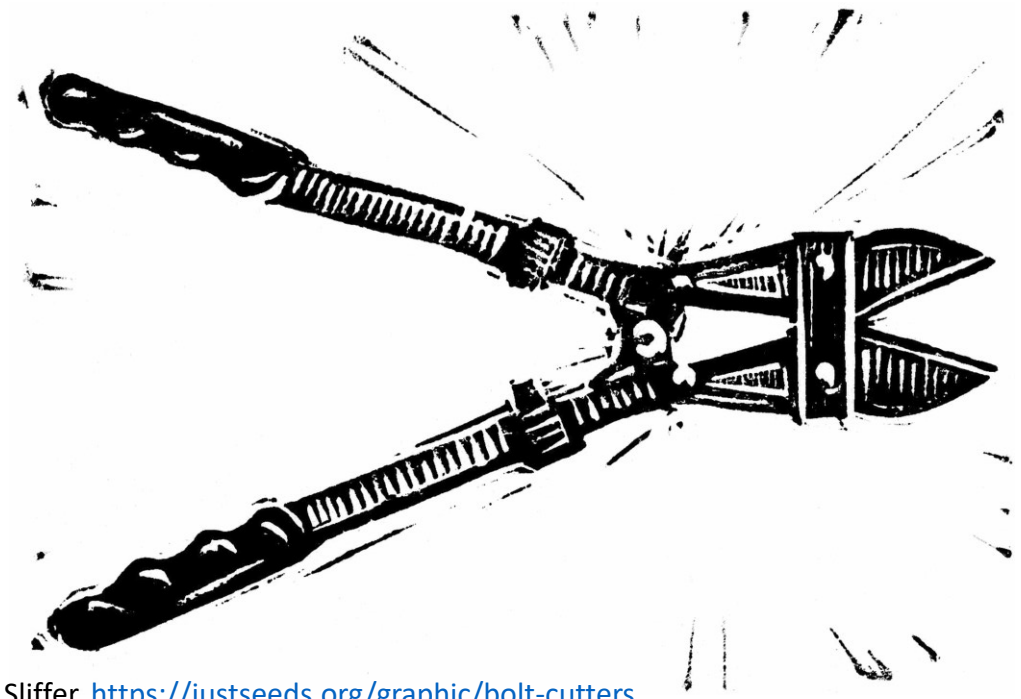
A Critical Look at the Digital Scholarship Corpus

How Access Influences the Questions We (Can) Ask



Shaun Sliffer, https://justseeds.org/graphic/bolt-cutters

What might you do with a large text corpus?

- Corpus linguistics
- Natural language processing
- Textual analysis/text mining
- Related: metadata studies

What might large text corpora collect?

- Books
- Newspaper/magazine content
- Journal articles
- Web content
- Government documents
- Legal documents
- Archival content (personal documents, correspondence, etc.)
- …and plenty of mixture of these formats

English	# words	language/dialect	time period
iWeb: The Intelligent Web-based Corpus **NEW!**	14 **billion**	US/CA/UK/IE/AU/NZ	2017
News on the Web (NOW)	6.04 **billion+**	20 countries / Web	2010-Aug 2018
Global Web-Based English (GloWbE)	1.9 **billion**	20 countries / Web	2012-13
Wikipedia Corpus	1.9 **billion**	English	-2014
Hansard Corpus	1.6 **billion**	British (parliament)	1803-2005
Early English Books Online	755 million	British	1470s-1690s
Corpus of Contemporary American English (COCA)	**560** million	American	1990-2017
Corpus of Historical American English (COHA)	400 million	American	1810-2009
Corpus of US Supreme Court Opinions	130 million	American (law)	1790s-present
TIME Magazine Corpus	100 million	American	1923-2006
Corpus of American Soap Operas	100 million	American	2001-2012
British National Corpus (BYU-BNC)*	100 million	British	1980s-1993
Strathy Corpus (Canada)	50 million	Canadian	1970s-2000s
CORE Corpus	50 million	Web registers	-2014

https://corpus.byu.edu

Start with which corpus?

Corpus	Size (words)
American	155 billion
British	34 billion
Spanish	45 billion

“HathiTrust Research Center Extends Non-Consumptive Research Tools to Copyrighted Materials: Expanding Research through Fair Use”

“HTRC now provides access to the text of the complete 16.7-million-item HathiTrust corpus for non-consumptive research, such as data mining and computational analysis, including items protected by copyright.”

September 20, 2018

Access to these large text corpora allows us to ask new questions!

But what if you’re looking for something a little more…specific?

What questions remain difficult to ask using text analysis, even with unprecedented access?

“One only needs to review the current work in digital literary studies to see that we have not escaped the traditional canon by turning to new methods of publication…” (Earhart, 2012)

What's missing?

- Through digitization, "we have given to the oldest of Western canons a new hyper-availability, and a new authority" (Hitchcock, 2012).
- But we all know the problems with the traditional English-language canon…
  - "the universe of digitized text is anything but representative of the temporal and geographic* contours of human life in the past" (Putnam, 2014, p. 14).

** [and linguistic, and cultural, and socioeconomic, and racialized, and gendered…]*

Who’s missing?

- “the noncanonical, the non-Western, the non-elite and the quotidian” (Hitchcock, 2012)
- Individuals and groups whose experiences, histories, and voices—when they were recorded at all—were not “recorded on their own terms” (Putnam, 2014, p. 16).

Reading “against the grain” using large text corpora

What questions can we ask of a large text corpus which is not designed for our approach?

decar66, https://www.flickr.com/photos/decar66/7197666906

What approaches remain challenging?

- Feminist, sexuality, and gender studies
- Place-based studies
- Critical race studies (generally, and/or specific populations)
- Postcolonial/decolonial studies
- Indigenous studies
- 20th and 21st century studies of materials under copyright
- Study of handwritten materials
- Study of certain languages
- Literature in translation

Reading “against the grain”

- Close reading has traditionally required that scholars studying marginalized voices unearth those experiences using textual data in creative ways
- Interacting with large text corpora will require a similar approach
- Does the claim to near comprehensiveness of certain large text corpora risk eliding these individuals once again?
  - “The bit players can finally seize center stage, and it turns out they have so much to say! The optic of the digitizable world captures history made not from the top down but from the *bottom of the top* and the *top of the bottom*” (Putnam, 2014, p. 16; emphasis mine)

Researching “around” a topic

- Looking for coded references to the population of study
- Looking for gaps and elisions
- Finding occasions when the population of study is spoken about from “normalized” perspectives within the record
- Seeking out the occasions when individuals from the studied population get to speak (and how that speech is filtered)

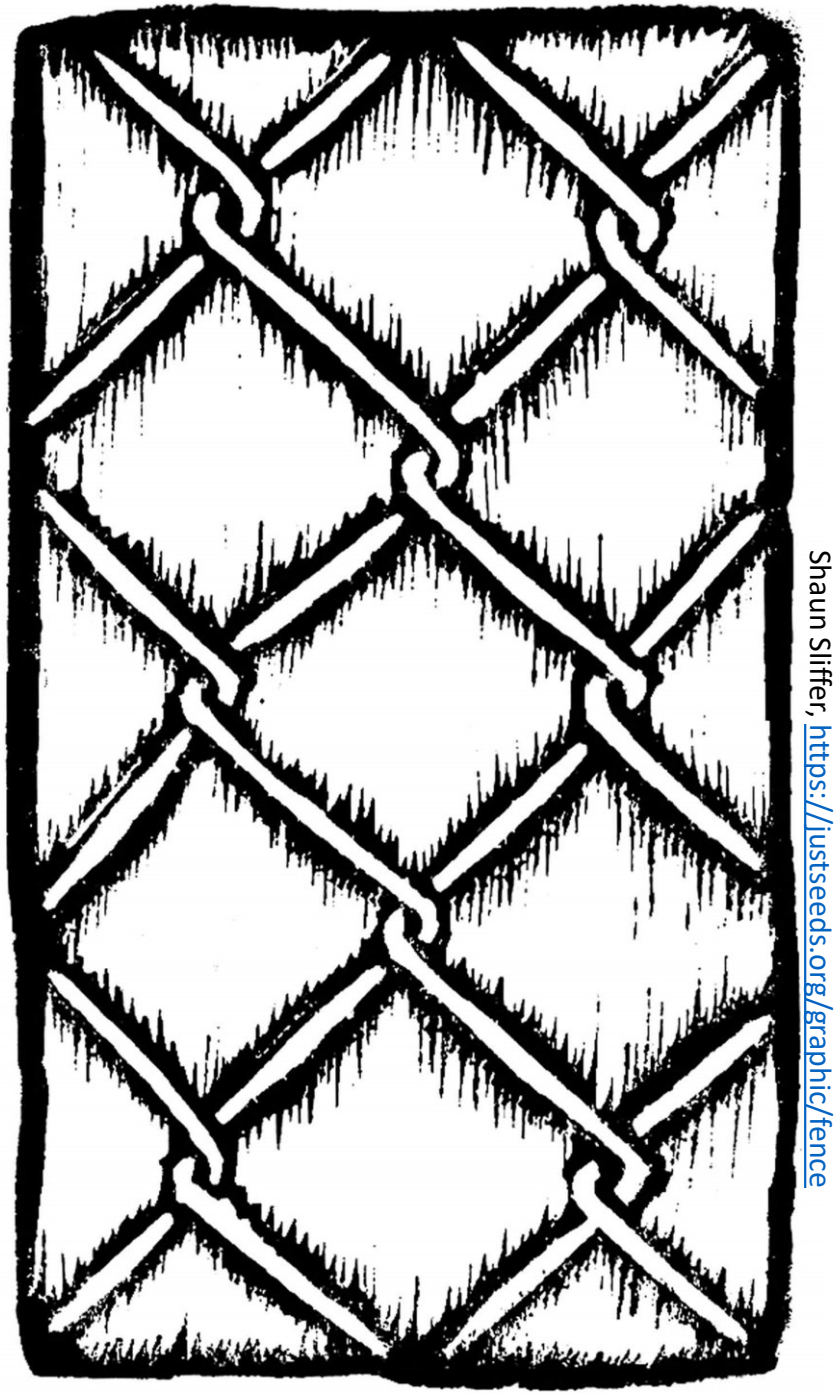
Researching “around” a topic

- A feminist approach to text analysis posted by Lisa Marie Rhody includes:
  - “exposing implicit and explicit choices that influence the construction of textual corpora,
  - articulating the rationale for their selection,
  - and carefully scoping the claims they make in deference to the representative limitations of their datasets”
    (2016; bullet points mine)

Assuming the primacy of text

- Textual data will not encompass all experiences
- Missing or not directly represented within these corpora are:
  - Oral & folk traditions
  - Songs and music
  - The non-literate
  - Those without the socioeconomic ability (time, materials, etc.) to write their own stories, and without the newsworthiness to be written about

Challenges to text analysis using large text corpora



Shaun Sliffer, https://justseeds.org/graphic/fence

Challenges to scholars studying marginalized voices and experiences

- What material was created and preserved
- What material is digitized
- Copyright and licensing
- Creation of more specific corpora

Copyright and licensing

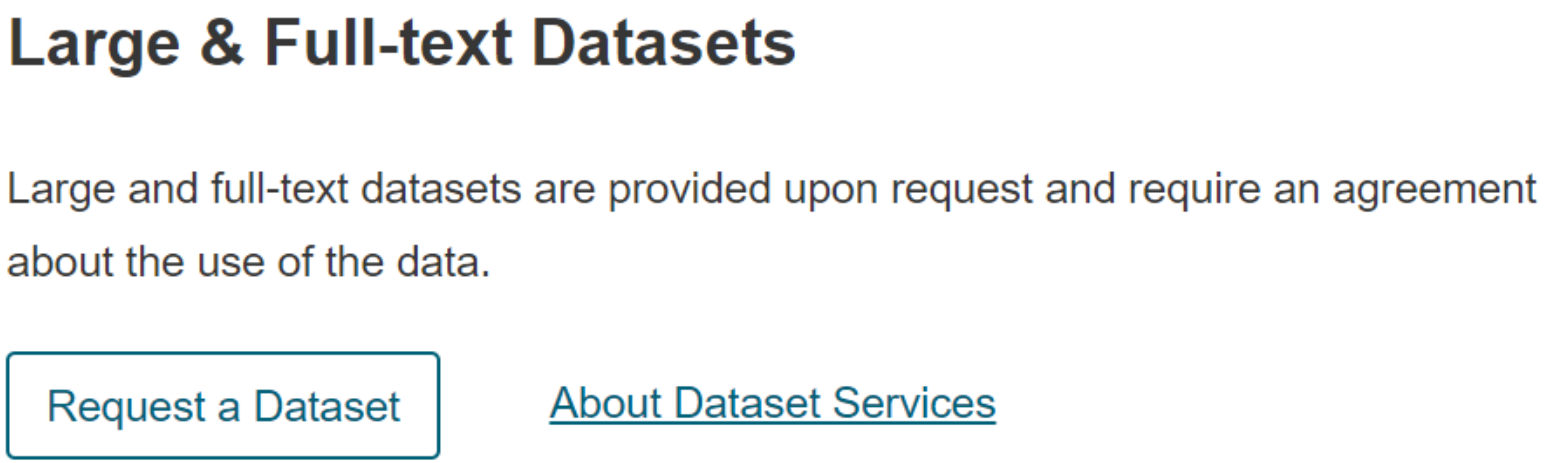
- Even among public domain works, copyright is a challenge
- When a work is out of copyright, the digitizer may still leverage copyright to restrict use—despite unclear legal basis to do so (Crews, 2012)
- Restricted access to out of copyright works can further be controlled through licensing

Fair use

- Scholars must often access works under copyright through systems that prevent: downloading, format changing, text extraction, or systematic text analysis
- A user may break the terms of use of a legitimately purchased digital work by seeking to transform it
- Creating a corpus of works that are still under copyright may therefore be labor-intensive and/or legally fraught (Senseney et al., 2018)

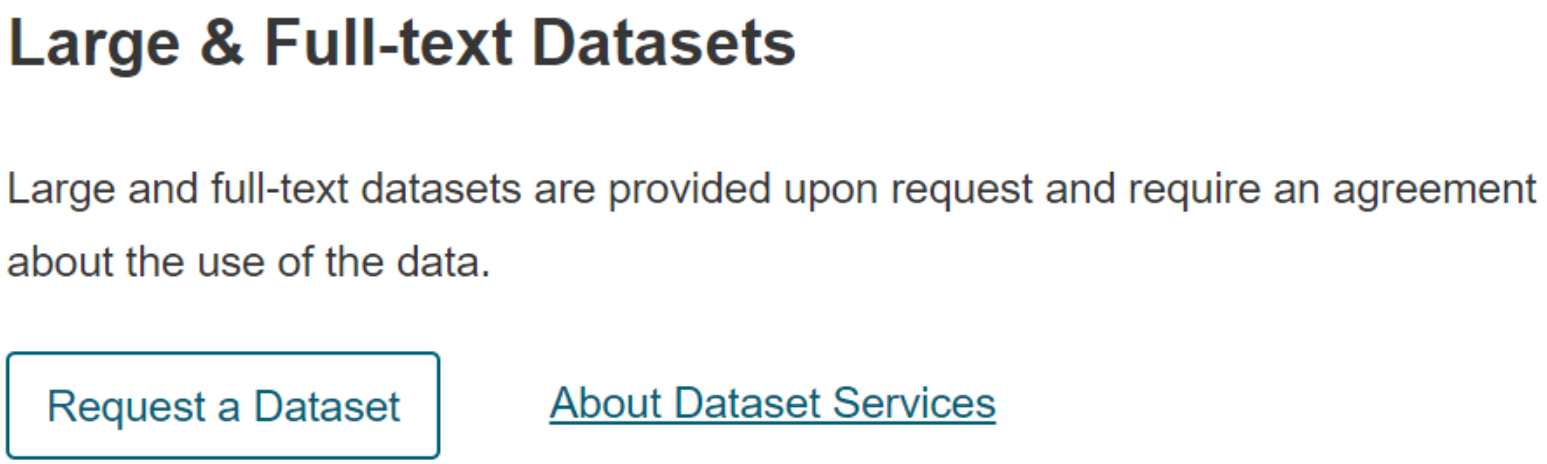
Vendor-owned content

- It is possible to negotiate access to text mine database content
- Often this requires a license



Vendor-owned content

- What barriers might stand in the way of such a license?
  - Lack of expertise on the part of the researcher
  - Lack of library access to the full dataset in question
  - Lack of librarian time or expertise in such a licensing endeavor



Creation of a corpus

- Time-consuming
  - “Manual entry is time-consuming and costly, and therefore unsuitable for the creation of very large corpora.
  - OCR output can be similarly costly if it requires substantial post-processing to validate the data.
  - Data acquired in electronic form from other sources will almost invariably contain formatting codes and other information that must be discarded or translated to a representation that is processable for linguistic analysis” (Ide, 2004).

Let's talk about skills

- Plenty of humanities researchers have digital skills
- How many humanities researchers have both the skill and the time to locate materials for a new corpus, digitize them if necessary, run OCR (what if they're handwritten?) on them, remove headers/footers/etc., check and validate the data…
  - May be possible if you have: a grant, a team, and/or an understanding department head (or dissertation chair!)
- Even if more of these techniques can be done in a machine-assisted manner, will the financial and labor cost be low enough to allow more practitioners to choose this path?

Let’s talk about power

- Scholars studying marginalized populations may already be pushing on the boundaries of their academic discipline before they consider diving in to digital scholarship
- Who should do the work of exposing underrepresented voices? Is it the sole province of these scholars of difference (Risam, 2015) to remediate the canon?

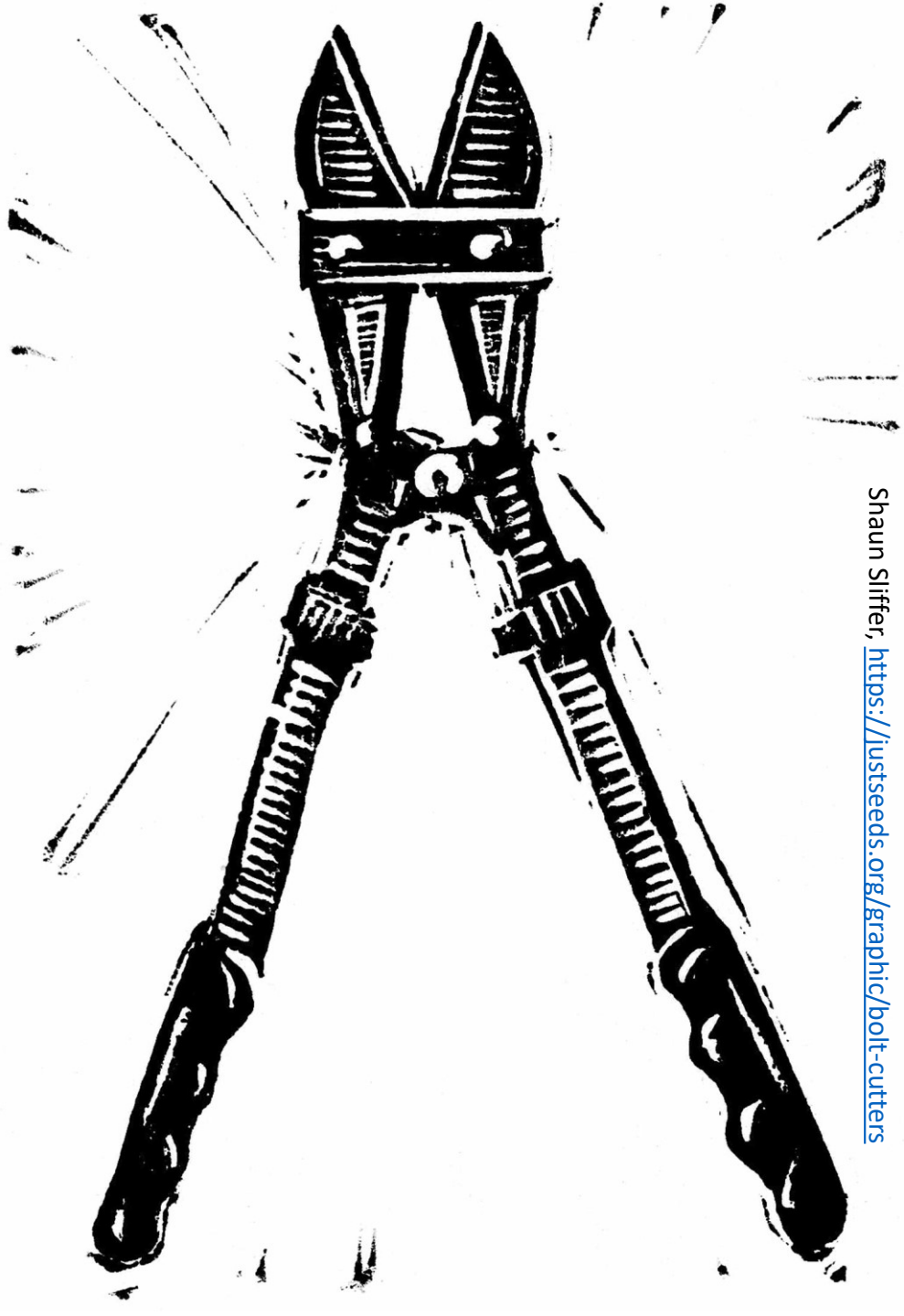
Who can do this work?

Among those studying topics not facilitated by free, existing corpora, whose access may be curtailed?

- Researchers at smaller institutions (or between institutions, or working beyond the college or university)
- Researchers who lack support to negotiate for text-mining access to closed databases
- Researchers without the time or expertise to compile (and clean!) a new specialty corpus
- Researchers who due to proximity or funding are unable to found or join a research team to compile a new specialty corpus

How do digital scholarship methods replicate existing patterns if important sites of research remain unfeasible?

And how can we work toward a new, liberatory scholarly paradigm?



Shaun Sliffer, https://justseeds.org/graphic/bolt-cutters

References

Crews, K. D. (2012). Museum Policies and Art Images: Conflicting Objectives and Copyright Overreaching. *Fordham Intellectual Property, Media & Entertainment Law Journal*, *22*, 795. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2120210

Earhart, A. E. (2012). Can information be unfettered? Race and the new digital humanities canon. In M. K. Gold (Ed.), *Debates in the Digital Humanities*. http://dhdebates.gc.cuny.edu/debates/text/16

Eveleth, R. (2017, November 28). Our father, who art in algorithm [Podcast episode]. *Flash Forward*. https://www.flashforwardpod.com/2017/11/28/our-father-who-art-in-algorithm

Hitchcock, T. (2012). A five minute rant for the Consortium of European Research Libraries. https://historyonics.blogspot.com/2012/10/a-five-minute-rant-for-consortium-of.html

Ide, N. (2004). Preparation and analysis of linguistic corpora. In S. Schreibman, R. Siemens, J. Unsworth (Eds.), *A Companion to Digital Humanities*. Oxford: Blackwell. http://www.digitalhumanities.org/companion

Putnam, L. (2014). The transnational and the text-searchable: Digitized sources and the shadows they cast [Post-print]. http://d-scholarship.pitt.edu/21663/1/PutnamDigitalShadowsPrePrint.pdf (Published 2016, *The American Historical Review*, *121*(2), 377-402. https://doi.org/10.1093/ahr/121.2.377)

Rhody, L. M. (2016). Why I dig: Feminist approaches to text analysis. In M. K. Gold & L. F. Klein (Eds.), *Debates in the Digital Humanities*. http://dhdebates.gc.cuny.edu/debates/text/97

Risam, R. (2015). Beyond the margins: Intersectionality and the digital humanities. *Digital Humanities Quarterly*, *9*(2). http://www.digitalhumanities.org/dhq/vol/9/2/000208/000208.html

Senseney, M., Dickson, E., Namachchivaya, B., & Ludäscher, B. (2018). Data mining research with in-copyright and use-limited text datasets: Preliminary findings from a systematic literature review and stakeholder interviews. Paper presented at the 13th International Digital Curation Conference, February 19-22, 2018, Barcelona, Spain. https://www.ideals.illinois.edu/handle/2142/99026

Smith, M. N. (2014). Frozen social relations and time for a thaw: Visibility, exclusions, and considerations for postcolonial digital archives. *Journal of Victorian Culture*, *19*(3), 403-410. doi: 10.1080/13555502.2014.947189