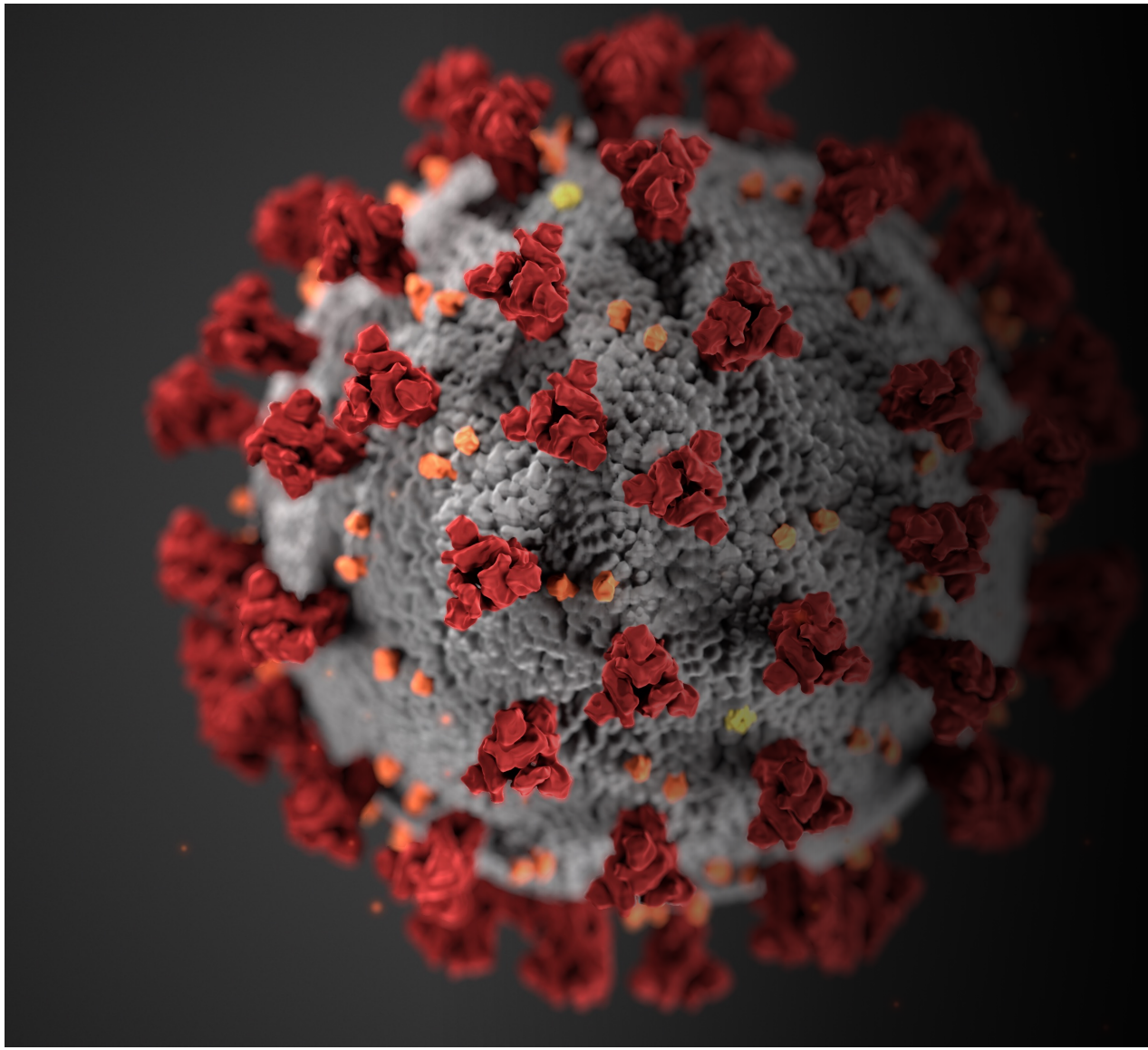
Digitized Newspapers as Everyday Interdisciplinarity: The Transformation of Historical Scholarship

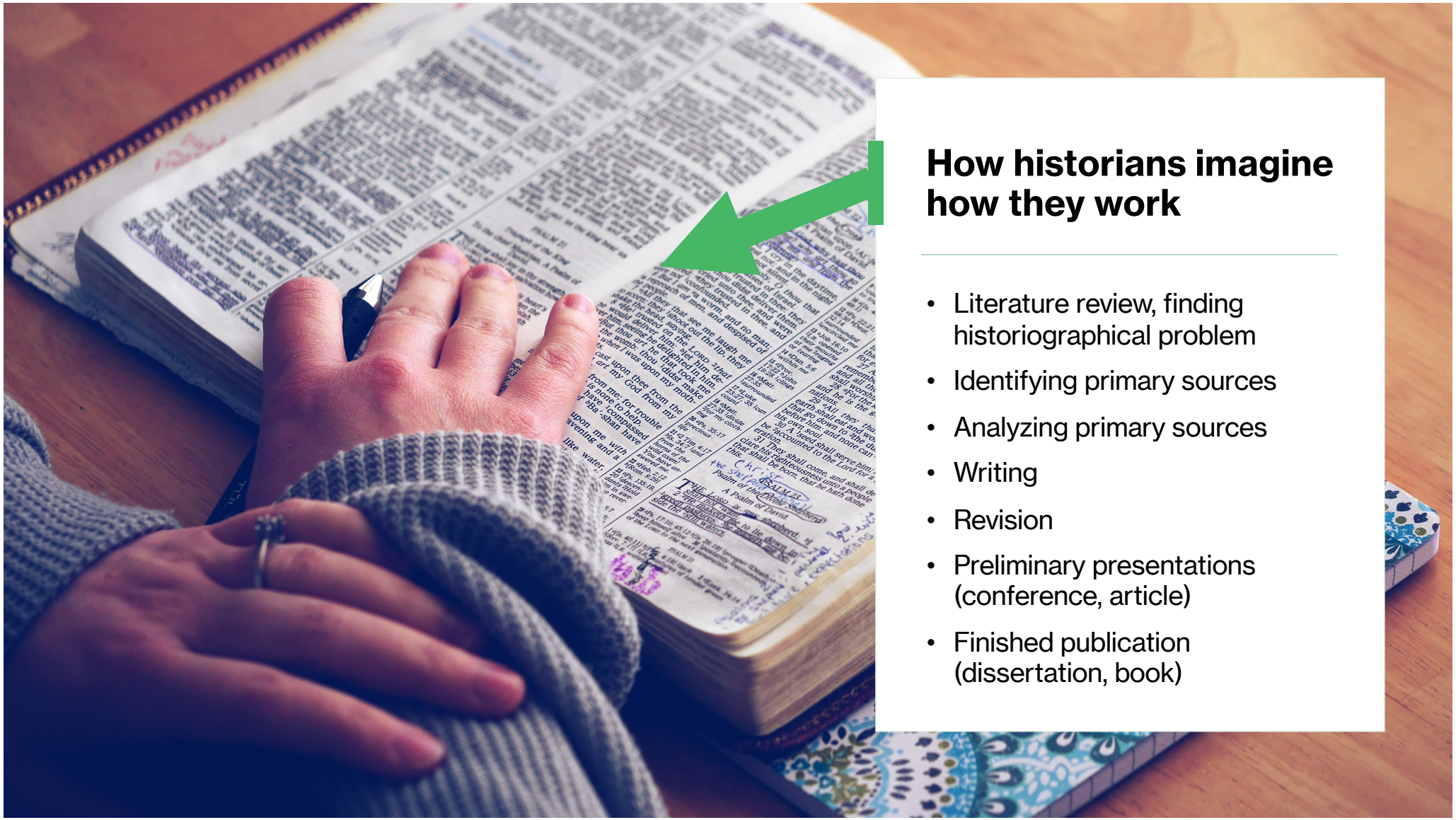
Ian Milligan

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*)


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

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

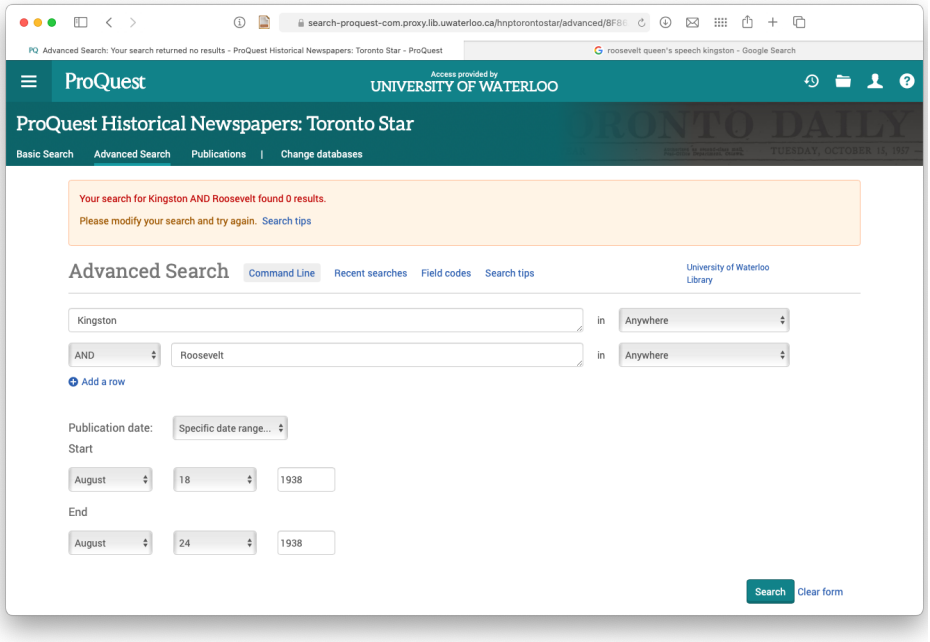
How did historians previously use newspapers?

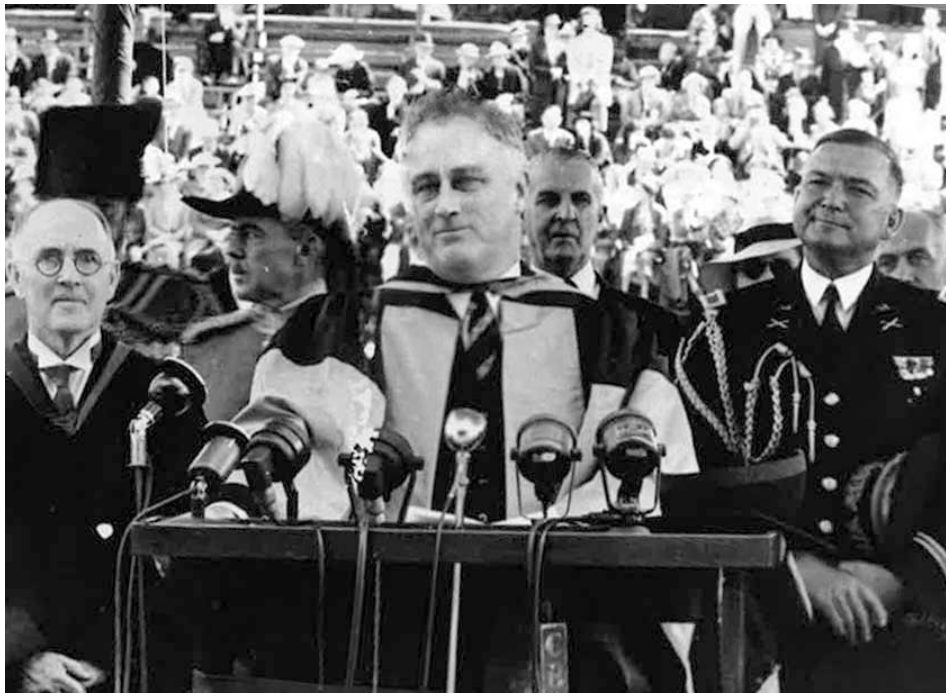
- Microfilm reels
- Sit in a dark basement, crank forward through pages looking for relevant documents
- Time consuming, boring (good chance to listen to music)
- But you learn a *lot* about context
    - i.e. a global event that happens, the tenor of advertisements, the relative placement of columnists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

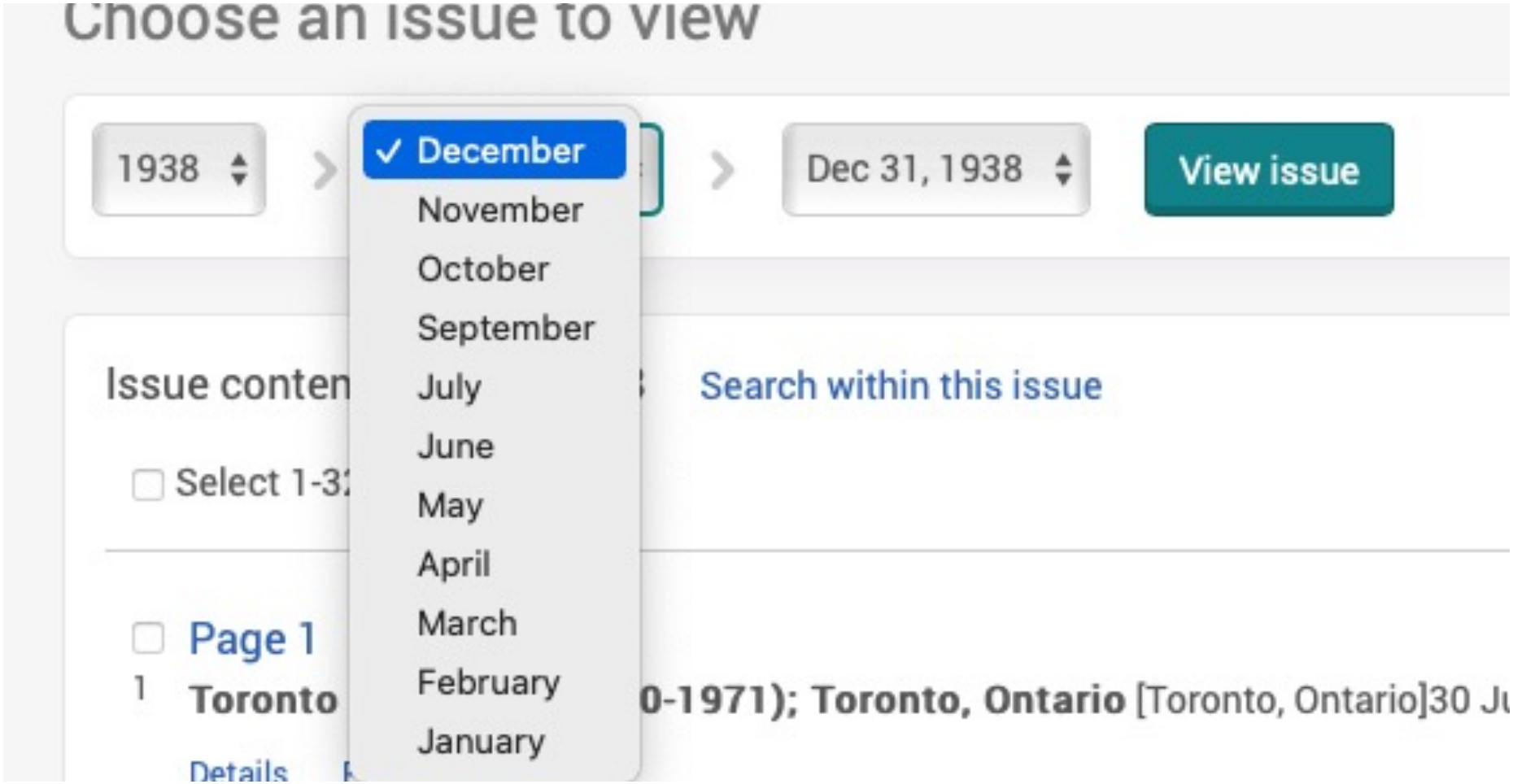
Instead, historians now do most of their work through keyword searching.


But we uncritically use interfaces.

Let me use a Canadian example.






Choose an issue to view
1938
December
November
October
September
July
June
May
April
March
February
January
Dec 31, 1938
View issue
Issue conten
Search within this issue
Select 1-3
Page 1
1
Toronto
0-1971); Toronto, Ontario [Toronto, Ontario]30 Ju
Details

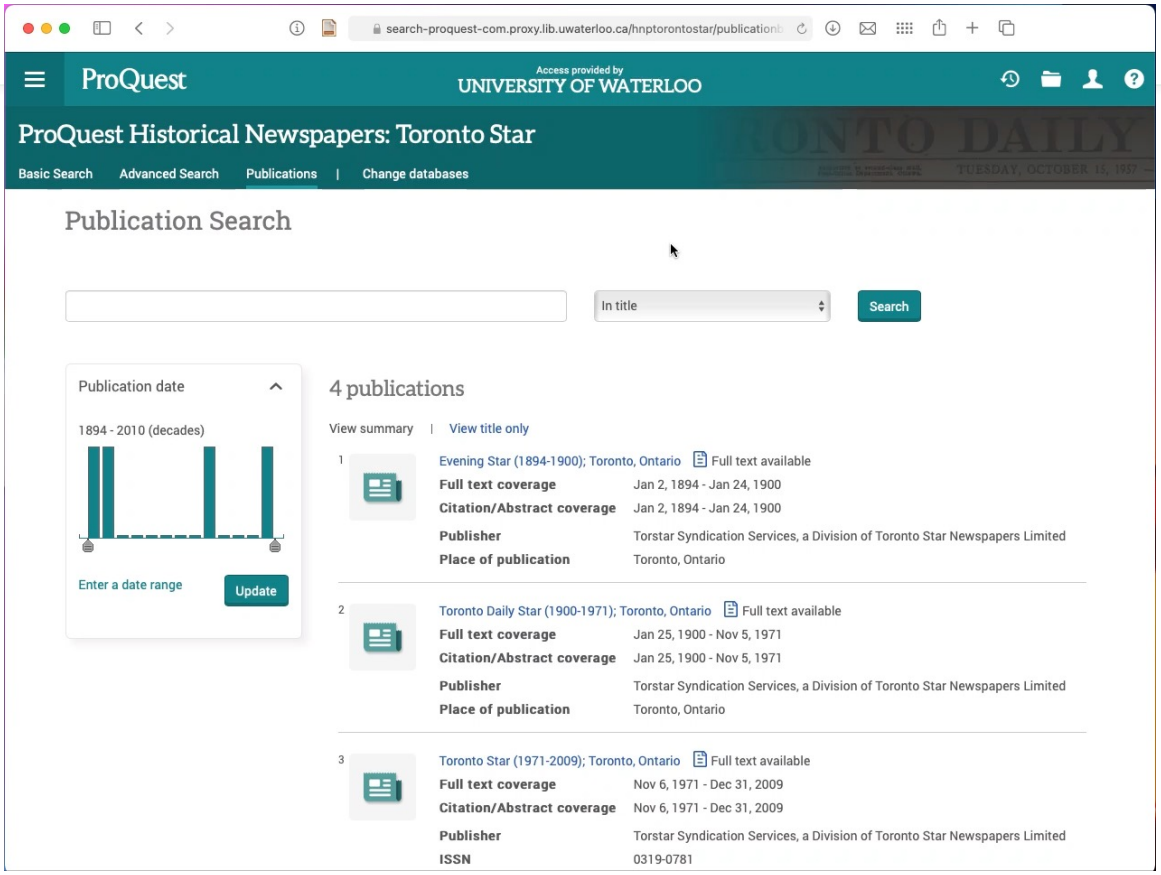
Missing

- 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

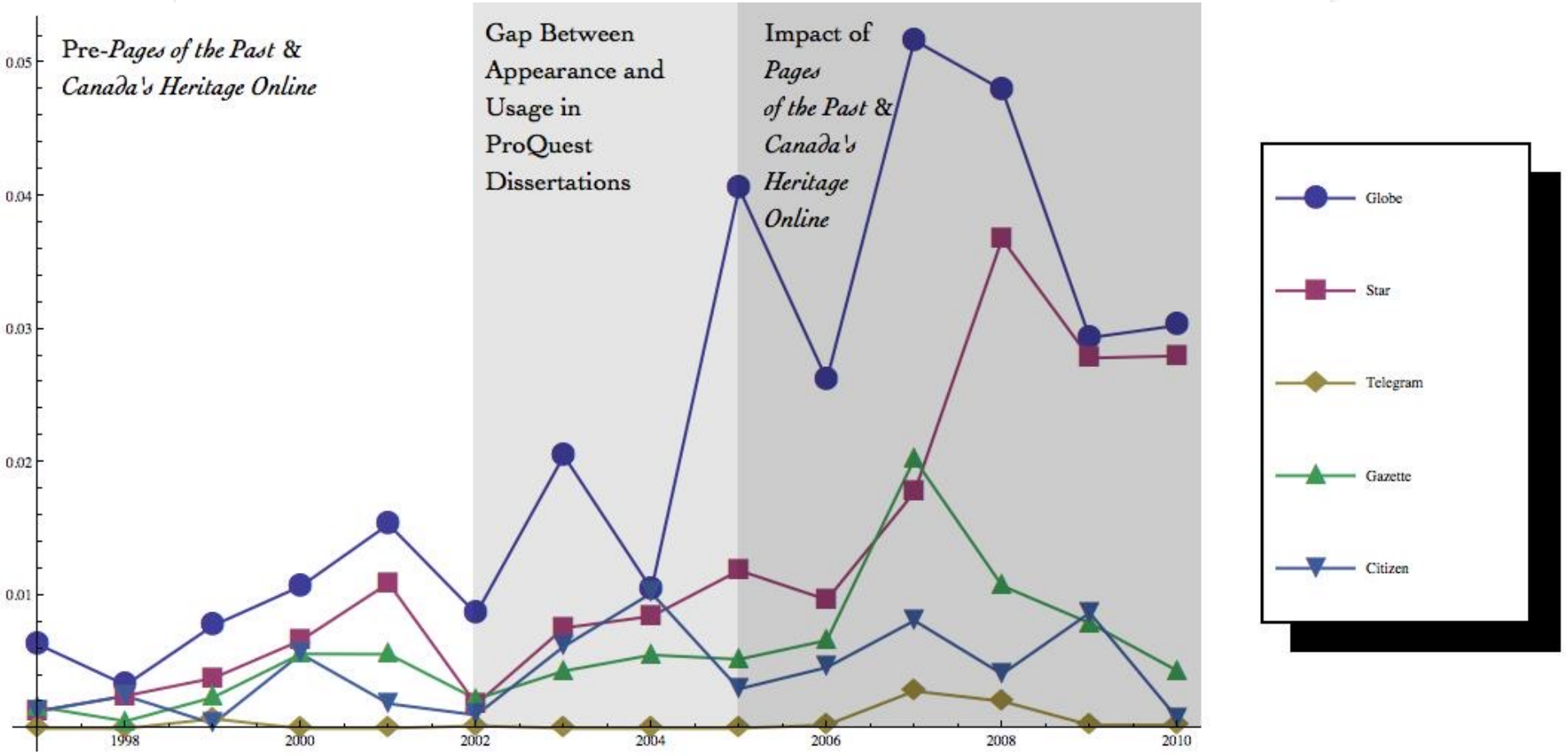
- The text that is being searched is created using **optical character recognition**, or OCR
  - ProQuest's implementation stems from *Pages of the Past*, an innovative project that saw the *Toronto Star* the first fully digitized newspaper in the world
  - But it's a commercial platform, so correcting 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




And
digitization is
uneven

In Canada


Pre-*Pages of the Past* & *Canada's Heritage Online*
Gap Between Appearance and Usage in ProQuest Dissertations
Impact of *Pages of the Past* & *Canada's Heritage Online*
0.05
0.04
0.03
0.02
0.01
1998
2000
2002
2004
2006
2008
2010
Globe
Star
Telegram
Gazette
Citizen

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

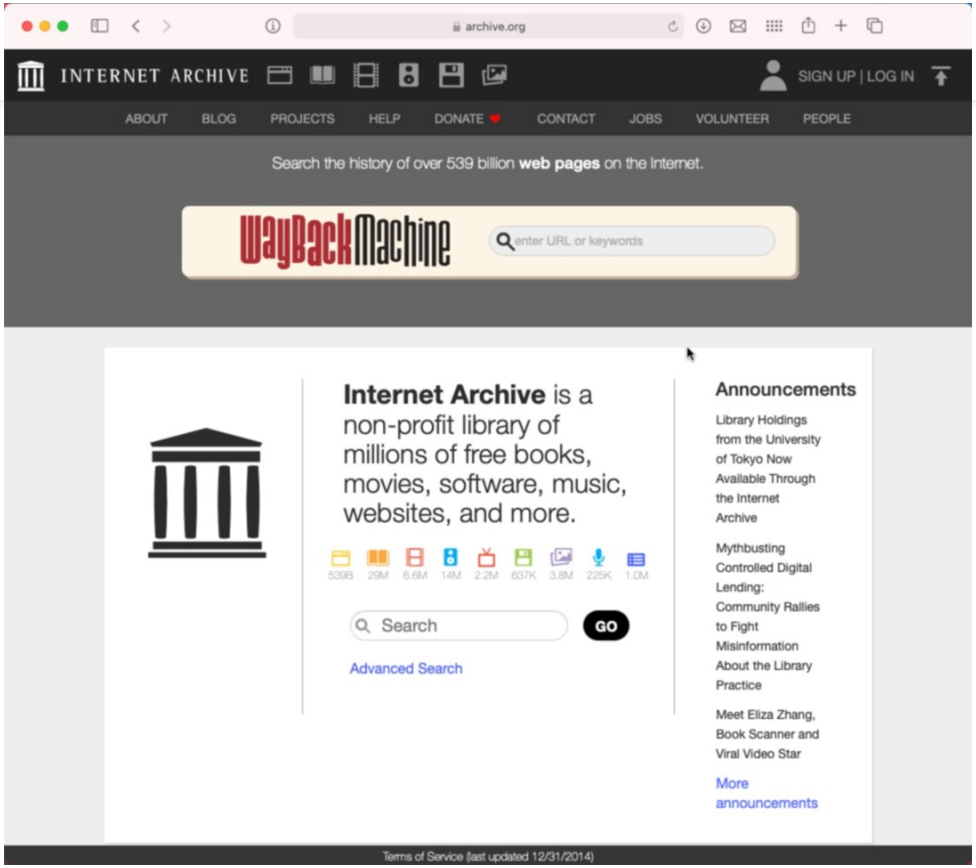
- We now interact primarily through keyword search (i.e. the system forces us more or less to do this)
- We don't fully understand the construction of this database.
- The text is inaccessible to do transformative digital scholarship with.
- **Yet we still cite it all the same: Pages of the Past, ProQuest, 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?

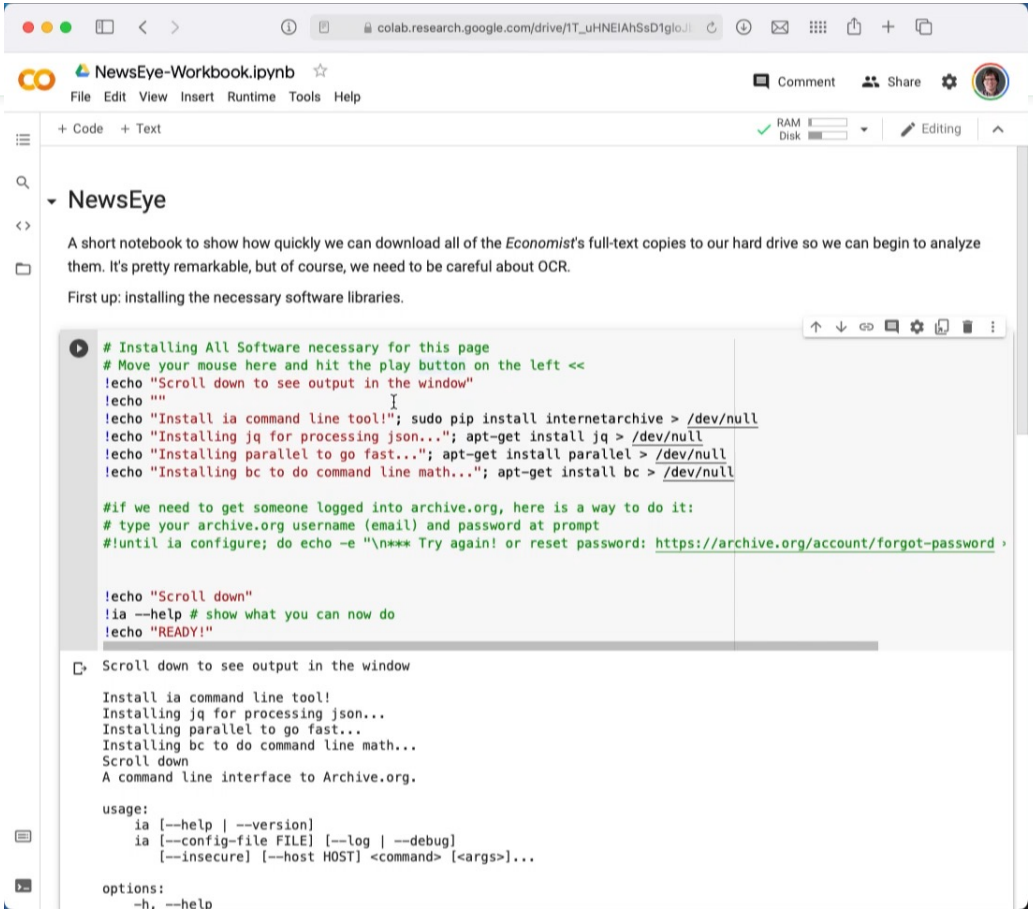
So what can we do?

Come to the NewsEye Conference!

Embrace Open Platforms and Tools



Embrace Open Platforms and Tools



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.

Looking forward to the discussion!