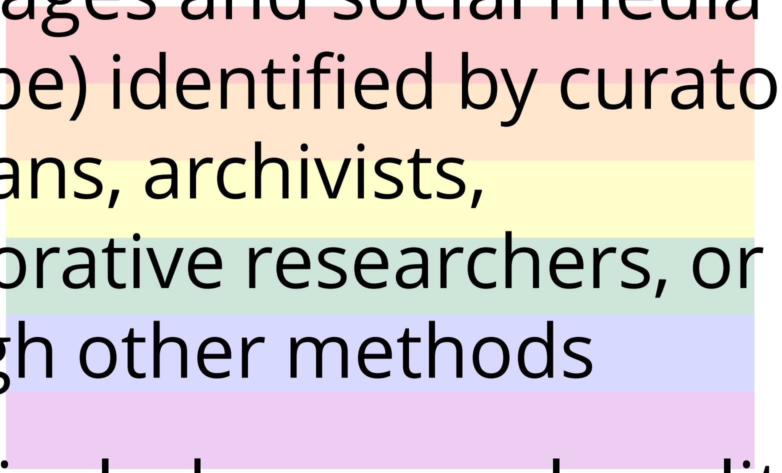


Event or Theme Driven



- Events such as national, regional, or local elections, global events, tragedies and disasters

- Themes such as community or political causes (e.g. LGBTQ+ rights, labor unions, environmental advocacy, or racial equality) or cultural endeavors (e.g. literature, art, food)



- Web pages and social media (if in scope) identified by curators, librarians, archivists, collaborative researchers, or through other methods
- Often includes manual quality assurance
- Often more detailed metadata and description

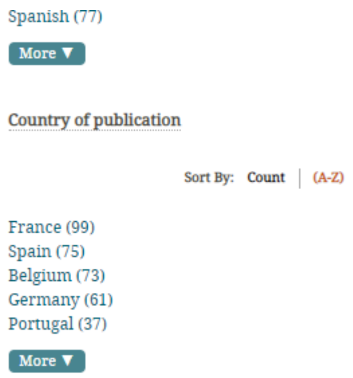
Example: IIPC Climate Change Collection

INTERNATIONAL INTERNET PRESERVATION CONSORTIUM

netpreserve.org

DigitalPreservationCoalition

- Collaborative collection from IIPC Content Development Group
- Aiming to gather multinational and multilingual content on Climate Change
- Looks to address topics that fall beyond a single organization's remit
- Also collections on: Olympic games, the COVID Pandemic, and the European Refugee Crisis



Spanish (77)

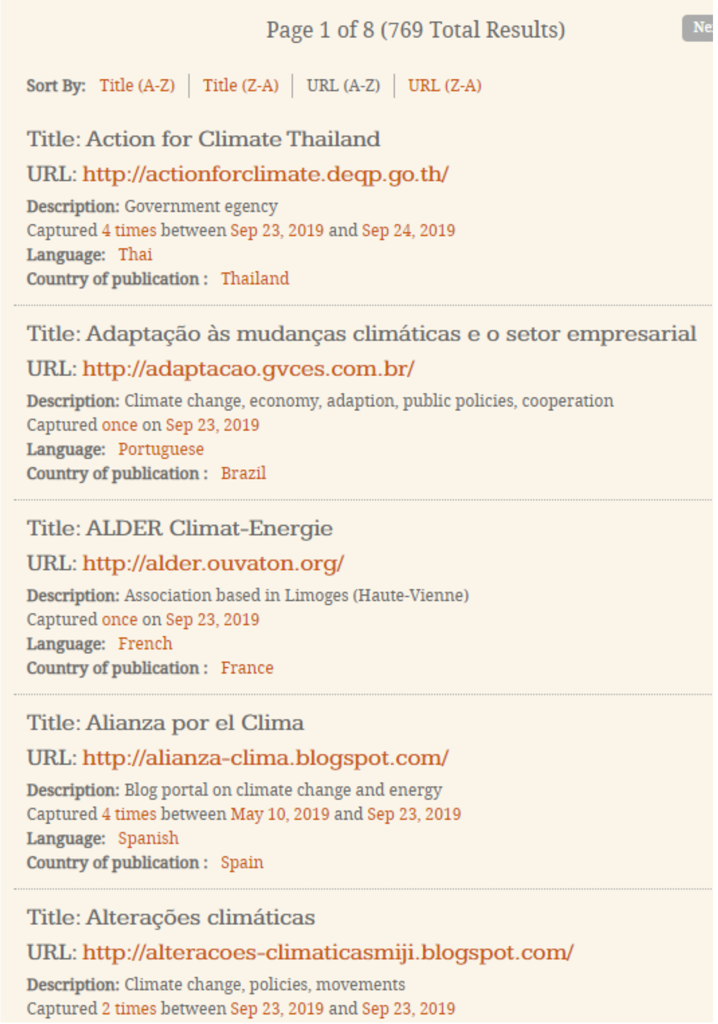
More ▼

Country of publication

Sort By: Count | (A-Z)

France (99)
Spain (75)
Belgium (73)
Germany (61)
Portugal (37)

More ▼



Page 1 of 8 (769 Total Results)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Title: Action for Climate Thailand
URL: http://actionforclimate.deqp.go.th/
Description: Government egency
Captured 4 times between Sep 23, 2019 and Sep 24, 2019
Language: Thai
Country of publication : Thailand

Title: Adaptação às mudanças climáticas e o setor empresarial
URL: http://adaptacao.gvces.com.br/
Description: Climate change, economy, adaption, public policies, cooperation
Captured once on Sep 23, 2019
Language: Portuguese
Country of publication : Brazil

Title: ALDER Climat-Energie
URL: http://alder.ouvaton.org/
Description: Association based in Limoges (Haute-Vienne)
Captured once on Sep 23, 2019
Language: French
Country of publication : France

Title: Alianza por el Clima
URL: http://alianza-clima.blogspot.com/
Description: Blog portal on climate change and energy
Captured 4 times between May 10, 2019 and Sep 23, 2019
Language: Spanish
Country of publication : Spain

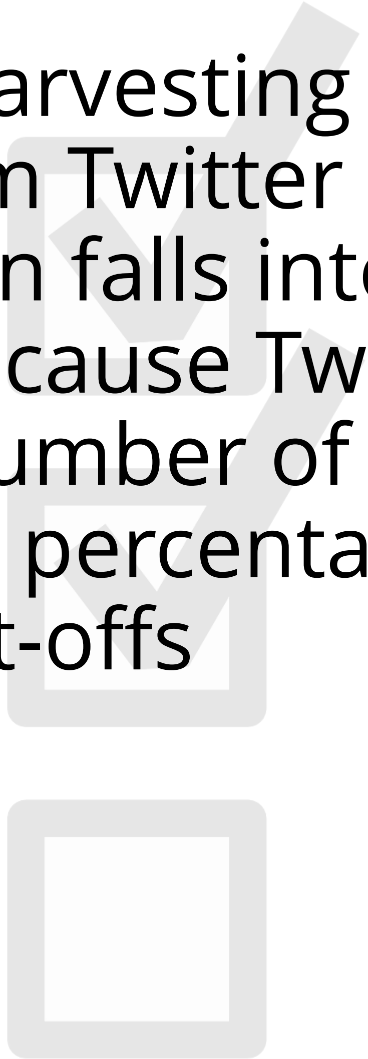
Title: Alterações climáticas
URL: http://alteracoes-climaticasmiji.blogspot.com/
Description: Climate change, policies, movements
Captured 2 times between Sep 23, 2019 and Sep 23, 2019

https://archive-it.org/collections/12116

Selective or Representative

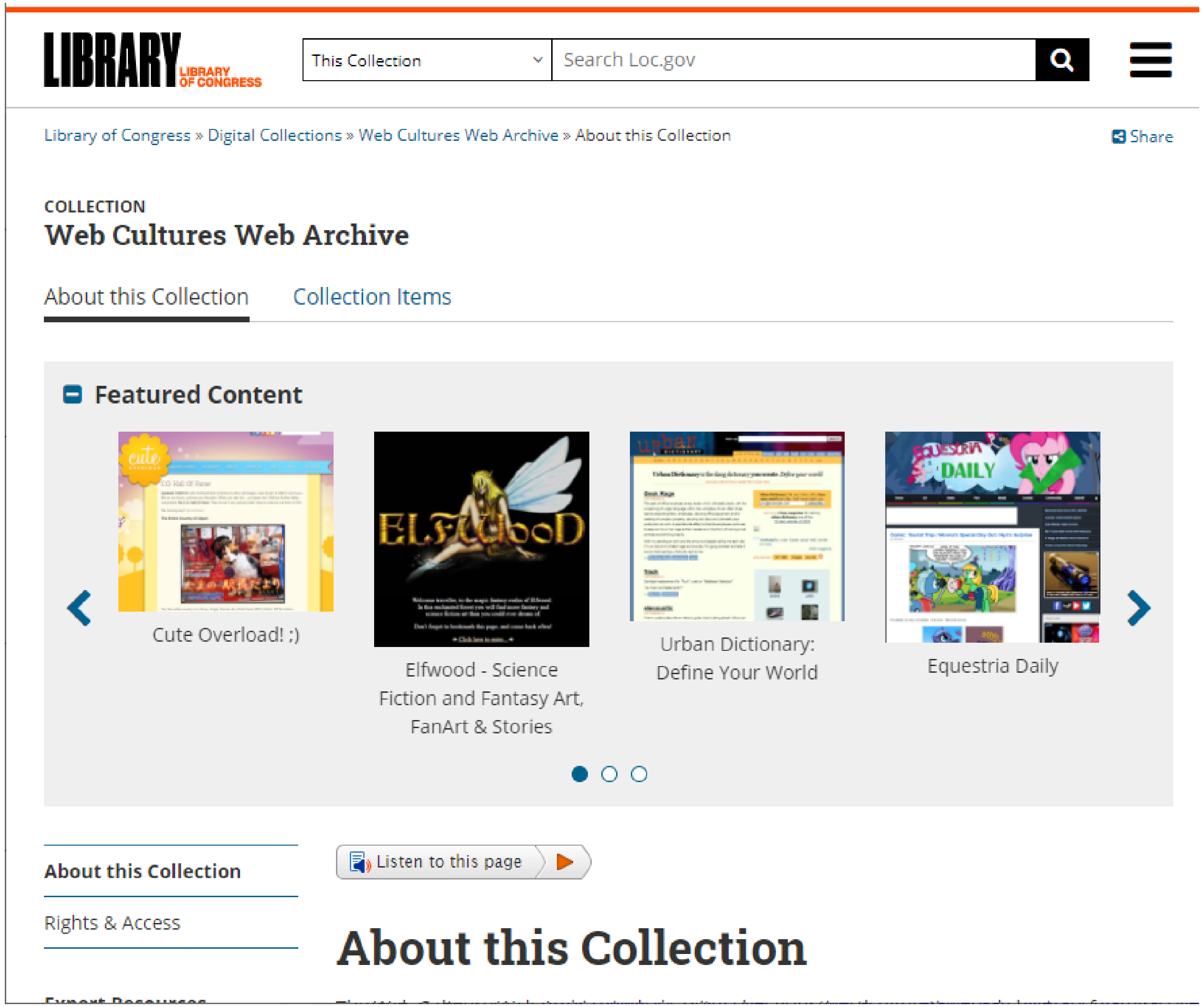
- Sampling or “Top 100” of web pages on a particular topic
  - e.g. News outlets or genres of publications like online periodicals or blogs
- Provides wide, but not exhaustive coverage of a particular topic identified by a curator, librarian, or archivist

- Example: Harvesting tweets directly from Twitter (using an API) often falls into this category because Twitter limits the number of tweets by a certain percentage or by other cut-offs



Example: "Web Cultures"

- Initiated 14 April 2014
- Collects sites documenting the creation and sharing of emergent cultural traditions on the web
- Content includes reaction GIFs, memes, icon-based communications (e.g. emojis), fan fiction, & more



Library of Congress » Digital Collections » Web Cultures Web Archive » About this Collection

COLLECTION

Web Cultures Web Archive

About this Collection | Collection Items

Featured Content

Cute Overload! :)

Elfwood - Science Fiction and Fantasy Art, FanArt & Stories

Urban Dictionary: Define Your World

Equestria Daily

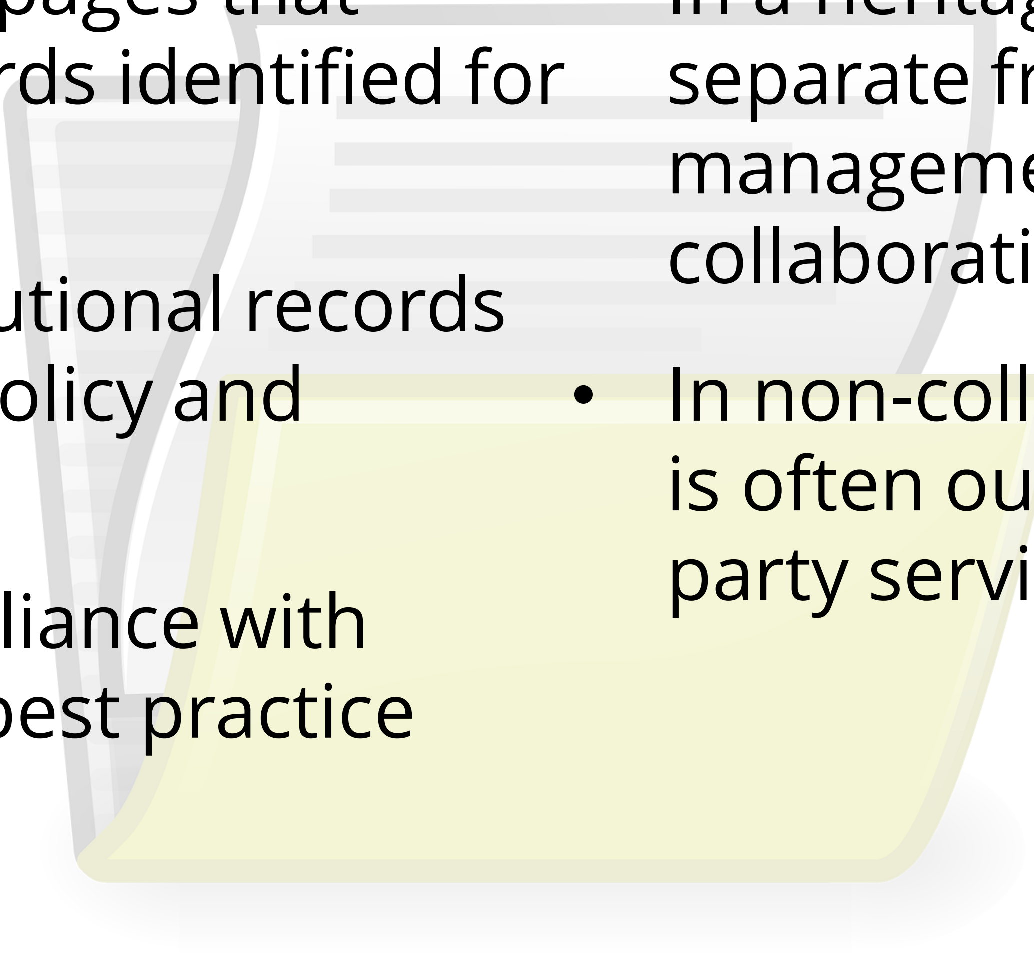
About this Collection

Rights & Access

About this Collection

https://www.loc.gov/collections/web-cultures-web-archive/about-this-collection/

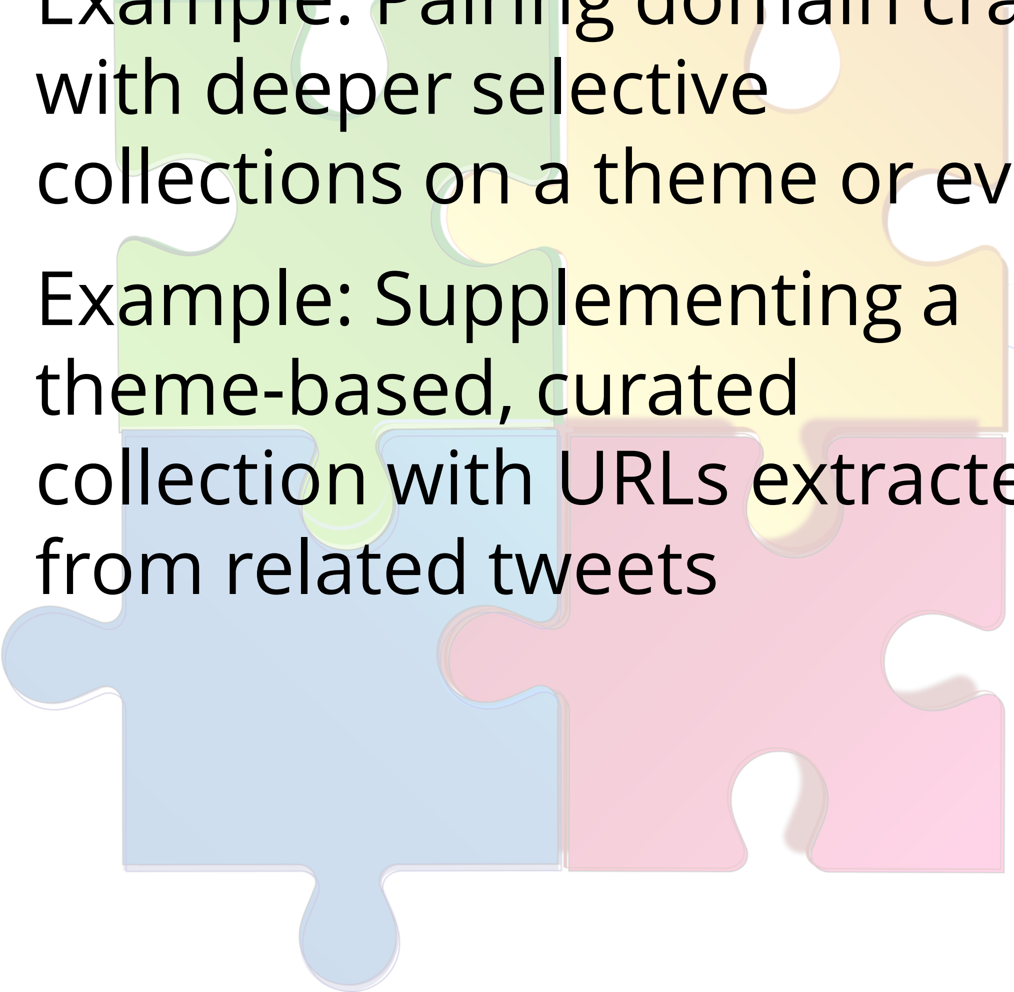
Records Management



- Capturing web pages that constitute records identified for retention
- Based on institutional records management policy and guidelines
- Supports compliance with regulations or best practice
- In a heritage institution, often separate from collections management, but could be collaborative
- In non-collecting institutions, it is often outsourced to a third party service

Hybrid

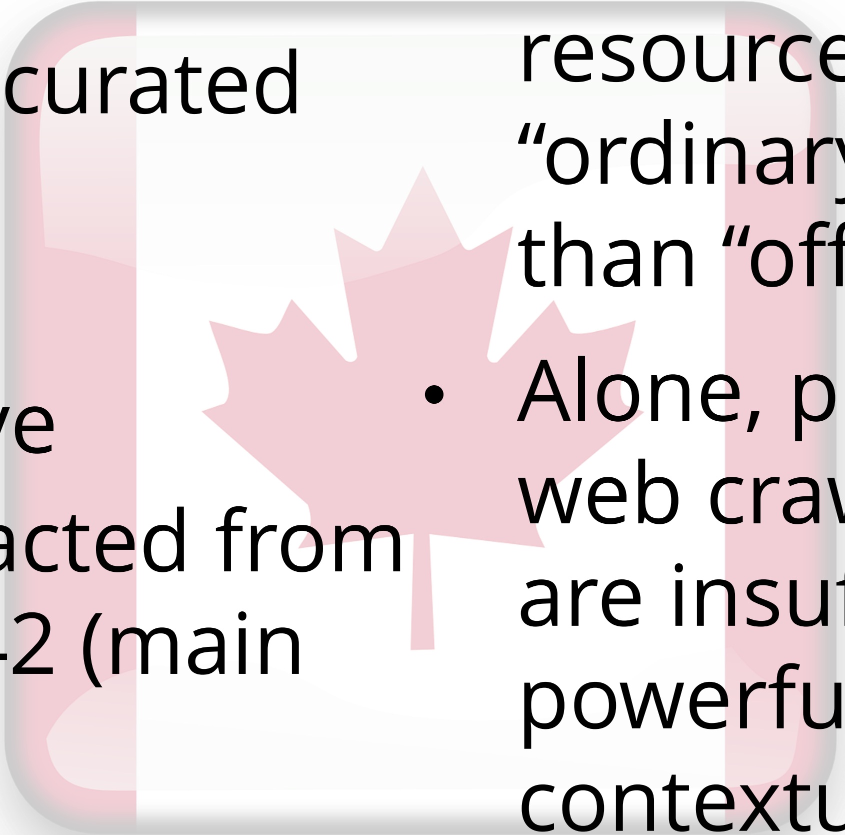
- No single approach can collect a replica of the whole web
- Each approach has strengths and weaknesses
- Combining approaches can strengthen overall collections
- Collaborating across institutions (and web archives) can improve access for researchers & other end users



- Example: Pairing domain crawls with deeper selective collections on a theme or event
- Example: Supplementing a theme-based, curated collection with URLs extracted from related tweets

Example: 2015 Canadian Federal Elections, University of Toronto

- Combines:
  1. Professionally curated collection
  2. Twitter links
  3. Internet Archive
- Twitter links extracted from harvest of #elxn42 (main election hashtag)



- Twitter links represent resources shared by “ordinary” people rather than “official” resources
- Alone, popularly-curated web crawls using Twitter are insufficient, but can be powerful alongside contextual archives

Tips for Selection

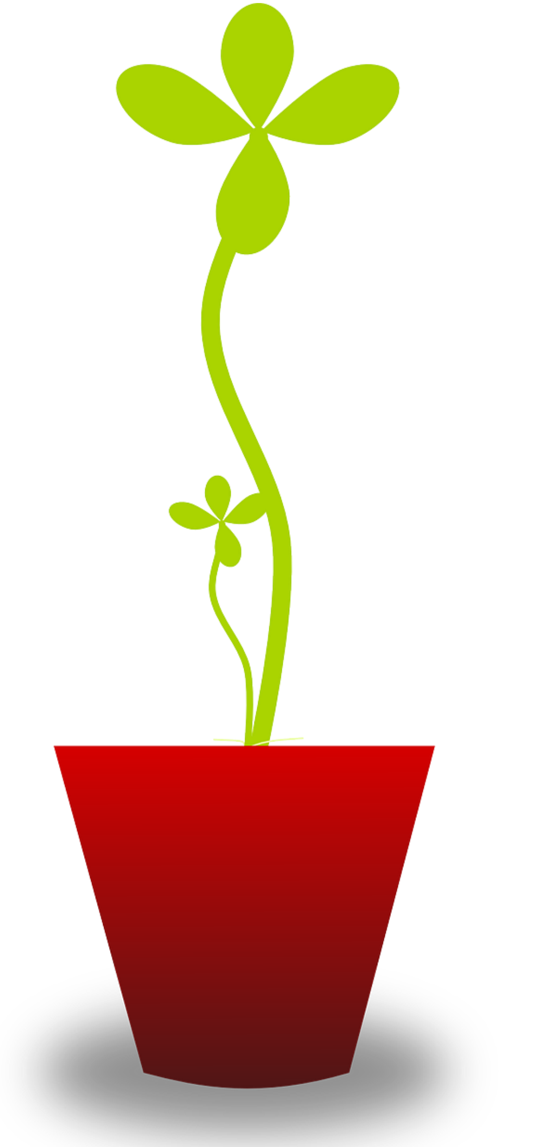
IIPC netpreserve.org INTERNATIONAL INTERNET PRESERVATION CONSORTIUM

DigitalPreservationCoalition

www.digitalbevaring.dk

Selection in Web Archiving

- Selection is how curators, librarians or archivists choose what web pages to capture & preserve
- Selection involves creating a list of URLs (called a ‘seed list’)
- A seed list represents websites or specific documents identified for archiving
- This list becomes the starting point for the crawler or capture tool



Selection Decisions Are Based On…

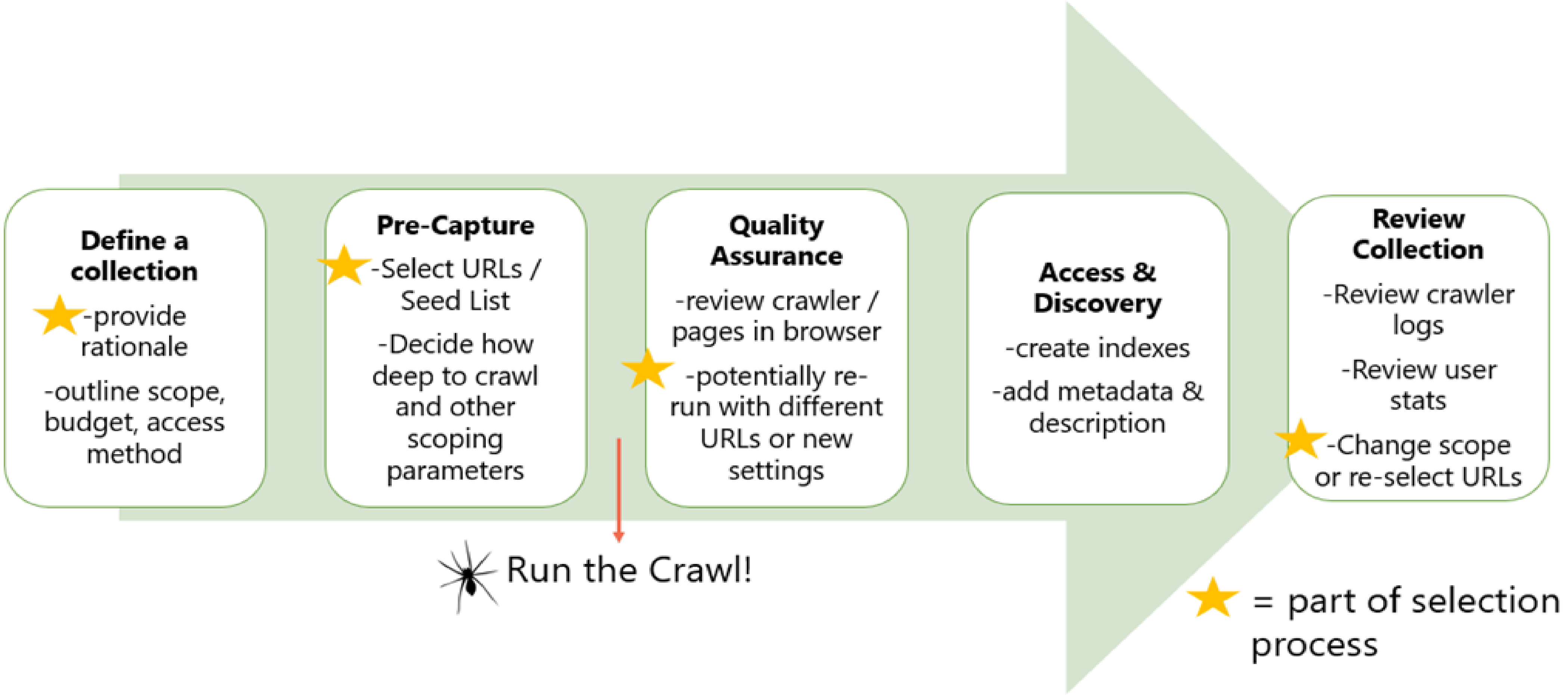
- An institution’s remit and objectives
- Collection Policy
- Expertise of curator
- Recommendations from collaborative researchers
- End user needs
  - Sometimes captured through website nomination forms submitted by users
- Financial budget, or data budget with subscription service, or local storage capacity
- Legal mandates or restrictions
- Review of other similar or potentially overlapping collections
- Functionality or limitations of technology used to capture
- Resources and staff

Selection vs Scoping

- Selection = WHAT to crawl & what NOT to crawl
- Scoping = HOW MUCH to crawl
- Scoping involves deciding:
  - How often to capture a web page
  - How much of an entire website to collect (e.g. number of "hops")
  - How long a single crawl session can last
  - What files formats to capture or exclude

- Selection and scoping are often iterative, occurring at several stages of the process as curators learn more about the size of their target web pages and ability of their tools to capture them

Example Workflow Based on an Archive-It User



Define a collection

★-provide rationale

-outline scope, budget, access method

Pre-Capture

★-Select URLs / Seed List

-Decide how deep to crawl and other scoping parameters

Run the Crawl!

Quality Assurance

-review crawler / pages in browser

★-potentially re-run with different URLs or new settings

Access & Discovery

-create indexes

-add metadata & description

Review Collection

-Review crawler logs

-Review user stats

★-Change scope or re-select URLs

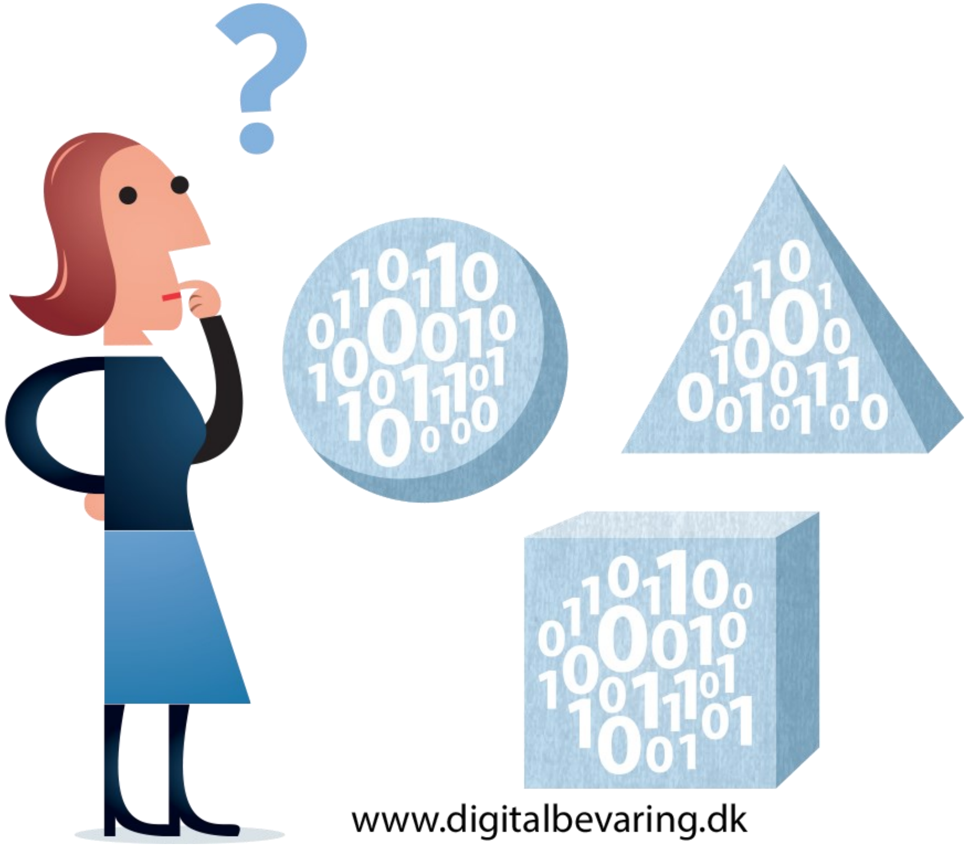
★ = part of selection process

Based on similar diagram from: ‘If these crawls could talk: Studying and documenting web archives provenance’ by Emily Maemura, Nicholas Worby, Ian Milligan, Christoph Becker (May 2018): https://doi.org/10.1002/asi.24048. Fig 1 General Workflow for creating a web archives collection at University of Toronto Library

Group Discussion

In your groups discuss your organization’s:

1. Key motivations for web archiving
2. Which selection methods might help to meet your aims


www.digitalbevaring.dk