Supporting Computational Research on Large Digital Collections

Internet Archive & Archives Unleashed

Jefferson Bailey & Nick Ruest

CNI Fall 2022

Presentation Outline

- General Challenges Supporting Computational Research on Large Digital Collections
- ARCH (Archives Research Compute Hub) Background, Goals, Status
- ARCH Technical Overview & Walkthrough
- Supporting Scholarly Use

The Challenge

How to understand and address the technical, conceptual, and practical issues inherent in supporting data-driven uses of large, complex, heterogeneous (born) digital collections

Technical Issues in Research Use

- Data format complexities (WARC, data, codecs)
- Data volume complexities (transfer, store)
- Data processing complexities (local v. cluster)
- Data “visibility” complexities (outputs)
- Data derivation complexities (pipelines)

Conceptual Issues in Research Use

- Provenance complexities
- Acquisition complexities
- Border/boundary complexities
- Breadth complexities
- Ellipses (elisions)

Practical Issues in Research Use

- Where is the front door? (requests)
- Scoping interviews (support)
- Research Agreements (paperwork)
- Time/cost modeling (budgeting)
- Responsible parties (staffing)
- Program Development (organization)

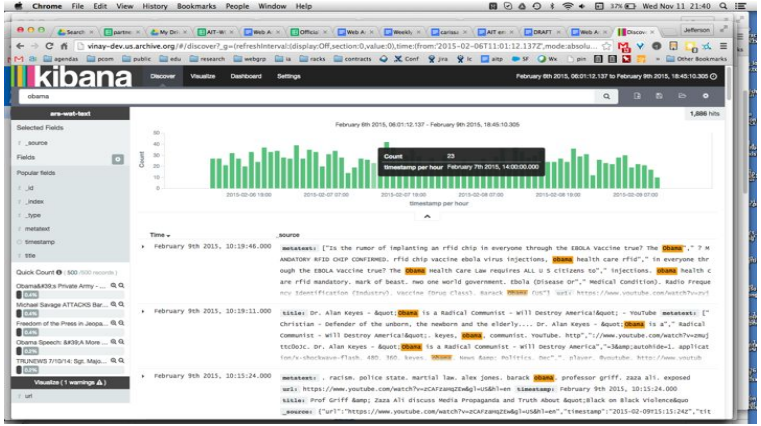
Typology of Computational Research Services Models

- Bulk Data Model
- Cyberinfrastructure Model
- Roll Your Own Model
- Middleware Model
- Prepackaged Model
- Support & Community Model

Practical Lessons

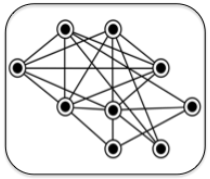
- Large digital collections:
  - can be unkind to traditional methods of scholarly inquiry
    - Create guides or expertise to help users know the datasets, subsets, extractions to suite their research
  - impose many technical and conceptual issues
    - Be prepared to discuss/document these at the beginning of research request process
  - are alluring but can be deceitful
    - More data isn't better research; informed computational research services are critical
  - may be part of an even larger research corpus aggregation
    - Research support will require cross-functional skills

Background: IA Computational Research Services



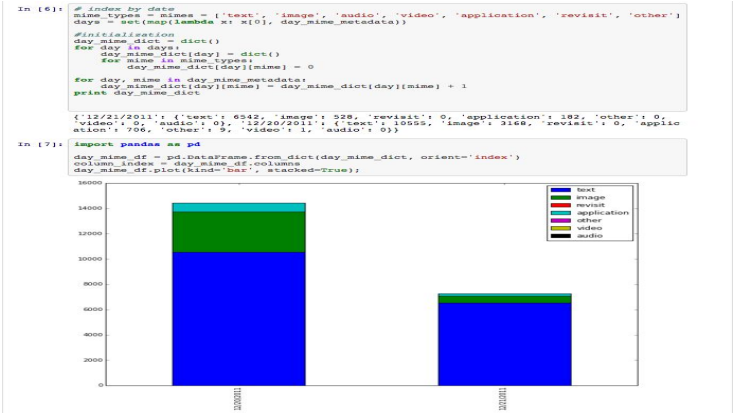

```
"id": 5260,
"account": 421,
"created_by": "sbreen",
"created_date": "2015-02-03T00:51:21Z",
"last_updated_by": "shallcro",
"last_updated_date": "2015-03-02T16:38:43Z",
"name": "U.S. Presidential Election 2016",
"tag": "",
"state": "INACTIVE",
"publicly_visible": true,
"one_hop_off": false,
"topics": "government-USFederal;politicsAndElections;government-Nation
"oai_exported": false,
"metadata": {
    "Contributor": [
        {
            "id": 928936,
            "value": "Shallcross, Michael"
        }
    ],
    "Description": [
        {
            "id": 928937,
            "value": "The 2016 U.S. Presidential Election web archive
        }
    ],
    "Title": [
        {
            "id": 928939,
            "value": "U.S. Presidential Election 2016"
        }
    ],
    "Creator": [
        {
            "id": 928938,
            "value": "Breen, Sarah; Nofziger, Cinda; and Thomas, Rob"
        }
    ],
    "Date": [
        {
            "id": 928935,
            "value": "2015-02-03"
        }
    ],
    "Type": [
        {
            "id": 928934,
            "value": "Web Archive"
        }
```


WAT Datasets
(Web Archive Transformation)
Key Metadata from Every Resource



LGA Datasets
(Longitudinal Graph Analysis)
What Links to What over Time

WANE Datasets
(Web Archive Named Entities)
Names of People, Places, Organizations



Background: The “Archives Unleashed” Project

- A long-running project (est. 2017) that “*aims to make petabytes of historical internet content accessible to scholars and others interested in researching the recent past*,” primarily supported by The Andrew W. Mellon Foundation.
- In 2020, Archives Unleashed partnered with the Internet Archive to develop “ARCH”.
- Three main partners: Internet Archive, University of Waterloo (hi Ian!), and York University.

Archives Unleashed

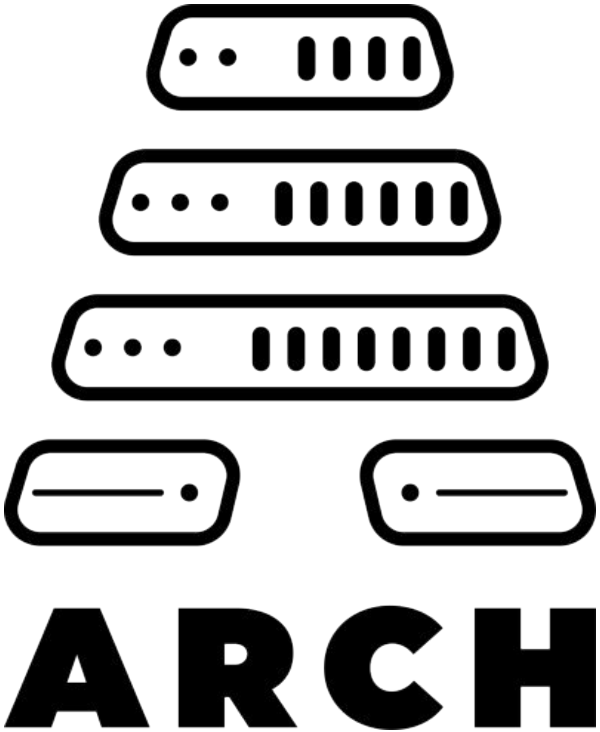
UNIVERSITY OF WATERLOO

YORK UNIVERSITÉ UNIVERSITY

INTERNET ARCHIVE

ARCH Goals, Development, Roadmap

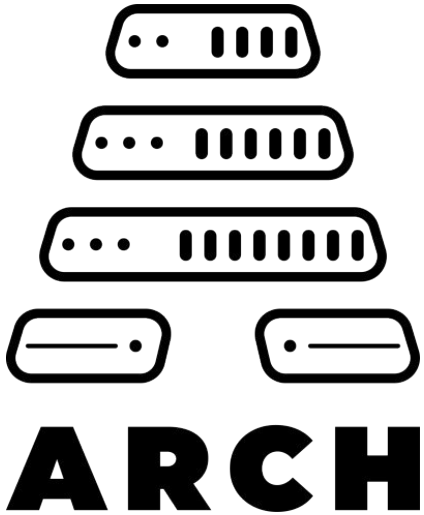
- Goal: merge and standardize our tools and services into one web/text/data mining platform in a OSS SaaS product model
- Goal: colocate data & compute in centralized IA-run infrastructure and data centers (eventually in US & Canada)
- Goal: embed researcher support into the platform dev process with funded cohorts teams and invited scholars
- Development: Engineering by AU & IA teams 2022-current
- Development: Supporting AU cohorts & IA pilot partners; 20+ LAMs & 50+ scholars, 16 countries; 200+TB processed
- Roadmap: Add text/image to collection types; query filtering; aggregated/uploaded collections; small/mid LAM outreach
- Roadmap: Production release Q1 2023



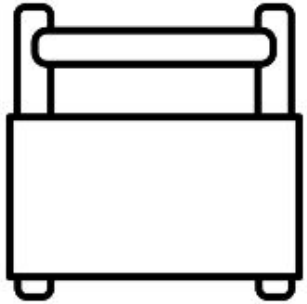
Archives Research Compute Hub

ARCH Platform Details

- Interactive web application interface both for collection curators and scholarly researchers
- Generate and download over 20 derivative datasets, and connect to Google Colab
- Three rounds of UI/UX testing
- In-browser visualizations and data previews that presents a glimpse into collection content
- Located in the Internet Archive data center, ARCH has quick access to the petabytes of content collected



ARCH Stack



Scalatra

Archives Research Compute Hub

Scala version 2.12.8 Scalatra version 2.5.4 license MIT

About

Web application for distributed compute analysis of Archive-It web archive collections.

Building

README.md

- `sbt "prod/clean" "prod/assembly" "prod/assemblyPackageDependency"`

Docker

1. Create a config (`config/config.json`) for your Docker setup, e.g., by copying the included template: `cp config/docker.json config/config.json`
2. `docker build --no-cache -t arch .`
3. `docker run -it --rm -p 12341:12341 -p 54040:54040 -v /home/nruest/Projects/au/sample-data/arch:/data -v /home/nruest/Projects/au/arch:/app arch`

Web application will be available at: http://localhost:12341/ait, and Apache Spark interface will be available at http://localhost:54040.

Citing ARCH

How to cite ARCH in your research:

> Helge Holzmann, Nick Ruest, Jefferson Bailey, Alex Dempsey, Samantha Fritz, Peggy Lee, and Ian Milligan. 2022. ABCDEF: the 6 key features behind scalable, multi-tenant web archive processing with ARCH: archive, big data, concurrent, distributed, efficient, flexible. In Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries (JCDL '22). Association for Computing Machinery, New York, NY, USA, Article 13, 1–11. https://doi.org/10.1145/3529372.3530916

Your citations help to further the recognition of using open-source tools for scientific inquiry, assists in growing the web archiving community, and acknowledges the efforts of contributors to this project.

License

MIT

Open-source, not open-contribution

Similar to SQLite, ARCH is open source but closed to contributions.

The level of complexity of this project means that even simple changes can break a lot of other moving parts in our production environment. However, community involvement, bug reports and feature requests are warmly accepted.

Acknowledgments

This work is primarily supported by the Andrew W. Mellon Foundation. Other financial and in-kind support comes from the Social Sciences and Humanities Research Council, Compute Canada, York University Libraries, Start Smart Labs, and the Faculty of Arts at the University of Waterloo.

Any opinions, findings, and conclusions or recommendations expressed are those of the researchers and do not necessarily reflect the views of the sponsors.

ABCDEF - The 6 key features behind scalable, multi-tenant web archive processing with ARCH: Archive, Big Data, Concurrent, Distributed, Efficient, Flexible

Helge Holzmann[1] , Nick Ruest[2] , Jefferson Bailey[1] , Alex Dempsey[1] , Samantha Fritz[3] , Peggy Lee[1] , and Ian Milligan,[3]

[1] Internet Archive
[2] Digital Scholarship Infrastructure Department, York University
[3] Department of History, University of Waterloo

ABSTRACT

Over the past quarter-century, web archive collection has emerged as a user-friendly process thanks to cloud-hosted solutions such as the Internet Archive's Archive-It subscription service. Despite advancements in collecting web archive content, no equivalent has been found by way of a user-friendly cloud-hosted analysis system. Web archive processing and research require significant hardware resources and cumbersome tools that interdisciplinary researchers find difficult to work with. In this paper, we identify six principles - the ABCDEFs (Archive, Big data, Concurrent, Distributed, Efficient, and Flexible) - used to guide the development and design of a system. These make the transformation of, and working with, web archive data as enjoyable as the collection process. We make these objectives – largely common sense – explicit and transparent in this paper. They can be employed by every computing platform in the area of digital libraries and archives and adapted by teams seeking to implement similar infrastructures. Furthermore, we present ARCH (Archives Research Compute Hub)[1], the first cloud-based system designed from scratch to meet all of these six key principles. ARCH is an interactive interface, closely connected with Archive-It, engineered to provide analytical actions, specifically generating datasets and in-browser visualizations. It efficiently streamlines research workflows while eliminating the burden of computing requirements. Building off past work by both the Internet Archive (Archive-It Research Services) and the Archives Unleashed Project (the Archives Unleashed Cloud), this merged platform achieves a scalable processing pipeline for web archive research. It is open-source and can be considered a reference implementation of the ABCDEF, which we have evaluated and discussed in terms of feasibility and compliance as a benchmark for similar platforms.

[1]https://github.com/internetarchive/arch

JCDL '22, June 20–24, 2022, Cologne, Germany

ACM ISBN 978-1-4503-9345-4/22/06…$15.00
https://doi.org/10.1145/3529372.3530916

CCS CONCEPTS

• **Information systems** → **Digital libraries and archives**; **Data extraction and integration**.

KEYWORDS

web archives, big data, data processing, distributed computing

ACM Reference Format:
Helge Holzmann[1] , Nick Ruest[2] , Jefferson Bailey[1] , Alex Dempsey[1] , Samantha Fritz[3] , Peggy Lee[1] , and Ian Milligan,[3] . 2022. ABCDEF - The 6 key features behind scalable, multi-tenant web archive processing with ARCH: Archive, Big Data, Concurrent, Distributed, Efficient, Flexible. In *The ACM/IEEE Joint Conference on Digital Libraries in 2022 (JCDL '22), June 20–24, 2022, Cologne, Germany.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3529372.3530916

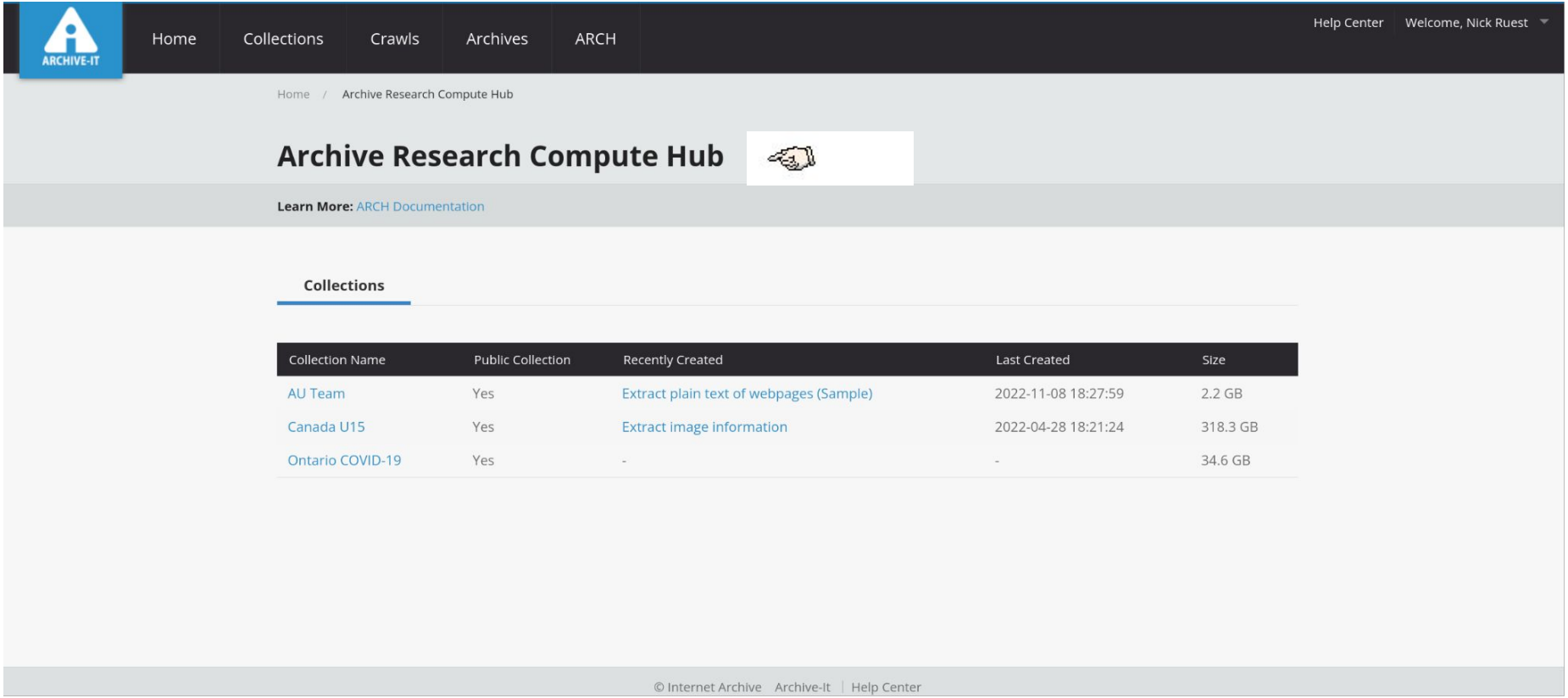
1 INTRODUCTION

Web archiving is an important component of modern digital libraries. It is essential for enabling future research into contemporary history and ensuring the long-term preservation of our documentary heritage [11] [3]. Yet while collecting web archive content has matured into a user-friendly process, thanks in no small part to cloud-hosted solutions such as the Internet Archive's Archive-It service, this ease-of-use has not been matched on the analysis side. We accordingly need a user-friendly system that can enable the creation of research datasets from web archives so that researchers can work with material at scale.

In this paper, we present the Archives Research Compute Hub (ARCH), a production system tightly integrated with the Internet Archive infrastructure and services. ARCH grew out of the Archives Unleashed Cloud: a proof-of-concept platform that demonstrated the ability of a web browser-based system to power backend Apache Spark-driven jobs on web archival datasets [12]. Powered by the Archives Unleashed Toolkit and the Internet Archive's Sparkling data processing library, the ARCH platform will become a complementary component of the Internet Archive's Archive-It system. ARCH is built around six key principles: archive, big data, concurrent, distributed, efficient, and flexible. We present these principles as considerations for projects and teams developing similar systems.

2 RELATED WORK AND PROJECT CONTEXT

Established in 2017, the Archives Unleashed project recognizes the collective need among researchers, librarians and archivists for analytical tools, community infrastructure, and accessible web archival


ARCHIVE-IT
Home
Collections
Crawls
Archives
ARCH
Help Center
Welcome, Nick Ruest
Home / Archive Research Compute Hub
Archive Research Compute Hub
Learn More: ARCH Documentation
Collections
Collection Name | Public Collection | Recently Created | Last Created | Size
AU Team | Yes | Extract plain text of webpages (Sample) | 2022-11-08 18:27:59 | 2.2 GB
Canada U15 | Yes | Extract image information | 2022-04-28 18:21:24 | 318.3 GB
Ontario COVID-19 | Yes | - | - | 34.6 GB
© Internet Archive Archive-It | Help Center

ARCHIVE-IT | Home | Collections | Crawls | Archives | ARCH | Help Center | Welcome, Nick Ruest

Home / Archive Research Compute Hub / U.S. LGBTQ Web Collection

U.S. LGBTQ Web Collection Analysis

Learn More: ARCH Documentation

Job Summary | Generate Datasets

Collection Overview

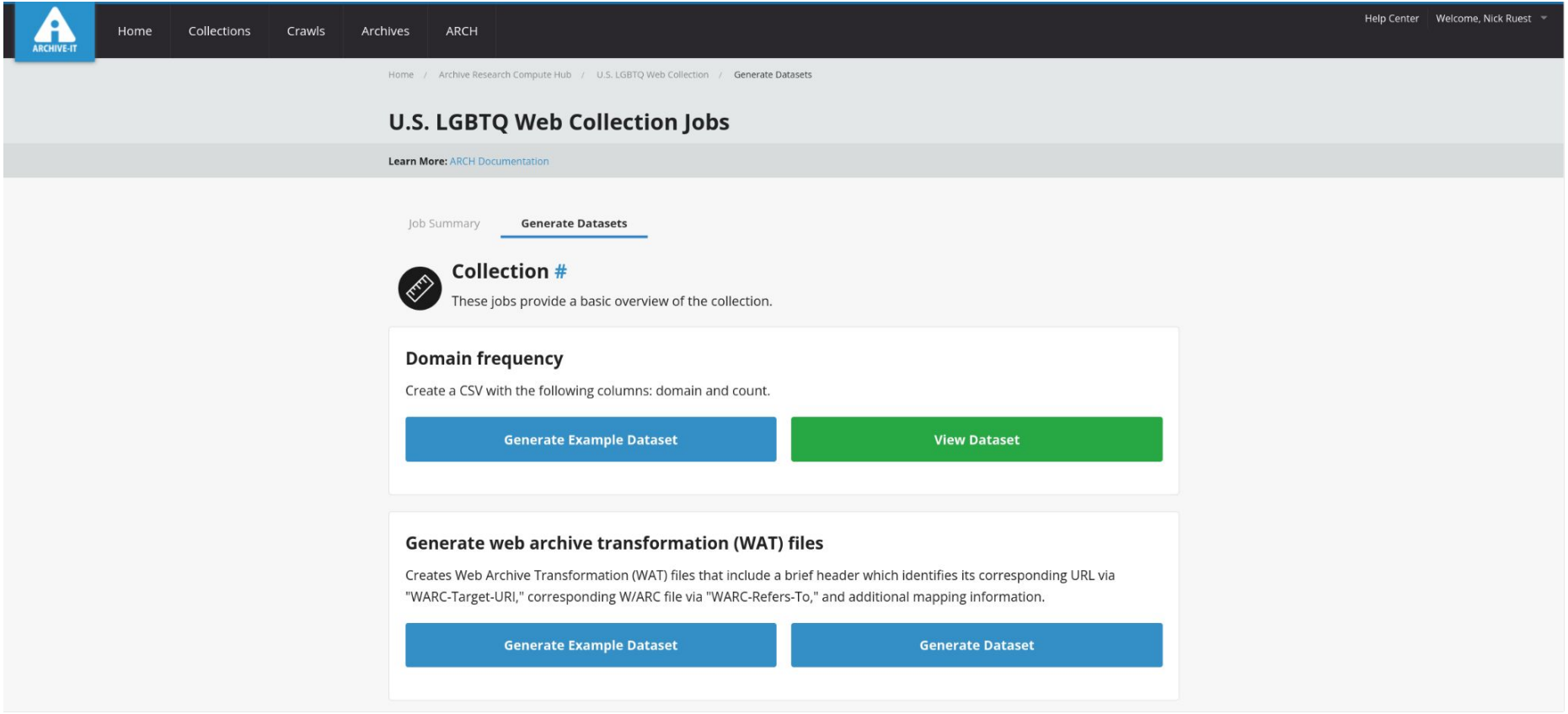
257 seeds	Crawled Jan 23, 2019	1.1 TB	Public Collection

Public Collection Link: https://archive-it.org/collections/02778

There are currently no active jobs, **generate a new dataset**.

Completed Jobs

Job	Finished
Domain frequency	2022-10-13 08:33:03
Extract domain graph	2022-10-13 08:52:04
Extract domain graph (Example)	2022-09-23 00:16:40
Extract image graph (Example)	2022-09-23 00:04:17
Extract plain text of webpages	2022-10-13 13:58:02
Extract plain text of webpages (Example)	2022-09-21 14:03:54
Extract web graph (Example)	2022-09-23 00:03:49


ARCHIVE-IT
Home
Collections
Crawls
Archives
ARCH
Help Center
Welcome, Nick Ruest
Home / Archive Research Compute Hub / U.S. LGBTQ Web Collection / Generate Datasets
U.S. LGBTQ Web Collection Jobs
Learn More: ARCH Documentation
Job Summary
Generate Datasets
Collection #
These jobs provide a basic overview of the collection.
Domain frequency
Create a CSV with the following columns: domain and count.
Generate Example Dataset
View Dataset
Generate web archive transformation (WAT) files
Creates Web Archive Transformation (WAT) files that include a brief header which identifies its corresponding URL via "WARC-Target-URI," corresponding W/ARC file via "WARC-Refers-To," and additional mapping information.
Generate Example Dataset
Generate Dataset

Network

These jobs produce files that provide network graphs for analysis, and offer an opportunity to explore the way websites link to each other.

Extract domain graph

Create a CSV with the following columns: crawl date, source domain, target domain, and count.

View Example Dataset | View Dataset

Extract image graph

Create a CSV with the following columns: crawl date, source of the image (where it was hosted), the URL of the image, and the alternative text of the image.

View Example Dataset | Generate Dataset

Extract longitudinal graph

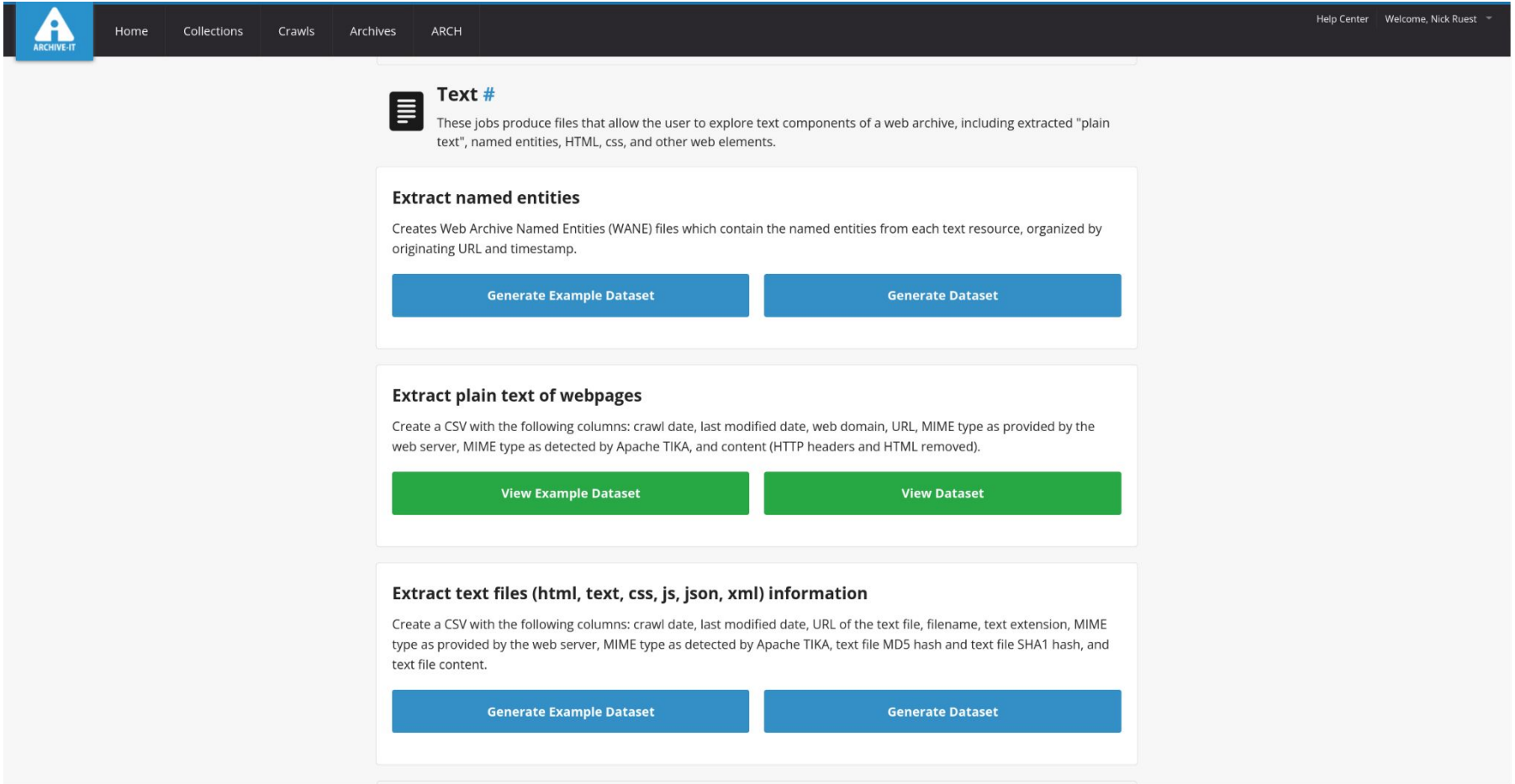
Creates Longitudinal Graph Analysis (LGA) files which contain a complete list of what URLs link to what URLs, along with a timestamp.

Generate Example Dataset | Generate Dataset

Extract web graph

Create a CSV with the following columns: crawl date, source, target, and anchor text. Note that this contains all links and is not aggregated into domains.

View Example Dataset | Generate Dataset


ARCHIVE-IT
Home
Collections
Crawls
Archives
ARCH
Help Center
Welcome, Nick Ruest
Text #
These jobs produce files that allow the user to explore text components of a web archive, including extracted "plain text", named entities, HTML, css, and other web elements.
Extract named entities
Creates Web Archive Named Entities (WANE) files which contain the named entities from each text resource, organized by originating URL and timestamp.
Generate Example Dataset
Generate Dataset
Extract plain text of webpages
Create a CSV with the following columns: crawl date, last modified date, web domain, URL, MIME type as provided by the web server, MIME type as detected by Apache TIKA, and content (HTTP headers and HTML removed).
View Example Dataset
View Dataset
Extract text files (html, text, css, js, json, xml) information
Create a CSV with the following columns: crawl date, last modified date, URL of the text file, filename, text extension, MIME type as provided by the web server, MIME type as detected by Apache TIKA, text file MD5 hash and text file SHA1 hash, and text file content.
Generate Example Dataset
Generate Dataset

File Formats

These jobs produce files that contain information on certain types of binary files found within a web archive.

Extract audio information

Create a CSV with the following columns: crawl date, last modified date, URL of the audio file, filename, audio extension, MIME type as provided by the web server, MIME type as detected by Apache TIKA, audio MD5 hash and audio SHA1 hash.

Generate Example Dataset | Generate Dataset

Extract image information

Create a CSV with the following columns: crawl date, last modified date, URL of the image, filename, image extension, MIME type as provided by the web server, MIME type as detected by Apache TIKA, image width, image height, image MD5 hash and image SHA1 hash.

Generate Example Dataset | Generate Dataset

Extract PDF information

Create a CSV with the following columns: crawl date, last modified date, URL of the PDF file, filename, PDF extension, MIME type as provided by the web server, MIME type as detected by Apache TIKA, PDF MD5 hash and PDF SHA1 hash.

Generate Example Dataset | Generate Dataset

Extract PowerPoint (e.g., .ppt, .odp, .key) information

Create a CSV with the following columns: crawl date, last modified date, URL of a PowerPoint or similar file, filename, PowerPoint or similar file extension, MIME type as provided by the web server, MIME type as detected by Apache TIKA, PowerPoint or similar file MD5 hash and PowerPoint or similar file SHA1 hash.

Generate Example Dataset | Generate Dataset

Extract spreadsheet information

Create a CSV with the following columns: crawl date, last modified date, URL of the spreadsheet file, filename, spreadsheet extension, MIME type as provided by the web server, MIME type as detected by Apache TIKA, spreadsheet MD5 hash and spreadsheet SHA1 hash.

Generate Example Dataset | Generate Dataset

Extract video information

Create a CSV with the following columns: crawl date, last modified date, URL of the video file, filename, video extension, MIME type as provided by the web server, MIME type as detected by Apache TIKA, video MD5 hash and video SHA1 hash.

Generate Example Dataset | Generate Dataset

Extract Word Documents (e.g., .doc, .odt, .rtf, .wpd) information

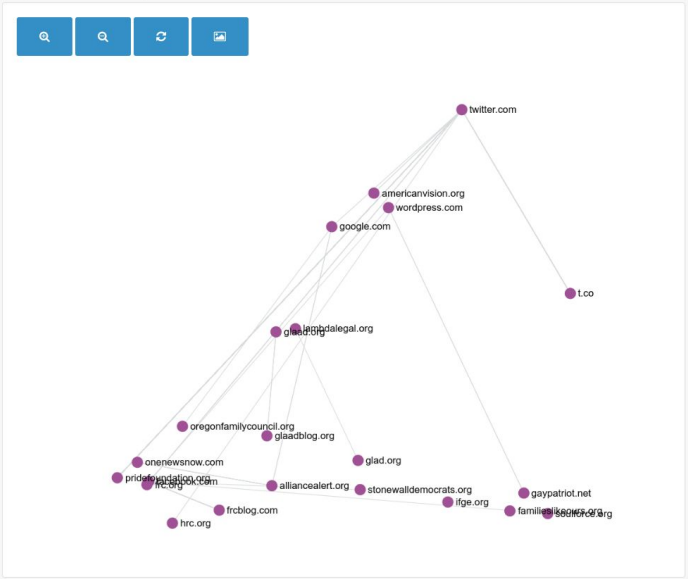
Create a CSV with the following columns: crawl date, last modified date, URL of the word document or similar file, filename, word document or similar file extension, MIME type as provided by the web server, MIME type as detected by Apache TIKA, word document or similar file MD5 hash and word document or similar file SHA1 hash.

Generate Example Dataset | Generate Dataset

U.S. LGBTQ Web Collection: Extract domain graph

Learn More: ARCH Documentation

Top Hosts Sample



This graph shows up to the top 100 nodes with highest degree (number of inlinks + outlinks) along with their edges.

Dataset(s)

A CSV with the following columns: crawl date, source domain, target domain, and count.

File name: domain-graph.csv.gz
File size: 42.6 MB
Result count: 6,327,594 lines
Date completed: 2022-10-13
Checksum(s): md5:d542459b517b1fc90c1ff01c69a132b4

Download

Preview

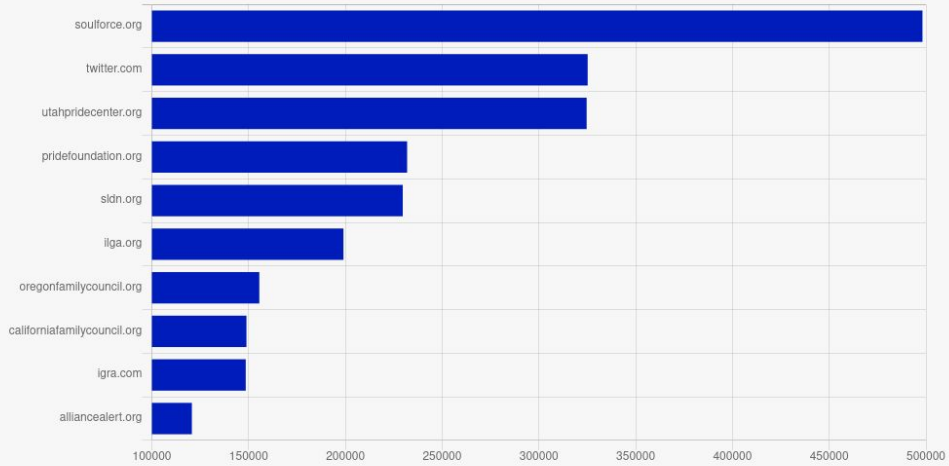
crawl_date	source	target	count
20180327152758	porco.ga	porco.ga	12081
20180327153659	susi.ml	susi.ml	12081
20180327152801	porco.gq	porco.gq	12081
20130408205620	facebook.com	facebook.com	11851
20130408203658	facebook.com	facebook.com	11254
20160616014151	tubepornstars.com	tubepornstars.com	9290
20170609194110	annakooiman.com	annakooiman.com	8961
20170609194142	squarespace.com	squarespace.com	8904
20130409013244	tubepornstars.com	tubepornstars.com	8806
20130408194055	facebook.com	facebook.com	8118
20130408194040	facebook.com	facebook.com	8063
20160629165957	facebook.com	facebook.com	7758
20130408190950	facebook.com	facebook.com	7720

Download Preview Data

U.S. LGBTQ Web Collection: Domain frequency

Learn More: ARCH Documentation

Domains



This graph shows the top 10 domains and the number of times that they occur in this collection.

Dataset(s)

A CSV with the following columns: domain and count.

File name: domain-frequency.csv.gz
File size: 170.4 KB
Result count: 23,910 lines
Date completed: 2022-10-13
Checksum(s): md5:e618bb22efe9f7ff2a85d28798d5b45f

Download

Preview

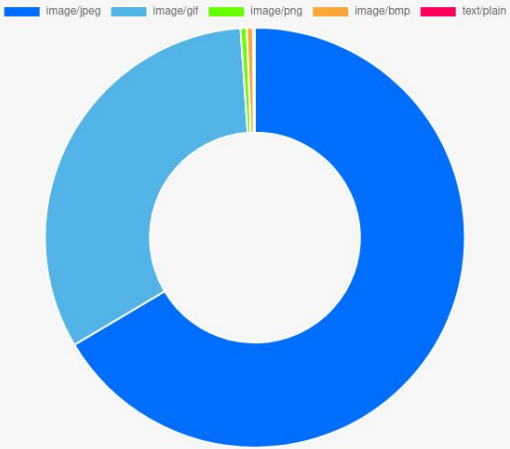
domain	count
soulforce.org	498191
twitter.com	325181
utahpridecenter.org	324651
pridefoundation.org	231888
sldn.org	229624
ilga.org	198982
oregonfamilycouncil.org	155503
californiafamilycouncil.org	148862
igra.com	148490
alliancealert.org	120648
gaygames.net	104389
narth.com	99862
gaypatriot.net	93066

Download Preview Data

Canada U15: Extract image information

Learn More: ARCH Documentation

File Format Distribution



This graph shows the distribution of the various file formats in the web archive collection.

Dataset(s)

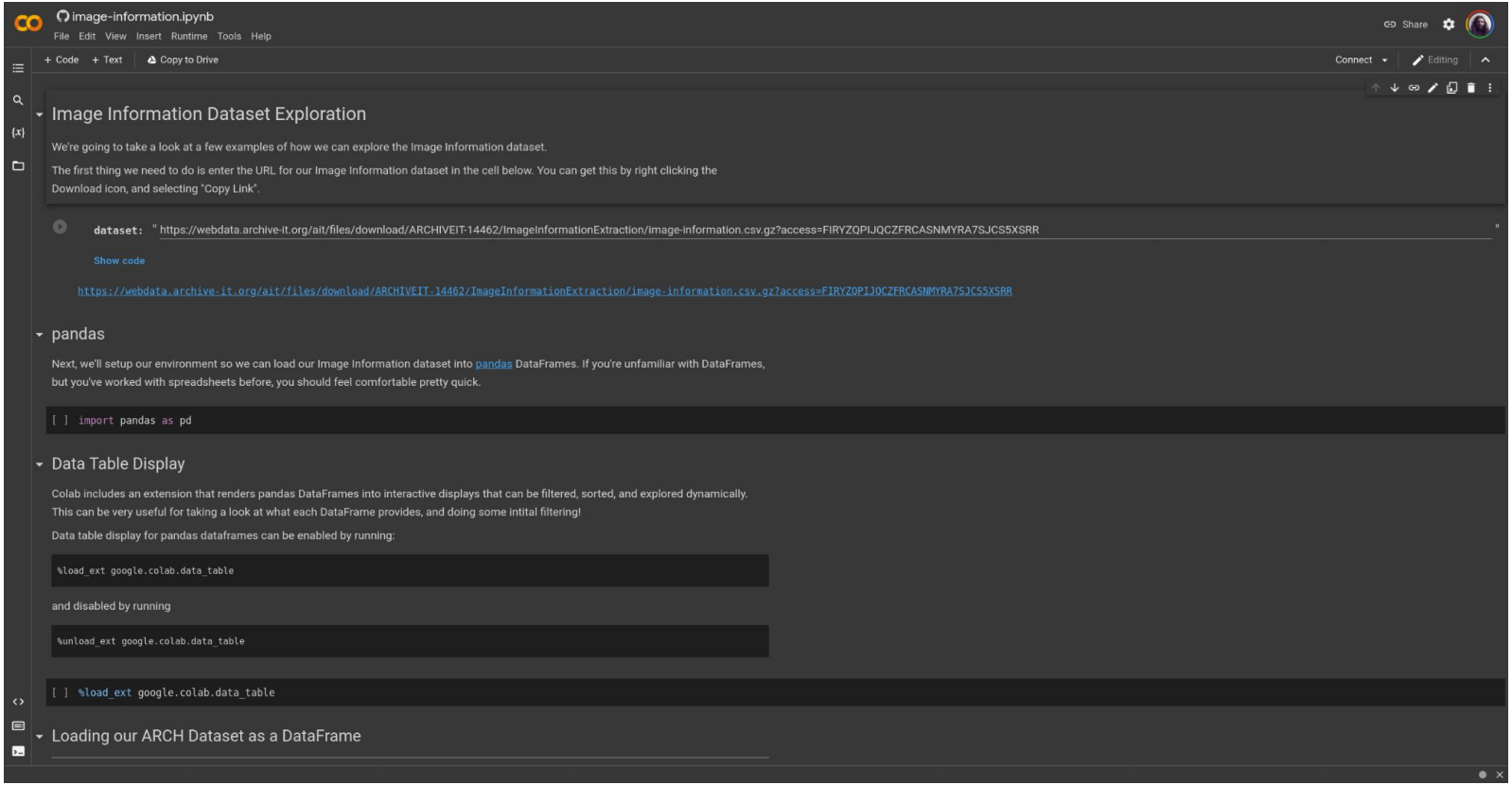
A CSV with the following columns: crawl date, last modified date, URL of the image, filename, image extension, MIME type as provided by the web server, MIME type as detected by Apache TIKA, image width, image height, image MD5 hash and image SHA1 hash.

File name: image-information.csv.gz
File size: 1.6 MB
Result count: 24,705 lines
Date completed: 2022-04-28
Checksum(s): md5:191494a2a761f203f1ec4123e751b334

Download

Open in Colab

My Research Page



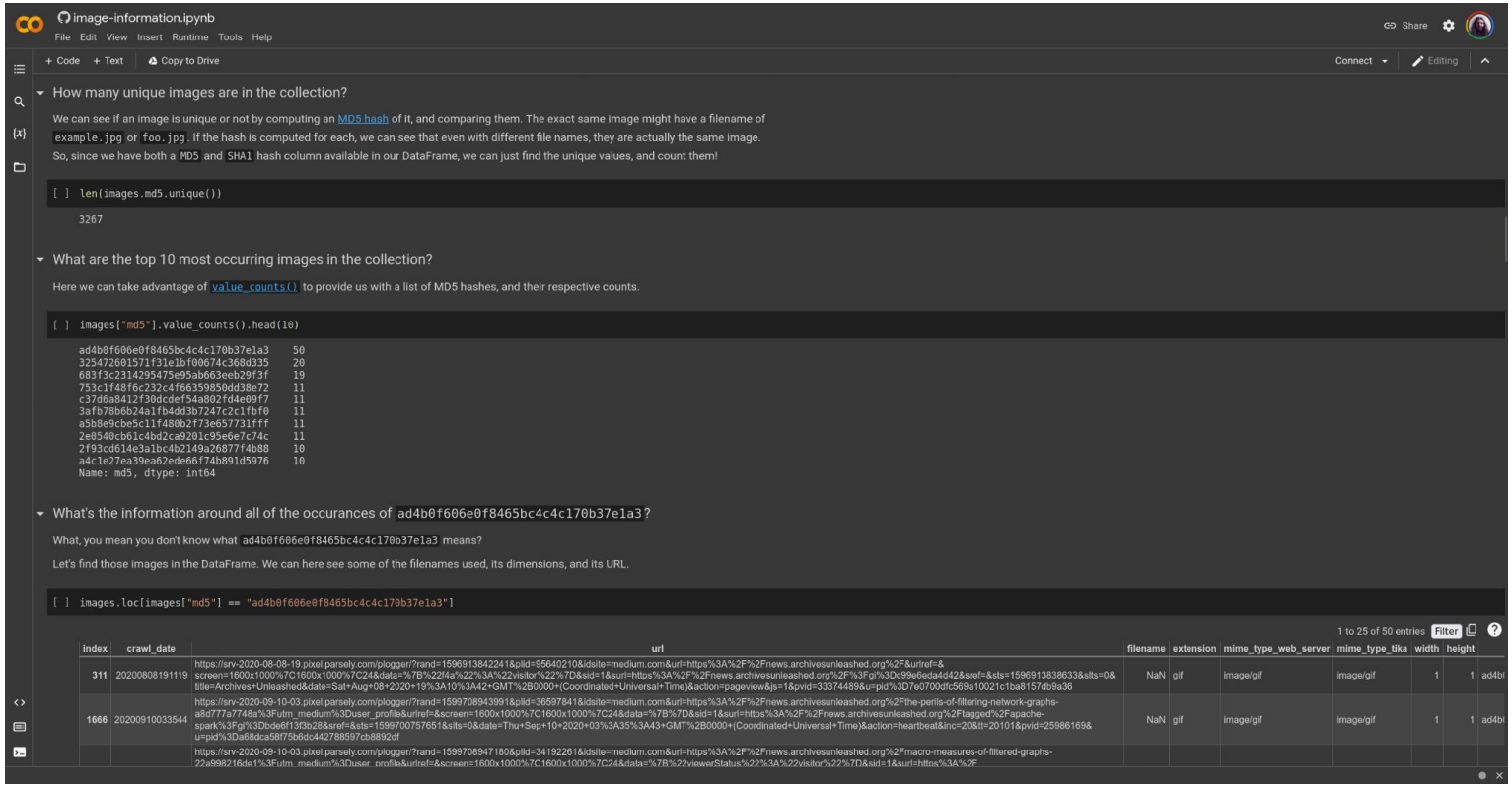


image-information.ipynb

File Edit View Insert Runtime Tools Help

Share

+ Code + Text Copy to Drive

Connect Editing

How many unique images are in the collection?

We can see if an image is unique or not by computing an MD5 hash of it, and comparing them. The exact same image might have a filename of `example.jpg` or `foo.jpg`. If the hash is computed for each, we can see that even with different file names, they are actually the same image.
So, since we have both a `MD5` and `SHA1` hash column available in our DataFrame, we can just find the unique values, and count them!

```
[ ] len(images.md5.unique())
```

```
3267
```

What are the top 10 most occurring images in the collection?

Here we can take advantage of `value_counts()` to provide us with a list of MD5 hashes, and their respective counts.

```
[ ] images["md5"].value_counts().head(10)
```

```
ad4b0f606e0f8465bc4c4c170b37e1a3    50
3254726015711f31e1bf00674c368d335   20
683f3c2314295475e95ab663eeb29f3f    19
753c1f48f6c232c4f66359850dd38e72    11
c37d6a8412f30dcdef54a802fd4e09f7    11
3afb78b6b24a1fb4dd3b7247c2c1fbf0    11
a5b8e9cbe5c11f480b2f73e657731fff    11
2e0540cb61c4bd2ca9201c95e6e7c74c    11
2f93cd614e3a1bc4b2149a26877f4b88    10
a4c1e27ea39ea62ede66f74b891d5976    10
Name: md5, dtype: int64
```

What's the information around all of the occurances of `ad4b0f606e0f8465bc4c4c170b37e1a3`?

What, you mean you don't know what `ad4b0f606e0f8465bc4c4c170b37e1a3` means?

Let's find those images in the DataFrame. We can here see some of the filenames used, its dimensions, and its URL.

```
[ ] images.loc[images["md5"] == "ad4b0f606e0f8465bc4c4c170b37e1a3"]
```

1 to 25 of 50 entries Filter

index	crawl_date	url	filename	extension	mime_type_web_server	mime_type_tika	width	height	
311	20200808191119	https://srv-2020-08-08-19.pixel.parsely.com/plogger/?rand=1596913842241&plid=95640210&idsite=medium.com&url=https%3A%2F%2Fnews.archivesunleashed.org%2F&urlref=&screen=1600x1000%7C1600x1000%7C24&data=%7B%226a%22%3A%22visitor%22%7D&sid=1&surl=https%3A%2F%2Fnews.archivesunleashed.org%2F%3Fgi%3Dc99e6eda4d42&sref=&sts=1596913838633&slts=0&title=Archives+Unleashed&data=Sat+Aug+08+2020+19%3A10%3A42+GMT%2B0000+(Coordinated+Universal+Time)&action=pageview&js=1&pvid=33374489&urpid%3D7e0700dfc569a10021c1ba8157db9a36	NaN	gif	image/gif	image/gif	1	1	ad4b0
1666	20200910033544	https://srv-2020-09-10-03.pixel.parsely.com/plogger/?rand=1599708943991&plid=36597841&idsite=medium.com&url=https%3A%2F%2Fnews.archivesunleashed.org%2Fthe-perils-of-filtering-network-graphs-a8d777a7748a%3Futm_medium%3Duser_profile&urlref=&screen=1600x1000%7C1600x1000%7C24&data=%7B%7D&sid=1&surl=https%3A%2F%2Fnews.archivesunleashed.org%2Ftagged%2Fapache-spark%3Fgi%3Dbde6ff13f3b28&sref=&sts=1599700757651&slts=0&date=Thu+Sep+10+2020+03%3A35%3A43+GMT%2B0000+(Coordinated+Universal+Time)&action=heartbeat&inc=20&tt=20101&pvid=25986169&u=pid%3Da68dca58f75b6dc442788597cb8892df	NaN	gif	image/gif	image/gif	1	1	ad4b0
		https://srv-2020-09-10-03.pixel.parsely.com/plogger/?rand=1599708947180&plid=34192261&idsite=medium.com&url=https%3A%2F%2Fnews.archivesunleashed.org%2Fmacro-measures-of-filtered-graphs-22a998216da1%3Futm_medium%3Duser_profile&urlref=&screen=1600x1000%7C1600x1000%7C24&data=%7B%22viewerStatus%22%3A%22visitor%22%7D&sid=1&surl=https%3A%2F							

BRAIN CLUB

image-information.ipynb

File Edit View Insert Runtime Tools Help

Share

+ Code + Text Copy to Drive

Connect Editing

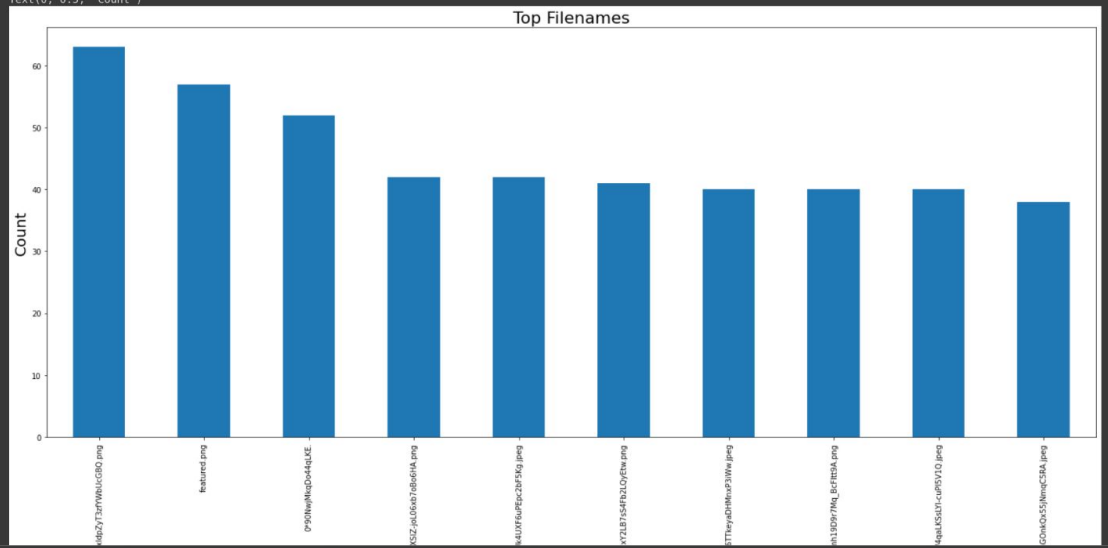
Let's create our first graph!

We'll plot the data first with pandas plot functionality, and then plot the data with Altair.

```
[ ] top_filenames_chart = top_filenames.plot.bar(figsize=(25, 10))

    top_filenames_chart.set_title("Top Filenames", fontsize=22)
    top_filenames_chart.set_xlabel("Filename", fontsize=20)
    top_filenames_chart.set_ylabel("Count", fontsize=20)
```

```
Text(0, 0.5, 'Count')
```



ARCHIVES AND RECORDS
2022, VOL. 43, NO. 3, 316–331
https://doi.org/10.1080/23257962.2022.2100336

Creating order from the mess: web archive derivative datasets and notebooks

Nick Ruest[a], Samantha Fritz[b] and Ian Milligan[b]

[a]York University, Toronto, Canada; [b]University of Waterloo, Waterloo, Canada

ABSTRACT

For a quarter-century, memory institutions have been preserving web-based content. These web archives have been collected and stored in ARC and WARC (W/ARC) file formats and will form a basis for contemporary histories. Yet, these formats present significant challenges to researchers who wish to access and use web archival data. This is primarily due to the nature of collecting, storing, and providing access to these multifaceted digital objects. In other words, web archives are messy. Applying traditional archival methods of description to digital-born collections is complicated due to issues of provenance, original order, and scale. However, we believe that archival description offers a practical starting point for thinking about access. This paper argues a robust finding aid must extend beyond basic collection-level description to allow for more meaningful interactions with web archives. As such, we propose a reimagining of a traditional finding-aid model into a three-level mode of description to include computational methods, the generation of derivative datasets, and interactive code-rich notebooks. These three factors combine to ultimately contribute to the expanded access and use of web archives.

KEYWORDS

Web archives; big data; notebooks; finding aids; data science

Introduction

Since 1996, organizations such as the Internet Archive have been collecting web archives with an eye to making digital heritage accessible. A quarter-century later, attention is turning to the problem of analysis: what can we do to make the petabytes of information — over 100PB in the case of the Internet Archive — usable.

Web archives are messy. Collected using a variety of platforms, web resources are aggregated in standardized file formats: the ARC file, followed by the successor WARC format (proposed as an ISO standard in 2009, and a standard since 2017). Working with these files is difficult for users. Resources are co-mingled, meaning the text of a webpage is interspersed with multi-faceted digital objects and data such as HTML, CSS, JavaScript, binary encoded images, videos, documents, or legacy file formats. Not only do WARCs contain a diverse range of digital objects, but the order also ultimately eludes human readership. Since crawlers often work in parallel, files from dozens or hundreds of websites are woven together. While web archives are traditionally accessed through a Wayback Machine replay instance, this only allows one-page-at-a-time browsing,

CONTACT Nick Ruest ruestn@yorku.ca

Supporting Research

Archives Unleashed Cohorts

AU Cohorts

- Purpose:
  - Facilitate research engagement with computational use of large web/digital archives through use of the ARCH platform in funded scholarly research projects
- Scope:
  - 10 total research teams (5 each year)
- Research support:
  - Funding (~1$0K USD), technical support, bi-weekly calls
  - 2 in person events

AU Cohort Projects

- AWAC2 Analysing Web Archives of the COVID Crisis through the IIPC Novel Coronavirus dataset
- Everything Old is New Again: A Comparative Analysis of Feminist Media Tactics between the 2nd- to 4th Waves
- Mapping and tracking the development of online commenting systems on news websites between 1996–2021
- Crisis Communication in the Niagara Region during the COVID-19 Pandemic
- Viral health misinformation from Geocities to COVID-19

AU Cohort Projects

- Latin American Women's Rights Movements: Tracing Online Presence through Language, Time and Space
- Historicizing Aughts-Era Mormon Mommy Blogging Media Landscapes
- Web Archiving and the Saskatchewan COVID Archive: Expanding Coverage to Capture Social Media, Medical Misinformation, and Radicalization
- Querying Queer Web Archives
- Using Web Archives for Mapping the Use of Cultural Practices in Postconflict Societies and During Reconciliation Processes

Acknowledgements of Institutional Support

ARCH is a joint project between Internet Archive and Archives Unleashed. This work is primarily supported by the Andrew W. Mellon Foundation. Other financial and in-kind support has come from the Social Sciences and Humanities Research Council, Compute Canada, York University Libraries, Start Smart Labs, and the Faculty of Arts and David R. Cheriton School of Computer Science at the University of Waterloo.

Thanks!

Any questions ?

Connect with out project teams:

- arch@archive.org
- archivesunleashed@gmail.com
- jefferson@archive.org
- nick@archivesunleashed.org