

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->

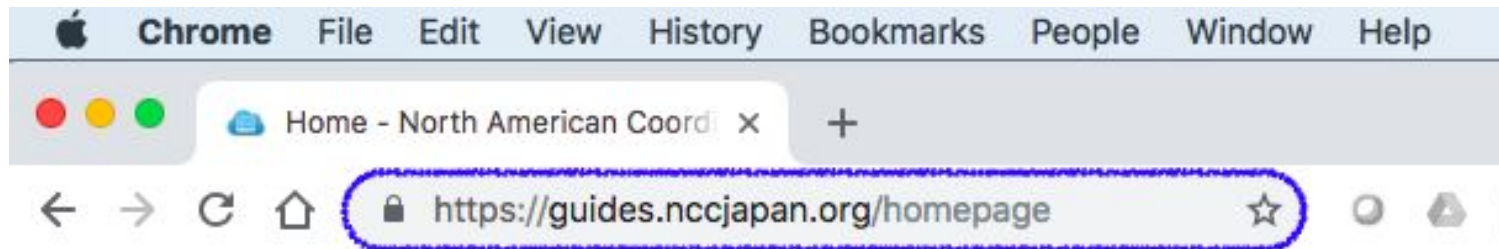
<https://physics.nist.gov/cuu/Units/binary.html>

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 Secure sada.la.coocan.jp Domain :http://sada.la.coocan.jp/ ☆

http://sada.la.coocan.jp/taketiti.JPG
is in scope

松浦武四郎案内処

松浦武四郎は八面六臂、多方面の活躍をした人物です。武四郎の足跡を追いながら、特に明治期の彼の業績と素顔に迫ってみようと思います。

2018年の 武四郎追跡	嘉永3(1850)年の稿本『鹿角日誌』『壺の碑考』の解読に取り組んでいます。10月 末に解読・入力を終え、今は解題・補注の執筆に取りかかっています。 なかなか進みませんが、11月末には終えるつもりです。 『奥州名山図譜』も同時収録予定です。
武四郎トピックス	都合により閉鎖しました。
お薦めサイト	松浦武四郎記念館→ https://takeshiro.net/ 高瀬元館長のサイト→ 松浦武四郎の足跡を訪ねて

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.google syndication.com

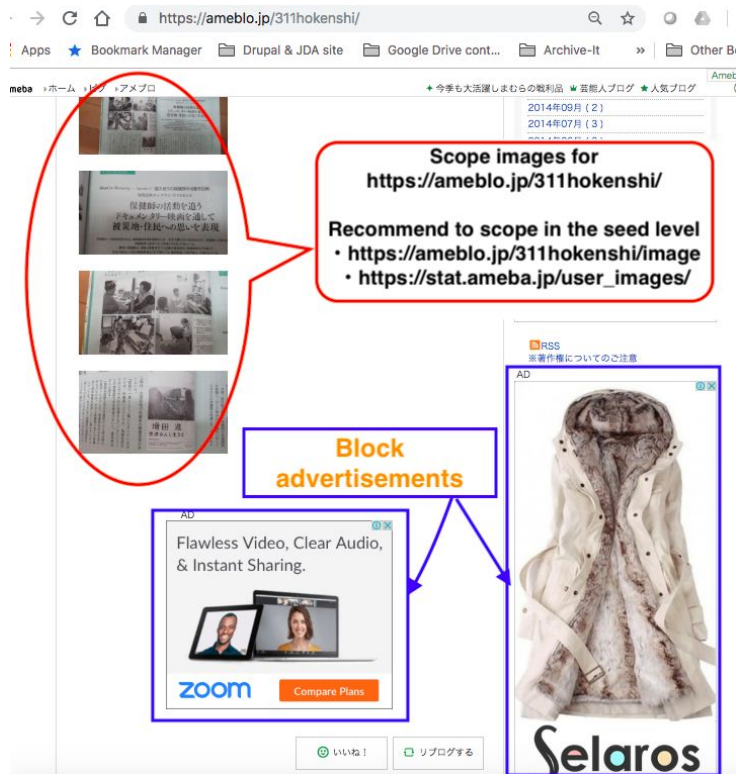
Result of capture

<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>

Run Crawl

Please select options below for a test crawl or one-time crawl of the selected seeds.

Crawl Type: ☒ One-Time Crawl ☐ Test Crawl

Doc. Limit: Documents

Data Limit: GB

Time Limit:

Crawl PDFs Only ☒

Crawling: ☒ Standard ☐ Brozzler (experimental)

Technology

1 selected seed will be crawled.

Not Secure | www.chosashi.or.jp/activity/publications/kaiho/kaiho2011.html

Bookmark Manager | Drupal & JDA site | Google Drive cont... | Archive-It | 1484

日本土地家屋調査士会連合会の活動

出版物のご紹介

会報の表紙は、写真コンクールの上位入賞作品から、季節にあわせた作品を紹介しています。

2012年3月号 (No.662)

- 02: 3.11 東日本大震災 被災地からのレポート
第3回 あの日あの日の後を想う
福島県土地家屋調査士会 小野田 幸一
- 05: 地籍学の法的側面・技術的側面について
後編 これからの都市計画・まちづくりにおける土地家屋調査士への役割
中央大学理工学部都市環境学科 谷下 雅哉
- 08: 事業所経営の必要知識—時代にあった事業所経営のために—
第12回 不動産の価値と調査・測量の成果
不動産鑑定士 田中 健

「もうすぐ改訂」
神田 女子大学
第26回写真コンクール金賞
賞状をダウンロードすると
会報表紙が印刷できます

2012年2月号 (No.661)

- 02: 地籍学の法的側面・技術的側面について
第6回 土地境界概念における対物性と観念性の相克
早稲田大学 山崎 孝夫
- 14: 事業所経営の必要知識
第10回 不動産市場を展望するためのポイント
株式会社 住友不動産研究所 上野 佳研究員 伊東 尚雄
- 16: 第26回日調連・第31回関東ブロック協議会合同
「東日本大震災」復興支援チャリティグループ大賞発表
東京土地家屋調査士会 後藤新長 木下 浩

「いざりがっこ作り」
佐藤 一太郎
第26回写真コンクール銀賞
賞状をダウンロードすると
会報表紙が印刷できます

PDFs are linked

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.

The screenshot shows the Partner Archive-It web interface. The address bar displays the URL <https://partner.archive-it.org/1484/collections/11371/crawls/test>. The navigation menu includes Home, Collections, Crawls (highlighted with a yellow circle), Archives, and ARS. The page title is "NCCworkshop111518". Below the title, it says "Created: Nov 8, 2018 by snoguchisnoguchi" and "Updated: Nov 8, 2018 by snoguchisnoguchi". A prompt "Select two crawls to compare." is visible. The horizontal navigation menu includes Overview, Seeds, Crawls (highlighted with a blue circle), Collection Scope, Metadata, and Wayback QA. Below this, there are four tabs: Crawl Reports (highlighted with a blue box), Current Crawls, Test Crawls (highlighted with a red box), and Scheduled Crawls. A red arrow points from the "Crawl Reports" tab to the "Test Crawls" tab. Below the tabs, there is a search bar labeled "Type to Filter Test Crawls" and a "Download Test Crawl List" link. A table of test crawl results is displayed below the search bar.

Crawl ID	Started	Completed	Status	New Data	Docs
709055	Nov 9, 2018	Nov 10, 2018	Finished: Saved	910.5 MB	15,235
709016	Nov 8, 2018	Nov 8, 2018	Finished: Saved	32.7 MB	1,778
709014	Nov 8, 2018	Nov 8, 2018	Finished: 55 Days Left to Save	522.9 MB	3,654

Crawl report table using Google Sheets (option)

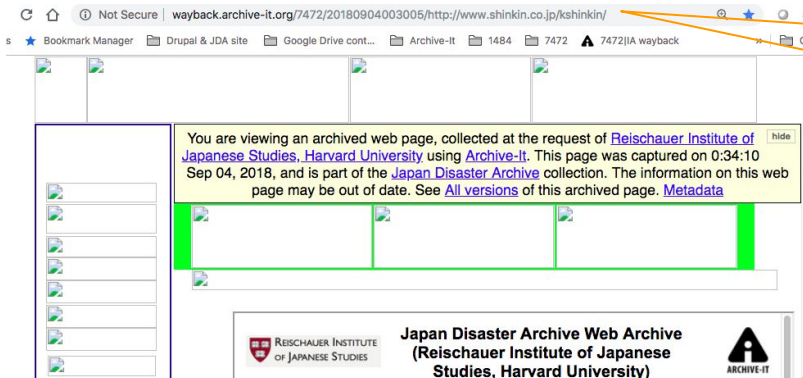
Nov 2, 2018	696897	Test	https://partner.archive-it.org/1131/collections/7472/crawl/696897/	https://partner.archive-it.org/1131/collections/7472/crawl/696897/seeds	https://partner.archive-it.org/1131/collections/7472/crawl/696897/types	no data yet
Nov 1, 2018	696799	One-Time	https://partner.archive-it.org/1131/collections/7472/crawl/696799/	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/collections/7472/crawl/*	Crawl reports	https://partner.archive-it.org/1131/collections/7472/crawl/*	https://partner.archive-it.org/1131/collections/7472/crawl/*	https://partner.archive-it.org/1131/collections/7472/crawl/*	50kb/s
Update date	Crawled date	Crawl ID				
2018/11/03	Nov 2, 2018	696893 Patch	https://partner.archive-it.org/1131/collections/7472/crawl/696893/	https://partner.archive-it.org/1131/collections/7472/crawl/696893/seeds	https://partner.archive-it.org/1131/collections/7472/crawl/696893/types	no data yet
2018/11/03	Nov 1, 2018	696890 Patch	https://partner.archive-it.org/1131/collections/7472/crawl/696890/	https://partner.archive-it.org/1131/collections/7472/crawl/696890/seeds	https://partner.archive-it.org/1131/collections/7472/crawl/696890/types	verified
2018/11/03	Nov 1, 2018	696889 Weekly	https://partner.archive-it.org/1131/collections/7472/crawl/696889/	https://partner.archive-it.org/1131/collections/7472/crawl/696889/seeds	https://partner.archive-it.org/1131/collections/7472/crawl/696889/types	review
2018/11/03	Nov 1, 2018	696800 One-Time	https://partner.archive-it.org/1131/collections/7472/crawl/696800/	https://partner.archive-it.org/1131/collections/7472/crawl/696800/seeds	https://partner.archive-it.org/1131/collections/7472/crawl/696800/types	reviewing
2018/11/03	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
2018/11/03	Nov 1, 2018	696755 Patch	https://partner.archive-it.org/1131/collections/7472/crawl/696755/	https://partner.archive-it.org/1131/collections/7472/crawl/696755/seeds	https://partner.archive-it.org/1131/collections/7472/crawl/696755/types	verified
2018/11/03	Nov 1, 2018	696753 Patch	https://partner.archive-it.org/1131/collections/7472/crawl/696753/	https://partner.archive-it.org/1131/collections/7472/crawl/696753/seeds	https://partner.archive-it.org/1131/collections/7472/crawl/696753/types	verified
2018/11/03	Nov 1, 2018	696709 Patch	https://partner.archive-it.org/1131/collections/7472/crawl/696709/	https://partner.archive-it.org/1131/collections/7472/crawl/696709/seeds	https://partner.archive-it.org/1131/collections/7472/crawl/696709/types	verified
2018/11/01	Oct 31, 2018	696492 One-Time	https://partner.archive-it.org/1131/collections/7472/crawl/696492/	https://partner.archive-it.org/1131/collections/7472/crawl/696492/seeds	https://partner.archive-it.org/1131/collections/7472/crawl/696492/types	verified
2018/11/01	Oct 30, 2018	696239 Patch	https://partner.archive-it.org/1131/collections/7472/crawl/696239/	https://partner.archive-it.org/1131/collections/7472/crawl/696239/seeds	https://partner.archive-it.org/1131/collections/7472/crawl/696239/types	verified
2018/11/01	Oct 30, 2018	696191 Test	https://partner.archive-it.org/1131/collections/7472/crawl/696191/	https://partner.archive-it.org/1131/collections/7472/crawl/696191/seeds	https://partner.archive-it.org/1131/collections/7472/crawl/696191/types	reviewing
2018/11/01	Oct 30, 2018	696180 Patch	https://partner.archive-it.org/1131/collections/7472/crawl/696180/	https://partner.archive-it.org/1131/collections/7472/crawl/696180/seeds	https://partner.archive-it.org/1131/collections/7472/crawl/696180/types	verified

Unfortunate results

Incomplete capture examples

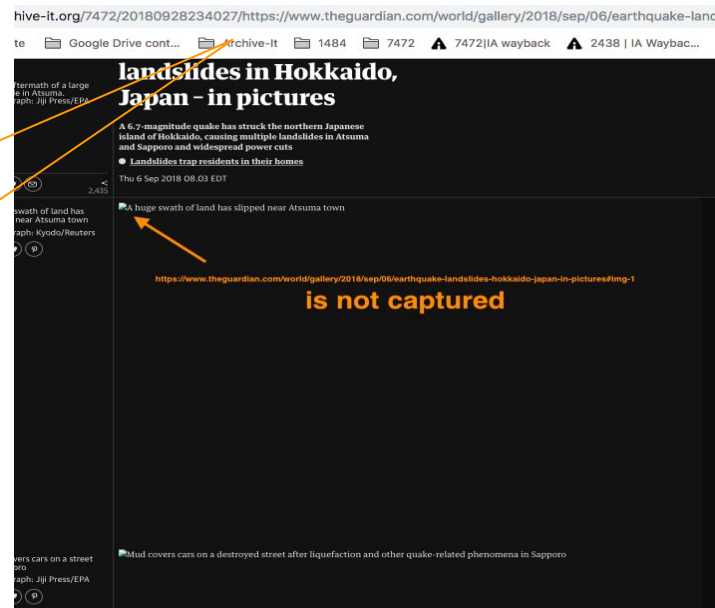


気仙沼信用金庫

<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>

Japan Disaster Archive Web Archive (Reischauer Institute of Japanese Studies, Harvard University)

Enter Web Address: All

Searched for <http://reroots.blog.shinobi.jp/Category/1/> 7 Results
[Look up URL](#) in general Internet Archive web collection [Proxy Mode Help](#)

* denotes when page was updated

Found 7 Captures between Jan 19, 2018 - Sep 13, 2018

2018
7 pages

[Jan 19, 2018](#) *
[Jan 30, 2018](#)
[Jan 30, 2018](#) *
[Feb 2, 2018](#) *
[Mar 13, 2018](#) *
[Aug 2, 2018](#) *
[Sep 13, 2018](#) *

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!