



Web Archives: A Doorway to Access and Usability

Samantha Fritz, *MLIS*

Project Manager

Archives Unleashed

sam.fritz@archivesunleashed.org

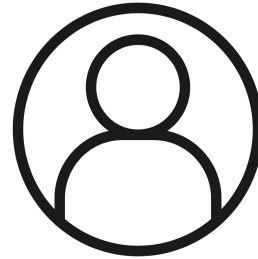
Emergence of the Web

- 1991 World Wide Web becomes publicly available
- The “fastest growing communications medium of all time” (British Council)
- Shaped our global climate of how we **connect with one another** and **interact with information**

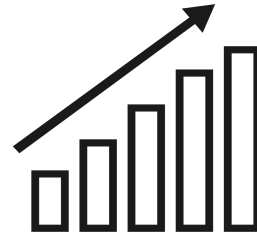


Information Age

- Influence the shape of our modern world and historical record
- Drastic shift in the way we **produce, interact** and **preserve** information



4.5 Billion
Internet Users



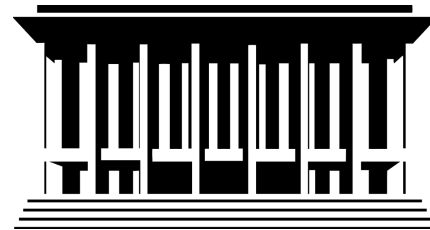
1.5 Billion
Current Webpages

Information Age

- Influence the shape of our modern world and historical record
- Drastic shift in the way we **produce, interact** and **preserve** information.
- The first large scale preservation project initiated by Brewster Khale (Internet Archive), and national libraries in Sweden & Australia in 1996



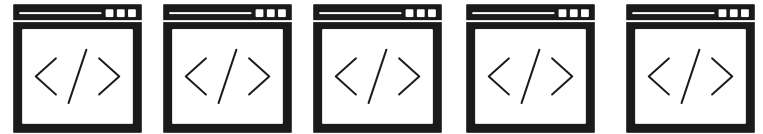
National Library
of Sweden



NATIONAL
LIBRARY
OF AUSTRALIA

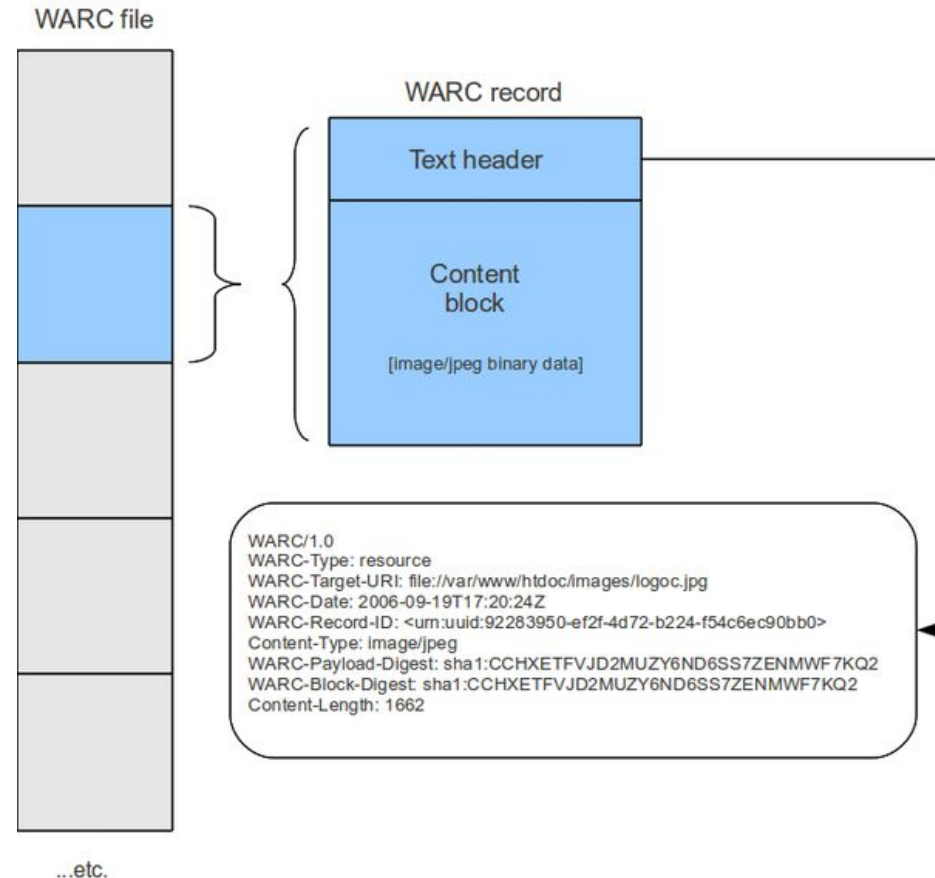
What is Web Archiving?

- **Web Archiving:** “any form of deliberate and purposeful collecting and preserving of web material” (Brugger, Niels)
- The web continues to grow, but it is also disappearing
- We are at risk of losing potentially significant information - there is no backup to the internet!



Using Web Archives

- Web archives are **essential** to studying post-1990s topics
- An **access point** for exploring our modern historical record
 - **Scope:** wider/diverse range of voices and perspectives.
 - **Scale:** shift from scarcity to abundance (Milligan, Ian)



Challenges


- **Abundance** of data is a challenge
- Overwhelming sense to cope with **data at scale**
- Access remains a significant **barrier** in the use of Web archives
- Computational access at scale requires an understanding of high-performance computing and the command line



How do we lower this barrier to
access and **use** of web archives?



Context

- 
- Definitions of Access, Accessibility, and Usability
 - General/large picture understanding
 - You cannot look at access without also acknowledging usability

Access

“freedom or ability to obtain or make use of something”

Accessibility

“capable of being reached; used or seen; of being understood...”

Usability

“the quality or state of being usable : ease of use”

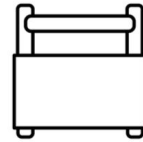
Archives Unleashed Project

- Archives Unleashed Team:
 - PI - Ian Milligan (*historian*)
 - Co-PI - Nick Ruest (*librarian*)
 - Co-PI - Jimmy Lin (*CS*)
- Tools developed to address scalable analysis of web archives
- Research needs of scholars in humanities and social sciences
- Spirit of Access + Usability to inform all project aspects



Archives Unleashed Project

Looking for a way to explore web archives through a user-friendly suite of tools?



AU Toolkit



AU Cloud

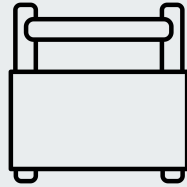


Warlight



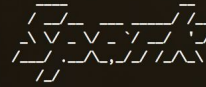
Notebooks

Code Base + Development



- Archives Unleashed Toolkit
 - Core code base
 - Flexible analysis solutions
 - Transparency
 - Publically available, free and open-source
 - Built on widely accepted practices

Welcome to



version 2.4.3

Using Scala version 2.11.12 (OpenJDK 64-Bit Server VM, Java 1.8.0_212)

Type in expressions to have them evaluated.

Type :help for more information.

```
scala> :paste
```

```
// Entering paste mode (ctrl-D to finish)
```

```
import io.archivesunleashed._
```

```
import io.archivesunleashed.matchbox._
```

```
val links = RecordLoader.loadArchives("/aut-resources/Sample-Data/*.gz", sc)  
  .keepValidPages()  
  .flatMap(r => ExtractImageLinks(r.getUrl, r.getContentString))  
  .countItems()  
  .take(20)
```

```
// Exiting paste mode, now interpreting.
```

```
import io.archivesunleashed._
```

```
import io.archivesunleashed.matchbox._
```

```
links: Array[(String, Int)] = Array((http://www.liberal.ca/shared/images/logo_footer.png,  
.gif,1966), (http://www.gca.ca/indexcms/img/Leer.gif,1780), (http://www.liberal.ca/images  
ca/images/section-headers/get-involved.png,1114), (http://www.plaxo.com/images/abc/button  
on-headers/newsroom.png,854), (http://www.davidsuzuki.org/files/dent.gif,764), (http://i.  
vote.ca/sites/fairvote.ca/themes/fvc_ruby/logo.png,465), (http://www.ndp.ca/sites/all/the
```

```
scala> :paste
```

```
// Entering paste mode (ctrl-D to finish)
```

```
import io.archivesunleashed._
```

```
import io.archivesunleashed.df._
```

```
val df = RecordLoader.loadArchives("/aut-resources/Sample-Data/*.gz", sc)  
  .extractValidPagesDF()
```

```
df.select(ExtractBaseDomain($"Url").as("Domain"))  
  .groupBy("Domain").count().orderBy(desc("count")).show()
```

```
// Exiting paste mode, now interpreting.
```

```
+-----+-----+  
|                Domain|count|  
+-----+-----+  
| www.equalvoice.ca| 4644|  
| www.liberal.ca   | 1968|  
| greenparty.ca    |  732|  
|www.policyalterna...|  601|  
| www.fairvote.ca  |  465|
```

Code Base + Development



- Archives Unleashed Cloud
 - Cloud-based
 - Links to Archive-It based WARC collections
 - Familiar interface type (GUI)
 - Click-to-results tasks

Archives Unleashed Cloud

Collections

Title	Status	Date Analyzed	Public	Files	Size
Hong Kong Politics			Yes	1106	1.04 TB
University of Toronto Archives Web Collection			Yes	10624	1.35 TB
University of Toronto Libraries Digital Collections			Yes	125	73.2 GB
Canadian Labour Unions			Yes	7757	1.03 TB
Canadian Political Interest Groups			Yes	100	8.75 GB
Canadian Political Parties and Political Interest Groups			Yes	6127	691 GB
Federal Election Candidate Sites 2015			Yes	310	206 GB
Global Summitry Archive			Yes	660	494 GB
Canadian Government Information			Yes	14358	4.66 TB
Aboriginal Canada Portal	Completed	July 11, 2019	Yes	10	426 MB
Ontario Provincial Election 2018	Completed	July 11, 2019	Yes	939	113 GB
Snowden Archive	Completed	July 11, 2019	Yes	42	7.16 GB
Toronto Municipal Election 2018	Completed	July 11, 2019	Yes	1106	34.3 GB
Toronto Mayoral Election 2014	Completed	July 27, 2019	Yes	292	292 GB

Jobs Run: 19
Disk Usage: 1.61 TB

MELLON UNIVERSITY OF WATERLOO YORK U
For more information on our project and sponsors, visit archivesunleashed.org/.
[About](#) | [Privacy Policy](#) | [Documentation](#) | [FAQ](#)

Archives Unleashed Cloud

Analyze Collection

Hyperlink Diagram

Domains


Domain	Count
archives.gov	10000
www.100.gov	8000
www.100.gov	6000
www.100.gov	4000
www.100.gov	3000
www.100.gov	2000
www.100.gov	1500
www.100.gov	1000
www.100.gov	800
www.100.gov	600
www.100.gov	400
www.100.gov	200
www.100.gov	100

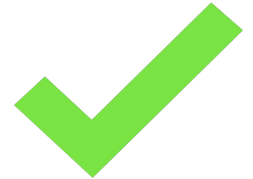
Download Collection Derivatives

High 3.56 MB | Raw Network 1.69 GB | Domains 40.9 KB | Full Text 921 MB | View By Domains 136 MB

MELLON UNIVERSITY OF WATERLOO YORK U
Learn more about these files here. We also have prototype Archives Unleashed Cloud Jupyter Notebooks available.
For more information on our project and sponsors, visit archivesunleashed.org/.
[About](#) | [Privacy Policy](#) | [Documentation](#) | [FAQ](#)

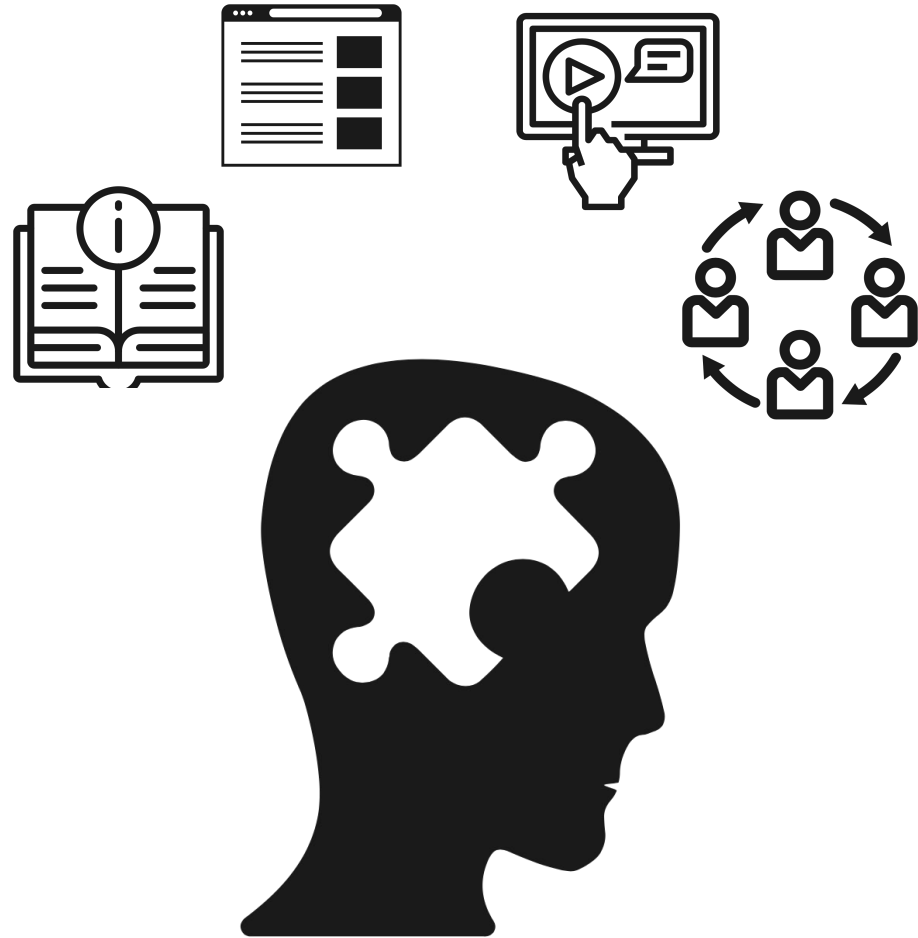
Learning Resources

-  Goals
 - Support
 - Encourage
 - Empower
 - Instil confidence
- Documentation = Usability



Learning Resources

- Goals
 - Support
 - Encourage
 - Empower
 - Instil confidence
- Documentation = Usability
- AU Documentation
 - Text based Instructions
 - Learning Resources
 - Videos
 - In-Person Training (Datathons)



Takeaways

- We actively participate in a content curation and preservation
- Look for opportunities to integrate access & usability in project cycles
- Value is derived from usability, but usability only exists if it is accessible
- Cooperative partnership between access and usability





Let's Connect

Samantha Fritz, *MLIS*

Project Manager

Archives Unleashed

sam.fritz@archivesunleashed.org

Archives Unleashed Project

<https://archivesunleashed.org>

@UnleashArchives

[Github](#) | [Slack](#)