



# Non-textual content in the **DC Punk** web archive

DC Punk Media Unleashed:  
Grace Thomas, James Jacobs  
Laura Wrubel, Oliver Kiechle

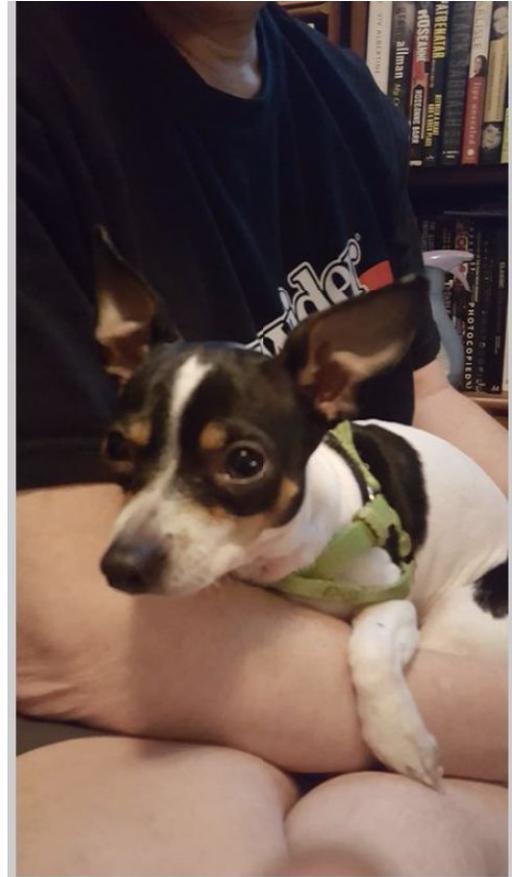
# Research Goals:

- Explore non-textual elements of the DC Punk web archive: 48 GB total
- Hypothesis was that there would be a lot of audio and video objects & we would like to explore that metadata



# Questions:

- What were the seeds?
- How did the crawler find these sites (*bandcamp*, in particular)?
- What is the distribution of MIME types across the archive?
- Are DC punkers dog people or cat people on average?



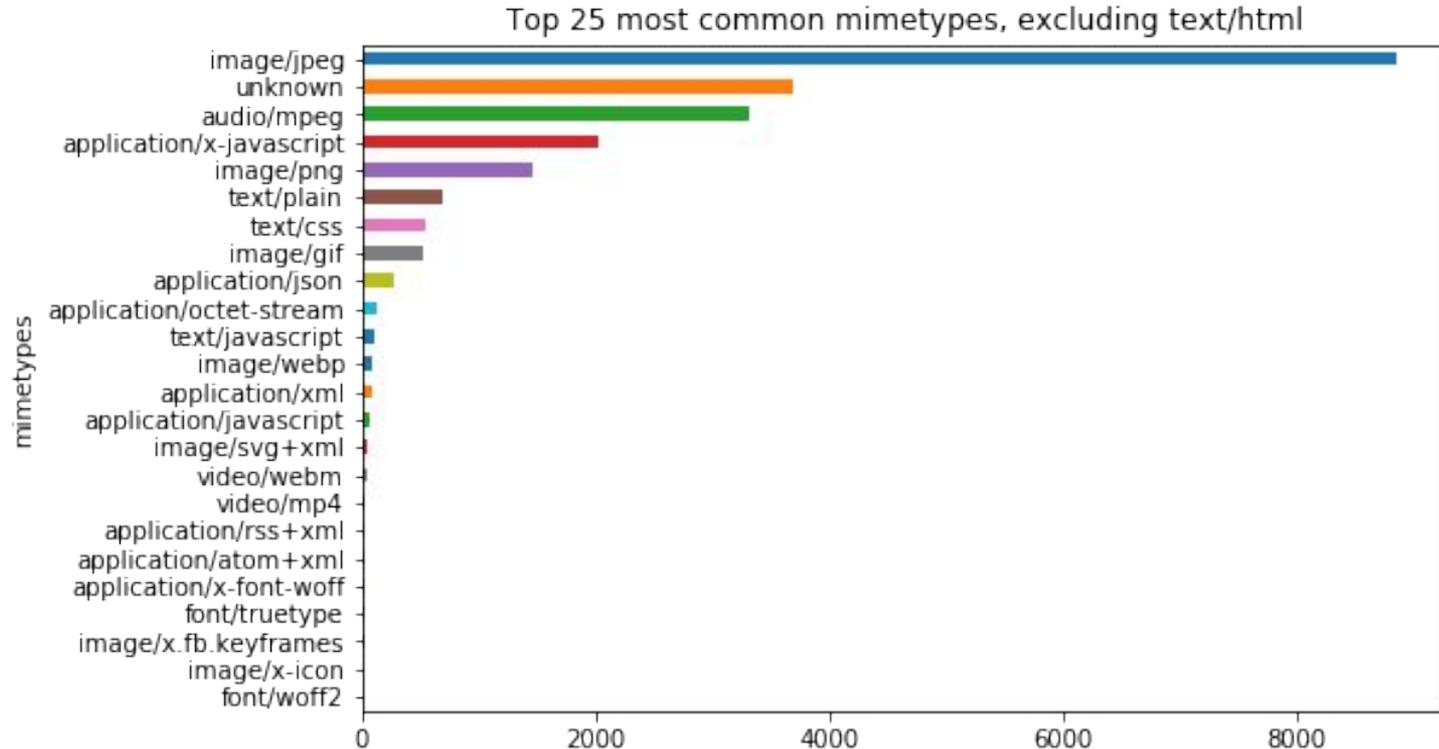
# Processing Steps:

1. Explored the frequency of top-level subdomains in the collection (derivative exported by AU Team)
  - a. Noticed 33 unique bandcamp subdomains
  - b. These unique bands were unrecognized by the local (GWU) punk expert
2. Ran AUT job to extract images from the WARCs
  - a. 10,620 total images: jpeg, gif, png most frequent
  - b. Observed the collection of images - album images, tickets, posters, photographs, general web resources
3. Ran AUT job to count the frequencies of MIME types across the archive
4. Ran AUT job to export all resource URLs and their corresponding MIME types from the WARCs
  - a. Next step focused in on resources reported as audio MIME types, specifically

# Creation of cool things:

- Created a network graph from the Gephi 10427-gephi.gexf file (derivative exported by AU team)
- Created a histogram of format types using Ryan's notebook as a starting point
- Created an image collage using Juxta tool, as recommended by Nick
- Determined dimensions of extracted images using Python PIL library:
  - Identified whether they were squares, horizontal rectangles or vertical rectangles.
  - Based on the ratio of width to height for a sample ticket, guessed at other images which might be tickets.
  - Guessed at album covers based on dimensions & created a chart of frequencies

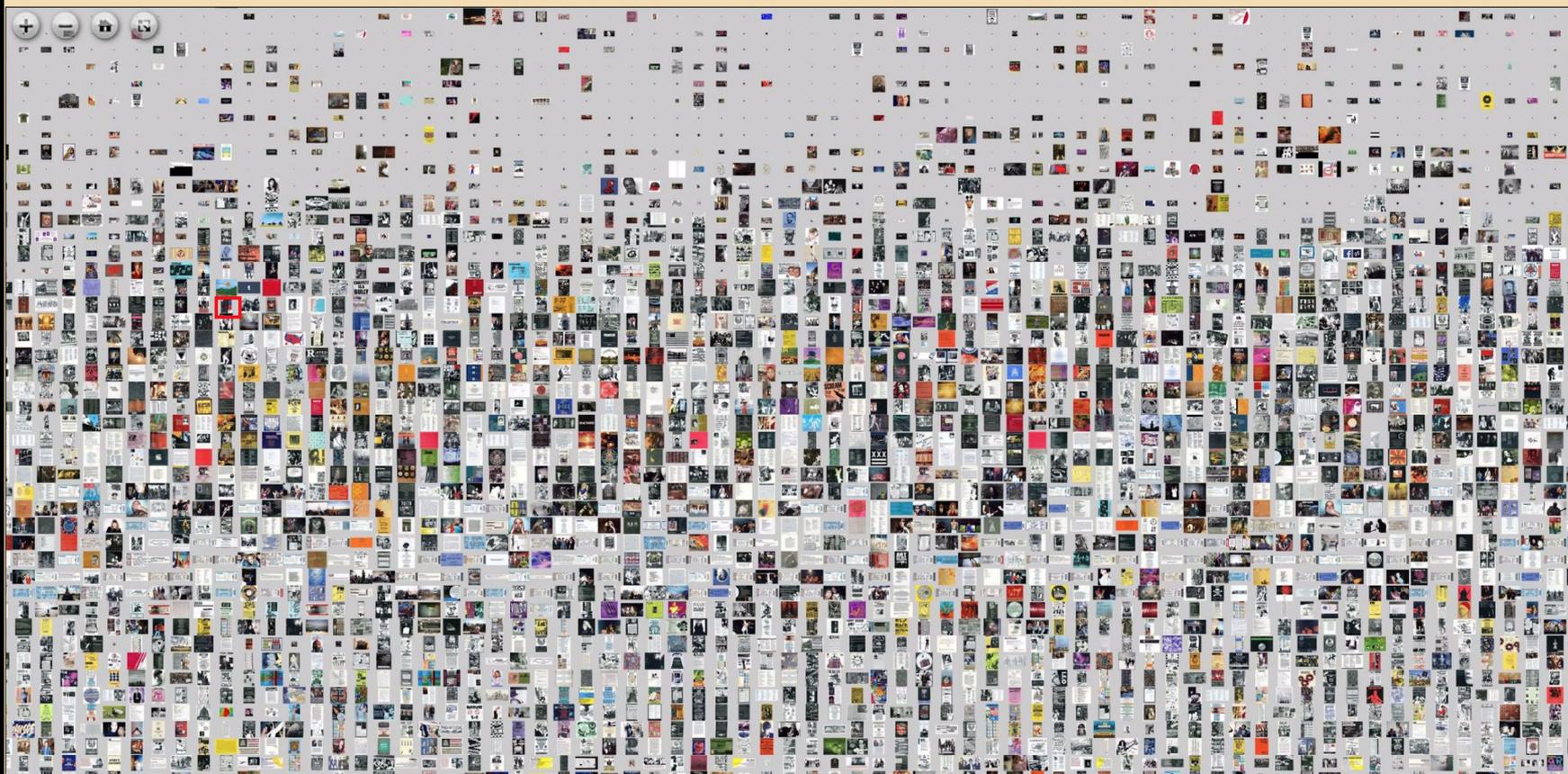
# MIME Type distribution:





# Juxta

Sample page with 10620 images for a total of 8399 MPixels. See [the Juxta GitHub page](#) for details.



# MISLED YOUTH



**"EXCUSE FOR EXISTENCE"**



# Example ticket

**ticketfly**

Sec:General Admission  
General Admission

Richard Lloyd (of Television)

Black Cat  
1811 14th St. NW Washington, DC

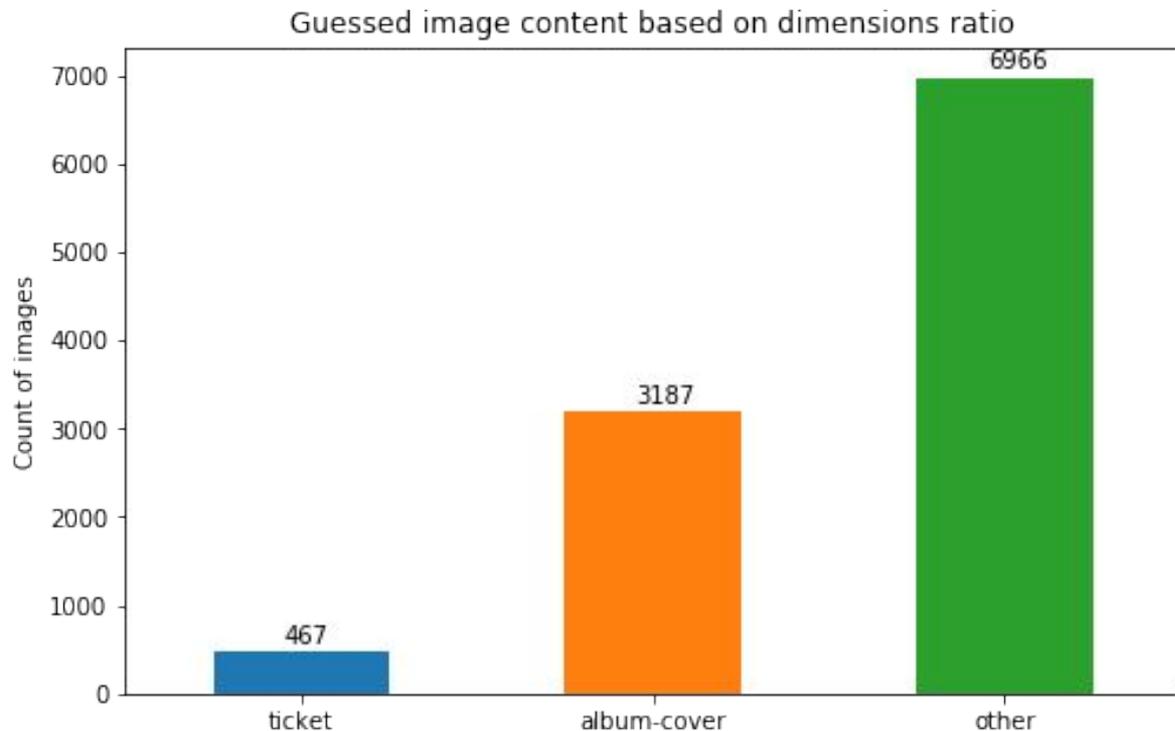
Thu Jun 1, 2017 7:30 PM  
Price:\$12.00

Order:094433288523      Purchased By:Order Box Office

Sec:General...  
Jun 1 2017 7:30 PM \$12.00  
00985914473968



# Further analysis of the images:



# Party Off The Pounds!

TUESDAY MARCH 12

## THE SNIFFS

(rock'n'roll rejects)

## WAILIN STORMS

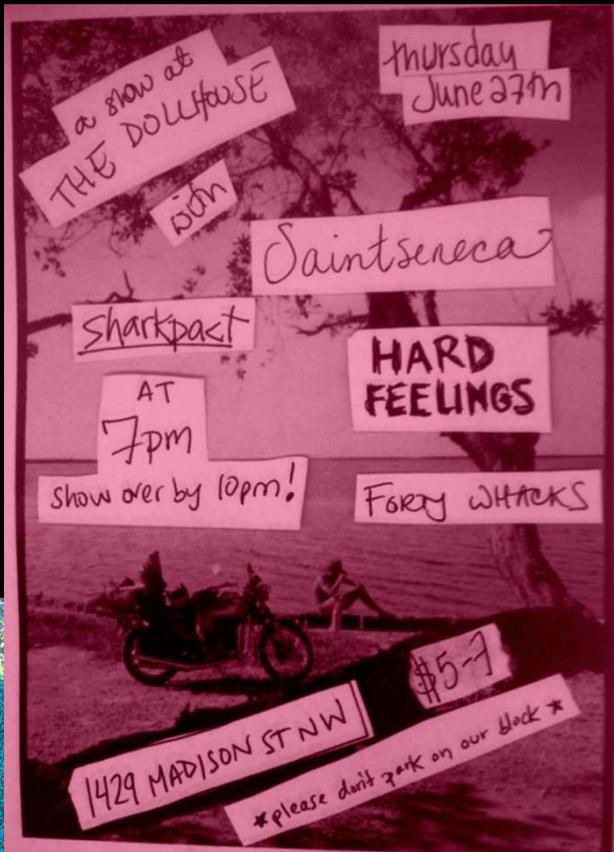
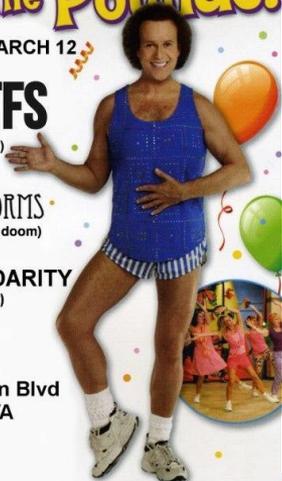
(Brooklyn garage doom)

## JAIL SOLIDARITY

(DC noisy sludge)

CD Cellar  
2607 Wilson Blvd  
Arlington VA

8 PM \$5



THE  
**HOWARD**  
THEATRE  
EST. 1910

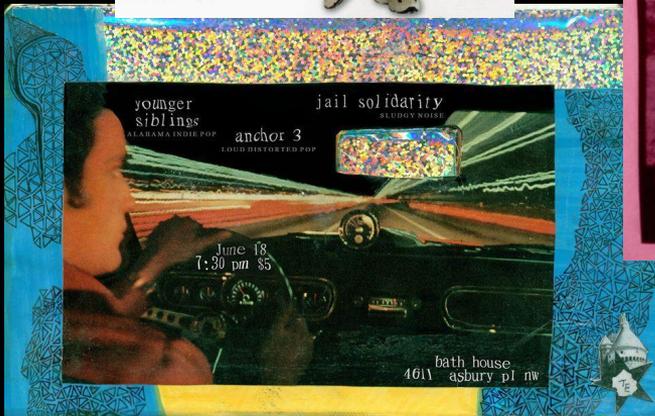
# BUDDY GUY

## WITH QUINN SULLIVAN



THURSDAY, OCTOBER 25TH &  
FRIDAY, OCTOBER 26TH

620 T ST. NW WASHINGTON D.C. 20001 : [WWW.THEHOWARDTHEATRE.COM](http://WWW.THEHOWARDTHEATRE.COM)



# Which sites have audio files?

domain	count
<a href="http://mrjimijam.com">mrjimijam.com</a>	3000
<a href="http://t4.bcbits.com">t4.bcbits.com</a>	284
<a href="http://static.xx.fbcdn.net">static.xx.fbcdn.net</a>	21
<a href="http://p.scdn.co">p.scdn.co</a>	13



# Next up / further study:

- What (kinds of) resources are labeled with the “unknown” MIME type?
- Analysis of audio files:
  - Melodies
  - Composition
  - Purported gender of vocalists
  - Instruments
  - Genre & comparison with other genres
  - Extraction & analysis of lyrics
  - Earworms
- Color analysis of images
- OCR of posters/tickets: do cover charge, venues indicate a certain level of cultural saturation / popularity / audience?
- Network analysis of MrJimiJam’s site (hub of the archive)
- We still don’t know if DC punkers are cat people or dog people...

# Takeaways

- Scope notes and communication with collection-builders are important for framing the provenance of the data. However, we were still able to glean these decisions from the network graph - can see that the boards are linked to bandcamp.
- Even though audio and video seemed immediately important to us, someone studying this particular subculture could find the images very useful since they included tickets, posters/flyers, cover art, photographs of events and bands, lyrics, popular GIFs found on the web.
- Issues about copyright and unauthorized recordings/sharing of albums come to the forefront.
- It takes a lot of time and computing power to analyze non-textual resources in web archives - thank you AU team for working to break down this barrier!!

**Thank you to Paul Kelly (DCPL) for curating this collection and providing it for our use & Tina Plottel (GWU) for sharing her expertise on the DC punk scene.**

**Thank you to the Archives Unleashed Team & Compute Canada for providing hand-holding & fabulous tools!**

