Archives Unleashed New York Datathon

Introduction

THE ANDREW W. MELLON FOUNDATION

COLUMBIA UNIVERSITY LIBRARIES
library.columbia.edu

Ivy Plus Libraries Confederation

compute canada | calcul canada

UNIVERSITY OF WATERLOO

START SMART LABS

YORK UNIVERSITÉ UNIVERSITY

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Welcome

- A warm welcome for the Archives Unleashed Team, Columbia University, and Ivy Plus Libraries Confederation!

- Thank you for joining us online for this two day datathon.

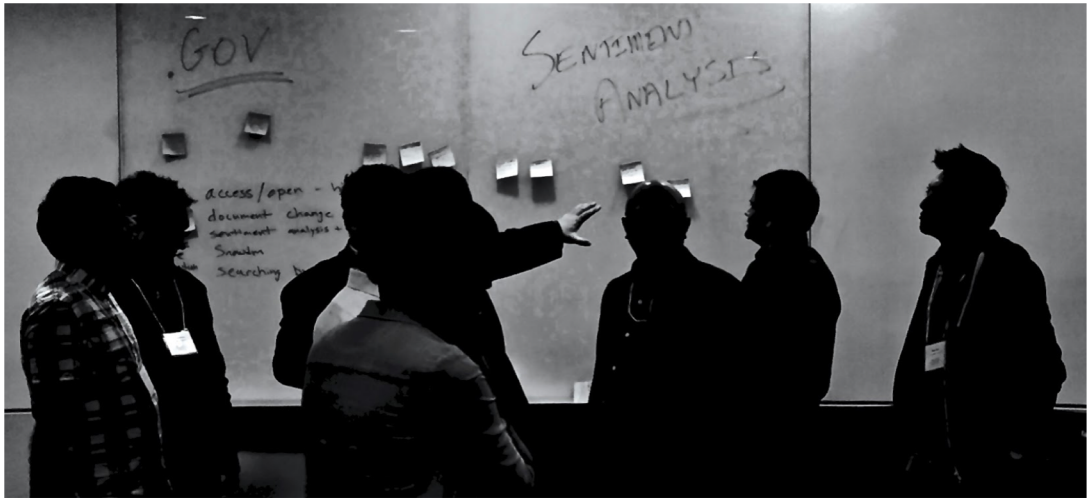
COLUMBIA UNIVERSITY LIBRARIES
library.columbia.edu

Ivy Plus Libraries Confederation

The name of the game is **flexibility**!

Overview

- Team Introductions
- Goals/Objectives
- Web Archiving Context
- Event Logistics
  - Communication
  - Computing Resources
  - Datasets
  - Schedule
- Participant Introductions



Organizing Team

Ian Milligan

Primary Investigator
University of Waterloo

Nick Ruest

Co-Investigator
York University

Jimmy Lin

Co-Investigator
University of Waterloo

Samantha Fritz

Project Manager
University of Waterloo

Pamela Graham

Director of Global Studies; Director, Center for Human Rights Documentation & Research
Columbia University

Alex Thurman

Web Resources Collection Coordinator
Columbia University

Samantha Abrams

Web Resources Collection Librarian
Ivy Plus Libraries Confederation

What do we want to accomplish

- Community building
- Common vision of web archiving development and tool development
- Avoiding the black boxes of search engines we don’t understand
- Equipping us as a collective to work with born-digital cultural resources!

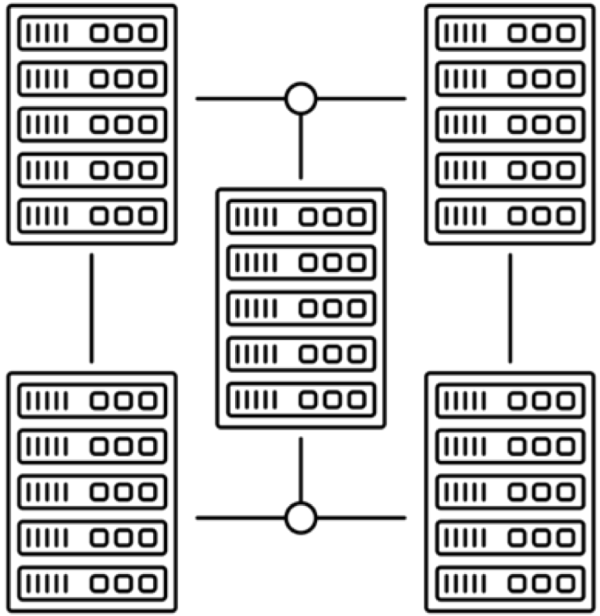
Why should we care about web archives?

Why we should care about web archives?

- How we preserve and disseminate cultural information has dramatically changed;
- Since 1996, and the advent of web archiving at the Internet Archive and national libraries, how we remember has dramatically altered:
  - In scope, speed, scale,
  - In speed
  - In scale
  - And beyond…

Why we should care about web archives?

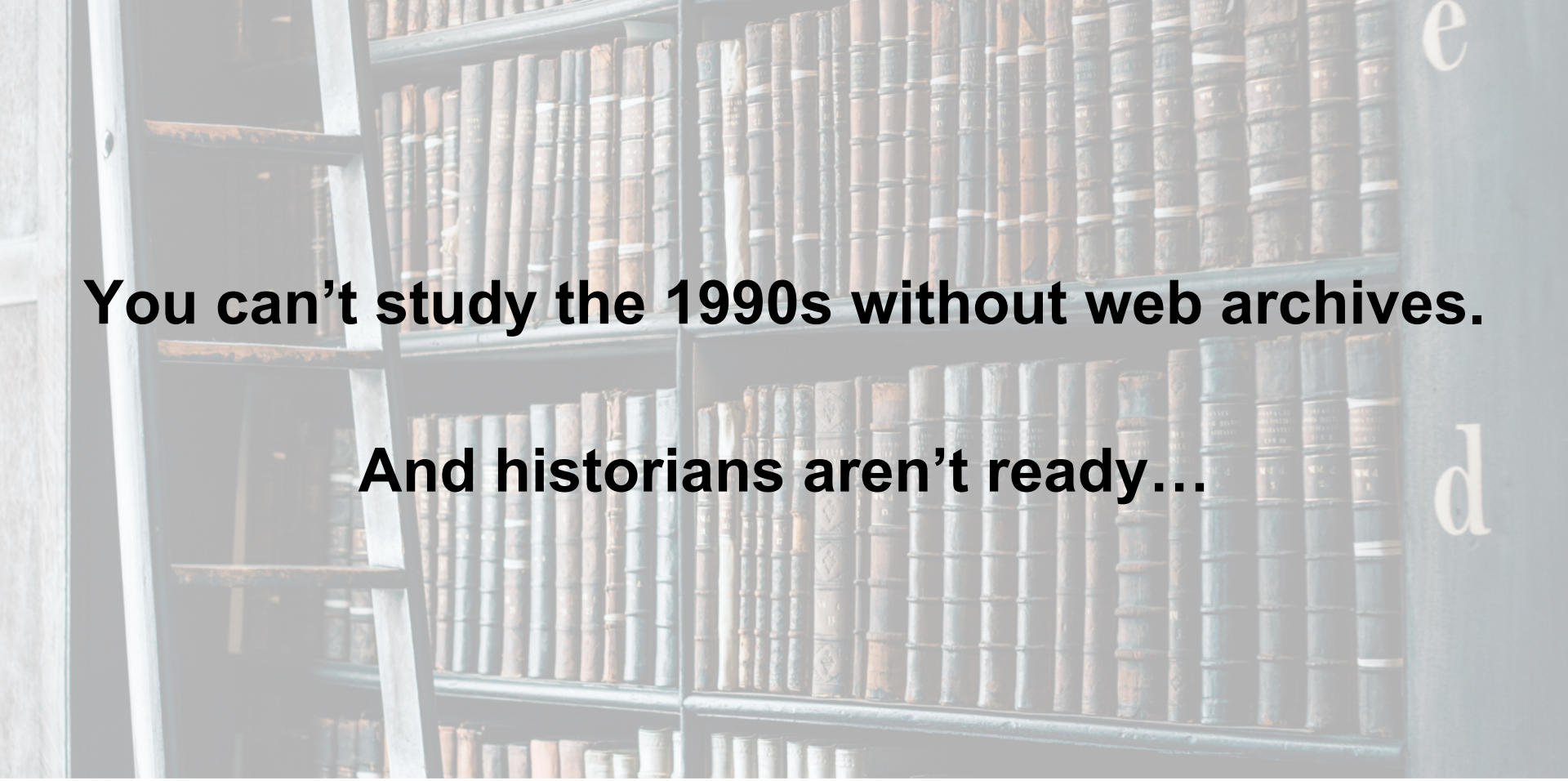
- More data than ever before is being preserved
- It'll be saved delivered to us in very different ways



Why we should care about web archives?

- **All** historians who want to study periods after the 1990s will have to use web archives;
- And the 1990s are now… history!

Web archives have been in existence since 1996.



You can’t study the 1990s without web archives.

And historians aren’t ready…

Why not?

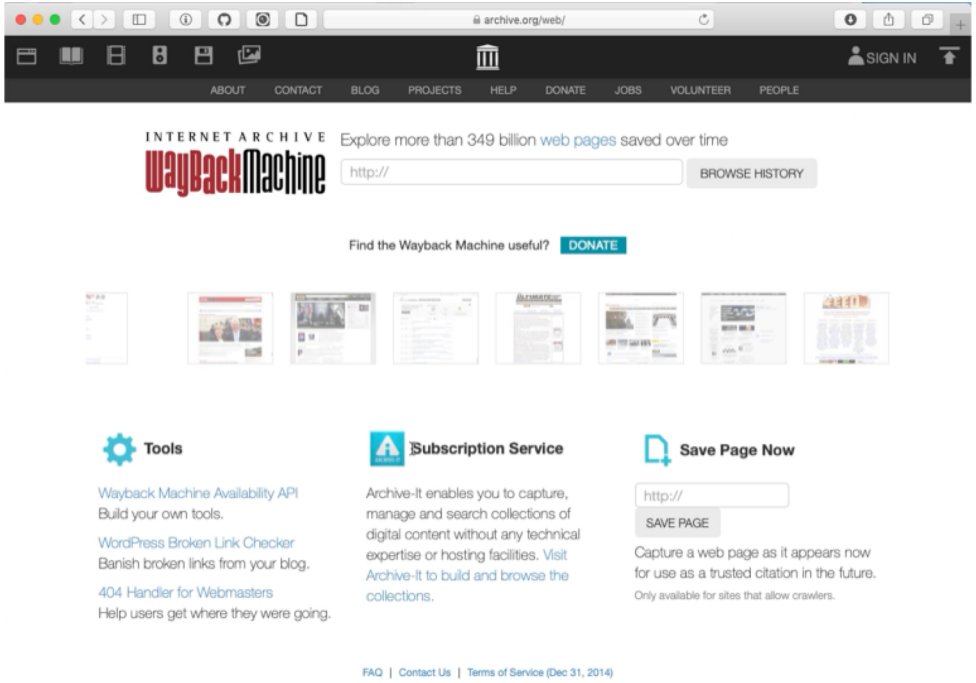
In part, it’s because access has lagged…

So how can we access?

Option 1: Wayback Machine

- Wayback Machine is **great** if you know what you're looking for;
  - Ever-improving keyword search functionality

- But not great for more detailed research queries:
  - You may want to do complicated queries (i.e. websites that say X and link to Y)
  - Exploratory text mining;
  - Image work;
  - etc.



Option 2: The WARC File

- The **W**eb**ARC**hive (ISO 28500:2009) file
- Pictured at right
- Hard to use and a bit idiosyncratic, with a smaller user base, so the first step is to usually transform the data into something that's a bit more common.

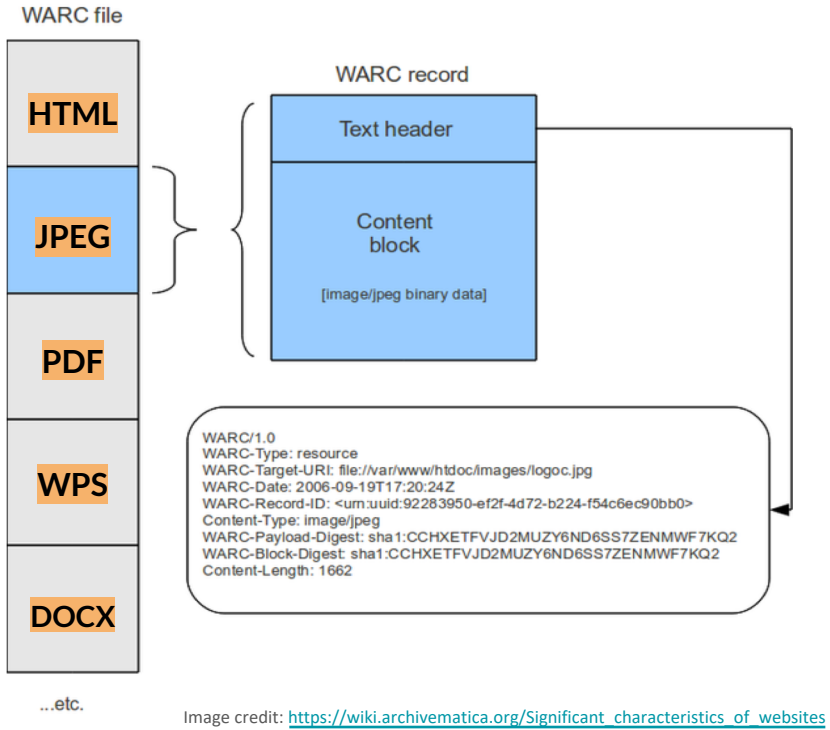
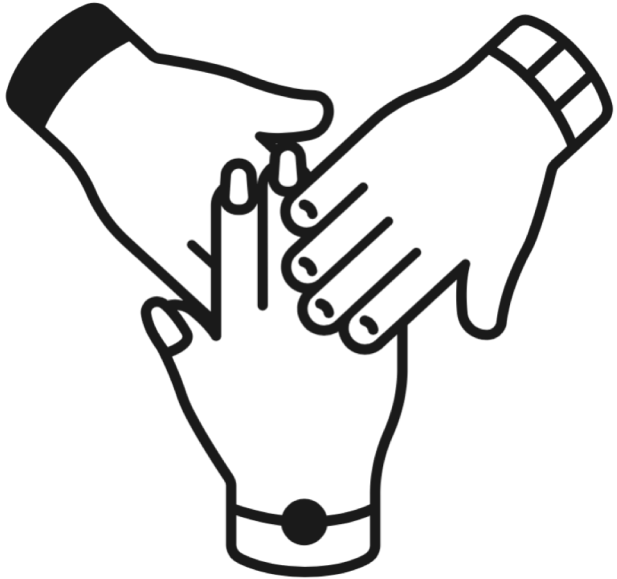


Image credit: https://wiki.archivematica.org/Significant_characteristics_of_websites

So what can we do?

That’s where all of you come in!!



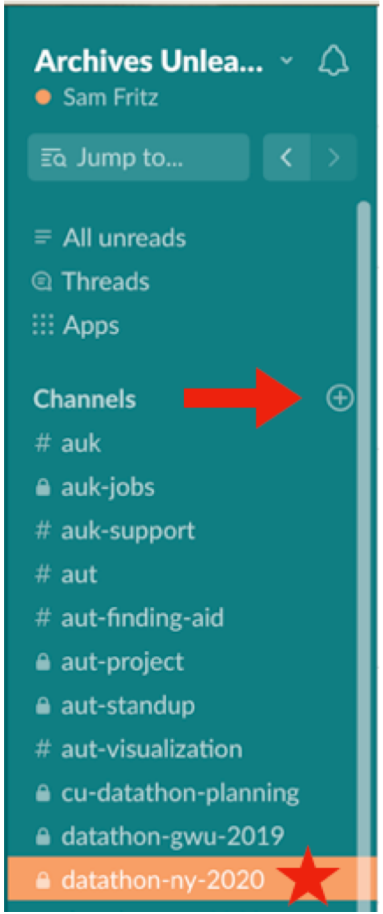
Event Logistics - Communication

Since we are all working remotely, get comfy!

Zoom: broadcasting presentations from AUT team & check-ins.

Slack: communication between AU team & participants.

- You should already be in Slack! If not, email Sam Fritz (sam.fritz@archivesunleashed.org)
- Feel free to create your own channels if collaborating.
- If you need anything, DM me (Ian Milligan) or any other team member.
- We'll do reminders over Slack.



Event Logistics - Communication

And special round of thanks to **Samantha Fritz** who has made all this possible (and, in a better world, had such a phenomenal event planned with our hosts at Columbia – I got to have a sneak peak at the menus…)

Event Logistics - Computing Resources

- **Google Colab Notebooks and Compute Canada VMs** for you to use.
- **VM = Virtual Machine, a sort of powerful cloud-based server that you can use** (8 virtual cores; 30GB RAM machines)
- **Nick and/or Ian have the keys and will provide over Slack**

compute canada | calcul canada

Event Logistics - Datasets

- Many thanks to **Columbia University Libraries** and **Ivy Plus Library Confederation** for sharing their collections for this event.
- Special thanks to **Nick Ruest** for all of the technical set up (notebooks, documentation, datasets, and VMs) and helping to make sure datasets are accessible via Zenodo + Dataverse!

Event Logistics - Schedule

- Mostly asynchronous
- Individuals are welcome to work alone or collaborate
- Check-ins throughout working sessions (via Zoom)
- AUT team members are around for troubleshooting

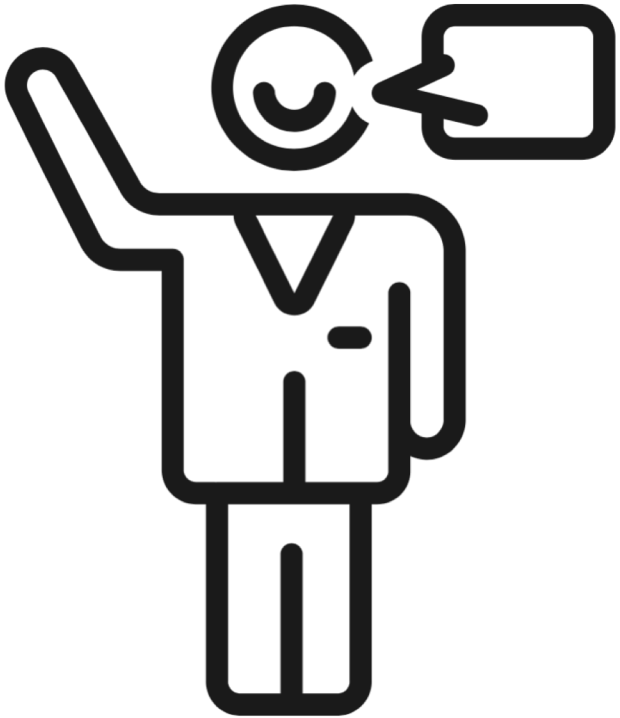
Thursday March 26	
9:15 AM	Participants connect in with Zoom
9:30 AM	Opening, Welcome, Introduction Presentations
10:15 AM	Idea Formation
10:30 AM	Demo of Notebooks + Google Colab Setup
10:45 AM	Let the hacking begin! (via Slack)
4:30 PM	Check-in with participants (via Zoom)
Friday March 27	
9:30 AM	Check-in with participants (via Zoom)
2:00 PM	Check-in with participants (via Zoom)
3:30 PM	Final Presentations (via Zoom)

Group Introductions

Let’s get to know each other! Or at least what we each sound like.

- What’s your name?
- Where are you from
- And in one sentence, what you find interesting about web archives

Planning Team + Participants (alphabetically by first name)



So what tools will we be using?

The Archives Unleashed Project

Our goal is to make petabytes of historical internet content accessible to scholars and others interested in researching the recent past.

Tools Development

- Objectives
  - First, relatively easy-to-use tools;
  - Second, transparent tools that are understandable – no black boxes;
  - Third, tools that can push forward research agendas in history, library/archives, and computer science

Archives Unleashed Toolkit

- An open-source platform for analyzing web archives with Apache Spark;
- Is the underlying code for the Cloud
- Scalable, and can work on:
  - a powerful cluster
  - a single-node server
  - a laptop (on MacOS, Linux, or Windows with a Linux VM)
  - a Raspberry Pi for all your personal web archiving analysis needs ☺

Using the Archives Unleashed Toolkit

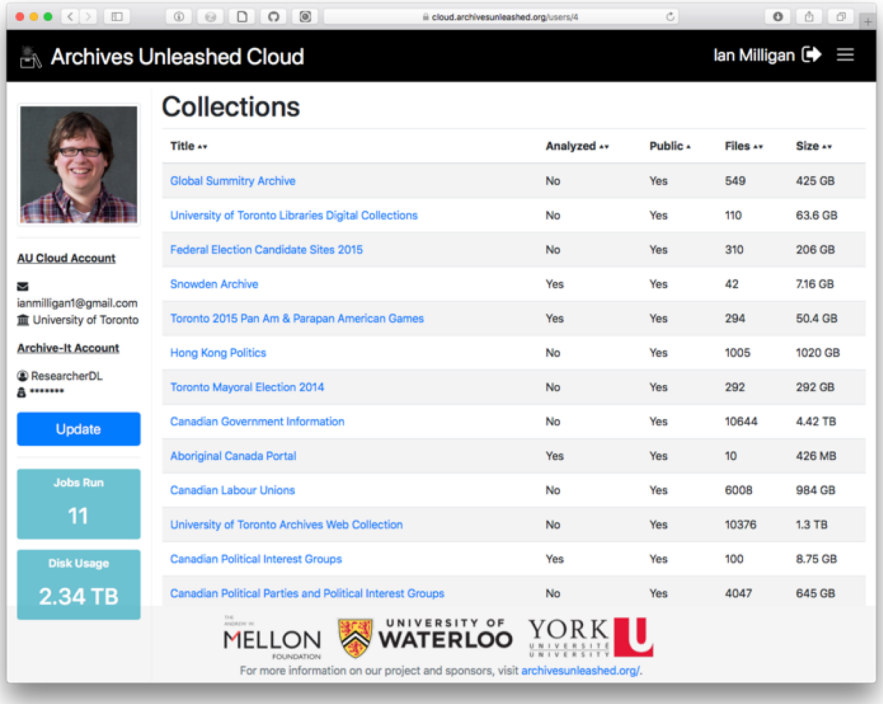
- Dozens of recipes that extract and filter
- Able to handle text, links, images, and binary analysis on various file types
- RDD + DataFrame output available

Filter examples:

- Filter by domain (i.e. all pages in “greenparty.ca”)
- Filter by URL pattern (i.e. all pages in “greenparty.ca/vegetables/*”)
- Filter out boilerplate (i.e. advertisements, navigational elements, templates, etc.)
- Filter by date (i.e. all pages on July 4th, 2015)
- Filter by languages (i.e. only French language pages from greenparty.ca)

Archives Unleashed Cloud

- A GUI-based front end to work with Archives Unleashed Toolkit;
- Runs on our central servers or you can run one yourself;
- Uses WASAPI – Web Archives Systems API – to transfer data
  - Currently only Archive-It supported;
  - We are working on integration with WebRecorder.io and other WASAPI endpoints
- Idea is to generate a basic set of **research derivatives** for scholars to work with so that they can incorporate WARCs into a digital humanities workflow



Archives Unleashed New York Datathon

Datasets & Set Up

THE ANDREW W. MELLON FOUNDATION

COLUMBIA UNIVERSITY LIBRARIES
library.columbia.edu

Ivy Plus Libraries Confederation

compute canada | calcul canada

UNIVERSITY OF WATERLOO

START SMART LABS

YORK UNIVERSITÉ UNIVERSITY

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Datasets - Ivy Plus

- National Statistical Offices and Central Banks Web Archive
- Contemporary Composers Web Archive (CCWA)
- #MeToo and the Women's Rights Movement in China Web Archive
- Geologic Field Trip Guidebooks Web Archive
- Literary Authors from Europe and Eurasia Web Archive
- Web Archive of Independent News Sites on Turkish Affairs
- State Elections Web Archive
- Brazilian Presidential Transition (2018) Web Archive
- Collaborative Architecture, Urbanism, and Sustainability Web Archive (CAUSEWAY)
- Global Webcomics Web Archive
- Queer Japan Web Archive
- Extreme Right Movements in Europe
- Latin American and Caribbean Contemporary Art Web Archive
- Popline and K4Health Web Archive
- Eastern Europe and Former Soviet Union Web Archive
- Independent Documentary Filmmakers from China, Hong Kong, and Taiwan Web Archive

Datasets - Columbia University

- General
- Resistance
- Stonewall 50 Commemoration
- Freely Accessible eJournals
- Avery Library Historic Preservation and Urban Planning
- Rare Book and Manuscript Library
- Burke Library New York City Religions

Archives Unleashed New York Datathon

Team Formation Activity

THE ANDREW W. MELLON FOUNDATION

COLUMBIA UNIVERSITY LIBRARIES
library.columbia.edu

Ivy Plus Libraries Confederation

compute canada | calcul canada

UNIVERSITY OF WATERLOO

START SMART LABS

YORK UNIVERSITÉ UNIVERSITY

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Teams vs. Individuals

Some of you may have kids at home (hello from some of us!), or other constraints. **The world is an interesting place today**.

In general, **teams** are great - you can bounce ideas off each other and connect. But you may also want to work as an **individual** if circumstances mean you cannot collaborate.

For teams we encourage:

- Be **considerate** of the constraints we are under;
- If people have to disappear to take care of kids, dogs, health, etc.: life comes first;
- A **mix** of synchronous and asynchronous work can be fun;
- Use Slack voice calls when you want, text otherwise.

Teams

If you do elect to join a team, we in general suggest a team of **three to five people**.

We can be flexible, of course. For example, if some of you have kids or other constraints, joining a large team might let you take on a smaller part of a bigger project.

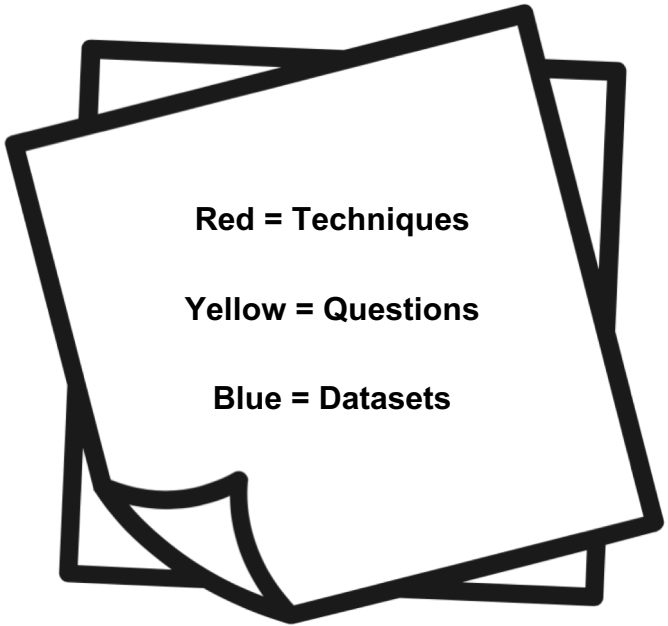
Flexibility is the name of the game!

Team Formation

Sticky Notes Activity - *to promote spontaneous idea formation and organization.*

Team Formation Spreadsheet

- Open Google sheet
- There are three categories we will organize around
- Write in each idea in a separate cell under the corresponding column/theme
- Add you name to the end of your note



Archives Unleashed New York Datathon

Notebooks + Gooble Colab

THE ANDREW W. MELLON FOUNDATION

COLUMBIA UNIVERSITY LIBRARIES
library.columbia.edu

Ivy Plus Libraries Confederation

compute canada | calcul canada

UNIVERSITY OF WATERLOO

START SMART LABS

YORK UNIVERSITÉ UNIVERSITY

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Notebooks

- Function: Analysis
- Open Source
- New Toolkit derivatives
- Web archive collections as data
- Tighter data science tool integration
  - Python 3.6+
  - pandas
- Moving closer to the user

Open in Colab

Working with Archives Unleashed Parquet Derivatives

In this notebook, we'll setup an environment, then download a dataset of web archive collection derivatives that were produced with the Archives Unleashed Toolkit. These derivatives are in the Apache Parquet format, which is a columnar storage format. These derivatives are generally small enough to work with on your local machine, and can be easily converted to Pandas DataFrames as demonstrated below.

This notebook is useful for exploring the following derivatives.

Binary Analysis

- Audio
- Images
- PDFs
- Presentation program files
- Spreadsheets
- Text files
- Word processor files

Web Pages

```
.webpages().select($"crawl_date", $"url", $"mime_type_web_server", $"mime_type_tika",
RemoveHTMLDF(RemoveHTTPHeaderDF(($"content"))).alias("content"))
```

Produces a DataFrame with the following columns:

- crawl_date
- url
- mime_type_web_server
- mime_type_tika
- content

As the webpages derivative is especially rich - it contains the full text of all webpages - we have a separate notebook for text analysis here.

Web Graph

```
.webgraph()
```

Produces a DataFrame with the following columns:

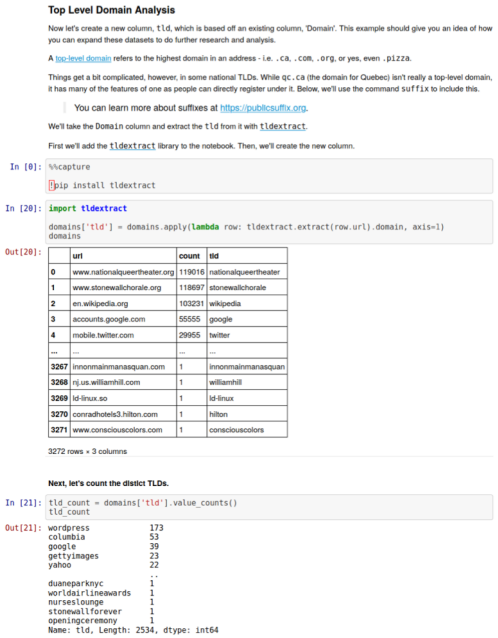
- crawl_date
- src
- dest
- anchor

Image Links

```
.imageLinks()
```

Produces a DataFrame with the following columns:

- src
- image_url



Top Level Domain Analysis

Now let's create a new column, `tld`, which is based off an existing column, 'Domain'. This example should give you an idea of how you can expand these datasets to do further research and analysis.

A top-level domain refers to the highest domain in an address - i.e. `.ca`, `.com`, `.org`, or yes, even `.pizza`.

Things get a bit complicated, however, in some national TLDs. While `qc.ca` (the domain for Quebec) isn't really a top-level domain, it has many of the features of one as people can directly register under it. Below, we'll use the command `suffix` to include this.

> You can learn more about suffixes at https://publicsuffix.org.

We'll take the `Domain` column and extract the `tld` from it with `tldextract`.

First we'll add the `tldextract` library to the notebook. Then, we'll create the new column.

In [0]:
```
%%capture
!pip install tldextract
```

In [20]:
```
import tldextract

domains['tld'] = domains.apply(lambda row: tldextract.extract(row.url).domain, axis=1)
domains
```

Out[20]:

	url	count	tld
0	www.nationalqueertheater.org	119016	nationalqueertheater
1	www.stonewallchorale.org	118697	stonewallchorale
2	en.wikipedia.org	103231	wikipedia
3	accounts.google.com	55555	google
4	mobile.twitter.com	29955	twitter
...	...	...	...
3267	innonmainmanasquan.com	1	innonmainmanasquan
3268	nj.us.williamhill.com	1	williamhill
3269	ld-linux.so	1	ld-linux
3270	conradhotels3.hilton.com	1	hilton
3271	www.conscioucolors.com	1	conscioucolors

3272 rows × 3 columns

Next, let's count the distict TLDs.

In [21]:
```
tld_count = domains['tld'].value_counts()
tld_count
```

Out[21]:
```
wordpress              173
columbia                53
google                  39
gettyimages             23
yahoo                   22
                      ...
duaneparknyc             1
worldairlineawards       1
nurseslounge             1
stonewallforever         1
openingceremony          1
Name: tld, Length: 2534, dtype: int64
```

Live demonstration

Hello Datathon Teams!

You will need:

- Github Account
- Google Account

Getting Started:

- ❑ Finalize your team members
- ❑ Choose a team name
- ❑ Set up a **public** Slack channel
  - ❑ E.g. name: ny2020-teamname *[**NOTE**: 21 character limit]*
- ❑ Set up your Colab Notebook (Nick will demo; a quick guide is also available on our YouTube channel https://youtu.be/BykdMm0BhUU
- ❑ Start Hacking
  - ❑ GitHub Notebooks read.me

TIPS

- ★ You can make **free** calls via Slack + screen share
- ★ When working with others, share Colab notebooks as a **view only** to to reduce edit overrides
- ★ AU team members are more familiar with **Mac/Linux** environment
- ★ Checkout projects from previous datathons:
  - Toronto | Vancouver | Washington

Final Presentations: Friday March 27th @3:30pm EST

- ❑ Share your Google Slides with us so we can queue things up
  - ❑ Naming Convention: AU-NY2020-TeamName
- ❑ Approx. 5 min to present
- ❑ **Content:** focus on sharing your project/work with the group; question(s) and dataset explored, methods and tools used, analysis results, & lessons learned.

Hello Datathon Teams!

Colab Notebooks Set Up (REVIEW)

You have access to **2** notebook templates:

Parquet_text_analyis_popline.ipynb | parquet_pandas_stonewall.ipynb

1. Click “Open in Colab”
2. Click “Copy to Drive”
3. Open the newly copied notebook (you may have to allow pop-ups)

In the copy you just created:

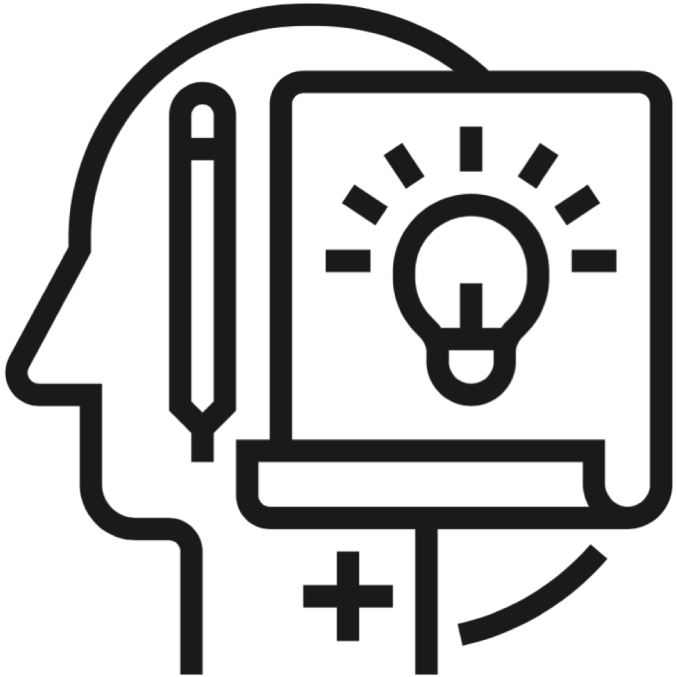
1. **Rename notebook**: add your name at the beginning so your teammates can identify quickly with multiple tabs open.
2. **Click “Share”** > “get sharable link” → make sure setting is selected to “**view only**”
3. **Share link** in your team slack channel
4. You will see a folder in your drive called Colab Notebooks, and inside will be the copied notebook that was renamed

Quick Guide to Setting up Colab Notebooks:https://www.youtube.com/watch?v=JDXQRUp_Tx4

Hello Datathon Teams!

Feel free to use the questions posed below as a starting point for developing the specifics of your project:

- Are you looking at a specific dataset? Does your research question lend itself to a particular dataset?
- What types of data do you want to extract and explore?
- What types of tools do you, (and/or your teammates), have experience with?
- Do you have an end goal in mind with the dataset you're exploring?
- Are the analysis methods chosen and/or depth we want to explore feasible in the time given?



Thanks again to our sponsors!

THE ANDREW W. MELLON FOUNDATION

COLUMBIA UNIVERSITY LIBRARIES
library.columbia.edu

Ivy Plus Libraries Confederation

compute canada | calcul canada

UNIVERSITY OF WATERLOO

START SMART LABS

YORK UNIVERSITÉ UNIVERSITY

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Let the hacking begin!

Archives Unleashed New York Datathon

Event Wrap Up & Final Presentations

THE ANDREW W. MELLON FOUNDATION

COLUMBIA UNIVERSITY LIBRARIES
library.columbia.edu

Ivy Plus Libraries Confederation

compute canada | calcul canada

UNIVERSITY OF WATERLOO

START SMART LABS

YORK UNIVERSITÉ UNIVERSITY

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Final Presentations

- **Latin American and Caribbean Contemporary Art Web Archive**
  Samantha Abrams, Sumitra Duncan, Mary Nakija, Jim Kammerer
- **Contemporary Composers Web Archives**
  Pamela Graham, Giulia Occhini, Nicole Greenhouse
- **Global Web Comics Web Archives**
  Kae Bara Kratcha, Kritika Garg, Wei Yin, Francis Kayiwa
- **Stonewall**
  Sarah McTavish, Alex Thurman, Dan Royles, Brian M. Watson

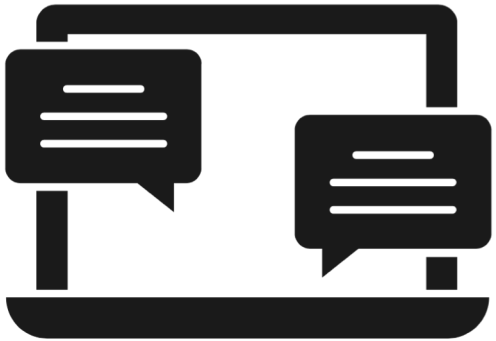
So Long, Farewell

slack

zoom

Final Thoughts

- ★ Thank you for your contributions, collaboration, and creativity!
- ★ This was our first online datathon, and while we had a few hiccups (curse you Google!), we hope you've enjoyed trying out new ways of working with web archives!
- ★ Very, ***very*** impressive work (compared to the dark fears we had about whether Slack + Zoom + Google Colab could do this!)



Many Thanks!

NY Datathon Planning Team







Sponsors

THE ANDREW W. MELLON FOUNDATION

COLUMBIA UNIVERSITY LIBRARIES library.columbia.edu

Ivy Plus Libraries Confederation

compute canada | calcul canada

UNIVERSITY OF WATERLOO

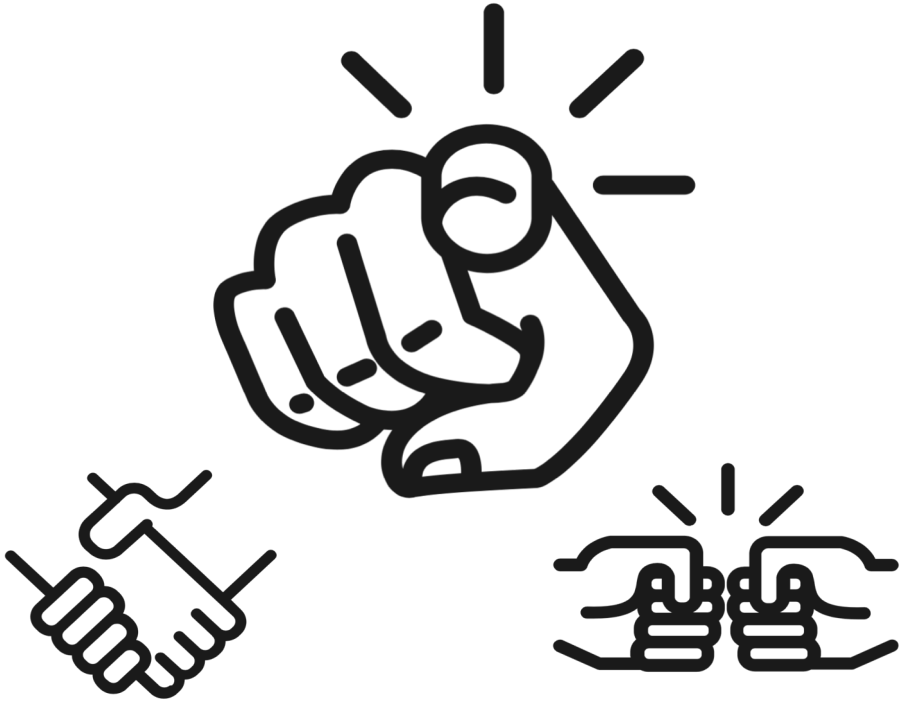
YORK UNIVERSITÉ UNIVERSITY

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Many Thanks and Round of Applause to **YOU**, our Participants!!



Special Thanks!

★ To Ian and Nick for all of the technical support and troubleshooting over the past 2 days!

“I guarantee you that you’ll be a pro with working with web archives by Friday afternoon!”

– *Professor Milligan’s Famous Last Words*

Post-Datathon

Keep the Energy Going

- Let's keep the conversation going however we can!
- **Stay on the Slack**! Invite your friends/colleagues.
- If you're an Archive-It subscriber, check out the **Cloud** (https://cloud.archivesunleashed.org)
- You can follow the Archives Unleashed project by subscribing to our newsletter: http://archivesunleashed.org/subscribe/#

Group Presentations

- We'd like to share and promote the work teams have completed!
- It's a great opportunity to share with the community examples of what's possible when working with WARCS.
- If you would prefer not to share, just let us know. Otherwise, we will post a PDF of your deck to our event page.

Post-datathon survey

- Our project thrives on feedback, and we'd love to know about your experience during this event
- **https://forms.gle/A9kJ6U5dJb1f3Ue8A**

Resources available

- **NY Datathon Github page**
- **Archives Unleashed Cloud Learning Guides -** great examples of how to use derivatives with other tools
- **YouTube Videos**
  - AU Datathon - Gephi Walkthrough (by Ian)
  - AU Quick Guide: Setting up Colab Notebooks
- **Connect w/ Archives Unleashed**
  - **Website:** https://archivesunleashed.org
  - **Github:** https://github.com/archivesunleashed
  - **Slack:** http://slack.archivesunleashed.org/
  - **Twitter:** https://twitter.com/unleasharchives
  - **Newsletter:** Archives Unleashed Newsletter Subscription

:*(

BUT LIKE ANY GAME.. it comes to an end, but you can play it again!

We are not sure about any future datathons, unfortunately, but we are planning to hold **future Archives Unleashed activities**. Please stay tuned (we'll announce via Slack, our newsletter, etc.).

Stay safe and healthy and we look forward to seeing you all in person at some point in our wonderful web archiving community!

Archives Unleashed New York Datathon

THE ANDREW W. MELLON FOUNDATION

COLUMBIA UNIVERSITY LIBRARIES
library.columbia.edu

Ivy Plus Libraries Confederation

compute canada | calcul canada

UNIVERSITY OF WATERLOO

START SMART LABS

YORK UNIVERSITÉ UNIVERSITY

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada