NCC Digital Scholarship Workshop

Thursday, November 15, 2018

Introduction to web archiving using Archive-It

https://archive-it.org/

Instructor: Koko Fujita Howell
Japan Disasters Digital Archive http://jdarchive.org/en
Digital Content Lead/Web Archive Curator

What Archive-it can do

It is a powerful tool to save live websites, including YouTube, which eventually disappear from the Internet at some point.

Websites usually last 1000 days to 5 years.

For example:

http://wayback.archive-it.org/2438/20130412201108/http://crms-setagaya.jimdo.com/

https://crms-setagaya.jimdo.com/

Archive-It Video Curriculum – Archive-It Help Center

Please watch these videos to understand how Archive-It works
https://support.archive-it.org/hc/en-us/articles/216489103-Archive-It-Video-Curriculum-

- Getting Started
  - Navigating Archive-It
  - Administrative Functions
  - Pre-crawl Scoping
  - Test Crawls

- Post Crawl Analysis
  - Getting the most from your post crawl reports
  - Understanding your Hosts Report
  - Quality Assurance
- Advanced Training Webinars
  - Advanced Scoping
  - Archiving Video Content
  - Archiving Social Media
  - Advanced Quality Assurance
  - Access to Archive-It Collections

Data Budget

Unit	Abbreviation	Storage
Byte/Octet	B	8 bits
Kilobyte	KB	1024 bytes
Megabyte	MB	1024KB
Gigabyte	GB	1024MB
Terabyte	TB	1024BG

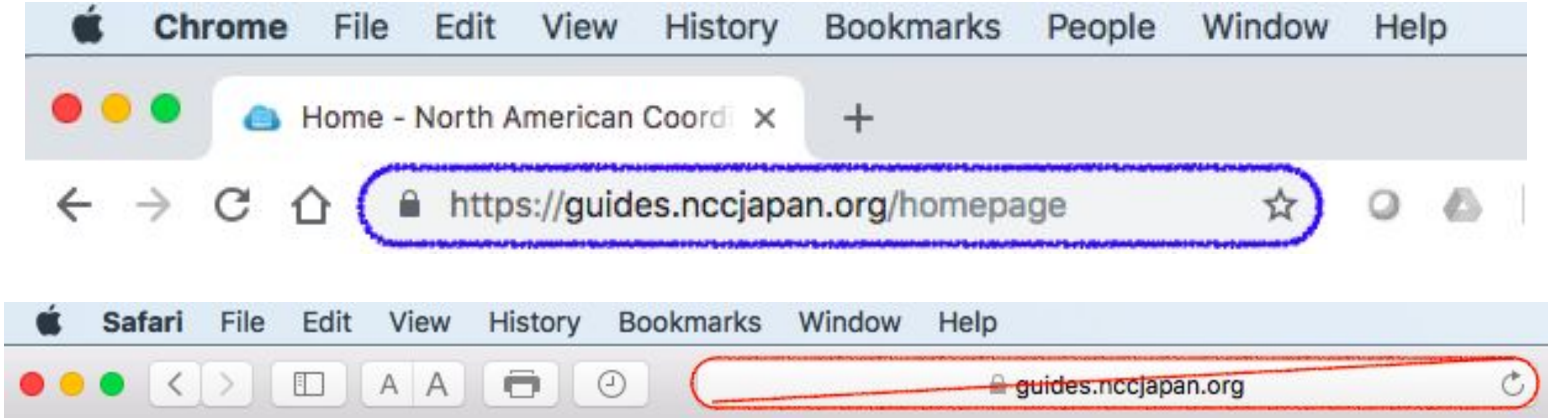
http://m3rabc.blogspot.com/ is about 1.8GB.
https://ameblo.jp/takahashi-photo/ is about 20~30GB

Monitor data budget usage in your account in order to prevent overspending.
https://partner.archive-it.org/1484/collections/11371/crawl/709014/hosts

Reference:
https://support.archive-it.org/hc/en-us/articles/208000096-Monitor-your-data-budget-
https://physics.nist.gov/cuu/Units/binary.html
https://www.kingston.com/us/flash/storage_chart

Tools for Archive-it

Web Browsers: Chrome and its bookmark bar
Safari is not suited because the address bar (location bar or URL bar) does not show full a URL.



Reference:
https://doepud.co.uk/blog/anatomy-of-a-url
https://support.archive-it.org/hc/en-us/community/posts/360012216926-2018-PM-in-DC-Breakout-Group-3-Collaborative-web-archiving-

Tools for Archive-it (cont'd)

Google Sheets to store your seeds (URLs)

It is suited for

- Easy export to your Archive-it collection seed page
- Easy Quality Assurance (QA) through relatively simple programming
- Easy search function for your task
- Easy filtering system

A seed

A seed is any URL that tells the Internet Archive web crawler to capture information on the web.

Examples of seeds

- **an entire website (domain): http://www.soumu.go.jp/**
- **a specific part (directory) of a website: http://www.soumu.go.jp/h30_hokkaido_iburitobu/**
- **specific documents:**
- **Html:http://www.soumu.go.jp/menu_kyotsuu/important/kinkyu01_000152.html**
- **PDF: http://www.soumu.go.jp/main_content/000573592.pdf**
- **Image: http://www.soumu.go.jp/main_content/000000014.gif**
- **CSS: http://www.soumu.go.jp/main_content/aly.css**
- **etc**

References

https://support.archive-it.org/hc/en-us/articles/208331753-How-to-select-seeds-#What-exactly-is-a-seed
https://en.wikipedia.org/wiki/Heritrix
https://en.wikipedia.org/wiki/Web_archiving

Scope

- Scope is what the crawler will capture and what it won't.
- Scoping refers to options for telling the crawler how much or how little of a seed to capture.
- Archive-It includes default scopes and seed- and collection-level scoping options

Reference:
https://support.archive-it.org/hc/en-us/articles/216489103-Archive-It-Video-Curriculum-
https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms
https://support.archive-it.org/hc/en-us/articles/208332843-Assign-and-edit-a-seed-type-

Scope (cont'd)

You have to determine which scope needed for needs.

- The crawler can capture your seed URL and links within the seed site when you set the scope for your crawl.

Archive-it default scoping rules

- **Standard**: the crawler can capture URLs which have the same domain of the website
- **Standard +**: the crawler can capture URLs which have the same domain of the website and one page of external links
- **One page:** the crawler can capture in one page of a website
- **One page +**: the crawler can capture in one page of a website and the external links.

Reference:
https://support.archive-it.org/hc/en-us/articles/208332843-Assign-and-edit-a-seed-type-
https://support.archive-it.org/hc/en-us/articles/208001076-How-our-crawler-determines-scope

Pre-crawl Scoping

Determining scope of crawl for your selected seedURL(s)

an entire website: http://sada.la.coocan.jp/

Link: https://takeshiro.net/ is **not in scope** (domain is different from the example URL above)

Link: http://www.mctv.ne.jp/~takase/index.htm is **not in scope** (domain is from the example URL above)

Link: http://sada.la.coocan.jp/ta/ansei/tkhp02.htm is **in scope** (domain is the same as the example URL above)

Embedded image: http://sada.la.coocan.jp/taketiti.JPG is **in scope** (domain is the same as the example URL above)

References:

http://wayback.archive-it.org/11371-test/20181109103249/http://sada.la.coocan.jp/

https://support.archive-it.org/hc/en-us/articles/208001076-How-our-crawler-determines-scope

https://support.archive-it.org/hc/en-us/sections/201864583-Scoping-Crawls

https://support.archive-it.org/hc/en-us/articles/360015086931-Modify-your-collection-or-seed-scope

Pre-crawl Scoping (cont'd)

Determining scope of crawl for your selected seedURL(s)

Not in scope

http://www.mctv.ne.jp/~takase/index.htm

Adding **Scope Rules** to the collection and seed scope rule pages

Some seeds need to scope to capture images and block unwanted advertisements and videos.

Do Test Crawls!(do not cost data budget unless you save the data)

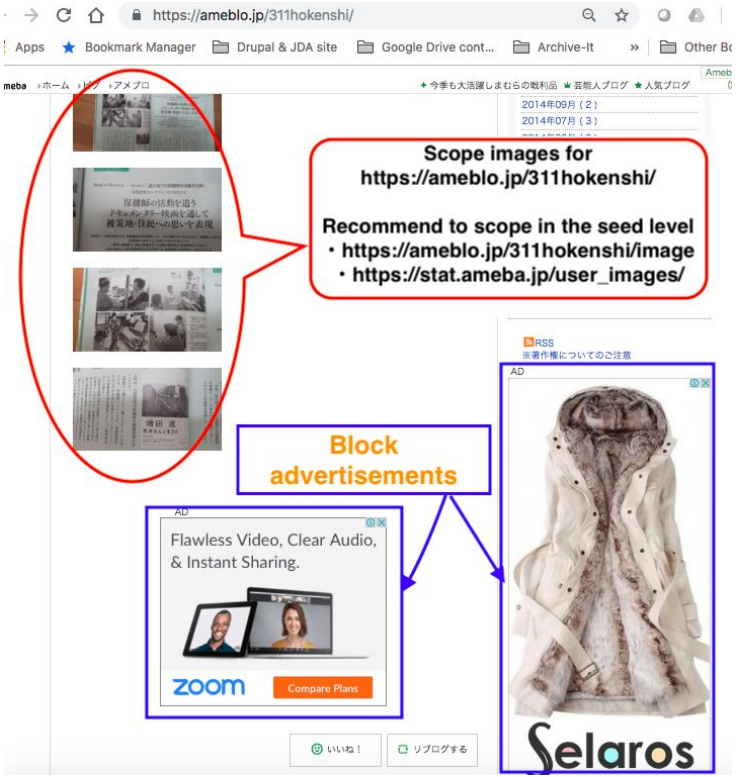
https://ameblo.jp/311hokenshi/

Needs special scope rule in order to capture images
https://ameblo.jp/311hokenshi/image
https://stat.ameba.jp/user_images/

Block images
S0.2mdn.net
Tpc.googlesyndication.com

Result of capture
http://wayback.archive-it.org/7472/20180524145206/https://ameblo.jp/311hokenshi/

Reference:
https://support.archive-it.org/hc/en-us/articles/208001106-Expand-the-scope-of-your-crawl



Selecting seeds

https://partner.archive-it.org/1484/collections/7472/crawl/647275/seeds

発行物・提言書 | NGOを支援するNGO　国際協力NGOセンター（JANIC）
http://wayback.archive-it.org/7472/20180716055132/http://janic.org/activ/earthquake/issue/index.php

Seed: http://janic.org/activ/earthquake/issue/index.php
Scope → One page +
Links within the web page
http://www.janic.org/MT/img/activity/JANIC_fukushima_kiroku.pdf
http://www.janic.org/MT/img/activity/ngo_recommendation.pdf
http://www.janic.org/MT/img/activity/positionpaper_nuclear.pdf
http://www.janic.org/MT/img/activity/statement_oda_japanearthquake.pdf
https://www.janic.org/MT/img/activity/JANIC_fukushima_kiroku.pdf
http://kodomira.com/news/HSK-release-all.html
http://www.shinhyoron.co.jp/978-4-7948-0913-1.html

Links are not in scope: (Links below have **different directory** from http://janic.org/activ/earthquake/issue/)
You can also submit link URLs as separate seeds as One page/once (option private setting).

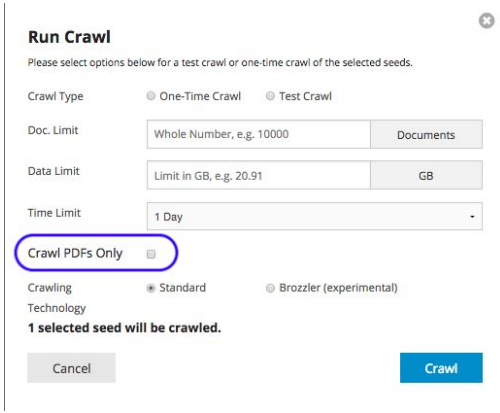
Seed scope rule (PDFs)

When you would like to capture pdfs within a web page you can also add scope rule and do patch crawl after the initial crawl

2011年度:日調連会報:出版物のご紹介:日本土地家屋調査士会連合会の活動【日本土地家屋調査士会連合会】
http://www.chosashi.or.jp/activity/publications/kaiho/kaiho2011.html

Seed level scope
http://www.chosashi.or.jp/activity/publications/kaiho/img/kaihou

Or do patch crawl for this page
http://wayback.archive-it.org/7472/20181022125137/http://www.chosashi.or.jp/activity/publications/kaiho/kaiho2011.html



Seed scope rule

The websites listed below have specific scoping rules.

Facebook, Twitter, Instagram, Wix and Wixsite base website, Blogspot, & YouTube videos

Please read the page

Reference:
https://support.archive-it.org/hc/en-us/sections/201841373-Scoping-crawls-for-specific-types-of-sites

How to look at Archive-it account and collection pages

Examples:
https://partner.archive-it.org/1484
https://partner.archive-it.org/1484/collections/11371

NCCworkshop111518 has the account number 1484 following the domain
https://partner.archive-it.org/

You have to login to get your account URL
When you subscribe, you get your account page with specific number
https://partner.archive-it.org/xxxx

When you make collection you also get another specific number following /xxxx/collections/
https://partner.archive-it.org/xxxx/collections/++++

Archive-it collection pages

Frequently visited pages

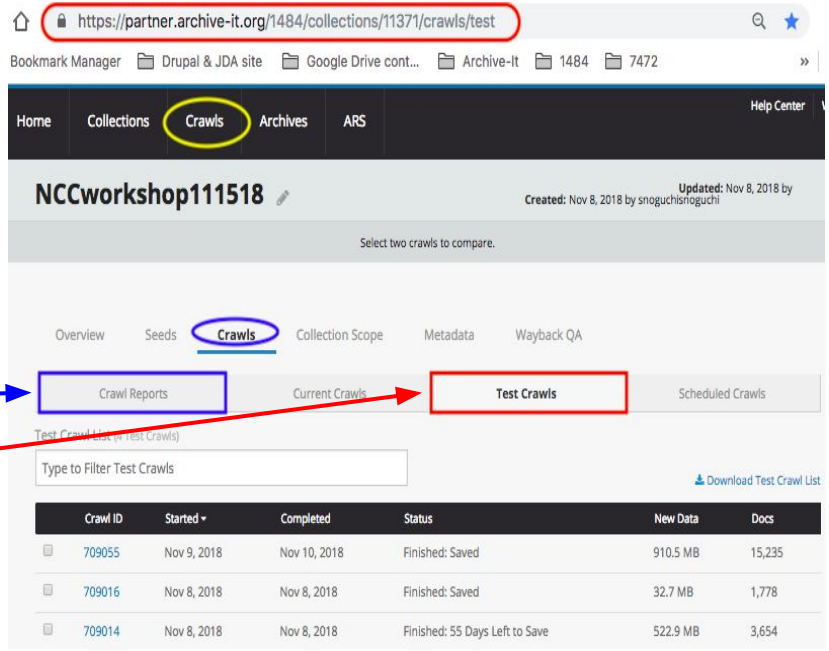
Seeds: add seed URL
https://partner.archive-it.org/1484/collections/11371/seeds

Crawls: see non-test crawl and patch crawl results through Crawl IDs
https://partner.archive-it.org/1484/collections/11371/crawls

Collection Scope: add collection scope rules throughout the collection.
https://partner.archive-it.org/1484/collections/11371/scope

Wayback QA: find missing URLs for crawl patch. Usually URLs listed in the page are found Under Queued or Out of Scope column in a host page within the Crawl ID page
https://partner.archive-it.org/1484/collections/11371/qa

Test crawl results



- Test crawl results are available only through "**Crawl Reports**". https://partner.archive-it.org/1484/collections/11371/crawls
- Go to crawl page and click "Test crawls" among horizontal navigation menus, https://partner.archive-it.org/1484/collections/11371/crawls/test
- Click test crawl id number, then the crawl overview page opens.
- Then, click "seed" to see your seed lists.
- Then, click under Wayback links.

You have to click many times to get test crawl results, so I recommend that you make a crawl report spreadsheet on your own.

Crawl report table using Google Sheets (option)

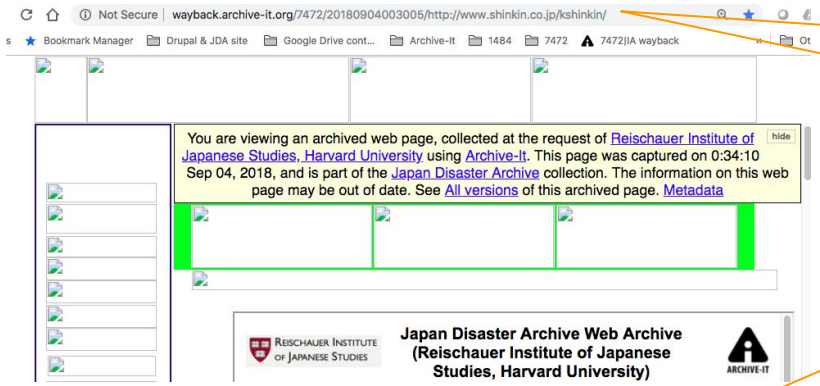
Nov 2, 2018	696897	Test	https://partner.archive-it.org/1131/collections/7472/crawl/696897/	https://partner.archive-it.org/1131/collections/7472/crawl/696897/seeds	https://partner.archive-it.org/1131/collections/7472/crawl/696897/types	no data yet
Nov 1, 2018	696799	One-Time	https://partner.archive-it.org/1131/collections/7472/crawl/696799/	https://partner.archive-it.org/1131/collections/7472/crawl/696799/seeds	https://partner.archive-it.org/1131/collections/7472/crawl/696799/types	verified

Note: you have to program the google sheet to open the desired page

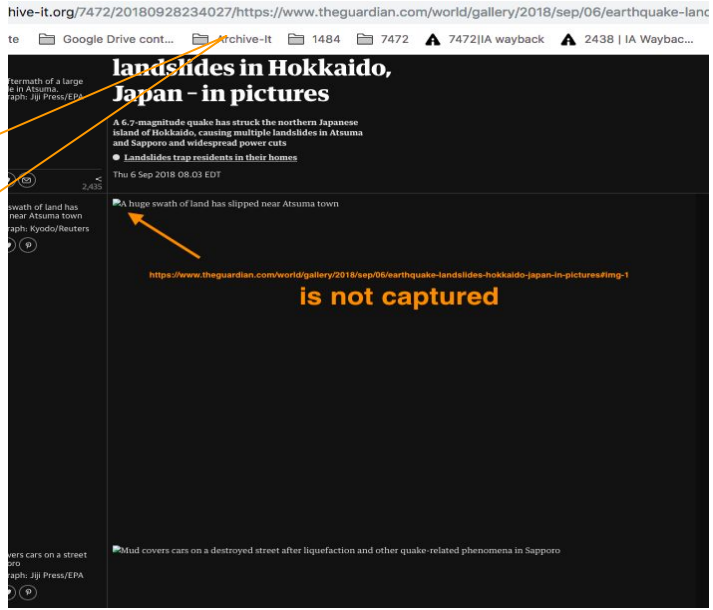
Add Seeds	https://partner.archive-it.org/1131/c	Crawl reports	https://partner.ar	https://partner.ar	https://partner.archive-it.org/1131/collections/7472/crawl/*/types		50kb/s
Update date	Crawled date	Crawl ID					
2018/11/03	Nov 2, 2018	696893 Patch	https://partner.ar	https://partner.ar	https://partner.ar	no data yet	
2018/11/03	Nov 1, 2018	696890 Patch	https://partner.ar	https://partner.ar	https://partner.ar	verified	GLIDE GARAGE
2018/11/03	Nov 1, 2018	696889 Weekly	https://partner.ar	https://partner.ar	https://partner.ar	review	
2018/11/03	Nov 1, 2018	696800 One-Time	https://partner.ar	https://partner.ar	https://partner.ar	reviewing	
2018/11/03	Nov 1, 2018	696799 One-Time	https://partner.ar	https://partner.ar	https://partner.ar	verified	なのはな生協　食の安全・安心を玄関先までお届け
2018/11/03	Nov 1, 2018	696755 Patch	https://partner.ar	https://partner.ar	https://partner.ar	verified	支援P 公式サイト – 災害ボランティア活動支援プロジェ
2018/11/03	Nov 1, 2018	696753 Patch	https://partner.ar	https://partner.ar	https://partner.ar	verified	ホーム - 宗教施設を地域資源とした地域防災のアクショ
2018/11/03	Nov 1, 2018	696709 Patch	https://partner.ar	https://partner.ar	https://partner.ar	verified	Yahoo!ネット募金
2018/11/01	Oct 31, 2018	696492 One-Time	https://partner.ar	https://partner.ar	https://partner.ar	verified	いのちを守る森の防潮堤
2018/11/01	Oct 30, 2018	696239 Patch	https://partner.ar	https://partner.ar	https://partner.ar	verified	Yahoo!ネット募金
2018/11/01	Oct 30, 2018	696191 Test	https://partner.ar	https://partner.ar	https://partner.ar	reviewing	ask anna-san about proxymode and Brozzler activation t
2018/11/01	Oct 30, 2018	696180 Patch	https://partner.ar	https://partner.ar	https://partner.ar	verified	一般社団法人 日本原子力産業協会 » 福島第一原子力発電

Unfortunate results

Incomplete capture examples



気仙沼信用金庫
http://wayback.archive-it.org/7472/20180904003005/http://www.shinkin.co.jp/kshinkin/



Earthquake triggers landslides in Hokkaido, Japan – in pictures | World news | The Guardian

http://wayback.archive-it.org/7472/20180928234027/https://www.theguardian.com/world/gallery/2018/sep/06/earthquake-landslides-hokkaido-japan-in-pictures

Unfortunate results & the Wayback calendar page

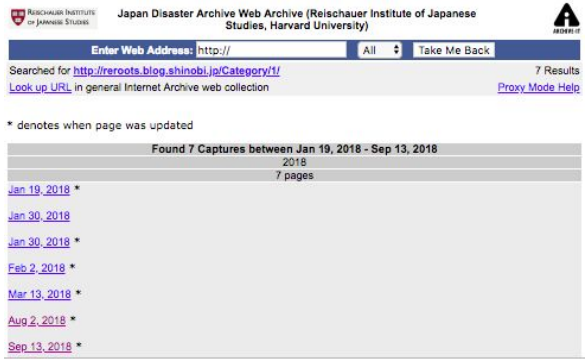
(Blocked by robots)

Seed URL	Seed Status	Docs	New Docs	Data ▾	New Data	Wayback Link
http://reroots.blog.shinobi.jp/Category/1/102/	Crawled	3	2	24.5 KB	24.1 KB	Wayback >
http://reroots.blog.shinobi.jp/Date/201803/1.html	Not crawled (blocked by robots)	0	0	0 bytes	0 bytes	
http://reroots.blog.shinobi.jp/Date/201802/1.html	Not crawled (blocked by robots)	0	0	0 bytes	0 bytes	
http://reroots.blog.shinobi.jp/Date/201801/1.html	Not crawled (blocked by robots)	0	0	0 bytes	0 bytes	

Wayback> is connected to the URL's calendar page, which shows the URL's captured dates.
http://wayback.archive-it.org/7472/*/http://reroots.blog.shinobi.jp/Category/1/

Unblock the domain in the collection or seed scope pages

Reference:
https://support.archive-it.org/hc/en-us/articles/208001096-Avoid-robots-txt-exclusions



The Main Point of This Presentation

Prior to Final Crawl

- Check your seed formatting (how a web page is constructed).
- Consider your seed types and decide the scopes: Standard, Standard +, One Page, One Page +
- Do Test crawls
  (Test crawls do not use up the data budget unless you save the data)
- Check the test crawls and add scoping rules (collection- and seed-level scoping) if necessary. (Before you try to do a second test crawl for the same seed, I recommend that you delete the test crawl previously done.)
- Do the actual crawl for the seed.

Introduction to web archiving using Archive-It

Instructor: Koko Fujita Howell
Japan Disasters Digital Archive http://jdarchive.org/en
Digital Content Lead/Web Archive Curator

Thank you

Time for the Archive-it experience!