



Digitized Newspapers as Everyday Interdisciplinarity: The Transformation of Historical Scholarship

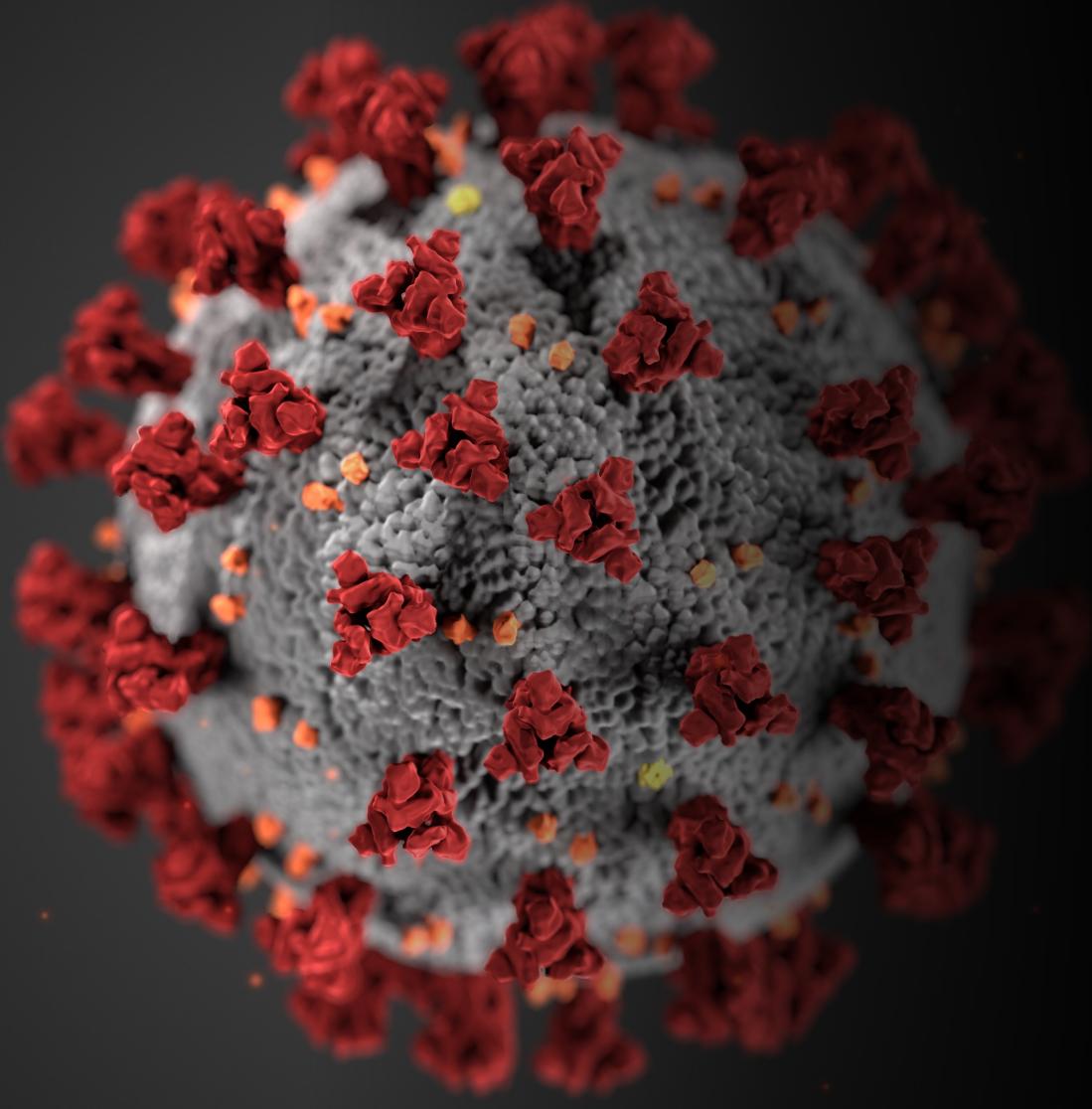
Ian Milligan

University of Waterloo, Canada





All historians are
digital historians



COVID is
accelerating
but not
inventing
trends



How historians imagine how they work

- Literature review, finding historiographical problem
- Identifying primary sources
- Analyzing primary sources
- Writing
- Revision
- Preliminary presentations (conference, article)
- Finished publication (dissertation, book)

How historians *really* work

- Literature review, finding historiographical problem (**shaped by digitized/non-digitized secondary resources**)
- Identifying primary sources (**online finding aids**)
- Analyzing primary sources (**doing keyword searches in databases that you barely understand of documents that you don't know provenance of**)
- Writing/Revision/Publishing





We know archives transform historical work...

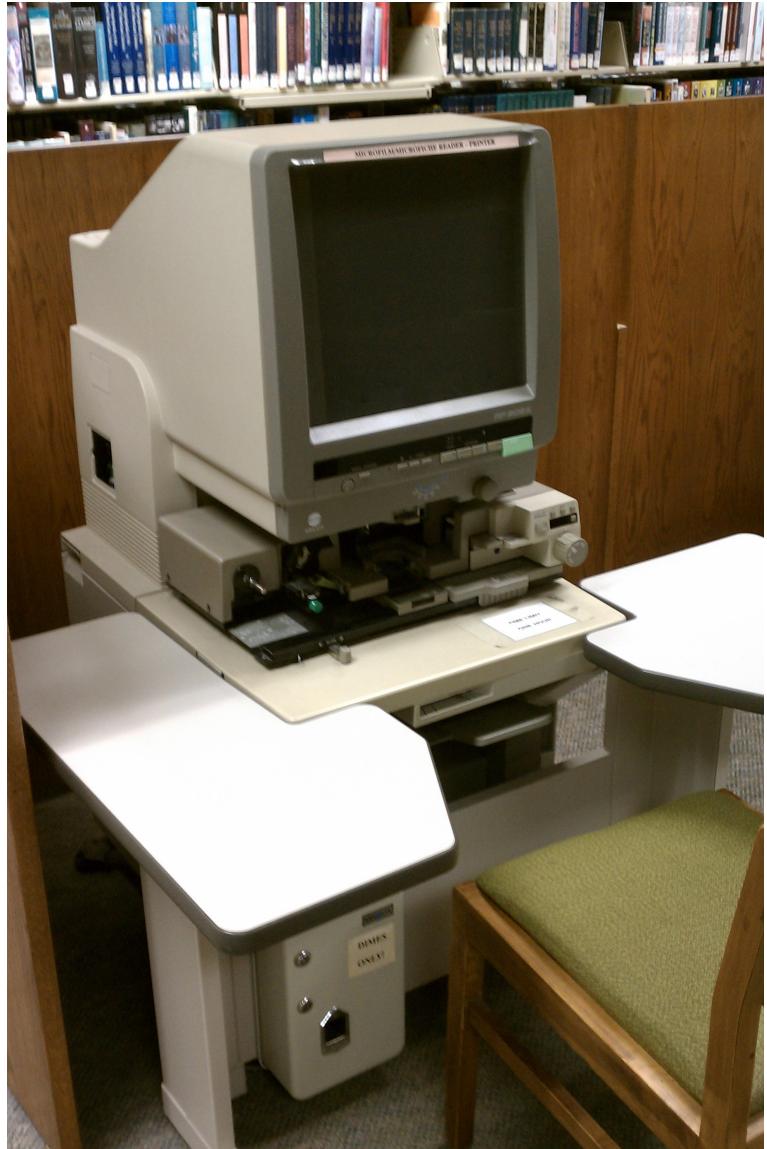
- We are often vaguely cognizant of the role that archives play in shaping our histories, but we still tend to treat them as “neutral and unproblematic reservoirs of historical fact.” (Walsham)
- Archivists and historians share common origins, but we have diverged
- “Any visit by a historian to an archival institution is now an exercise in interdisciplinarity.” (Blouin and Rosenberg, *Processing the Past*)

A close-up photograph of a person's hand holding a black magnifying glass over a dark laptop keyboard. The magnifying glass is focused on the keys, creating a bright, blurred reflection in its lens. The background is dark and out of focus.

**Just as archives mediate the
past, so do the workflows
that we use in the digital age.**

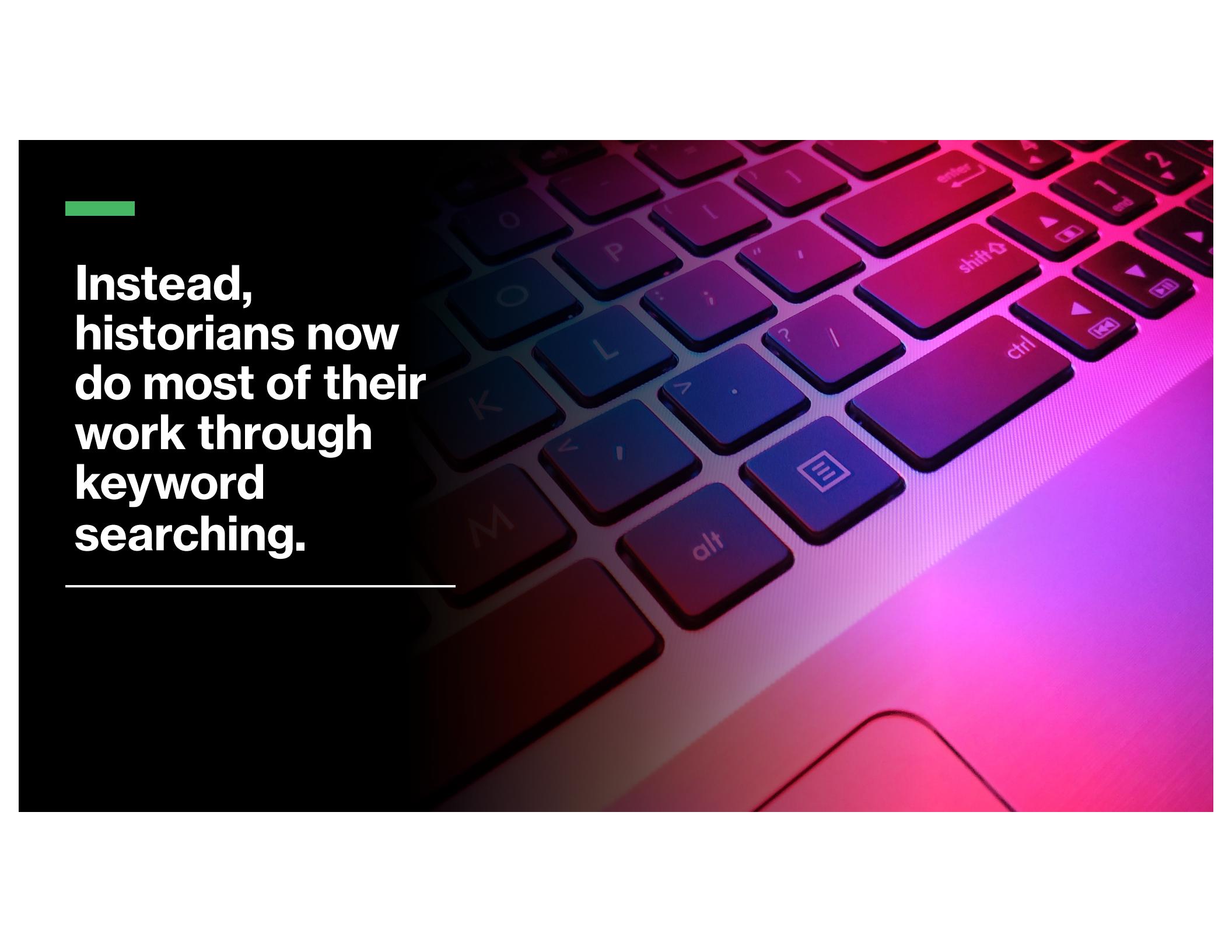
So, is using a digitized newspaper an exercise in interdisciplinarity?



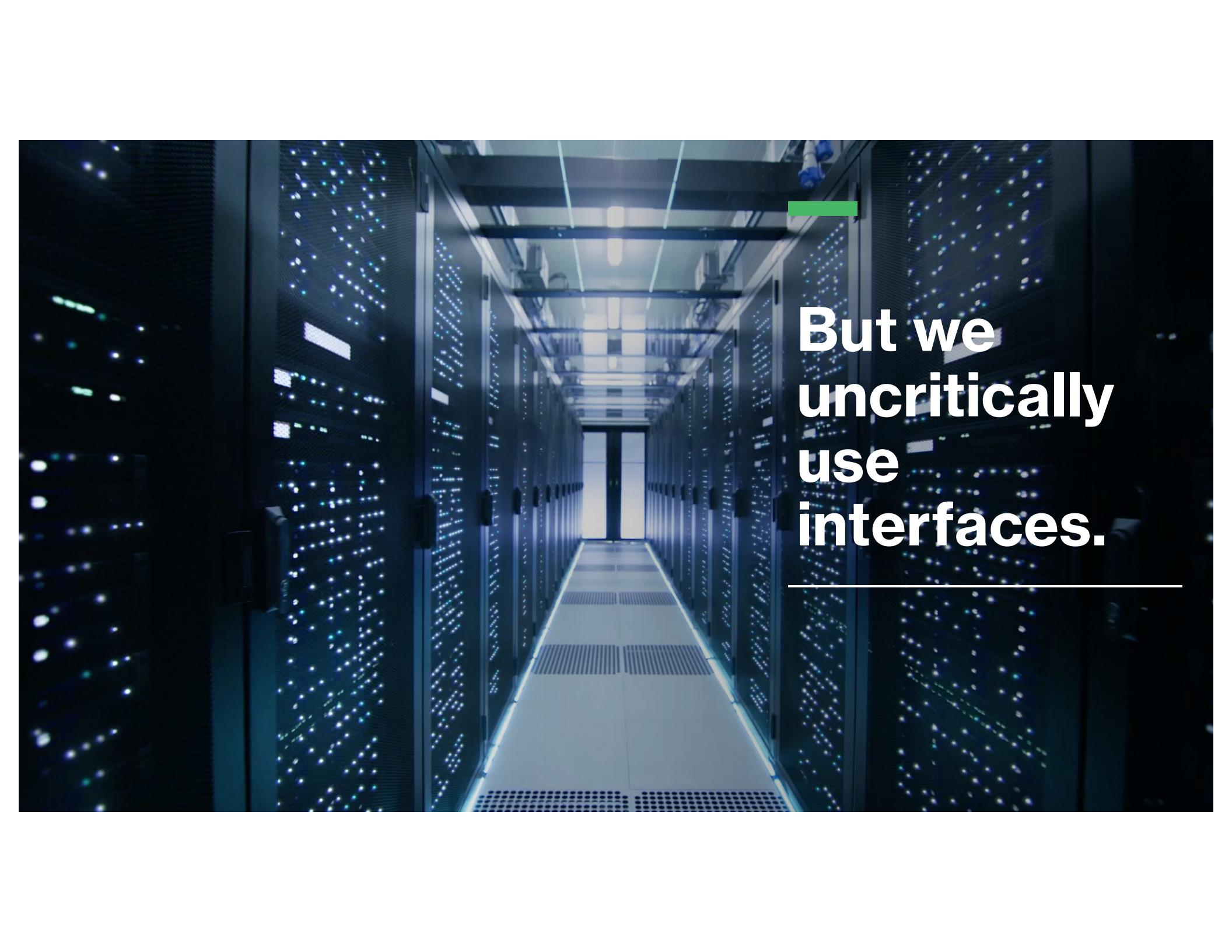


How did historians previously use newspapers?

- Microfilm reels
- Sit in a dark basement, crank forward through pages looking for relevant documents
- Time consuming, boring (good chance to listen to music)
- But you learn a *lot* about context
 - i.e. a global event that happens, the tenor of advertisements, the relative placement of columnists or articles.



**Instead,
historians now
do most of their
work through
keyword
searching.**



**But we
uncritically
use
interfaces.**

Let me use a Canadian example.

PQ Advanced Search: Your search returned no results - ProQuest Historical Newspapers: Toronto Star - ProQuest

Access provided by UNIVERSITY OF WATERLOO

ProQuest Historical Newspapers: Toronto Star

TUESDAY, OCTOBER 15, 1917

Your search for Kingston AND Roosevelt found 0 results.

Please modify your search and try again. [Search tips](#)

Advanced Search [Command Line](#) Recent searches Field codes Search tips University of Waterloo Library

Kingston in Anywhere

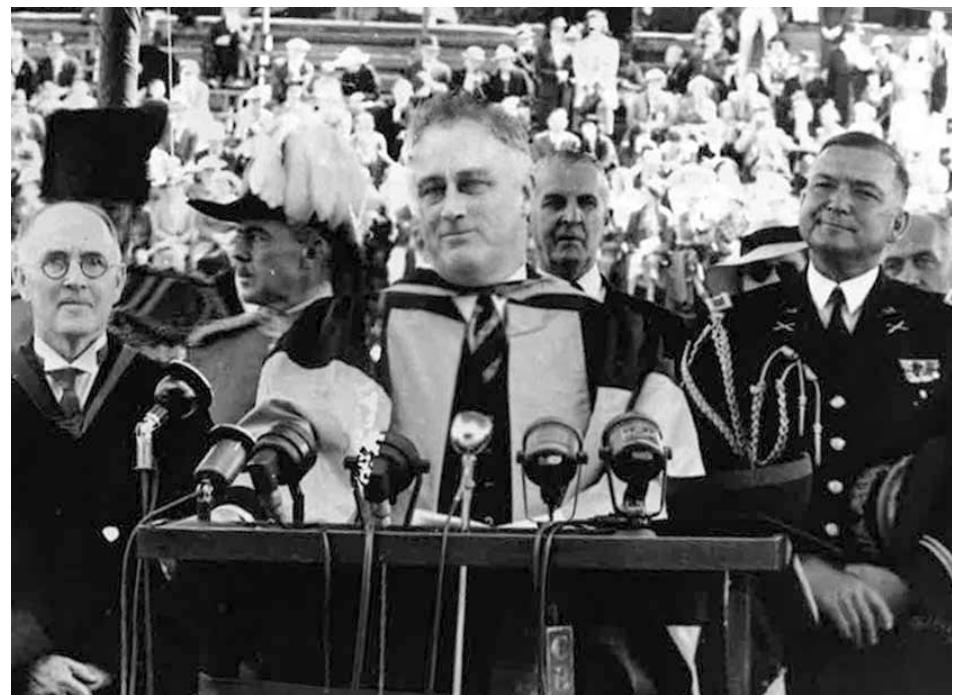
AND Roosevelt in Anywhere

[Add a row](#)

Publication date: Specific date range... Start 18 1938

End 24 1938

[Search](#) [Clear form](#)



Choose an issue to view

1938 ▾

✓ December

November

October

September

July

June

May

April

March

February

January

Dec 31, 1938 ▾

View issue

Issue content

Select 1-3:

Page 1

1 **Toronto**

[Details](#)

[Search within this issue](#)

0-1971); Toronto, Ontario [Toronto, Ontario]30 Ju



Missing

- For some reason, August 1938 is missing (the 1 September 1938 issue has reference in the “letters to the editors” about articles written in the past week, so there were apparently issues)
- Crucially, you wouldn’t know that if you just did keyword searches.
- This was the first month I went looking for, because of FDR, what other gaps are there?

Other limitations of search that historians may not know of

- The text that is being searched is created using **optical character recognition**, or OCR
 - ProQuest's implementation stems from *Pages of the Past*, an innovative project that saw the *Toronto Star* the first fully digitized newspaper in the world
 - But it's a commercial platform, so correcting OCR is difficult
 - Doesn't catch line-break hyphenation
 - **Cutting-edge OCR would have best-case scenario of 98%; even that leads to 50 incorrect characters on an average page of 500 words; word accuracy would be around 90%**

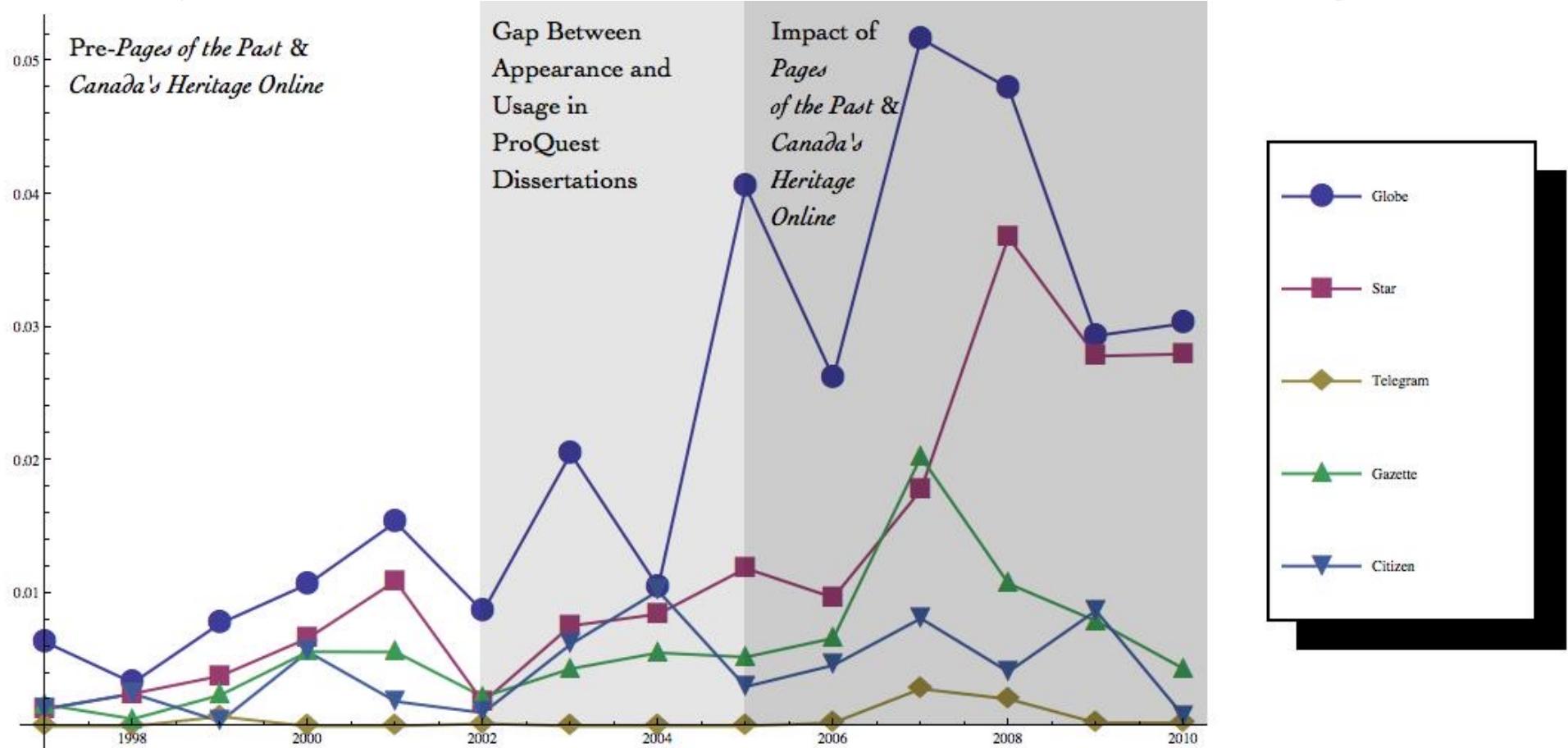
Skimming is nearly impossible

The screenshot shows a web browser window for the ProQuest Historical Newspapers: Toronto Star. The URL is search-proquest-com.proxy.lib.uwaterloo.ca/hnptorontostar/publicationsb. The page title is "ProQuest Historical Newspapers: Toronto Star". The top navigation bar includes links for Basic Search, Advanced Search, Publications, and Change databases. A green banner at the top right indicates "Access provided by UNIVERSITY OF WATERLOO". Below the banner, a newspaper masthead for "TORONTO DAILY" dated "TUESDAY, OCTOBER 15, 1957" is visible. The main content area is titled "Publication Search". A search bar contains the query "In title: Toronto Daily". To the left, a sidebar titled "Publication date" shows a histogram for the years 1894 - 2010 (decades). It includes a "Enter a date range" input field and an "Update" button. The main search results show "4 publications": 1. Evening Star (1894-1900); Toronto, Ontario (Full text available) 2. Toronto Daily Star (1900-1971); Toronto, Ontario (Full text available) 3. Toronto Star (1971-2009); Toronto, Ontario (Full text available)

The background of the image is a close-up, high-contrast photograph of a weathered metal surface. The texture is rough and uneven, with large, irregular patches of dark green and blue-green patina (mossy scale) overlying a darker, more metallic base. Some areas show signs of wear, with the underlying metal visible through the edges of the patina. The lighting is dramatic, highlighting the depth of the crevices and the variations in the color of the corrosion.

And
digitization is
uneven

In Canada



In other words

The more something is digitized the more it is used (i.e. the *Toronto Star* and the *Globe and Mail* are used far more than before; the *Toronto Telegram* is almost never used)

The mediation of a source impacts its use

The Impact of this Medium Shift

- We now interact primarily through keyword search (i.e. the system forces us more or less to do this)
- We don't fully understand the construction of this database.
- The text is inaccessible to do transformative digital scholarship with.
- **Yet we still cite it all the same: Pages of the Past, ProQuest, Clipping File, Microfilm; yet each system dramatically impacts our work and the way we understand the source.**



**If using an archive is an exercise
in interdisciplinarity, perhaps we
should think of using online
platforms the same way?**

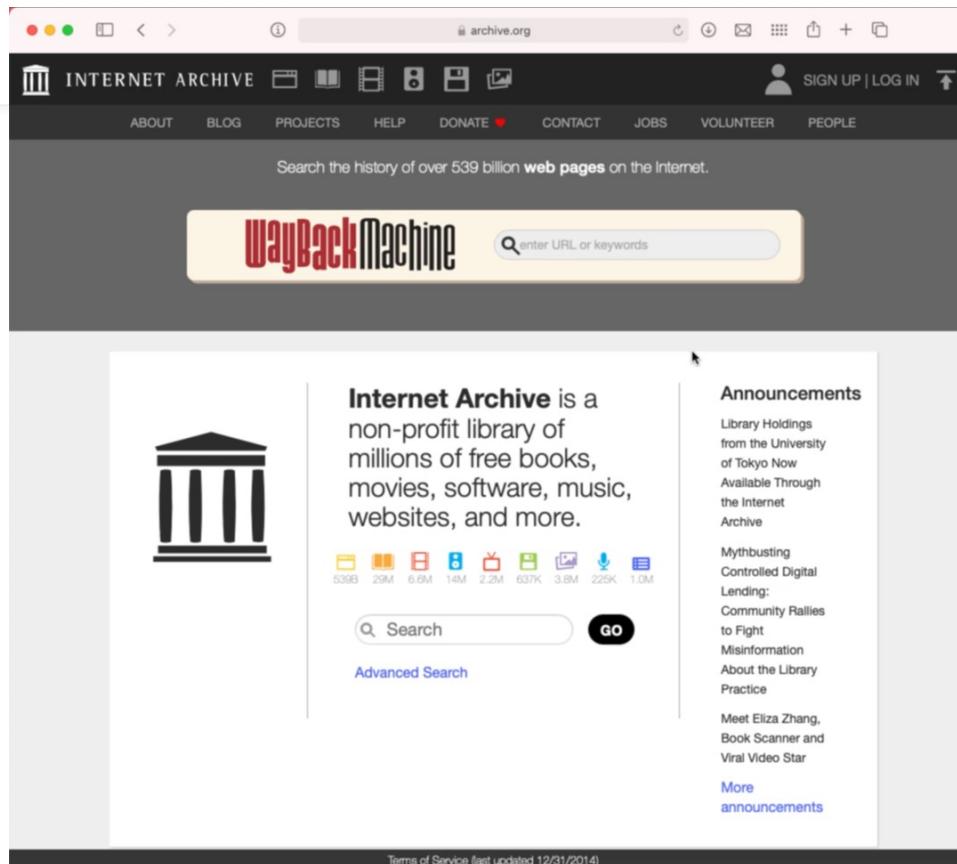


So what can we do?



Come to the NewsEye
Conference!

Embrace Open Platforms and Tools



Embrace Open Platforms and Tools

The screenshot shows a Jupyter Notebook interface with the title "NewsEye-Workbook.ipynb". The notebook contains a single cell with the following code:

```
# Installing All Software necessary for this page
# Move your mouse here and hit the play button on the left <<
!echo "Scroll down to see output in the window"
!echo ""
!echo "Install ia command line tool!"; sudo pip install internetarchive > /dev/null
!echo "Installing jq for processing json..."; apt-get install jq > /dev/null
!echo "Installing parallel to go fast..."; apt-get install parallel > /dev/null
!echo "Installing bc to do command line math..."; apt-get install bc > /dev/null

#if we need to get someone logged into archive.org, here is a way to do it:
# type your archive.org username (email) and password at prompt
#!until ia configure; do echo -e "\n*** Try again! or reset password: https://archive.org/account/forgot-password" > /dev/null; done

!echo "Scroll down"
!ia --help # show what you can now do
!echo "READY!"
```

Below the code cell, the output is displayed:

```
Install ia command line tool!
Installing jq for processing json...
Installing parallel to go fast...
Installing bc to do command line math...
Scroll down
A command line interface to Archive.org.

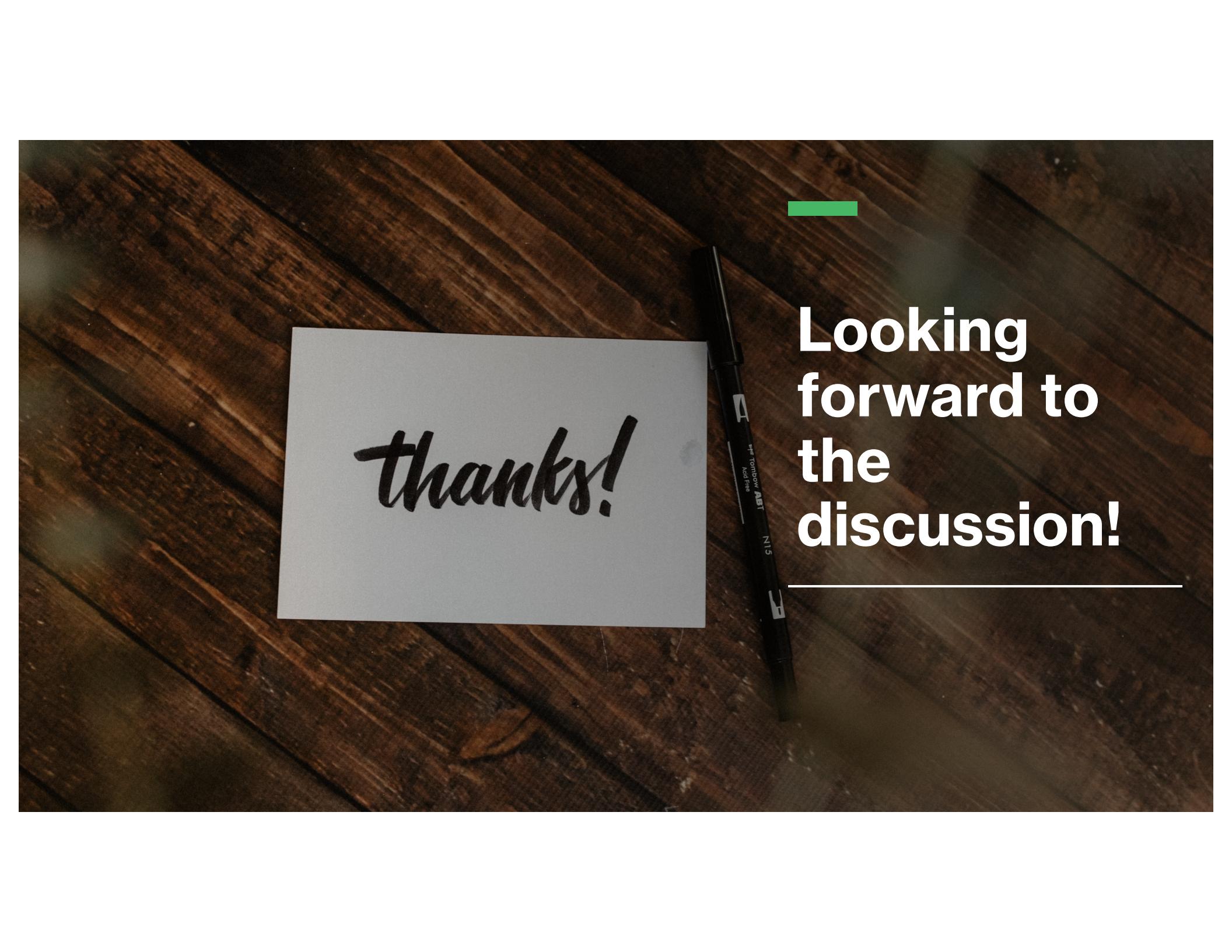
usage:
    ia [--help | --version]
        ia [--config-file FILE] [--log | --debug]
            [--insecure] [--host HOST] <command> [<args>]...

options:
    -h. --help
```



What does this mean?

- If working with an archive we don't really understand is an “exercise in interdisciplinarity,” **isn’t working with a newspaper that we don’t really understand one too?**
- **Open platforms as much as possible** (I know copyright is a thing!)
- **Finding ways to translate knowledge;** I think historians can get hit by a blast of the obvious when they realize that they need to think about their platforms. But they rarely do.



A white rectangular card with the word "thanks!" written in black cursive ink is positioned on a dark, horizontally-grained wooden surface. A black Tombow ABT N15 brush pen lies next to the card. The background is a dark wood texture. There are two small horizontal bars: a thin green bar above the text and a thin white bar below it.

Looking
forward to
the
discussion!