# INSTRUCTIONS:

Find your title slide below. Use the blank slide following your title slide for your 1-minute lightning talk.

- Do **not** include any audio, video, or clickable links in your slide.
- If you wish to use a different template for your slide, click “LAYOUT” in the toolbar to see your options.
- You may add other images, text boxes, or tables/diagrams if needed, but take care not to overload your slide.

If you don’t need a slide for your talk, we will use the slide with the title of your poster. Please remove your “blank” slide if you choose this option.

National Library of Norway

# IIPC GENERAL ASSEMBLY & WEB ARCHIVING CONFERENCE

**OSLO, 8–10 APRIL 2025** | NETPRESERVE.ORG/GA2025

IIPC

netpreserve.org

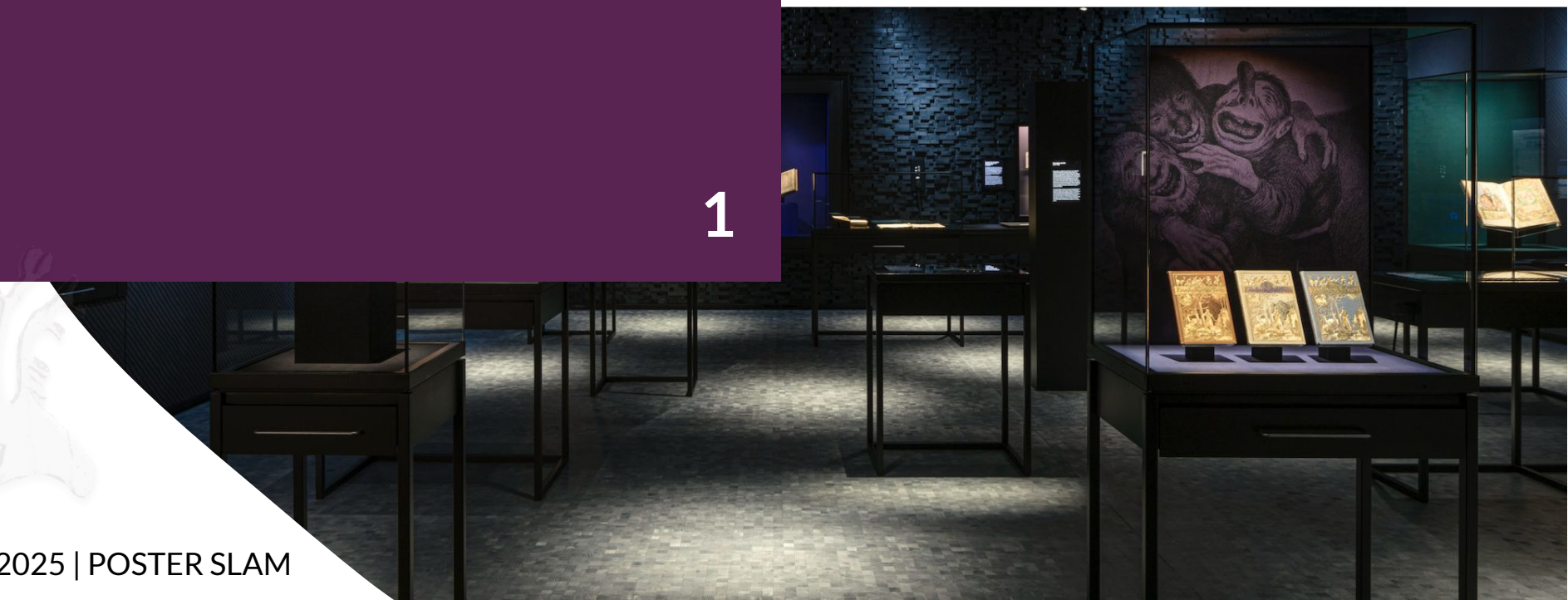
National Library of Norway

IIPC netpreserve.org

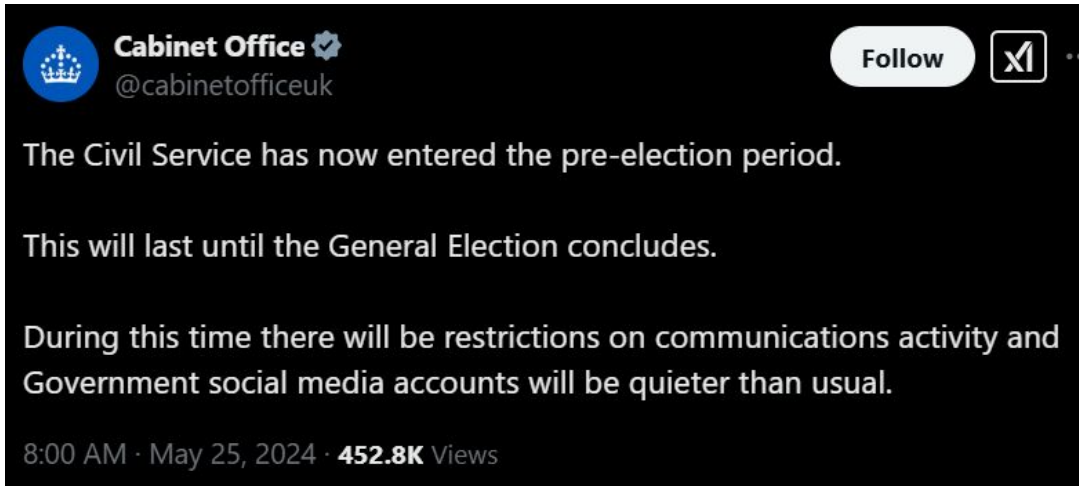
# ‘We Are Now Entering the Pre-election Period’: Experimental Twitter Capture at The National Archives

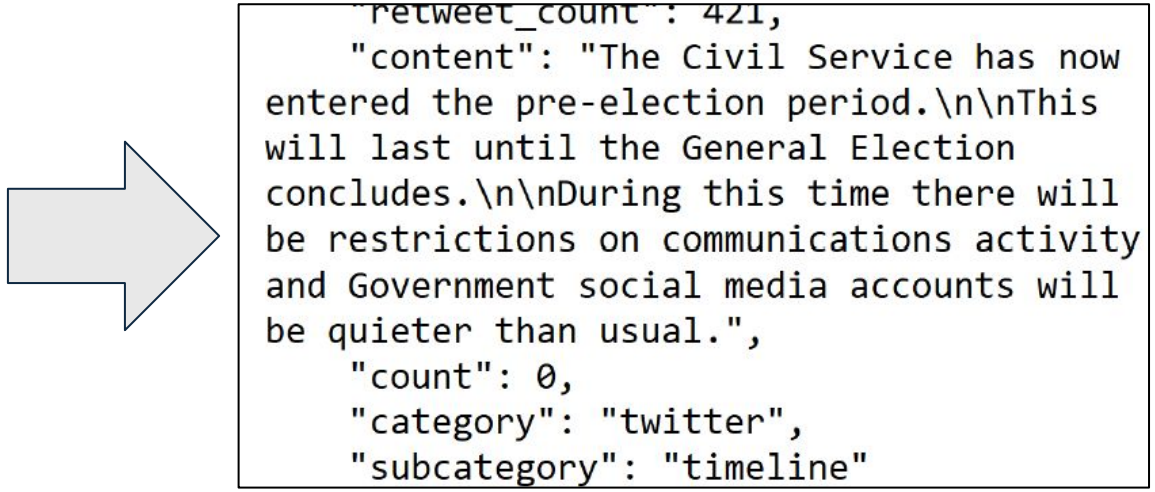
Jake Bickford
The National Archives (UK)

1



# ‘We Are Now Entering the Pre-election Period’: Experimental Twitter Capture at The National Archives (UK)




FOLLOWING THE UNEXPECTED ANNOUNCEMENT OF THE 2024 UK GENERAL ELECTION, WE USED CLIENT-SIDE CAPTURE WITH GALLERY-DL TO RAPIDLY FILL GAPS IN OUR SOCIAL MEDIA COLLECTION THAT HAD BEEN CAUSED BY CHANGES TO THE TWITTER API

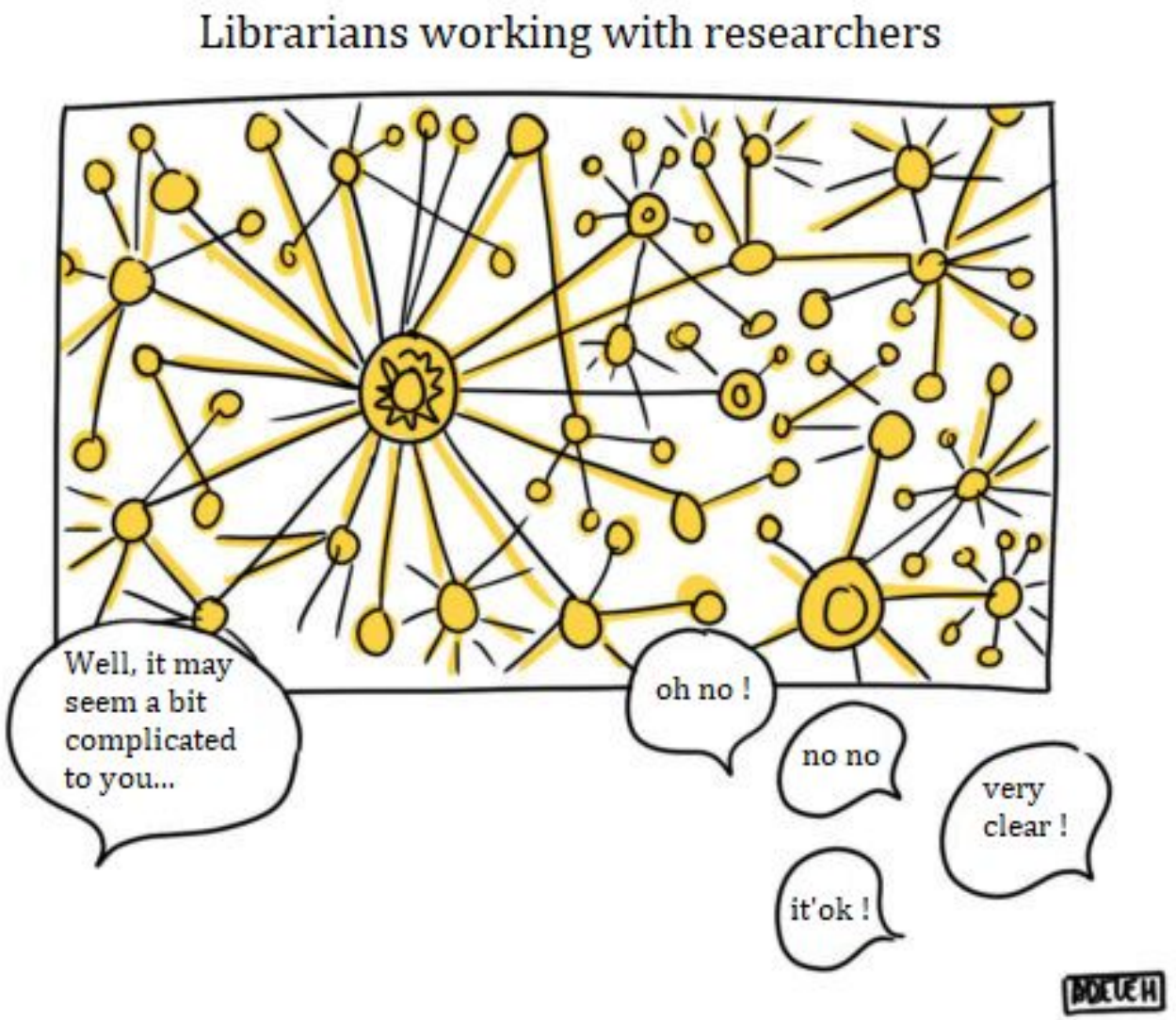
National Library of Norway

IIPC netpreserve.org

# The BnF DataLab Services and Tools for Researchers Working on Web Archives

Sara Aubry & Dorothée Benhamou-Suesser
National Library of France

2




Source :
https://ffl.hypotheses.org

National Library of Norway

IIPC netpreserve.org

# Designing Art Student Web Archives

Katherine Martinez
The New School

3

# Designing Art Student Web Archives

## RESULTS FROM A 2 YEAR STUDY OF WORKFLOWS

Staff support

Project Tracking Database

Archive-It organization

Finding Aid description

Accession documentation (accruals?!)

THE NEW SCHOOL

Katie Martinez
The New School Archives
New York, New York

National Library of Norway

IIPC netpreserve.org

# Next Steps Towards a Formal Registry of Web Archives for Persistent and Sustainable Identification

Eld Zierau
Royal Danish Library

4

# Next Steps Towards a Formal Registry of Web Archives for Persistent and Sustainable Identification

## Why do we need a registry?

A registry can bridge between web archives & current/future researchers


Sustainable access to resources

## Minimum information!

Example ISNI

Merge

New domain

Web Archives Registry

| Formal Id A-Z | Name A-Z | Domain A-Z | Valid from A-Z | Valid to A-Z |
|---|---|---|---|---|
| 1000 0000 0168 4170 | sh Web Archive | nli.ie | 01-01-2011 | |
| 1000 0000 0168 4177 | NLA Australia | pandora.nla.gov.au | 01-01-2004 | 01-01-2012 |
| 1000 0000 0168 4177 | Australian Web Archive | webarchive.nla.gov.au | 01-01-2009 | |
| 1000 0000 0168 4182 | Netarkivet | netarkivet.dk | 01-07-2005 | 01-07-2035 |
| 1000 0000 0168 4182 | Danish Web Archive | kb.dk | 01-07-2035 | |

## Other information elsewhere?

Alternative domains, e.g. for Archive.Today
- archive.ph
- archive.md
- …

Service path, https://arquivo.pt/wayback

Service type, e.g. Access

Trailing characters to date, e.g. im_ for image

Case sensitive lookup

URL-pattern, e.g. https://archive-it.org/home/nli/{4-14-digit YYYYMMDDhhmmss}/{url}

## What is the challenges?

- Web archives will merge National Library of Australia
- Web archives will change URL domain (National Library of Ireland 2019)
- Web archives will change access URLs (Royal Danish Library 2030)

## Where to use it?

- when a reference points to an archive URL that does not work
- to update resolver to PWIDs

## Next steps

- learn more at the poster
- give feedback
- come to IIPC webinar this year



National Library of Norway

IIPC netpreserve.org

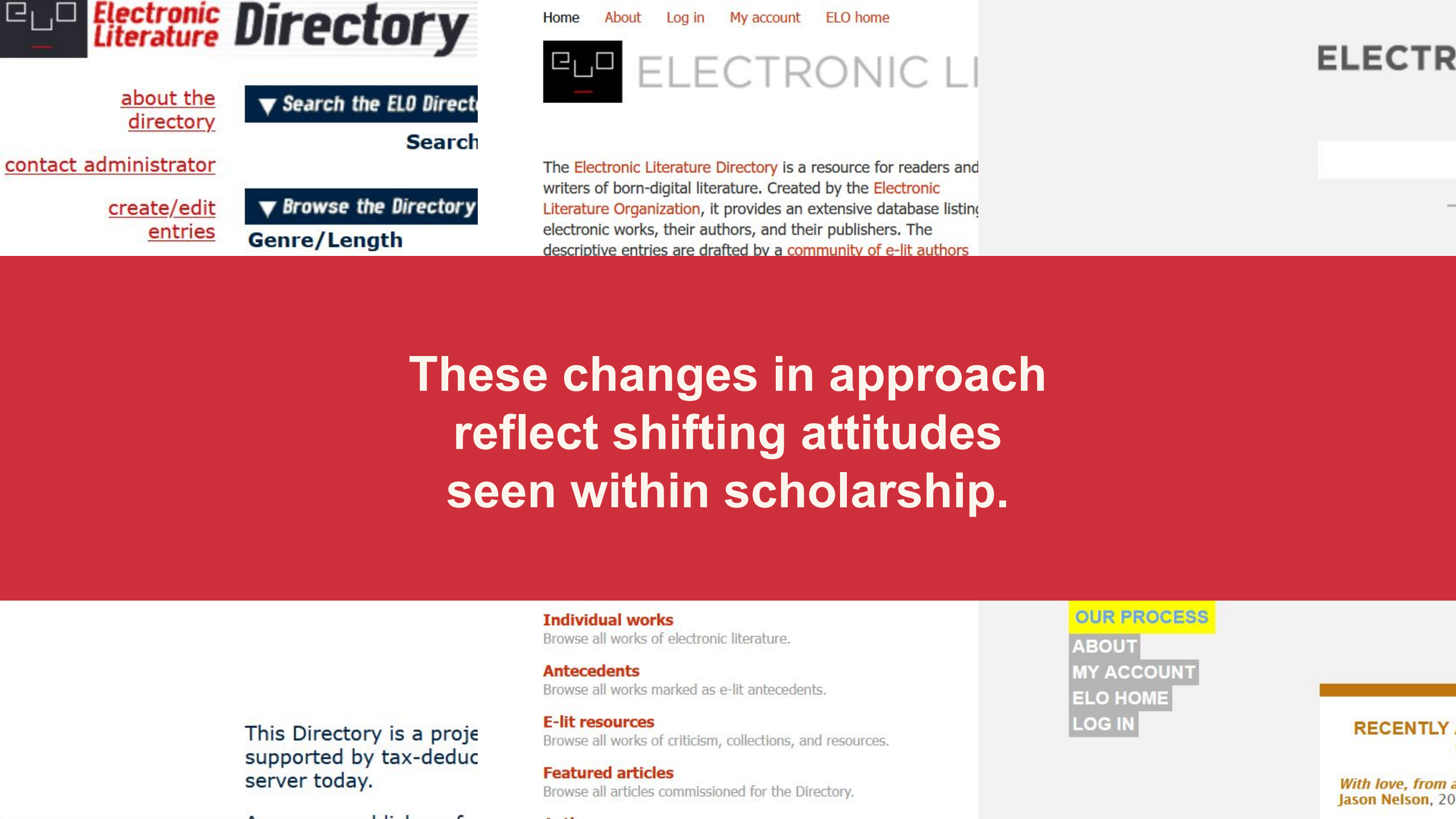
National Library of Norway

IIPC netpreserve.org

# Using Web Archives to Construct the History of an Academic Field

Tegan Pyke
University of Bergen

5


Electronic Literature Directory
about the directory
contact administrator
create/edit entries
Search the ELO Direct
Search
Browse the Directory
Genre/Length
Home
About
Log in
My account
ELO home
ELECTRONIC LI
The Electronic Literature Directory is a resource for readers and writers of born-digital literature. Created by the Electronic Literature Organization, it provides an extensive database listing electronic works, their authors, and their publishers. The descriptive entries are drafted by a community of e-lit authors
ELECTR
These changes in approach reflect shifting attitudes seen within scholarship.
This Directory is a proje
supported by tax-deduc
server today.
Individual works
Browse all works of electronic literature.
Antecedents
Browse all works marked as e-lit antecedents.
E-lit resources
Browse all works of criticism, collections, and resources.
Featured articles
Browse all articles commissioned for the Directory.
OUR PROCESS
ABOUT
MY ACCOUNT
ELO HOME
LOG IN
RECENTLY
With love, from a
Jason Nelson, 20

National Library of Norway

IIPC netpreserve.org

# Consortium on Electronic Literature (CELL)

Hannah Ackermans
University of Bergen

6


CELL
CONSORTIUM ON ELECTRONIC LITERATURE

National Library of Norway

IIPC netpreserve.org

# Arquivo.pt Annual Awards: A Glimpse

Daniel Gomes
Arquivo.pt

7

# Arquivo.pt Award 2025 (annual): open applications

10 000€ FOR THE WINNER FOT

WORKS THAT MAKE USE OF ARQUIVO.PT

APPLICATIONS UNTIL 6 MAY 2025

HIGH PATRONAGE OF PRESIDENT OF PORTUGAL

SHARE!

KNOW MORE: ARQUIVO.PT/AWARD

National Library of Norway

IIPC netpreserve.org

# Arquivo.pt Api/Bulk Access and Its Usage

Vasco Rato & Daniel Gomes
Arquivo.pt

8

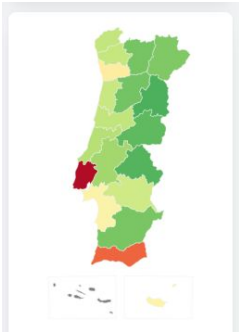
# Arquivo.pt Api/Bulk Access and Its Usage

- Data is only useful if it's accessible and searchable!

arquivo.pt

- Text Search API
- Image Search API
- CDX Server API
- Memento API
- CDX Bulk Download





**GlórIA LLM**



National Library of Norway

IIPC netpreserve.org

National Library of Norway

IIPC netpreserve.org

# Failed Capture or Playback Woes? A Case Study in Highly Interactive Web-Based Experiences

Mari Allison
Smithsonian Libraries and Archives

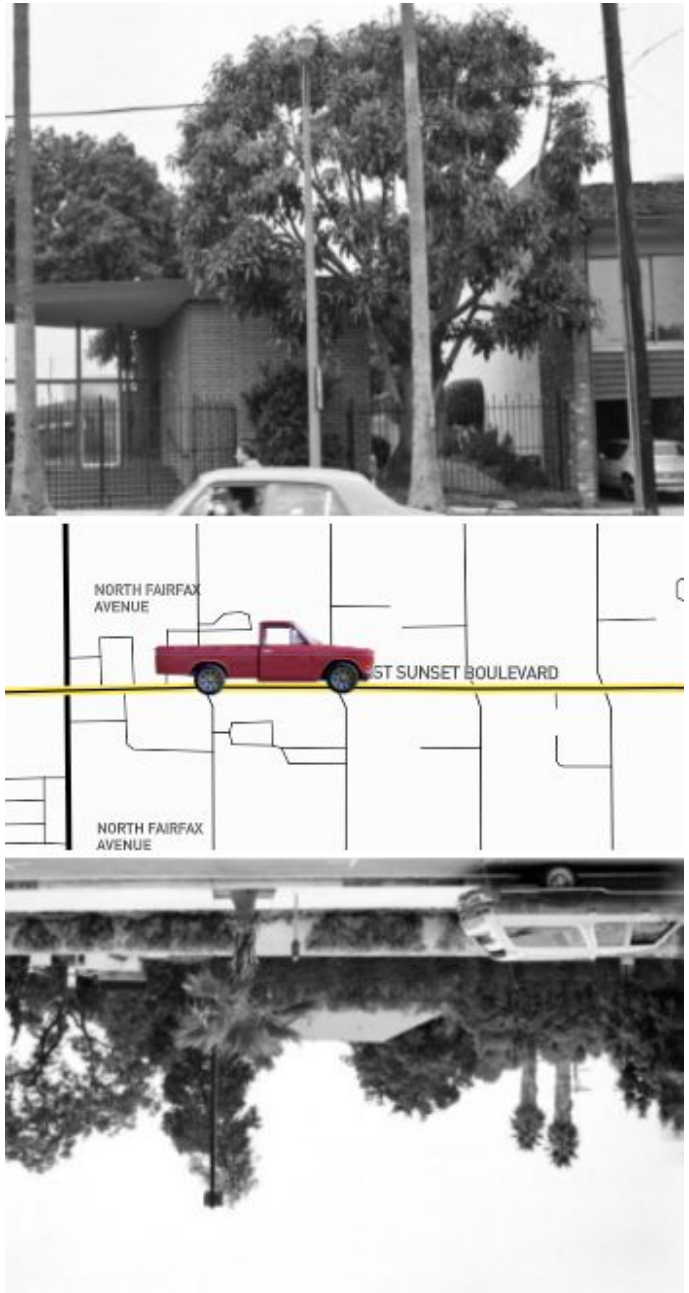
9

# Breaking the cycle

**Experiment with a new capture technology**

**Export WARC and upload it to our Archive-It**

**Capture does not appear to work when viewed**






National Library of Norway

IIPC netpreserve.org

National Library of Norway

IIPC netpreserve.org

# HAWathon: Participants Experience

Ingeborg Rudomino & Anamarija Ljubek
National and University Library in Zagreb

10

National Library of Norway

IIPC netpreserve.org

# Supporting Best Practices for Archiving Social Media by Heritage Institutions in Flanders (and Beyond)

Ellen Van Keer & Katrien Weyns
meemoo, Flemish Institute for Archives

11

KU LEUVEN

KADOC

Community of Practice
Social Media Archiving

### Praktijkwijzer Sociale Media Archiveren

Toolbox

https://kennisbank.meemoo.be/toolbox/praktijkwijzer-sociale-media-archiveren

Deze praktijkwijzer leidt naar praktische handleidingen en workflows voor socialemedia-archivering die nog relevant zijn en zo goed als mogelijk up-to-date worden gehouden door de archiverende community. We moedigen aan om eigen handleidingen of relevante links toe te voegen aan...

### Sociale media archiveren met Browsertrix

Toolbox

https://kennisbank.meemoo.be/toolbox/sociale-media-archiveren-met-browsertrix

In het kader van het project Best practices voor de archivering van sociale media in Vlaanderen en Brussel werden verschillende tools getest om socialemediaplatformen te archiveren. Deze handleiding beschrijft de tool Browsertrix die gebruikt kan worden voor Facebook en...

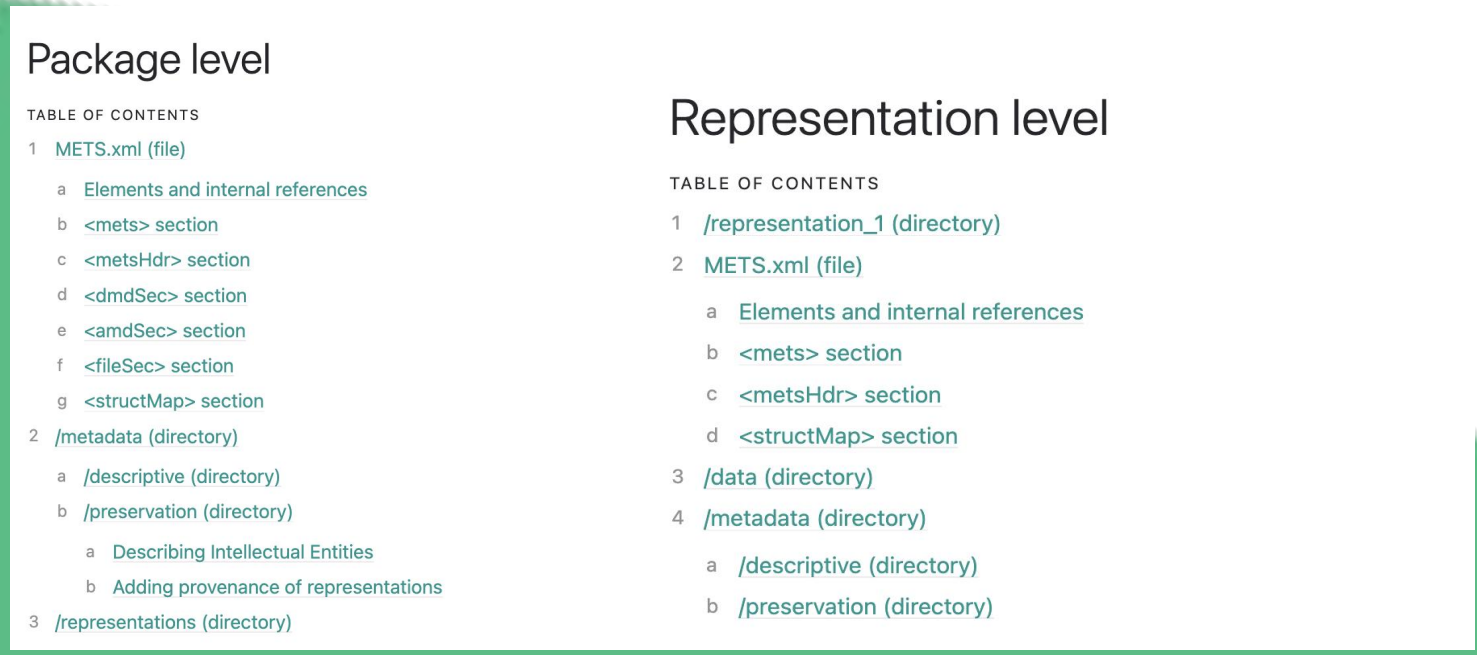
### Sociale media archiveren met Webrecorder

Toolbox

https://kennisbank.meemoo.be/toolbox/sociale-media-archiveren-met-webrecorder

In het kader van het project Best practices voor de archivering van sociale media in Vlaanderen en Brussel werden verschillende tools getest om socialemediaplatformen te archiveren. Deze handleiding beschrijft de tool Webrecorder voor het archiveren van sociale media.

tutorials on tools and workflows
https://meemoo.be/en/knowledge-base



### Package level

TABLE OF CONTENTS

1 METS.xml (file)
- a Elements and internal references
- b <mets> section
- c <metsHdr> section
- d <dmdSec> section
- e <amdSec> section
- f <fileSec> section
- g <structMap> section

2 /metadata (directory)
- a /descriptive (directory)
- b /preservation (directory)
  - a Describing Intellectual Entities
  - b Adding provenance of representations

3 /representations (directory)

### Representation level

TABLE OF CONTENTS

1 /representation_1 (directory)
2 METS.xml (file)
- a Elements and internal references
- b <mets> section
- c <metsHdr> section
- d <structMap> section

3 /data (directory)
4 /metadata (directory)
- a /descriptive (directory)
- b /preservation (directory)

e-ARK based SIP specification https://developer.meemoo.be

# Supporting best practices for archiving social media by heritage institutions in Flanders (and beyond)

meemoo
FLEMISH INSTITUTE FOR ARCHIVES

Ellen Van Keer ellen.vankeer@meemoo.be
Katrien Weyns katrien.weyns@kuleuven.be

National Library of Norway

IIPC netpreserve.org

# Planning Web Archiving Within a Four-Year Scope: Making the New Collection Plan for the Years 2025-2028 in the National Library of Finland

Sanna Haukkala
National Library of Finland

12

# Planning Web Archiving Within a Four-Year Scope

We have made a new collection plan for years 2025–2028

- covers both web archiving and e-legal deposit

🚧 Also, a new policy document (what we should collect) is in the works 🚧

Interested about the comparison results or the actual plan? Come chat with me!

National Library of Norway

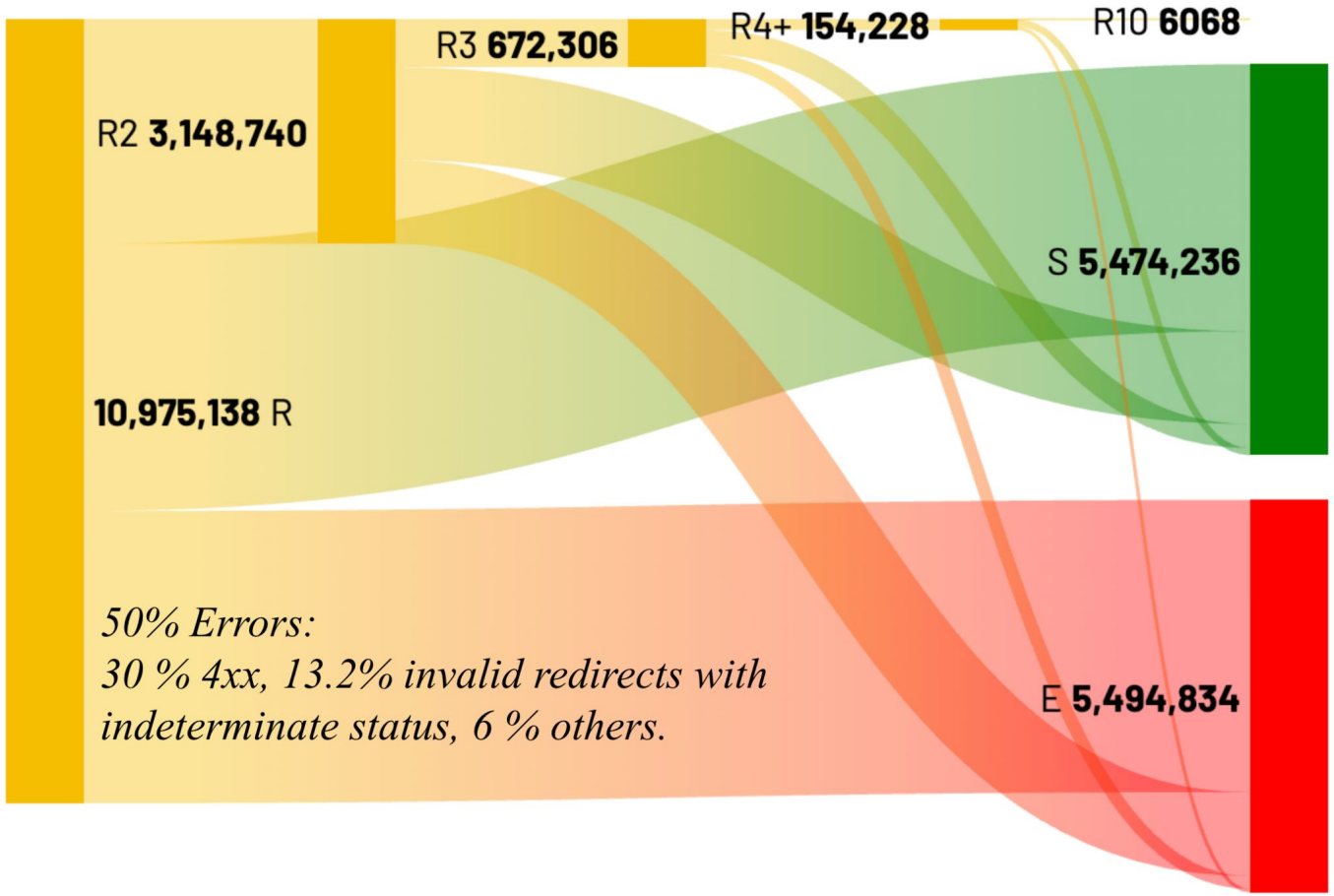
IIPC netpreserve.org

# Redirects Unraveled: From Lost Links to Rickrolls

Kritika Garg[1], Sawood Alam[2], Michele Weigle[1], Michael Nelson[1], Mark Graham[2], & Dietrich Ayala[3]
[1]Old Dominion University; [2]Internet Archive; [3]Filecoin Foundation

13

# Redirects Unraveled: From Lost Links to Rickrolls







National Library of Norway

IIPC netpreserve.org

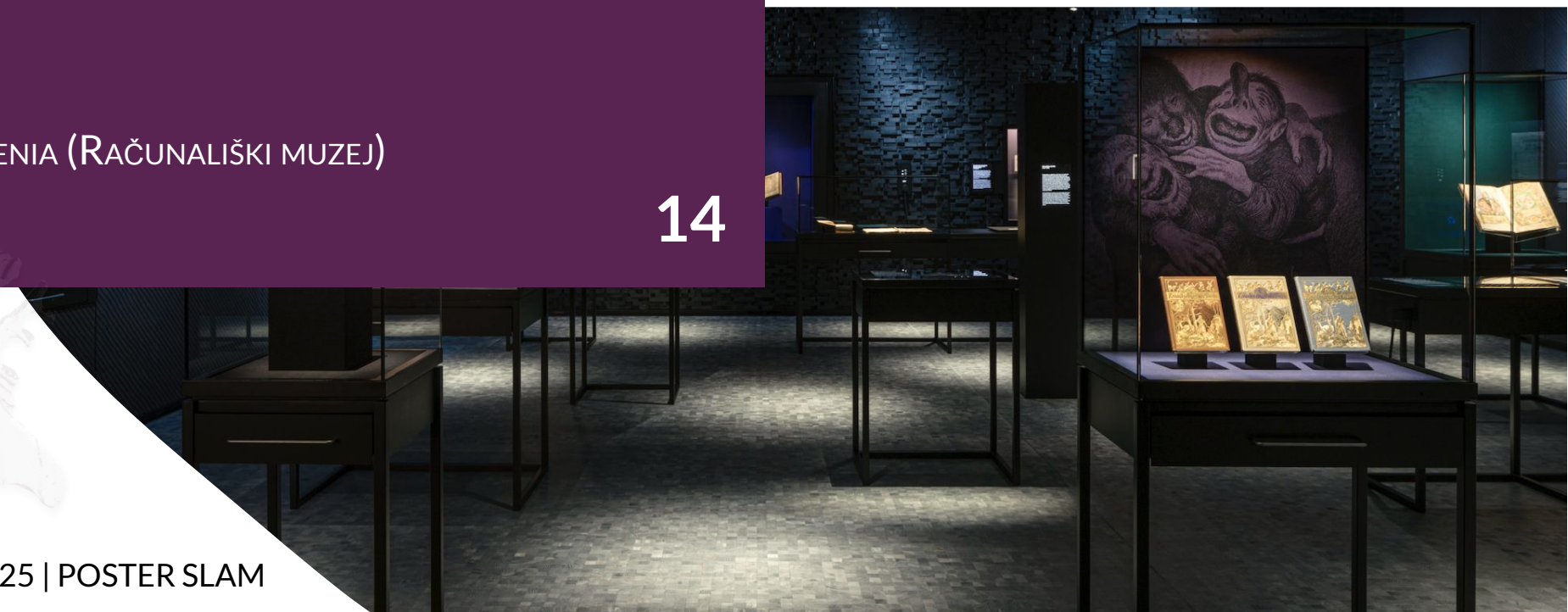
National Library of Norway

IIPC netpreserve.org

# Use of Screenshots as a Harvesting Tool for Dynamic Content and Use of AI for Later Data Analysis

Gaja Zornada & Boštjan Špetič
Computer History Museum Slovenia (Računališki muzej)

14



## Step 1

### Capture

- Screenshots are taken from dynamic or ephemeral content sources
- Ensuring contextual accuracy and relevance

## Step 2

### Data Extraction with AI

- Image Recognition: Identifying key visual elements
- Natural Language Processing (NLP): Extracting and interpreting text
- Optical Character Recognition (OCR): Converting embedded text into searchable data

## Step 3

### Data Utilization

Extracted data can be categorized and stored for:

- Metadata analysis
- User interaction studies
- Archival documentation

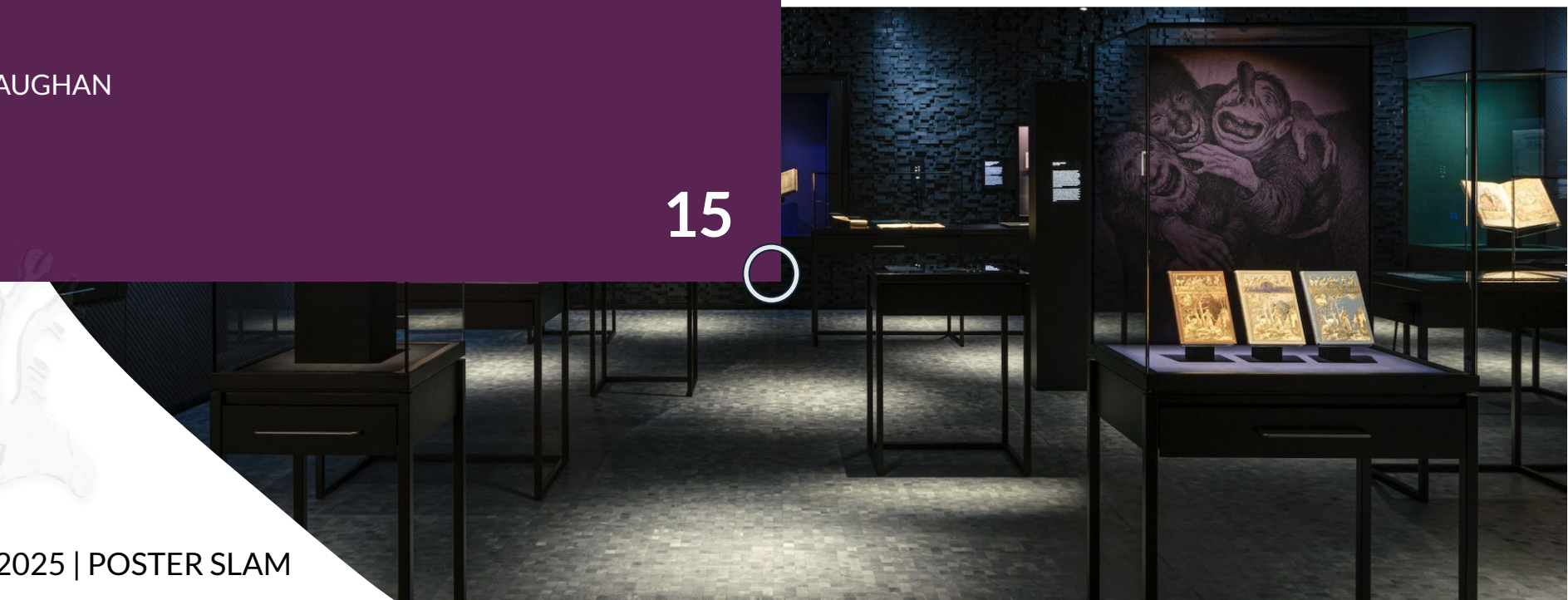
National Library of Norway

IIPC netpreserve.org

# Asynchronous and Modular Pipelines for Fast WARC Annotation

Pedro Ortiz Suarez & Thom Vaughan
Common Crawl Foundation

15



# A Prototype Pipeline for WARC Annotation

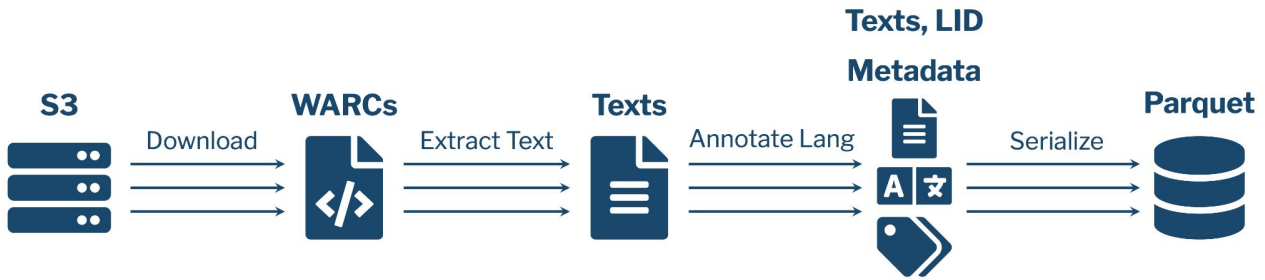



Figure 1. A schema of the first prototype for our pipeline. represents the AWS S3 bucket where the Common Crawl Data is stored, represents WARC files, represents text extracts of WARCs, represents language tags, represents metadata coming from WARC headers and represents Parquet files. Finally the arrows represent parallel processing.

- We inspire ourselves on existing pipelines from the NLP and the web archiving communities, and develop our own prototype of a pipeline for WARC annotation
- The first experimental prototype aims to be efficient, modular, open-source and user-friendly so that little knowledge of large-scale data processing is needed
- This first prototype focuses on organizing and compiling some of the hectic developments of data pipelines in the NLP/ML space and bringing them to the web archiving community
- Our prototype focuses on textual data, as Common Crawl is a text-only archive
- We note that the architecture of our prototype is extensible and borrows ideas from NLP pipelines that have already been extended to non-textual archives
- For this first prototype we only implement text extraction and language annotations.

National Library of Norway

IIPC netpreserve.org

# Politely Downloading millions of WARC Files Without Burning the Servers Down

Pedro Ortiz Suarez, Thom Vaughan, & Greg Lindahl
Common Crawl Foundation

16

# cc-downloader


- We use **Exponential Backoff** and **Jitter** to develop our own download client for **Common Crawl, cc-downloader.**
- We develop it in the **Rust Programming Language** and release pre-compiled binaries for Linux , Mac and Windows (x86-64).
- We release cc- downloader under Apache 2.0 and MIT licenses
- The default configuration can be tweaked by the user
- cc-downloader is distributed as a CLI tool, but we are planning on releasing it as a library and adding python bindings

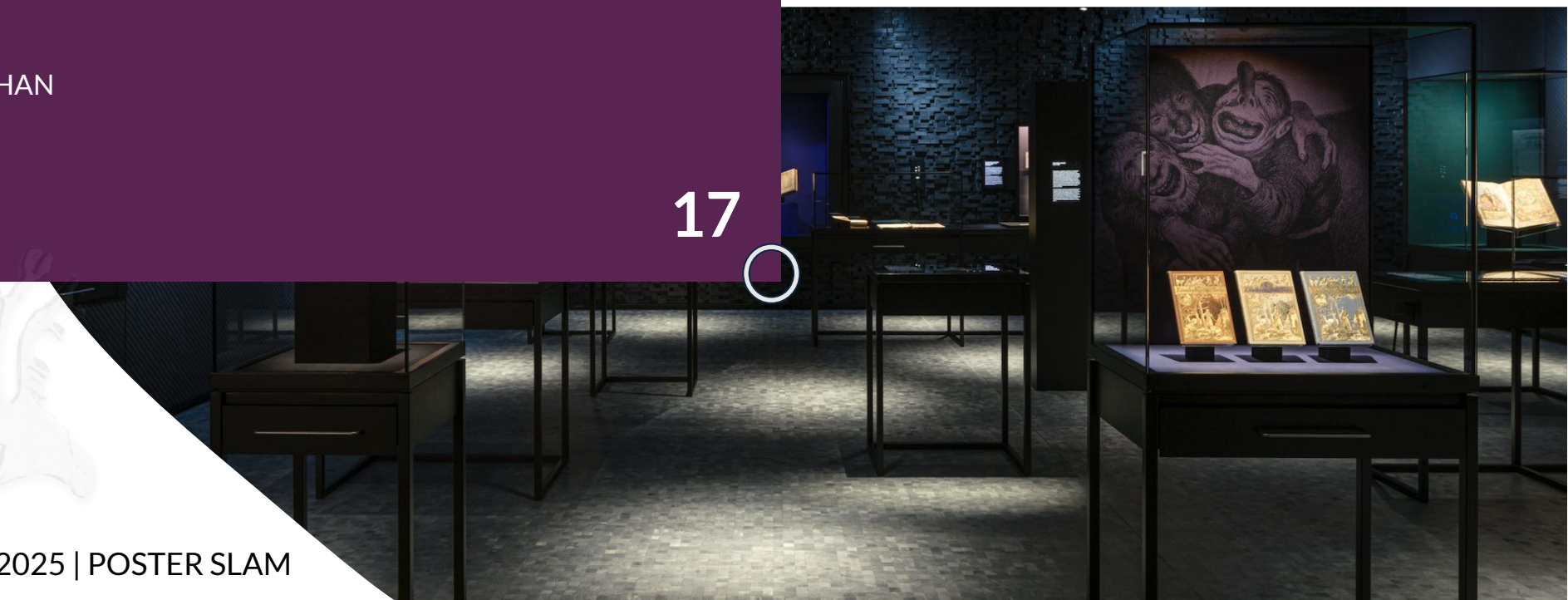
National Library of Norway

IIPC netpreserve.org

# Robots.txt and Crawler Politeness in the Age of Generative AI

Sebastian Nagel & Thom Vaughan
Common Crawl Foundation

17



# The Robots Exclusion Protocol

- 1994 Robots.txt discussed on mailing list
  1996 inofficial RFC proposal

- 2022 RFC 9309 standardizes the REP

- 2025 robots.txt is the most widely used protocol to block crawling by generative AI crawlers and agents
  - Which crawlers ("user-agents")?
  - Does this impact web archiving?

National Library of Norway

IIPC netpreserve.org

# Experiences Switching an Archiving Web Crawler to Support HTTP/2

Sebastian Nagel
Common Crawl Foundation

18

# HTTP/2 Support in a Web Archive Crawler

- HTTP/2 introduced in 2015 as an improvement over HTTP/1.1 (1997)

- 70% of the traffic in the web are already on HTTP/2

  High time to add support for it to our web crawler

  The next version - HTTP/3 - is out since 2022

- Implementation and performance of the upgraded crawler

  WARC captures of HTTP/2 requests and responses