Next Steps Towards A Formal Registry Of Web Archives For Persistent And Sustainable Identification

Why do we need a registry?

- **Web archives URLs will brake over time**
- **Preservation is about sustainable access resources also meaning persistent reference to resources**

What are the challenges?

Many reasons why archive URLs will stop resolving

- Web archives will change URL domain
  Example:
  National Library of Ireland 2019 (Aturban, 2019, Aturban et al., 2021).
  - from IMF platform ~~http://collection.europarchive.org/nli/~~
  - to Archive-IT service https://archive-it.org/home/nli

- Web archives will change access URLs
  Example:
  Royal Danish Library 2030
  - Old Wayback http://kb-prod-way-001.kb.dk:8081/query?type…
  - SOLRWayback https://solrwb.kb.dk:4000/solrwayback/services/web/…
  - Python Wayback https://pywb.kb.dk/…

- Web archives will merge
  Example:
  National Library of Australia
  - http://pandora.nla.gov.au/…
  - https://webarchive.nla.gov.au/…

What will it solve?

Finding sources for web materials with broken archive links

http://collection.europarchive.org/nli/

https://archive-it.org/home/nli

Update PWID Resolver

Web archive · Time of archiving · Precision · Archived URL

urn:pwid:archive.org:2024-03-12T03:15:44Z:page:https://netpreserve.org/

http://web.archive.org/web/20240312031544/https://netpreserve.org/

What do we have?

Archive services [edit]

Internet Archive Wayback Machine

- Article: Wayback Machine
- Domain: archive.org, waybackmachine.org
- Launched: 2001
- Date range: 1996-
- Hostname: <none>, web, wayback, liveweb, www, www.web, classic-web, web-beta, replay, replay.web, web.wayback
- Path: <none>, web
- Timestamp: Number 1 digit; 4–14 digits. Or "*". Or "?". Or combination. May also contain trailing chars like "re_" for (?), "if_" for frames and "im_" for images. If timestamp missing returns best available page.
- Examples:
  - http://www.web.archive.org/{14_digits}/http..
  - http://web.archive.org/web/{14_digits}/http..
  - http://wayback.archive.org/http..
  - http://web.waybackmachine.org/20081212010700/http..
- Oldest:
  - http://web.archive.org/web/0/http..
  - http://web.archive.org/web/1/http..
- Newest:
  - http://web.archive.org/web/2/http..
- Index:
  - http://web.archive.org/web/*/http..
- Submit page:
  - https://web.archive.org/save/
  - http://web.archive.org/record/
- Interval between manual captures: 1 hour.
- Prefix searching (except after the question mark in query strings)
- Case-insensitive lookup

Archive.Today

…

https://en.wikipedia.org/wiki/Wikipedia:List_of_web_archives_on_Wikipedia

Web archiving initiatives [edit]

Name	Country	Creation Year	Technologies	Number of Employees: Full-time	Number of Employees: Part-time	Comments
KBR web archive [15]	Belgium	2020		1		KBR[16] or the Belgian Royal Library is developing an operational web archive based on the findings of the PROMISE research project[7] (2017–2019). Operational policies and technical infrastructure will be developed based on the strategy outlined in the PROMISE project
OoCities …	Germany	2009				

https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives

None of these are in machine readable form

What are we missing?

- Formal part
  - What information to include and which identifier type?
  - Where to place formal registry? ISNI? IANA? Other?
- Additional information
  - What information to include?
  - On Wikipedia? Existing page? New page? Somewhere else?
  - How to make it computational readable?
- How to administrate?
  - Who can change/add/remove
  - How to authenticate
  - Who should be administrator
  - …

Minimum information in a formal registry

Formal registries should ***not*** contain information which changes often

Web Archive Registry

Formal Id	Name	Alias Id	Valid from	Valid to
1000 0000 0168 4170	Irish Web Archive	nli.ie	01-01-2011	
1000 0000 0168 4171	Portuguise Web Archive	arquivo.pt	08-11-2007	
1000 0000 0168 4172	York University Web Archive	yorku.ca	01-01-2010	
1000 0000 0168 4175	Archive.Today	archive.ph	01-01-2012	
1000 0000 0168 4177	NLA Australia	pandora.nla.gov.au	01-01-2004	01-01-2012
1000 0000 0168 4177	Australian Web Archive	webarchive.nla.gov.au	01-01-2009	
1000 0000 0168 4178	Internet Archive	archive.org	01-01-2001	
1000 0000 0168 4179	Web Citation (WebCite)	webcitation.org	01-01-2005	
1000 0000 0168 4182	Netarkivet	netarkivet.dk	01-07-2005	01-07-2035
1000 0000 0168 4182	Danish Web Archive	kb.dk	01-07-2035	

Example ISNI · New domain · Merge · Change of service is not visible

More information needed

Example: Change of service URL

Web Archive Information Formal id. 1000 0000 0168 4170

Name: Irish Web Archive · Case sensitive lookup · Alias/domain Id: nli.ie
Access type: Open
Homepage: www.nli.ie

Services

Type	Prim.	Specification	Valid from	Valid to	URL-Pattern
Page access	No	Timestamp/URL	01-01-2011	01-01-2019	http://collection.europarchive.org/nli/{4-14-digit YYYYMMDDhhmmss}/{url}
Page access	No	Timestamp/URL	01-01-2019		https://archive-it.org/home/nli{4-14-digit YYYYMMDDhhmmss}/{url}

Example: Several domain id's

Web Archive Information Formal id. 1000 0000 0168 4175

Name: Archive.Today · Case sensitive lookup
Access type: Open
Homepage: https://en.wikipedia.org/wiki/Archive.today

Alias/domain Id: archive.ph
2. Alias/domain Id: archive.md
3. Alias/domain Id: archive.fo
4. Alias/domain Id: archive.vn
5. Alias/domain Id: archive.li
6. Alias/domain Id: archive.is
7. Alias/domain Id: archive.today

Services

Type	Prim.	Specification	Valid from	Valid to	URL-Pattern
Page access	Yes	Timestamp/URL	01-01-2010		http://archive.today{4-14-digit YYYYMMDDhhmmss}/{url}
Page access	No	Timestamp/URL	01-01-2010		https://archive.today{4-14-digit YYYY.MM.DD-hh:mm:ss}/{url}
Page-index	No	Only URL	01-01-2013		http://archive.today{url}
Save-page	No	Only URL in path	01-01-2013		http://archive.today?run=1&url={url}

Example: Trailing characters

Web Archive Inf… · **Services** for archive: Portuguise Web Archive

Name: Portuguise Web Arc
Access type: Open
Homepage: https://arquivo.pt/

Services

Type	Prim.	Specification
Page access	No	Timestamp/UR…

Service type: Page access: Access to web pages content of web archive
Spec. pattern: Timestamp/URL: {4-14-digit YYYYMMDDhhmmss}/{url}
Primary service
https · http · www
Service path: arquivo.pt/wayback
Valid from: 11-11-2024 to

Trailing characters to the date

Chars	Description
if_	
im_	Snapshot

Examples with values and trailing characters:
https://arquivo.pt/wayback/20120326012340/https://netpreserve.org/
https://arquivo.pt/wayback/20120326012340if_/https://netpreserve.org/
https://arquivo.pt/wayback/20120326012340im_/https://netpreserve.org/

Other information

Web archive address · Contact information · …

Type of archive (e.g. citation service, government, …)

Next steps

IIPC Meeting /webinar
later this year

Web archive registry
a bridge between web archives and current/future researchers
"In this meeting we will discuss, and possible plan, how web archives can aid researchers make web archive references and resolve old web archive references by having a description/registration of web archives and their history. …"

Presenters:
Eld Zierau, Senior Researcher, Digital preservation
Caroline Nyvang, Senior Researcher, History

Any contributions
Please contact:
Eld Zierau, elzi@kb.dk