Archives Unleashed Cohort Program

Research with Web Archives and Digital Collections

https://bit.ly/AUCohortProjects

Cohort Program

Overview

The **Archives Unleashed Cohort Program** was launched in 2021, following the successes of the Archives Unleashed datathon series, to foster research engagement with web archives.

The program has seen two cycles of year-long intensive collaborations to support research with web archives as a primary data source and scholarly objects.

Ten research teams were provided technical and academic mentorship to steward projects through direct one-on-one consultation from Archives Unleashed, connections to field experts, and opportunities for peer-to-peer support.

The program also supported teams in publishing scholarly research and in highlighting use cases of working with web archival data.

Cohort Program Overview

50 Researchers

20 Institutions

9 Countries

Cohort Projects (2021-2023)

Cohort 1 Projects: https://archivesunleashed.org/cohorts2021-2022/

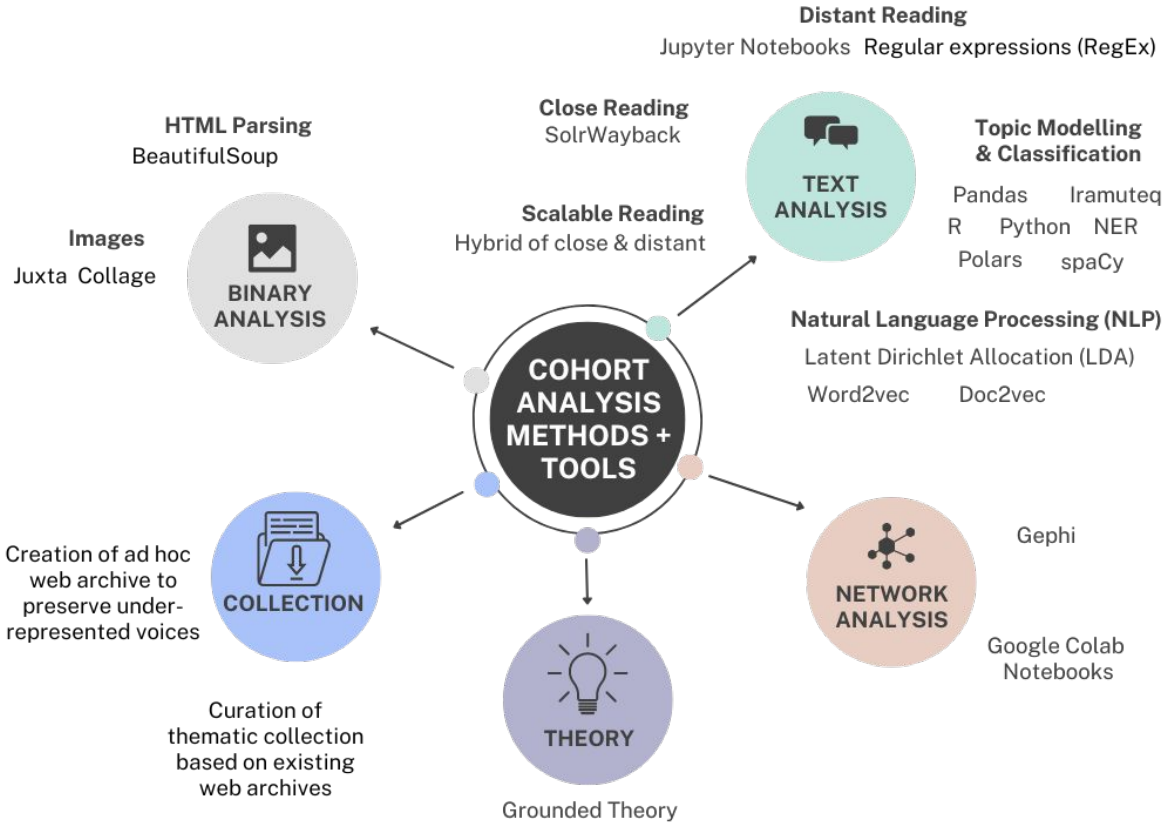
- AWAC2 Analysing Web Archives of the COVID Crisis through the IIPC Novel Coronavirus dataset
- Crisis Communication in the Niagara Region during the COVID-19 Pandemic
- Mapping and tracking the development of online commenting systems on news websites between 1996–2021
- Everything Old is New Again: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

Cohort 2 Projects: https://archivesunleashed.org/cohorts2022-2023/

- 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

Cohort Project Themes

Cohort Project Methods, Techniques, and Tools



AWAC2 Analysing Web Archives of the COVID Crisis through the IIPC Novel Coronavirus dataset

Valérie Schafer [1] Karin De Wild [2] Frédéric Clavert [1] Niels Brügger [3] Susan Aasman [4] Sophie Gebeil [5]Joshgun Sirajzade [1]

Project Description

Investigating transnational events through web archive collections, the AWAC2 team will focus on a distant reading of the IIPC COVID-19 web archival collection to understand actors, content types and interconnectivity throughout it.

Subject/Field of Study: COVID19, Pandemic, IIPC collection, multilingualism, topic modeling, distant reading, women and COVID19

Collections

- ➔ The IIPC Novel Coronavirus collection, made available on Archive-It (https://archive-it.org/collections/13529). It was created by the Content Development Group (CDG) of IIPC, thanks to the contribution of over 30 IIPC members and public nominations from over a hundred individuals or institutions.

Research Questions

- ➔ To deepen the approach of transnational events through web archives we had started through the WARCnet project. We were especially interested in a distant reading helping us in analyzing web archives representativeness (URLs in-/outside ccTLD? comparison of national collections and their selection for IIPC, categories of actors/websites, MIME types, etc.)
- ➔ A second set of research questions, that were less data-driven, were oriented towards Women, Gender and COVID within this IIPC collection (e.g., domestic violence, care and homeschooling, etc.) and the way it could be studied and analysed through the IIPC collection (creation of sub-corpora, topic modeling, etc).

[1] University of Luxembourg [2] Leiden University [3] Aarhus University [4] University of Groningen [5] University of Aix-Marseille

AWAC2 Analysing Web Archives of the COVID Crisis through the IIPC Novel Coronavirus dataset

Valérie Schafer [1] Karin De Wild [2] Frédéric Clavert [1] Niels Brügger [3] Susan Aasman [4] Sophie Gebeil [5] Joshgun Sirajzade [1]

Methods & Tools

- ➔ Scalable reading (hybridising close and distant reading)
- ➔ Topic modeling
- ➔ Panda
- ➔ Iramuteq
- ➔ Latent Dirichlet Allocation (LDA), Word2vec and Doc2vec

Research Insights

- We were underestimating the noises in web archives, and notably for news websites.
- Multilingualism is another key challenge and notably for topic modeling.
- We also realised how much contextual information is needed for the kind of transversal study we wanted to conduct (i.e., advanced knowledge of each webosphere and of each national context, in addition to computational skills, knowledge related to gender, etc.)
- One year is short for such a study. It allowed us to perform some analysis and conduct some case studies, but there is much more to explore and to achieve with such a dataset.

Outputs and Scholarship

- ➔ App. 10 presentations in seminars, international conferences, etc., including the WARCnet final conference and the IIPC conferences (2022 and 2023).
- ➔ 2 blogposts on netpreserve.org, the blog of the IIPC
- ➔ 2 articles mentioning the project are already published, a third one is under review. The three of them are taking the AWAC2 project as a case study within a broader approach of web archives' challenges. One chapter to be published in 2024, which is entirely related to the project and more precisely to the subproject on Women, Gender and COVID 19.

[1] University of Luxembourg [2] Leiden University [3] Aarhus University [4] University of Groningen [5] University of Aix-Marseille

AWAC2 Analysing Web Archives of the COVID Crisis through the IIPC Novel Coronavirus dataset

Valérie Schafer [1] Karin De Wild [2] Frédéric Clavert [1] Niels Brügger [3] Susan Aasman [4] Sophie Gebeil [5] Joshgun Sirajzade [1]

Reflections

Challenges/Obstacles + Solutions

The first main challenge was the size of the dataset.
The second one was its multilingualism.
We also faced computational issues and added a computer scientist to the team to go further in our analysis.
For the topic related to Women, Gender and COVID, we faced duplicates, noises in the results. We created sub corpora in French and in English but had again to reframe results and to mix distant reading with close reading, to better approach the results (i.e. occurrences of COVID and gender that are close in news website, as the metoo movement and COVID crisis were at the same time on top of the news , but without a link between both news)

Using ARCH / Cohort Program Experience

- *ARCH is a pedagogical interface and it allows to easily get an overview of the dataset and download metadata and content. The plain text is very useful. We faced some issues at the very beginning as it took quite a time to download this huge dataset (especially for plein text) and we could not proceed within the team at the same time. There were some trial and error but we benefited from updates, explanations and support from the AU team.*
- *The regular meeting with AU team were useful, supportive and motivating. We benefited from technical support but also from scientific input. Meetings with the other teams were also appreciated. The level of computational skills in the AU team is really key and useful. Their ability to listen and answer to questions is also very valuable.*

[1] University of Luxembourg [2] Leiden University [3] Aarhus University [4] University of Groningen [5] University of Aix-Marseille

Crisis Communication in the Niagara Region during the COVID-19 Pandemic

Tim Ribaric, David Sharron, Cal Murgu, Karen Louise Smith, and Duncan Koerber[1]

Project Description

Using web archives collected by Brock University, this project examined how organizations in the Niagara region communicated about bylaw changes, masking requirements, and the vaccine rollout in the first two years of the COVID-19 pandemic. Analysis was based on a close reading of web pages and social media posts as well as Big Data analysis of broader trends over two years. Findings from this research will inform future crisis communication organizational planning, specifically at the local and municipal level. The project also created several open computational notebooks to support teaching, learning, and research.

Subject/Field of Study: COVID19, Pandemic, Crisis Communication, Risk Communication, Health & Science Communication, Big Data

Collections

- COVID-19 in Niagara (https://archive-it.org/collections/13781) A weekly crawl of websites of major institutions, governments and organizations in the Niagara area focusing on the varied responses to the COVID-19 pandemic in 2020 and 2021.

Research Questions

1. How did Niagara Region organizations communicate about COVID-19 rules, plans, and vaccination in the first two years of the pandemic?
2. Did the organizations' public communication reflect best practices in the field of risk and crisis communication?
3. How did the organizations' communication focus and approach change over time?

[1] Brock University

Crisis Communication in the Niagara Region during the COVID-19 Pandemic

Tim Ribaric, David Sharron, Cal Murgu, Karen Louise Smith, and Duncan Koerber[1]

Methods & Tools

- ➔ Close reading of Archive contents using the SolrWayBack platform (https://github.com/netarchivesuite/solrwayback)
- ➔ Custom analysis dashboards written in Jupyter Notebooks (https://github.com/BrockDSL/ARCH_Data_Explore)

Research Insights

- All municipalities informed, educated, and engaged citizens to some degree
- Varying communication efficacy and scope led to differences in community resilience between municipalities; Niagara Region public communication patched over variances
- Web archives approach and data analysis tools enabled Big Data collection of the COVID-19 web pages over long period and facilitated complex analysis of page changes and the close reading of web pages

Outputs and Scholarship

- ➔ Journal article currently under peer review with the *Canadian Journal of Communication*.
- ➔ Paper presented at International Association for Media and Communication Research 2023 conference.
- ➔ "Brock examining Niagara's crisis communications during COVID-19. St. Catharines Standard (https://www.stcatharinesstandard.ca/news/niagara-region/2021/11/12/brock-examining-niagaras-covid-19-crisis-communicated.html)
- ➔ Project Website and blog: https://brockdsl.github.io/archives_unleashed/

[1] Brock University

Crisis Communication in the Niagara Region during the COVID-19 Pandemic

Tim Ribaric, David Sharron, Cal Murgu, Karen Louise Smith, and Duncan Koerber[1]

Reflections

Challenges/Obstacles + Solutions

One major challenge for us was moving back and forth between close reading and data analysis. Sometimes, observations from the close reading could not be substantiated in the broader data analysis. This forced us to shift our analysis in other directions.

Writing custom analysis notebook required having a programmer on the research team to translate questions asked by researchers into the equivalent Python code.

SaaS resources were required to properly bootstrap the SolrWayBack search engine; this created some complexity in scaffolding the project.

Using ARCH / Cohort Program Experience

“This project truly opened my eyes to using Big Data from web archives in crisis communication research. Typically, I have focused on small case studies, but the web archives facilitated a ‘big picture’ project I have not done before. This scope also increased the significance of the findings.” **– Duncan Koerber**

“ARCH represents a very interesting set of opportunities for researchers. Instead of cumbersome WARC files that are often gigabytes in size, small portable CSV’s can be used instead. There is a large audience out there that already knows how to use this type of data.” **– Tim Ribaric**

[1] Brock University

Mapping and tracking the development of online commenting systems on news websites between 1996–2021

Anne Helmond[1], Johannes Paßmann, Lisa Gerzen,[3] Martina Schories[2], Robert Jansma[3] with contributions from Luca Hammer[2], Dave Wahl[4], Steffen Reinhard[3], and Theresa Schulte[2]

Project Description

This project aims to reconstruct a history of online commenting by examining the role of commenting technologies in the popularisation of commenting practices. It will do so by examining the distribution and evolution of commenting technologies on the top 25 Dutch, German, and world news websites from 1996–2021, to understand how they have shaped the practices of users. This will allow them to explore the interplay between technologies and practices of the past and to investigate histories of natively-born technologies and practices.

Subject/Field of Study: commenting systems, online comments, news websites, web archives

Collections

➔ Three collections of the top 50 international, German, and Dutch archived news websites 2007–2021.

- Dutch: 9GB
- German: 22GB
- World: 32GB

Research Questions

1. How can we systematically study the dissemination of online commenting systems on news websites?
2. How can we understand the evolution of commenting systems and their practices over time?
3. What do changes in commenting systems reveal about the relationship between newspapers and their readers, about moderation practices, about the evolving web ecosystem, and about the professionalisation and commodification of commenting?

[1] Utrecht University [2] University of Siegen [3] Ruhr University Bochum [4] University of Amsterdam

Mapping and tracking the development of online commenting systems on news websites between 1996–2021

Anne Helmond[1], Johannes Paßmann, Lisa Gerzen,[3] Martina Schories[2], Robert Jansma[3] with contributions from Luca Hammer[2], Dave Wahl[4], Steffen Reinhard[3], and Theresa Schulte[2]

Methods & Tools

- ➔ Regular expressions (RegEx)
- ➔ Dedicated HTML parsers (BeautifulSoup)
- ➔ R
- ➔ Jupyter Notebooks
- ➔ Close reading

Research Insights

- Overall, we found that the “death of commenting” proclaimed by tech journalists in 2015 has been highly exaggerated.
- When commenting systems were missing a manual inspection revealed 1) new self-coded systems that we iteratively included, 2) that some news websites turned their commenting systems on and off (i.e. they shut down their commenting possibilities at night to prevent abuse), revealing unknown practices to protect public discourse.
- We observed the rise and fall of Disqus and Facebook over time, with newspapers turning to Livefyre, Gigya and ViaFoura. These newer commenting systems indicate a professionalisation of commenting as they are often part of larger customer relation suites and of larger moderation tools.

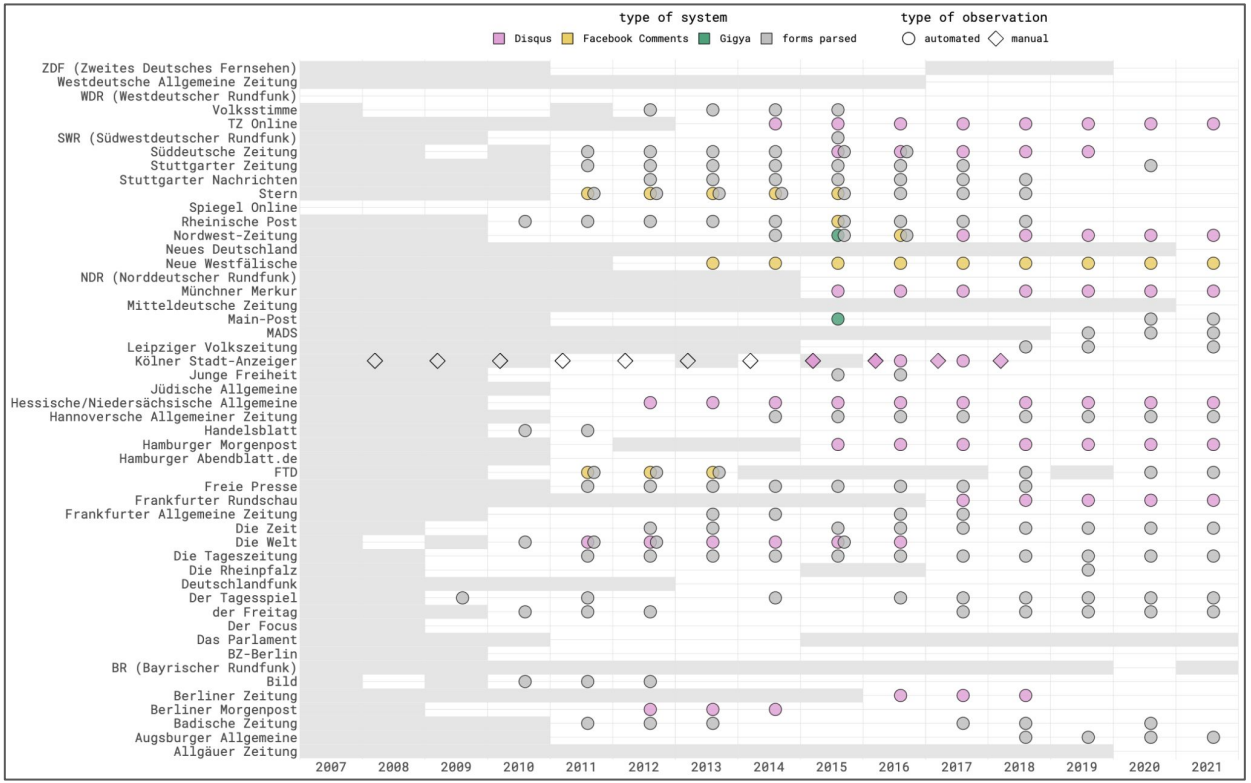
Outputs and Scholarship

- ➔ Conference presentations: Two presentations at the Association of Internet Researchers conference 2022 in Dublin, Ireland [Slides]. Presentation at the WARCnet closing event, 2022 in Aarhus, Denmark [Slides]. Presentation at the Internet Archive, 2021, online event [Slides]. Presentation at the Internet Archive, 2022, Canada.
- ➔ Workshops on ‘online technography’ using web archives at the University of Siegen.
- ➔ Development of Jupyter Notebooks for internal use.
- ➔ Development of ‘the Technograph’ tool (in-progress), an visualisation interface on the dataset to assist in the analysis.

[1] Utrecht University [2] University of Siegen [3] Ruhr University Bochum [4] University of Amsterdam

Mapping and tracking the development of online commenting systems on news websites between 1996–2021

Anne Helmond[1], Johannes Paßmann, Lisa Gerzen,[3] Martina Schories[2], Robert Jansma[3] with contributions from Luca Hammer[2], Dave Wahl[4], Steffen Reinhard[3], and Theresa Schulte[2]



[1] Utrecht University [2] University of Siegen [3] Ruhr University Bochum [4] University of Amsterdam

Mapping and tracking the development of online commenting systems on news websites between 1996–2021

Anne Helmond[1], Johannes Paßmann, Lisa Gerzen,[3] Martina Schories[2], Robert Jansma[3] with contributions from Luca Hammer[2], Dave Wahl[4], Steffen Reinhard[3], and Theresa Schulte[2]

Reflections

Challenges/Obstacles + Solutions

Tracing code snippets in archived HTML (cf. Lerner et al., 2016; Helmond, 2015; 2017; Nielsen, 2019; Owens and Thomas, 2019) to detect and examine the presence and removal of commenting systems proved to be more complicated than anticipated. To address this we made a database of code snippets that serves as an inventory of patterns to look for. The side effect was that we learned more about the commenting systems and how they function on websites.

The amount of data was initially overwhelming, and understanding what was included in our data set through the sampling method of including one subpage for each news website. We are still working through these issues. To navigate the data in a visual way, we built the Technograph tool, to assist in the analysis.

Using ARCH / Cohort Program Experience

ARCH and the Cohorts Program have enabled us to kickstart our project in multiple ways, e.g. by gaining access to the datasets we wanted to work with.

The AU team's exceptional computational skills and knowledge of web archive research in general played a pivotal role in our progress. Their expertise in this area was key and proved to be immensely useful throughout the project.

[1] Utrecht University [2] University of Siegen [3] Ruhr University Bochum [4] University of Amsterdam

Everything Old is New Again: A Comparative Analysis of Feminist Media Tactics between the 2nd- to 4th Waves

Shana MacDonald[1], Aynur Kadir[1], Brianna Wiens[1], Sid Heeg[1], Hannah Delamere[1]

Project Description

Project members explored web archive collections to conduct a comparative analysis of the history of feminist media practices across interdisciplinary multi-media sources. The team sought produce a timeline of issue responses from different historical moments and map different feminist media practices over this timeline to determine overlaps. The project's key outcome is to recover earlier feminist media practices and contextualize them in the digital present.

Subject/Field of Study: (Feminist Media, Feminist Archives, second wave feminism, digital feminism, feminist activism)

Collections

- ➔ Tamiment-Wagner Feminism and Women's Movements
- ➔ Tamiment-Wagner Other Left Activism
- ➔ Fales Library New York Feminist Art Institute
- ➔ Fales Library Guerrilla Girls
- ➔ Sallie Bingham Center for Women's History and Culture
- ➔ Contemporary Women Artists on the Web
- ➔ Capturing Women's Voices
- ➔ Schlesinger Library #MeToo Web Archives collection

Research Questions

- ➔ Examine feminist media tactics across different historical eras (1960s-present) to show continuities (and breaks) in approaches over time.
- ➔ Map and analyze the presence of 'feminist key concepts' in these archives.
- ➔ Challenge the current overemphasis on intergenerational conflicts within feminism and the narrative of feminist "waves" by mapping the histories or continuities of these terms

[1] University of Waterloo

Everything Old is New Again: A Comparative Analysis of Feminist Media Tactics between the 2nd- to 4th Waves

Shana MacDonald[1], Aynur Kadir[1], Brianna Wiens[2], and Sid Heeg[1]

Methods & Tools

Examples

- ➔ Keyword analysis
- ➔ Comparative analysis between collections
- ➔ Juxta Image collage
- ➔ Close reading/zoom in from big data
- ➔ New method of timeline close reading from Juxtas

Research Insights

We discovered many surprises in the data. The most significant was that there are limits to big data web crawls for certain topics that require greater nuance in their analysis because of the social complexity of the issues. For instance, our work with the Schlesinger #MeToo archive revealed the limits of institutional collections as it did not reflect the MeToo movement itself in any significant degree. To pivot from this we began working with Juxtas which offer a lot of promise for 'zooming' into specific cases of data from a big data set for close reading methods.

Outputs and Scholarship

- ➔ "Approaches to Archiving Feminist Memes." Preserving Digital Born Media by Women: methods for decolonial & feminist futures (Panel). Film and Media Studies Association of Canada annual conference, May 2023
- ➔ "Activists Archiving the Internet: Social Justice Informed Approaches to Digitally Born Content." Panel with Nick Ruest, Brianna Wiens, Shawn Walker, Mina Momeni. Shaking Up the Archive, Queen Margaret University, Edinburgh, June 23-25, 2023. Accepted February 2023.
- ➔ "Reconceptualizing Internet Archives: Feminist Memes as Repertoire" Canadian Association for Theatre Research, Halifax, June 9-12, 2023. Accepted February 2023.
- ➔ "From Placards to Memes: The Utopic Refusals of Feminist Media Techno-Imaginaries," co-author Brianna Wiens, *Feminist Encounters* (Invited for special issue Sept 2023)

[1] University of Waterloo [2] York University

Everything Old is New Again: A Comparative Analysis of Feminist Media Tactics between the 2nd- to 4th Waves

Shana MacDonald[1], Aynur Kadir[1], Brianna Wiens[2], and Sid Heeg[1]

Reflections

Challenges/Obstacles + Solutions

Main challenge was getting acclimated to working with big data and computational systems. This is a huge learning curve for those with no prior experience and led us to many points of stasis.

Solutions were to develop new methods that got us ‘closer’ to the data so we could ‘dwell’ with it and that opened us to new forms of scholarly insights about the process of working with data in general

Best outcome was working with Juxtas as they provide a means of big picture thinking through big data as well as the potential to ‘zoom in’ to specific data points for further analysis

Using ARCH / Cohort Program Experience

Really informed our methodological approach to digitally born archives in productive ways (ie helped articulate our need to ‘dwell’ which we have now published on)

Loved encountering and learning from the ARCH team and getting at some really crucial questions around the tensions and struggles for bringing humanities perspectives to digital archives

Really celebrate and appreciate how the ARCH team and Nick, Sam, Ian in general worked with us to find valuable solutions especially in our work with Juxtas

Juxtas are a huge tool for anyone doing visual cultural analysis and working with digitally-born artifacts

[1] University of Waterloo [2] York University

Viral health misinformation from Geocities to COVID-19

Shawn Walker, Michael Simeone, Kristy Roschke, Anna Muldoon, Major Brown[1]

Project Description

This project will examine and compare two case studies of health misinformation: HIV mis/disinformation circulating on Geocities in the mid-1990s to early 2000s with the role of official COVID-19 Dashboards in COVID mis/disinformation. This work contributes to our understanding of current and historical health misinformation as well as the connections between them, and will also garner insights into how historical narratives of health misinformation have been recycled and repurposed.

Subject/Field of Study: Misinformation, sensemaking, social media, communication, digital humanities, journalism, information science

Collections

- ➔ Internet Archive GeoCities Special Collection 2009
- ➔ OoCities GeoCities Web Archive
- ➔ IIPC Novel Coronavirus (COVID-19) Archive-It collection
- ➔ COVID-19 Twitter Stream

Research Questions

- ➔ What are the narratives and linking patterns of HIV mis/disinformation circulating on GeoCities?
- ➔ What are the narratives and linking patterns referencing official COVID-19 dashboards on the public web and Twitter?
- ➔ What information can we extract from official (national, state/province, and local) COVID-19 sources such as dashboards that were archived, but do and do not playback?

[1] Arizona State University

Viral health misinformation from Geocities to COVID-19

Shawn Walker, Michael Simeone, Kristy Roschke, Anna Muldoon, Major Brown[1]

Methods & Tools

- ➔ Network Analysis
- ➔ Close reading/qualitative analysis
- ➔ NLP (topic modeling)/Corpus linguistics
- ➔ Gephi

Outputs and Scholarship

- ➔ Viral health misinformation: From GeoCities to COVID-19, presented at the Association of Internet Researchers, Dublin
- ➔ Paper under review

Research Insights

- The predominant sentiment/motivation in the sample is one of support and sense-making. This is in keeping with what we know about Geocities spaces as replicating positive IRL communities in an online realm.
- Categories of (mis)information in Geocities:
  - **Supportive**. Information for treating HIV infection and symptoms.
  - **Preventative.** Information offering methods of preventing HIV infection – often a mix of accurate information (safer sex practices) and incorrect prevention methods
  - **Debunking.** Attempts to “debunk” or correct early misinformation or misconceptions about HIV.
  - **Hate**. Primarily homophobic and hate-filled content often framing HIV infection as a result of living a gay lifestyle.
- Supportive, debunking, and preventive are attempts to fill information voids.
- What we do not see are government conspiracies like we do in COVID-19. These may be other forums

[1] Arizona State University

Linked pairs in IA dataset

Geocities and Oocities: term frequencies overall, top 30 terms

8365	aids	
5174	hiv	
1323	people	
1166	health	
980	will	
931	information	
735	virus	
707	com	
684	one	
624	de	
623	sex	
597	new	
594	also	
580	disease	
531	women	
527	infected	
508	may	
497	africa	
489	children	
489	world	
469	medical	
451	like	
451	treatment	
450	hearing	
448	many	
446	time	
442	living	
442	prevention	
435	national	

Viral health misinformation from Geocities to COVID-19

Shawn Walker, Michael Simeone, Kristy Roschke, Anna Muldoon, Major Brown[1]

Reflections

Challenges/Obstacles + Solutions

Initially our challenge was to bring a team who had not worked with web archives (except for one member) up to speed. The mentorship and regular meetings with the AU Team were instrumental in this process.

We were surprised by the challenge of contextualizing the historical context of GeoCities and to not be biased by our present-day interpretations. At the time of GeoCities, HIV was still novel so what seemed like misinformation to us might have been attempts to navigate an information void vs a deliberate attempt to spread misinformation. We worked through this via thoughtful discussions within our team and multiple consultations with the AU Team.

Using ARCH / Cohort Program Experience

Feel free to use this space to share thoughts around questions like:

- *Our regular meetings with the AU Cohorts team helped our cohort to better understand the methodological and technical challenges of using web archives. It was invaluable for us to work with a group of scholars with a high level of rigor and deep methodological experience.*
- *Using ARCH allowed us to focus on our research instead of the technical details instead of developing our own toolchain to process web archives – we got to focus on our areas of expertise instead of reinventing the wheel.*
- *The AU team listened to our feedback and added features to ARCH to support our research.*

[1] Arizona State University

Latin American Women's Rights Movements: Tracing Online Presence through Language, Time and Space

Sylvia Fernandez[1], Rosario Rogel-Salazar[2], Verónica Benítez-Pérez[2], Alan Colín-Arce[2] and Abraham García-Monroy[2]

Project Description

In analyzing web archives related to human rights and feminist movements, we developed a topic analysis of human rights reports from organizations in Mexico and Latin America, particularly those focusing on eradicating femicides and gender violence. We also studied national similarities and differences in the issues addressed by human rights and feminist organizations in Spanish-speaking Latin American countries.

Subject/Field of Study: Digital Activism, Latin American Feminism, Cyberfeminism, Counterarchives

Collections

- ➔ Human Rights (Columbia University Libraries)
- ➔ Activismos Feministas en América Latina (Huellas Incómodas)

Research Questions

1. What are the differences and similarities between countries in the issues addressed by human rights and feminist organizations in Latin America?
2. Do Spanish and English human rights reports on Latin America address the same issues?
3. What are the most common places, organizations, and people mentioned in Latin American human rights reports?

[1] University of Texas at San Antonio [2] Universidad Autónoma del Estado de México, Toluca

Latin American Women's Rights Movements: Tracing Online Presence through Language, Time and Space

Sylvia Fernandez[1], Rosario Rogel-Salazar[2], Verónica Benítez-Pérez[2], Alan Colín-Arce[2] and Abraham García-Monroy[2]

Methods & Tools

Examples

- ➔ Topic modeling
- ➔ Named entity recognition (NER)
- ➔ Creation of an *ad hoc* web archive
- ➔ Curating a thematic collection from an existing web archive

Research Insights

In a sample of 13,000 human rights reports on Latin America, six main topics were covered in both Spanish and English: Climate change, Criminal justice, Gender-based violence, Human rights cases, Human rights provision, and Indigenous peoples. However, English reports primarily focused on human rights cases, while Spanish reports were more centered on human rights provision.

Regarding gender issues, the human rights reports mainly emphasized reproductive rights and sexual violence. Other important issues like feminicide/femicide, domestic work, the wage gap, and lesbian and transgender issues were seldom mentioned. Among the seven analyzed countries, four with recurrent reports on reproductive rights have either the most stringent abortion laws in the world (Nicaragua and El Salvador) or the most progressive laws (Argentina and Uruguay).

These findings indicate that gender studies utilizing large text analysis cannot generalize gender violence in Latin America due to variations in the issues addressed by each country.

When it comes to web archiving in Latin America, there are very few web archives created within the region. Only institutions from Colombia, Mexico, and Puerto Rico have collections on Archive-It, while Chile is the sole country with a national web archive. This situation creates a reliance on foreign institutions to preserve Latin American web history and may contribute to a digital divide between the Global North and South in terms of digital preservation.

Outputs and Scholarship

- ➔ Web Archive Collection, part of Huellas Incómodas Project: https://idrhku.org/huellasincomodas/webarchive
- ➔ Spanish Wikipedia entry for web archiving https://es.wikipedia.org/wiki/Archivado_web
- ➔ DH Workshop hosted at the Autonomous University of the State of Mexico, 2023, organized with a grant from the Science Council of the State of Mexico https://idrhku.org/huellasincomodas/investigaciondigital
- ➔ Two conference presentations
- ➔ (Forthcoming Article) Rosario Rogel-Salazar, Abraham García, Alan Colin-Arce, Verónica Benítez-Pérez. Preserving the memory of Latin American feminist movements on digital and web counterarchives
- ➔ $100,000 MXN grant from the Science Council of the State of Mexico under the program Research Funding for Women Scientists.

[1] University of Texas at San Antonio [2] Universidad Autónoma del Estado de México, Toluca

Latin American Women's Rights Movements: Tracing Online Presence through Language, Time and Space

Sylvia Fernandez[1], Rosario Rogel-Salazar[2], Verónica Benítez-Pérez[2], Alan Colín-Arce[2] and Abraham García-Monroy[2]

Reflections

Challenges/Obstacles + Solutions

1. Lack of infrastructure for processing large amounts of data.
    a. Solutions: extracting a thematic sample of all the data in a large (more than 16 TB) collection so that it addressed only the region we were interested in studying.
    b. Analyzing pdfs instead of the full-text of the websites. This allowed faster processing times and the data was easier to clean.

2. Lack of web archives with content in Spanish and created in Latin America.
    a. Creation of an *ad hoc* collection that preserved a selection of 76 feminist websites from Latin America.
    b. Promotion of web archiving in conferences, local universities, and the Mexican national archive.
    c. Development of workshops to promote digital skills among feminist collectives.

Using ARCH / Cohort Program Experience

- ARCH is a convenient tool to obtain several datasets from web archives. Creating a more extensive description of the derivatives could be useful for researchers beginning with the analysis and use of web archives.

- The cohort program introduced us to a completely new approach of studying feminist movements through web archives and pointed us to an issue that has been underexplored in Latin America and web history in general, which is the lack of web archiving initiatives in the Global South.

[1] University of Texas at San Antonio [2] Universidad Autónoma del Estado de México, Toluca

Historicizing Aughts-Era Mormon Mommy Blogging Media Landscapes

Emily Edwards[1], Robin Hershkowitz[2], and Lauren Andrikanich

Project Description

We explore Mormon Mommy Bloggers (MMBs) as significant creators who used the long-form blog to center topics of child-rearing, domesticity, and lifestyle themes, fostering community and transforming invisibilized domestic labor into monetizable content. Uniting feminist data visualization, network, and textual analysis, we situate MMBs as having critically influenced contemporary manifestations and trends of mommy influencing on social media platforms as both a lucrative and ideological form of digital labor.

Subject/Field of Study: Digital Labor, Motherhood, Blogging, Feminism, Domesticity, Network Analysis, Data Visualization

Collections

- ➔ Mormon Blogs Collection
- ➔ Mormon Journals and Magazine Collection
- ➔ Mormon Websites Collection

Research Questions

1. How was the informational infrastructure of the “Bloggernacle” organized? How did Mormon women contribute as key architects?
2. What were the topics, content, and discussions that animated MMBs?
3. How can we historicize the blogging practices of MMBs as forms of digital labor before influencing became a legible occupation?
4. What can this period of blogging tell us about contemporary mother-influencer trends and the ”end” of the blogger in a new media ecology?

[1] St. Francis College [2] Bowling Green State University

Historicizing Aughts-Era Mormon Mommy Blogging Media Landscapes

Emily Edwards[1], Robin Hershkowitz[2], and Lauren Andrikanich

Methods & Tools

Examples

- ➔ Network Analysis via Gephi
- ➔ Google Colab Notebooks
- ➔ Grounded Theory Methodology
- ➔ Close Textual Reading
- ➔ Distant Reading
- ➔ Juxta (Visual Analysis)

Research Insights

- Mormon Mommy Bloggers (MMBs) composed a strong and influential part of the larger male-dominated “Bloggernacle” and feminist voices were some of the most central within this landscape.
- MMBs transformed intimate, autobiographical forms of writing into monetizable forms of content, paving the way for present-day mommy-influencers.
- Answering critical, feminist research questions with web archives involves circular, iterative engagement with data and the reconstruction of media ecologies via visual, textual, and network methods.

Outputs and Scholarship

- ➔ Abstract accepted, "Digital Pioneers: Mormon Mommy Bloggers and Digital Domestics,” for a special issue of *Internet Histories* on “Gender and the Internet/Web History” forthcoming 2025
- ➔ Zine “Sliding Data: Feminist Methodological Pathways” accepted for publication in *DIY Methods* Low-Carbon Research Methods Initiative forthcoming 2023
- ➔ Submission as part of panel on “Digital scholarship and the web: Exploring new sources and emerging research methods” for the Digital Library Federation Forum Conference 2023

[1] St. Francis College [2] Bowling Green State University

Historicizing Aughts-Era Mormon Mommy Blogging Media Landscapes

Emily Edwards[1], Robin Hershkowitz[2], and Lauren Andrikanich

Reflections

Challenges/Obstacles + Solutions

Our foundational challenges were related to size and form, e.g., as feminist researchers without a programming background working with “large” amounts of data (multiple GB) in unfamiliar forms required technical learning and the creation of strategies to transform the data into useful materials to answer our RQs.

To address these challenges required both engaging in our own technical development with data analytic methods and “taking the hammer” to normative, linear expectations of data analysis.

We implemented strategies to make the data smaller and more approachable (learning the command line, “cleaning” and cutting CSV files, using Google Colab Notebooks & basic Python, exporting data into workable formats .gexf files, etc).

Once we developed our technical skills to transform the data, we were able to “read” the materials. Through this process we found a utility in uniting feminist values that privilege multiple perspectives, partiality, and sharing power with technical methodological tricks to make web archives legible for us as researchers without a quantitative background and traditional computational skill-set.

Using ARCH / Cohort Program Experience

ARCH was a critical intermediary platform for us to interface with forms of data less familiar to us as feminist, qualitative researchers. The ability to create and explore derivatives allowed us an immediate point of entry. Even when derivatives were too large for us to process ourselves, the Example Datasets gave us a place to begin and practice our skills.

The AU Program was extremely helpful for us as a smaller team with a quantitative background limited to network analysis. The mentorship provided was key for us as we developed our technical skills. We believe this model of sustained support and education in foundational concepts and programs (from the command line to working with Google Colab) via bi-weekly meetings with the AU team is necessary to foster an inclusive approach to the study and use of web archives that welcomes humanities researchers from a variety of backgrounds and skill-sets.

[1] St. Francis College [2] Bowling Green State University

Web Archiving and the Saskatchewan COVID Archive: Expanding Coverage to Capture Social Media, Medical Misinformation, and Radicalization

Jim Clifford, Derek Cameron, Erika Dyck, Craig Harkema, Patrick Chasse, and Tim Hutchinson[1]

Project Description

This is a part of the larger Remember Rebuild Saskatchewan project focused on documenting the COVID 19 pandemic in Saskatchewan. The project combines population survives, oral history interviews, creating a large Zotero database of COVID content and web archiving. The web archive project focused on processing an Archive-It collection and creating a new Twitter archive. We tried developing a Facebook archive as well, but struggled to work with the WARC files created by Webrecorder.

Subject/Field of Study: (COVID19, Pandemic, Twitter, Saskatchewan, Political history, Social History, Disinformation)

Collections

- ➔ Web archive created by the University of Saskatchewan library and archives.
- ➔ Twitter archive created during the project.

Research Questions

- ➔ What do we do with 2TB of archive it data?
- ➔ How do record the Twitter conversation about COVID 19 in Saskatchewan (particularly when the future of Twitter and academic API access was un in the air).
- ➔ Can we develop new ways for researchers to access and explore the web archive and Twitter archive?

[1] University of Saskatchewan

Web Archiving and the Saskatchewan COVID Archive: Expanding Coverage to Capture Social Media, Medical Misinformation, and Radicalization

Jim Clifford, Derek Cameron, Erika Dyck, Craig Harkema, Patrick Chasse, and Tim Hutchinson[1]

Methods & Tools

Examples

- ➔ Twitter API download, filtering and processing with basic text mining tools.
- ➔ Transformer models for sentiment analysis and topic analysis.
- ➔ Core content extraction from the HTML using Newfeed.

Research Insights

We ended up focused a lot on how to filter archives down to content that is actually about COVID 19 in one Canadian province. This is a mix of technical and humanist decisions to figure out how to filter without losing too much relevant material.

LLM transformers show a lot of potential. For example, we got good results processing all the Tweets for sentiment analysis. Unfortunately, for this project, it turns out people Tweeting about COVID 19 were overwhelmingly negative (and I don't know if we needed advanced computing to figure this out).

Outputs and Scholarship

- ➔ Nazeem Muhajarine et al., "Capturing and Documenting the Wider Health Impacts of the COVID-19 Pandemic Through the Remember Rebuild Saskatchewan Initiative: Protocol for a Mixed Methods Interdisciplinary Project," *JMIR Research Protocols* 12, no. 1 (June 6, 2023): e46643, https://doi.org/10.2196/46643.
- ➔ Derek Cameron, "Archiving Twitter During the Upheaval'" ActiveHistory.ca, https://activehistory.ca/2022/11/archiving-twitter-during-the-upheaval/

[1] University of Saskatchewan

Web Archiving and the Saskatchewan COVID Archive: Expanding Coverage to Capture Social Media, Medical Misinformation, and Radicalization

Jim Clifford, Derek Cameron, Erika Dyck, Craig Harkema, Patrick Chasse, and Tim Hutchinson[1]

Reflections

Challenges/Obstacles + Solutions

Scale is a problem. Even after spending weeks to develop filters to shrink the number of pages we're processing down to 80,000 and extracting only the core text, it remains too big for TD-IDF on the whole collection even when we access the large memory machines from Compute Canada.

Scale also makes it hard to know what to do with the content for historical research. We've made a lot of progress in using distant reading to better understand the collection and to identify gaps in the Twitter archive. But we still need to do more work to figure out how to use these collections to write a book about COVID 19 in Saskatchewan beyond close reading a small sample of the total.

Using ARCH / Cohort Program Experience

ARCH helped a lot by providing a giant CSV file with all the HTML plus metadata. Some of the other tools, like the domain list, was helpful at the start.

The Cohort Program gave us the opportunity to build the Twitter archive and start processing the web archive and create the SOLR index.

[1] University of Saskatchewan

Querying Queer Web Archives

Filipa Calado[1], Corey Clawson[2], Di Yoong[1], and Lisa Rhody[1]

Project Description

We are interested in studying collections of web archives relating to queer communities, organizations, and web spaces, specifically through *querying* both the material and methodology of web archival work---how queer identities and communities interact with, are informed by, and form web spaces, as well as critical methodologies for studying digital archives, including concerns relating to the ethics of web archiving as notions of web spaces transformed over time. We are interested in exploring concepts like utopia, radicalism, normativity, religion, and conversion, and how they affect queer identity and discourse formation over time.

Subject/Field of Study: queer spaces, queer identities, ethics, temporality

Collections

- ➔ Soc.Mots Document Collection;
- ➔ LGBTQ Web Collection;
- ➔ Interactive Fiction in Queer Literature

Research Questions

- ➔ How do queer identities and communities interact with and inform web spaces?
- ➔ How are web spaces themselves formed and cultivated by communities?
- ➔ How do concepts like utopia, radicalism, normativity, religion, and conversion affect queer identity and discourse formation over time?
- ➔ What vocabularies have been/are used to conceptualize homosexuality and other queer sexualities/identities by various social organizations (such as religious organizations) and communities?

[1] The Graduate Center, CUNY [2] Rutgers University - Newark

Querying Queer Web Archives

Filipa Calado[1], Corey Clawson[2], Di Yoong[1], and Lisa Rhody[1]

Methods & Tools

- Topic Modeling; Structured Topic Models
  - pandas, polars & spaCy in Python
  - STM in R
- Close reading

Research Insights

- Hard to employ a one-size-fits-all approach with cleaning and analyses with the vastly differently sized datasets
- Sense of privacy shifts over time; users divulge personal information (e.g. physical addresses) from datasets with webpages from the late 90s/early 00s
- Purpose of sites (e.g. personal social media v. organizational websites) offers different foci on queer discourse and identities

Outputs and Scholarship

- Yoong, D., Calado, F., & Clawson, C., (2023, May 3). *Querying Queer Web Archives* [Conference presentation]. IIPC WAC 2023, Hilversum, The Netherlands.

[1] The Graduate Center, CUNY [2] Rutgers University - Newark

Querying Queer Web Archives

Filipa Calado[1], Corey Clawson[2], Di Yoong[1], and Lisa Rhody[1]

Reflections

Challenges/Obstacles + Solutions

Limited computing power meant that we had to employ creative solutions to our analyses. We first took a randomized sample of the largest dataset (LGBTQ Web Collection) and proceeded to test our approaches. While this provided some workaround, it was still insufficient in producing adequate analyses. As a result, we began to run analyses based on the domains collected, further breaking down our collected dataset. This approach is inlined with our initial desires to understand overlapping conversations and topics across different organizations and groups.

The varied size of the datasets of the different collections and the time period of which these collections were produced also meant that the method we fine-tuned with the LGBTQ Web Collection could not be easily applied or translated. As a result we had to adopt a different approach to the soc.mots and interactive fiction collection. In addition, different notions of privacy and concerns for anonymity also factored into the process of cleaning and analyses. For example, the earliest dataset (soc.mots) contained personal information that was not easily removed with the approach we took with the LGBTQ Web Collection, and resulted in a more manual removal while we explored NER (Named Entity Recognition) approaches.

Using ARCH / Cohort Program Experience

This was a really supportive space for us to explore working with web archives and the methods of inquiries. Folx were really patient with us as we bring new inquiries, questions, and concerns to each meeting. Ranging from the technical to the theoretical and ethical concerns, folx at ARCH consistently offered help and support. The “mid-term” conversation with other teams in the cohort was also really helpful and it was great to hear what others were working on before the final meeting at the end of the experience.

[1] The Graduate Center, CUNY [2] Rutgers University - Newark

Using Web Archives for Mapping the Use of Cultural Practices in Postconflict Societies and During Reconciliation Processes

Ricardo Velasco Trujillo[1] and Luis Gomez[2]

Project Description

Despite the growing interest in the role of cultural practices during reconciliation processes, in societies trying to confront the legacies of troubled pasts and widespread human rights violations, there is a broad gap in the reliable data regarding the type of cultural practices that are employed, and they way they interact with or are entangled within processes of reconciliation across different geographical and cultural contexts. Using methods such as contextual search and data mining, the projects aims at making an initial assessment and mapping out the use of cultural practices in different reconciliation processes by different institutions and organizations working in these contexts.

Subject/Field of Study: Cultural Practices; Cultural Activism; Reconciliation Processes; Historical Redress; Human Rights Defense

Collections

- ➔ Human Rights Documentation Initiative, University of Texas Libraries.
- ➔ Digital Tools for Human Rights Awareness, University of Michigan School of Information.
- ➔ Human Rights, Columbia University Libraries.
- ➔ Truth and Reconciliation, University of Manitoba.
- ➔ Truth and Reconciliation Commission, Library and Archives Canada.
- ➔ National Centre for Truth and Reconciliation, University of Manitoba.

Research Questions

- ➔ What cultural practices and artistic formats (video, photography, or other expressive forms) are the most common in reconciliation processes or amongst human rights organizations working in this context ?
- ➔ What type of relationships or entanglements can we find in the use of cultural, digital media, and art practices during reconciliation processes ?

[1] University of Minnesota Twin Cities [2] The Australian Academy of Science

Using Web Archives for Mapping the Use of Cultural Practices in Postconflict Societies and During Reconciliation Processes

Ricardo Velasco Trujillo[1] and Luis Gomez[2]

Methods & Tools

- ➔ Data Mining
- ➔ Contextual Search

Outputs and Scholarship

- ➔ Digital publication in Cultural Ecologies of Memory (CEM) Platform: http://www.culturalecologies.com/case/DataMining

Research Insights

1. Cultural venues such as museums, art galleries, and local community and cultural centers provide key opportunities for official bodies to effectively disseminate information, resources, and to engage different sectors of the populations in reconciliation processes.

2. Cultural venues offered opportunities for these bodies to provide services to vulnerable populations, including mental health and wellness services, or capacitation through workshops in craft making and other artistic skills. These resources can lead to positive outcomes among these vulnerable communities.

3. Cultural Practices provide opportunities for actors from vulnerable communities, such as indigenous youth and women, to participate in community building activities.

4. Cultural venues also allow opportunities for different actors to tell their stories and testimonies through video recordings and to have these records archived.

5. Video and audio testimonies become a central vehicle by which the memory and history of communities affected by historical injustices are archived and secured for future generations.

[1] University of Minnesota Twin Cities [2] The Australian Academy of Science

Resources

Archives Unleashed Project

- Project Website: https://archivesunleashed.org/
- GitHub: https://github.com/archivesunleashed
- Twitter: @UnleashArchives

Blog Posts about Cohorts

- Introducing Archives Unleashed Cohorts, 2021
- Research Applications with Web Archives: Collaboration Among Archives Unleashed Cohorts, 2022
- Web Archives Research: Return of the Cohorts, 2022