Querying Queer Web Archives

Filipa Calado, PhD Candidate in English, the Graduate Center, CUNY
Corey Clawson, PhD Student in American Studies, Rutgers University - Newark
Di Yoong, PhD Candidate in Critical Social/Personality & Environmental Psychology, the Graduate Center, CUNY

Overview

- Goals of project and questions
- Contextualizing the dataset
- Current progress
- Ethical considerations
- Further questions:
    - Analog and digital archives

Querying the material and methodology of web archival work

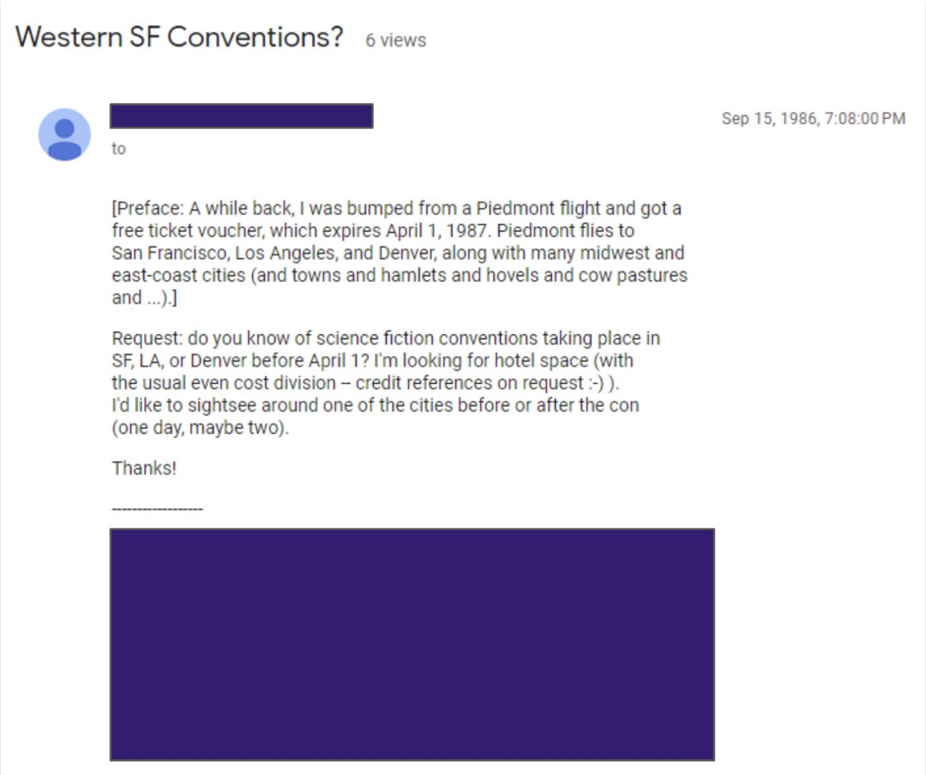
- How do queer identities and communities interact with and inform web spaces?
- How are web spaces themselves formed and cultivated by communities?
- How do concepts like utopia, radicalism, normativity, religion, and conversion affect queer identity and discourse formation over time?
- What vocabularies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

Defining “queerness”

“Queerness is a structuring and educated mode of desiring that allows us to see and feel beyond the quagmire of the present…. Queerness is that thing that lets us feel that this world is not enough, that indeed something is missing.”

- José Esteban Muñoz, *Cruising Utopia: The Then and There of Queer Futurity*.

Soc.motss (USENET)



Western SF Conventions?

[redacted]

to

Sep 15, 1986, 7:08:00 PM

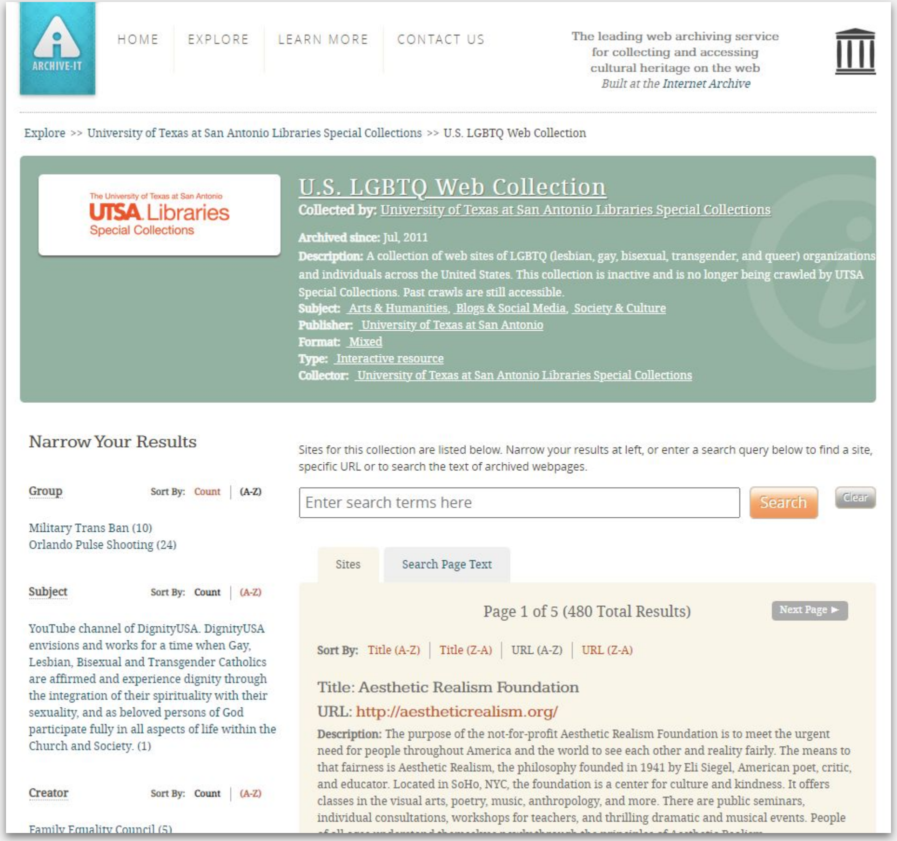
[Preface: A while back, I was bumped from a Piedmont flight and got a
free ticket voucher, which expires April 1, 1987. Piedmont flies to
San Francisco, Los Angeles, and Denver, along with many midwest and
east-coast cities (and towns and hamlets and hovels and cow pastures
and ...).]

Request: do you know of science fiction conventions taking place in
SF, LA, or Denver before April 1? I'm looking for hotel space (with
the usual even cost division -- credit references on request :-)).
I'd like to sightsee around one of the cities before or after the con
(one day, maybe two).

Thanks!

[redacted]

US LGBTQ Web Collection



The leading web archiving service
for collecting and accessing
cultural heritage on the web
Built at the Internet Archive

Explore >> University of Texas at San Antonio Libraries Special Collections >> U.S. LGBTQ Web Collection

U.S. LGBTQ Web Collection

Collected by: University of Texas at San Antonio Libraries Special Collections

Archived since: Jul, 2011
Description: A collection of web sites of LGBTQ (lesbian, gay, bisexual, transgender, and queer) organizations and individuals across the United States. This collection is inactive and is no longer being crawled by UTSA Special Collections. Past crawls are still accessible.
Subject: Arts & Humanities, Blogs & Social Media, Society & Culture
Publisher: University of Texas at San Antonio
Format: Mixed
Type: Interactive resource
Collector: University of Texas at San Antonio Libraries Special Collections

Narrow Your Results

Group Sort By: Count | (A-Z)

Military Trans Ban (10)
Orlando Pulse Shooting (24)

Subject Sort By: Count | (A-Z)

YouTube channel of DignityUSA. DignityUSA envisions and works for a time when Gay, Lesbian, Bisexual and Transgender Catholics are affirmed and experience dignity through the integration of their spirituality with their sexuality, and as beloved persons of God participate fully in all aspects of life within the Church and Society. (1)

Creator Sort By: Count | (A-Z)

Family Equality Council (5)

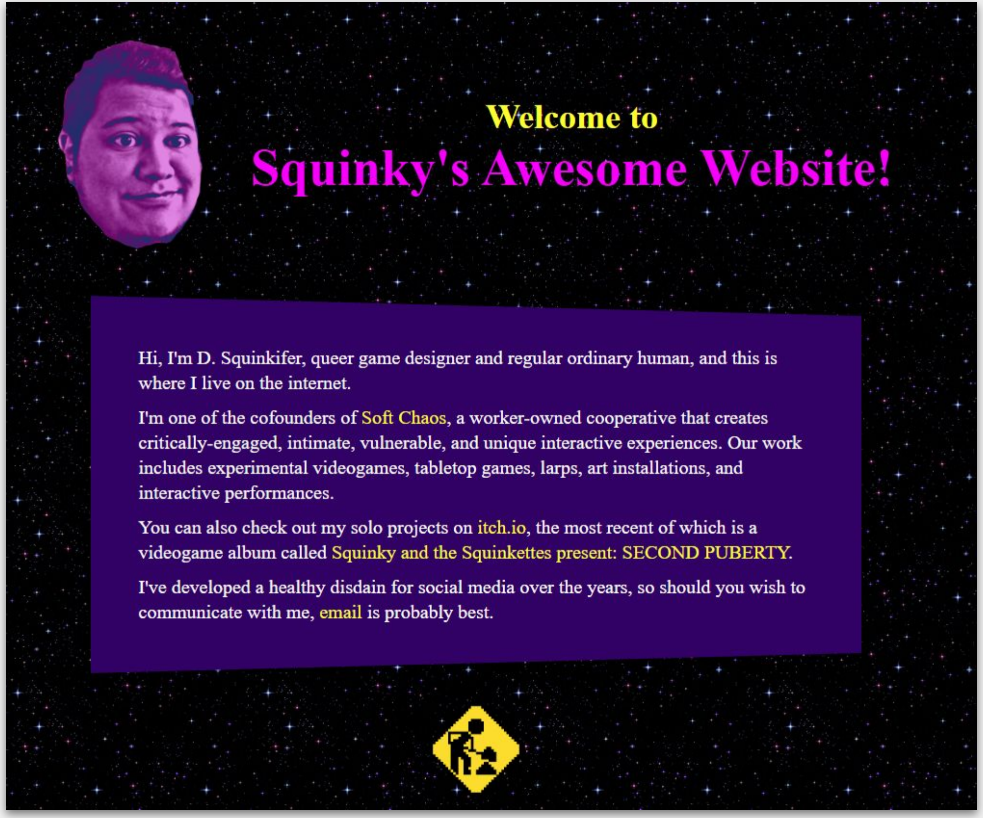
Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Page 1 of 5 (480 Total Results) Next Page ►

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Title: Aesthetic Realism Foundation
URL: http://aestheticrealism.org/
Description: The purpose of the not-for-profit Aesthetic Realism Foundation is to meet the urgent need for people throughout America and the world to see each other and reality fairly. The means to that fairness is Aesthetic Realism, the philosophy founded in 1941 by Eli Siegel, American poet, critic, and educator. Located in SoHo, NYC, the foundation is a center for culture and kindness. It offers classes in the visual arts, poetry, music, anthropology, and more. There are public seminars, individual consultations, workshops for teachers, and thrilling dramatic and musical events. People

Interactive Fiction in Queer Literature



Welcome to

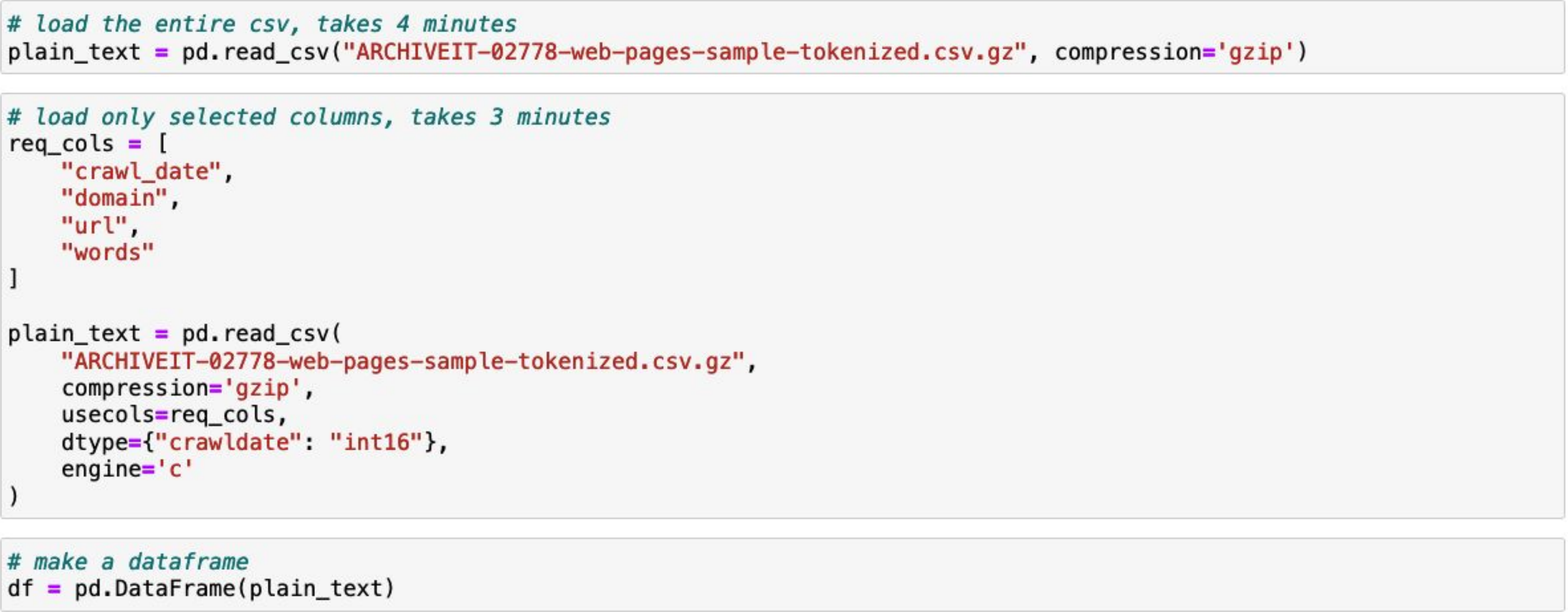
Squinky's Awesome Website!

Hi, I'm D. Squinkifer, queer game designer and regular ordinary human, and this is where I live on the internet.

I'm one of the cofounders of Soft Chaos, a worker-owned cooperative that creates critically-engaged, intimate, vulnerable, and unique interactive experiences. Our work includes experimental videogames, tabletop games, larps, art installations, and interactive performances.

You can also check out my solo projects on itch.io, the most recent of which is a videogame album called Squinky and the Squinkettes present: SECOND PUBERTY.

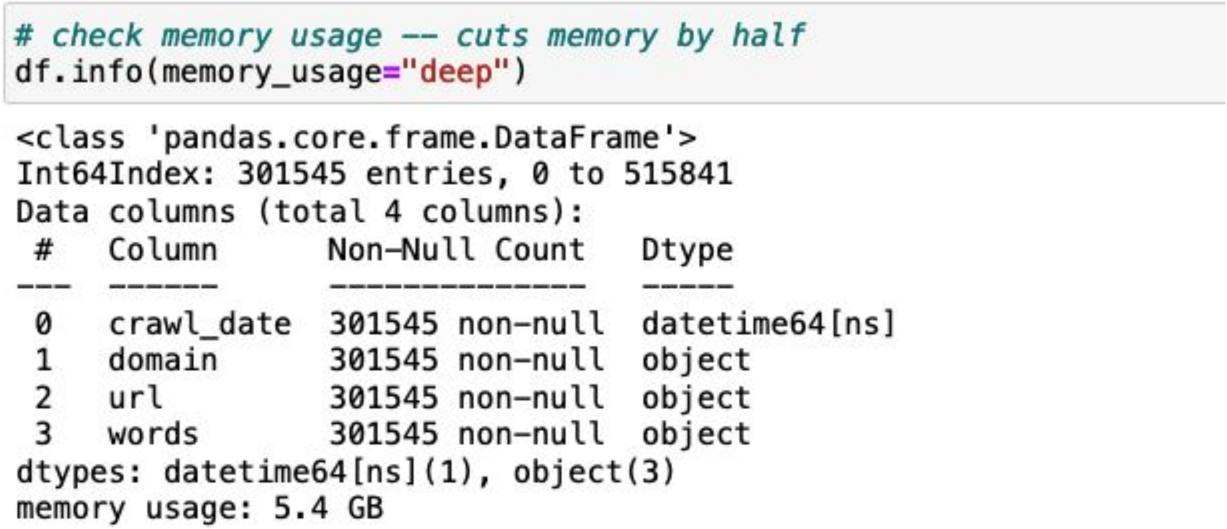
I've developed a healthy disdain for social media over the years, so should you wish to communicate with me, email is probably best.



```
# load the entire csv, takes 4 minutes
plain_text = pd.read_csv("ARCHIVEIT-02778-web-pages-sample-tokenized.csv.gz", compression='gzip')
```

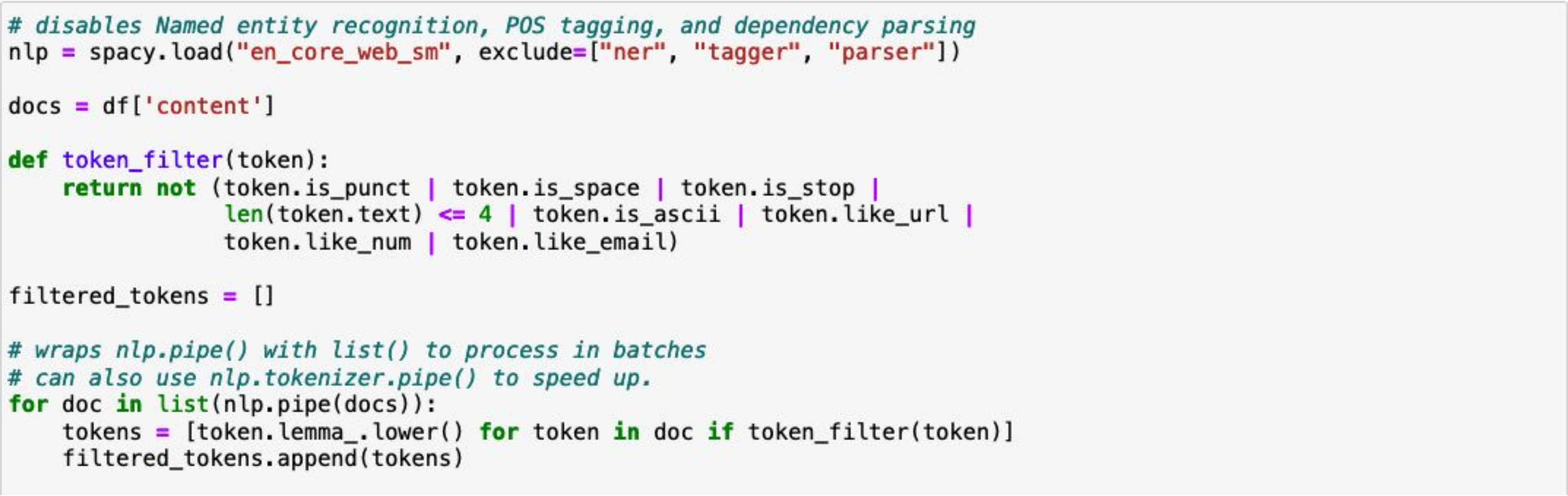
```
# load only selected columns, takes 3 minutes
req_cols = [
    "crawl_date",
    "domain",
    "url",
    "words"
]

plain_text = pd.read_csv(
    "ARCHIVEIT-02778-web-pages-sample-tokenized.csv.gz",
    compression='gzip',
    usecols=req_cols,
    dtype={"crawldate": "int16"},
    engine='c'
)
```

```
# make a dataframe
df = pd.DataFrame(plain_text)
```



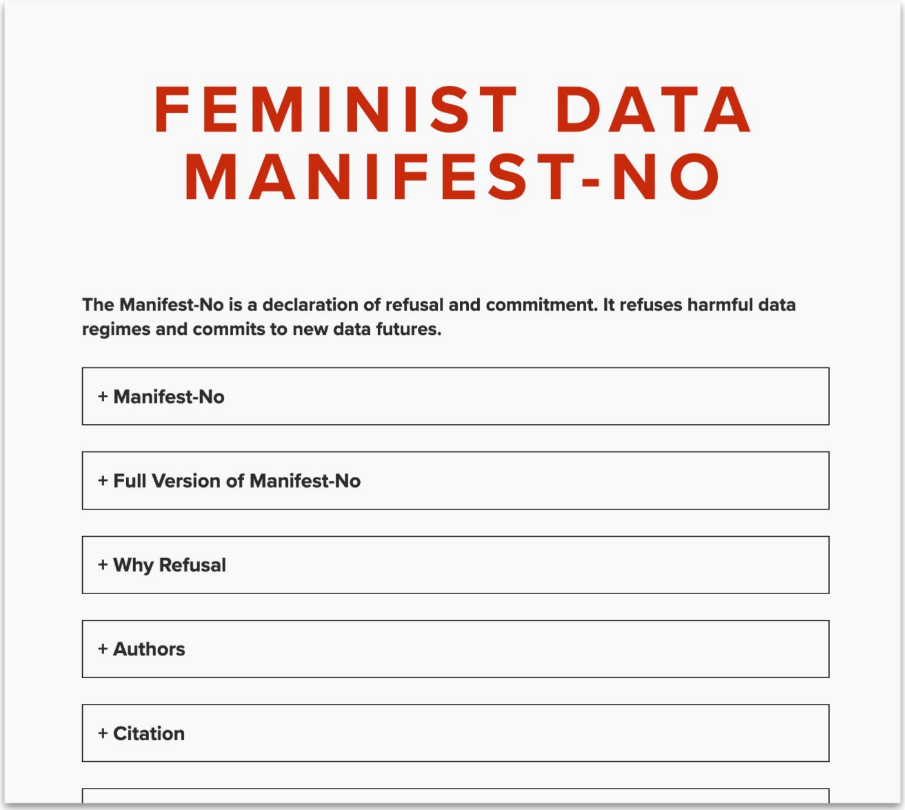
```python
# check memory usage -- cuts memory by half
df.info(memory_usage="deep")
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 301545 entries, 0 to 515841
Data columns (total 4 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   crawl_date  301545 non-null  datetime64[ns]
 1   domain      301545 non-null  object
 2   url         301545 non-null  object
 3   words       301545 non-null  object
dtypes: datetime64[ns](1), object(3)
memory usage: 5.4 GB
```



```
# disables Named entity recognition, POS tagging, and dependency parsing
nlp = spacy.load("en_core_web_sm", exclude=["ner", "tagger", "parser"])

docs = df['content']

def token_filter(token):
    return not (token.is_punct | token.is_space | token.is_stop |
                len(token.text) <= 4 | token.is_ascii | token.like_url |
                token.like_num | token.like_email)

filtered_tokens = []

# wraps nlp.pipe() with list() to process in batches
# can also use nlp.tokenizer.pipe() to speed up.
for doc in list(nlp.pipe(docs)):
    tokens = [token.lemma_.lower() for token in doc if token_filter(token)]
    filtered_tokens.append(tokens)
```

Feminist Data Manifest-No & Documenting the Now



FEMINIST DATA MANIFEST-NO

The Manifest-No is a declaration of refusal and commitment. It refuses harmful data regimes and commits to new data futures.

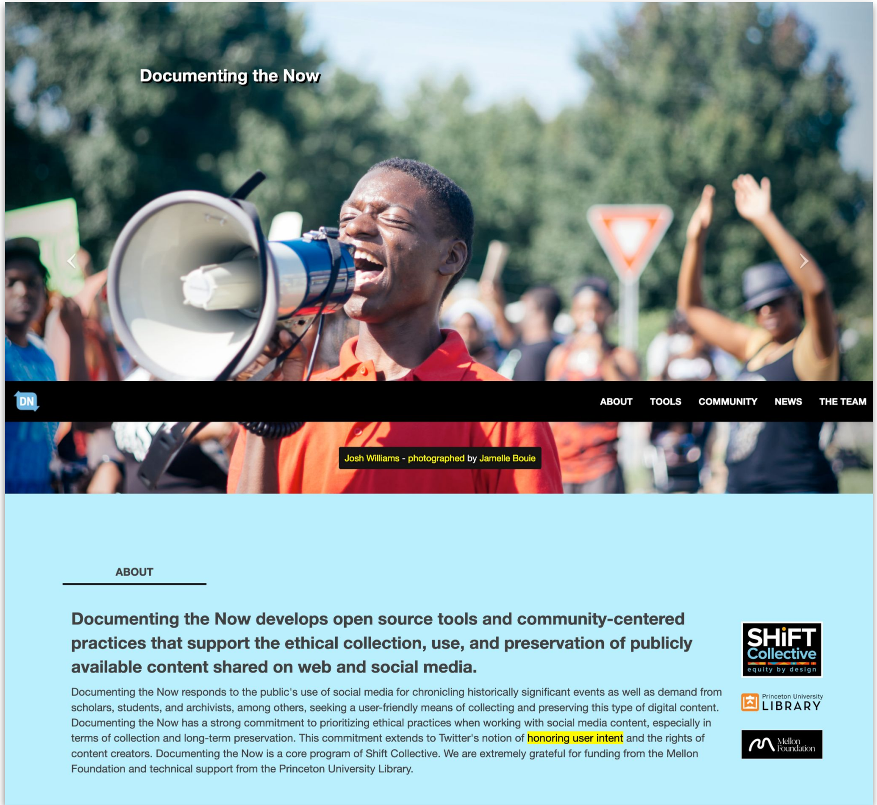
+ Manifest-No

+ Full Version of Manifest-No

+ Why Refusal

+ Authors

+ Citation



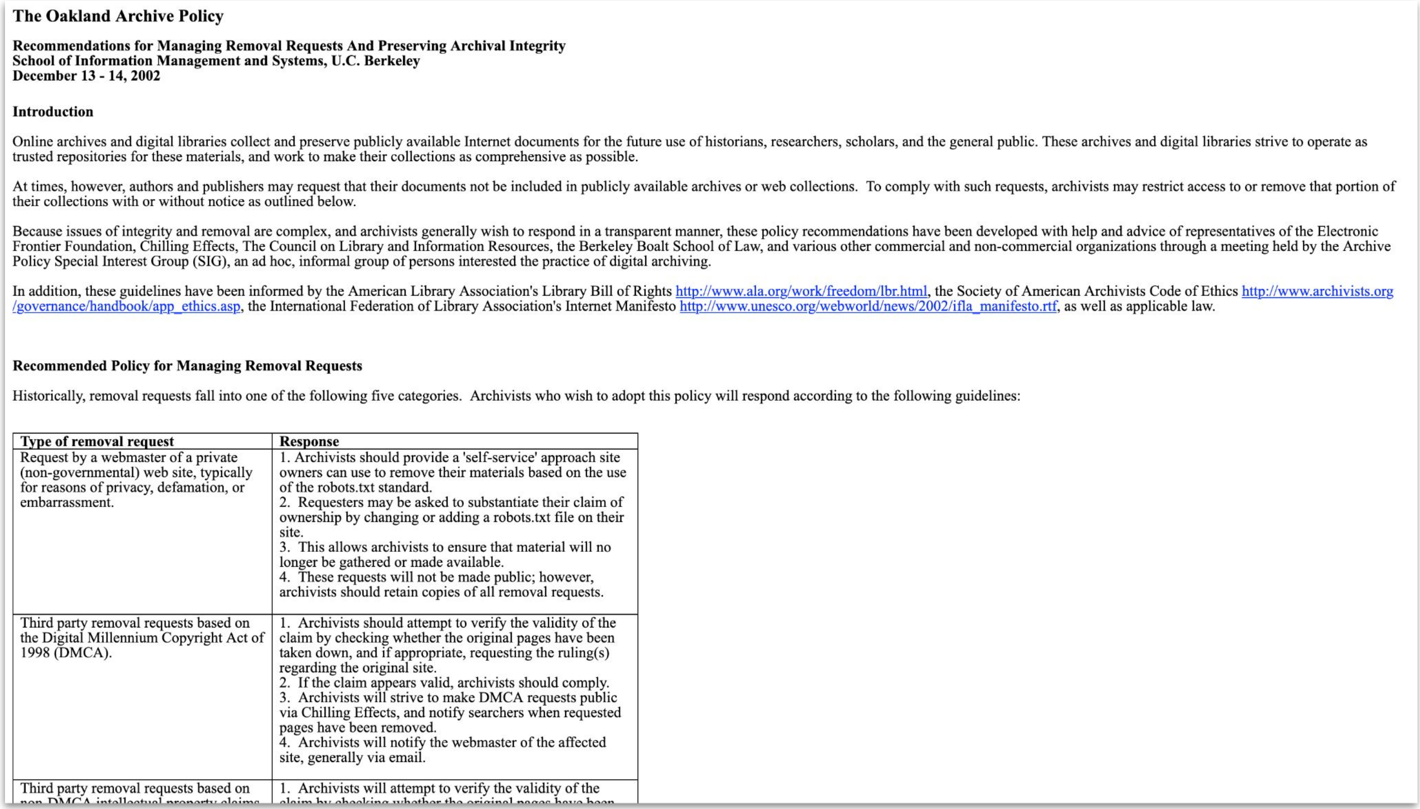
Documenting the Now
DN
ABOUT
TOOLS
COMMUNITY
NEWS
THE TEAM
Josh Williams - photographed by Jamelle Bouie
SHiFT Collective equity by design
Princeton University LIBRARY
Mellon Foundation

ABOUT

Documenting the Now develops open source tools and community-centered practices that support the ethical collection, use, and preservation of publicly available content shared on web and social media.

Documenting the Now responds to the public's use of social media for chronicling historically significant events as well as demand from scholars, students, and archivists, among others, seeking a user-friendly means of collecting and preserving this type of digital content. Documenting the Now has a strong commitment to prioritizing ethical practices when working with social media content, especially in terms of collection and long-term preservation. This commitment extends to Twitter's notion of honoring user intent and the rights of content creators. Documenting the Now is a core program of Shift Collective. We are extremely grateful for funding from the Mellon Foundation and technical support from the Princeton University Library.

The Oakland Archive Policy



The Oakland Archive Policy

Recommendations for Managing Removal Requests And Preserving Archival Integrity
School of Information Management and Systems, U.C. Berkeley
December 13 - 14, 2002

Introduction

Online archives and digital libraries collect and preserve publicly available Internet documents for the future use of historians, researchers, scholars, and the general public. These archives and digital libraries strive to operate as trusted repositories for these materials, and work to make their collections as comprehensive as possible.

At times, however, authors and publishers may request that their documents not be included in publicly available archives or web collections. To comply with such requests, archivists may restrict access to or remove that portion of their collections with or without notice as outlined below.

Because issues of integrity and removal are complex, and archivists generally wish to respond in a transparent manner, these policy recommendations have been developed with help and advice of representatives of the Electronic Frontier Foundation, Chilling Effects, The Council on Library and Information Resources, the Berkeley Boalt School of Law, and various other commercial and non-commercial organizations through a meeting held by the Archive Policy Special Interest Group (SIG), an ad hoc, informal group of persons interested the practice of digital archiving.

In addition, these guidelines have been informed by the American Library Association's Library Bill of Rights http://www.ala.org/work/freedom/lbr.html, the Society of American Archivists Code of Ethics http://www.archivists.org/governance/handbook/app_ethics.asp, the International Federation of Library Association's Internet Manifesto http://www.unesco.org/webworld/news/2002/ifla_manifesto.rtf, as well as applicable law.

Recommended Policy for Managing Removal Requests

Historically, removal requests fall into one of the following five categories. Archivists who wish to adopt this policy will respond according to the following guidelines:

Type of removal request	Response
Request by a webmaster of a private (non-governmental) web site, typically for reasons of privacy, defamation, or embarrassment.	1. Archivists should provide a 'self-service' approach site owners can use to remove their materials based on the use of the robots.txt standard. 2. Requesters may be asked to substantiate their claim of ownership by changing or adding a robots.txt file on their site. 3. This allows archivists to ensure that material will no longer be gathered or made available. 4. These requests will not be made public; however, archivists should retain copies of all removal requests.
Third party removal requests based on the Digital Millennium Copyright Act of 1998 (DMCA).	1. Archivists should attempt to verify the validity of the claim by checking whether the original pages have been taken down, and if appropriate, requesting the ruling(s) regarding the original site. 2. If the claim appears valid, archivists should comply. 3. Archivists will strive to make DMCA requests public via Chilling Effects, and notify searchers when requested pages have been removed. 4. Archivists will notify the webmaster of the affected site, generally via email.
Third party removal requests based on non-DMCA intellectual property claims	1. Archivists will attempt to verify the validity of the claim by checking whether the original pages have been

“Data is a thing, a process, and a relationship we make and put to use. We can make it and use it differently.”

- *Feminist Data Manifest-No*

On Analog vs. Digital Archives