Websites, Web Archives, WARCs, and Archived Web Objects

Grace Thomas

**Digital Collections Specialist,
Library of Congress
Web Archiving Team**

Designing Storage Architectures
September 10, 2019

SCALE

17,440

web archives available on loc.gov

3.9 PB

total size

1.85 PB

unique size

9 Million

total files on storage

4.3 Million

unique files on storage

18+ Billion

total archived web objects

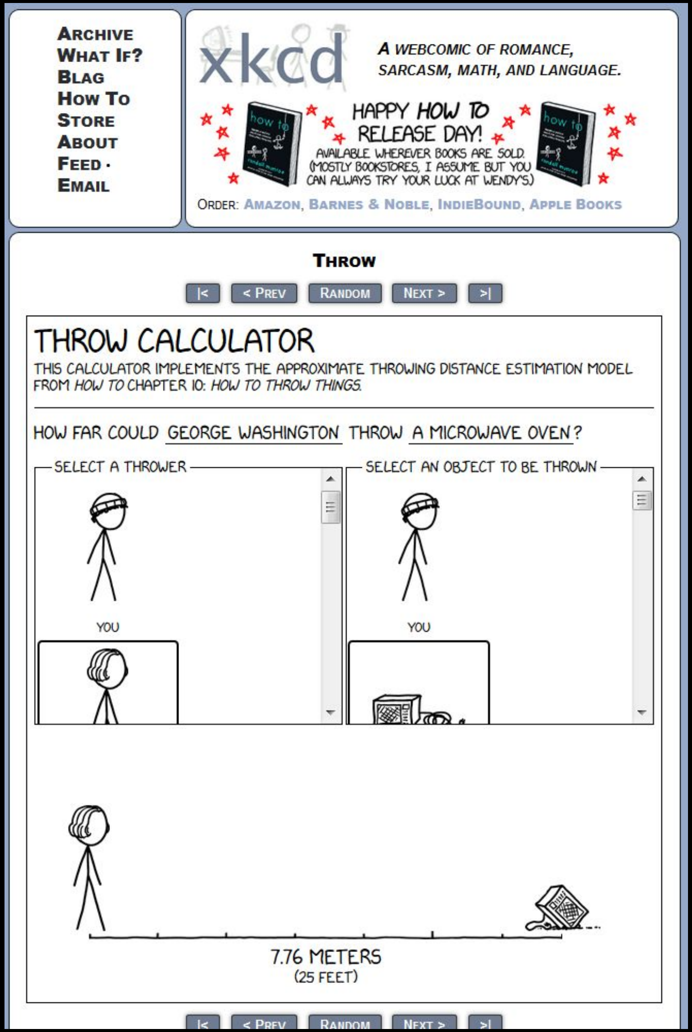
6.2+ Billion

unique archived web objects

Discussion Points

- Lifecycle of web archives data from live website to containers on storage to public-facing, replayed content

- Using Wayback Machine indexes to study individual resources and groups of individual resources in the archive



ARCHIVE
WHAT IF?
BLAG
HOW TO
STORE
ABOUT
FEED ·
EMAIL

xkcd

A WEBCOMIC OF ROMANCE, SARCASM, MATH, AND LANGUAGE.

HAPPY HOW TO RELEASE DAY!
AVAILABLE WHEREVER BOOKS ARE SOLD. (MOSTLY BOOKSTORES, I ASSUME. BUT YOU CAN ALWAYS TRY YOUR LUCK AT WENDY'S.)

ORDER: AMAZON, BARNES & NOBLE, INDIEBOUND, APPLE BOOKS

THROW

|< &lt; PREV RANDOM NEXT &gt; &gt;|

THROW CALCULATOR

THIS CALCULATOR IMPLEMENTS THE APPROXIMATE THROWING DISTANCE ESTIMATION MODEL FROM *HOW TO* CHAPTER 10: *HOW TO THROW THINGS.*

HOW FAR COULD GEORGE WASHINGTON THROW A MICROWAVE OVEN?

SELECT A THROWER

YOU

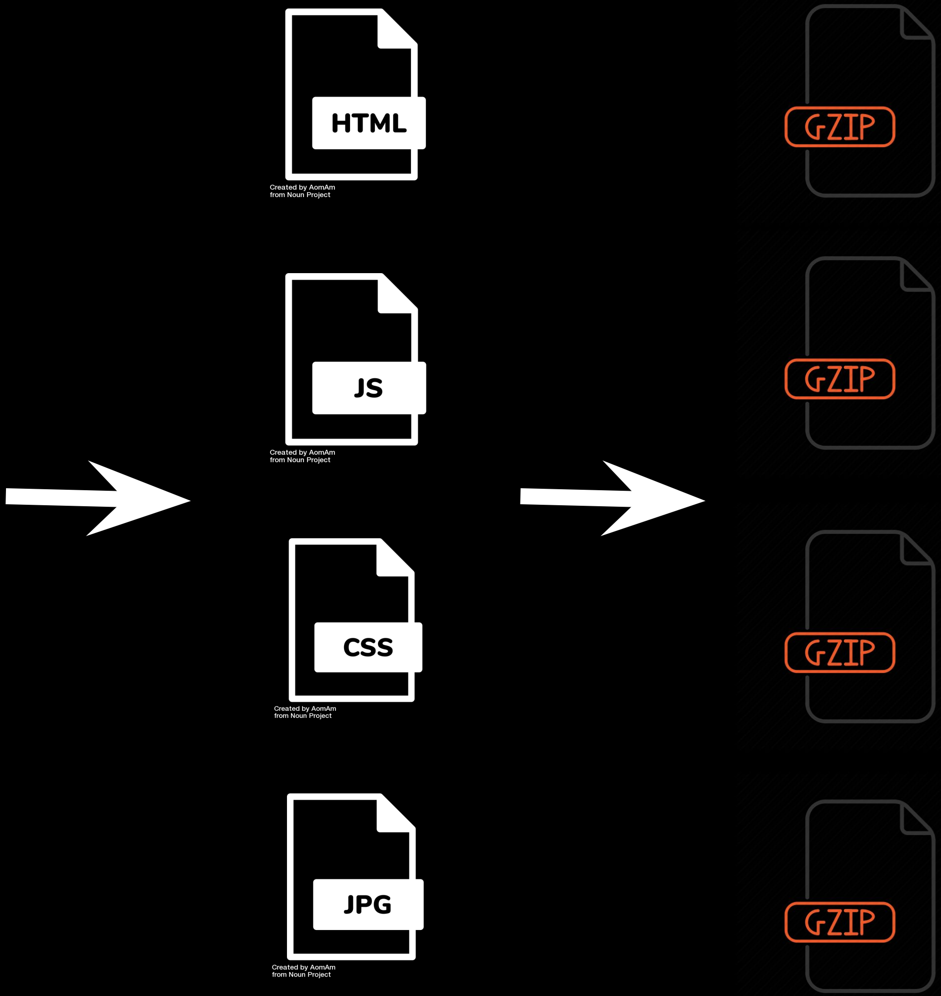
SELECT AN OBJECT TO BE THROWN

YOU

7.76 METERS
(25 FEET)

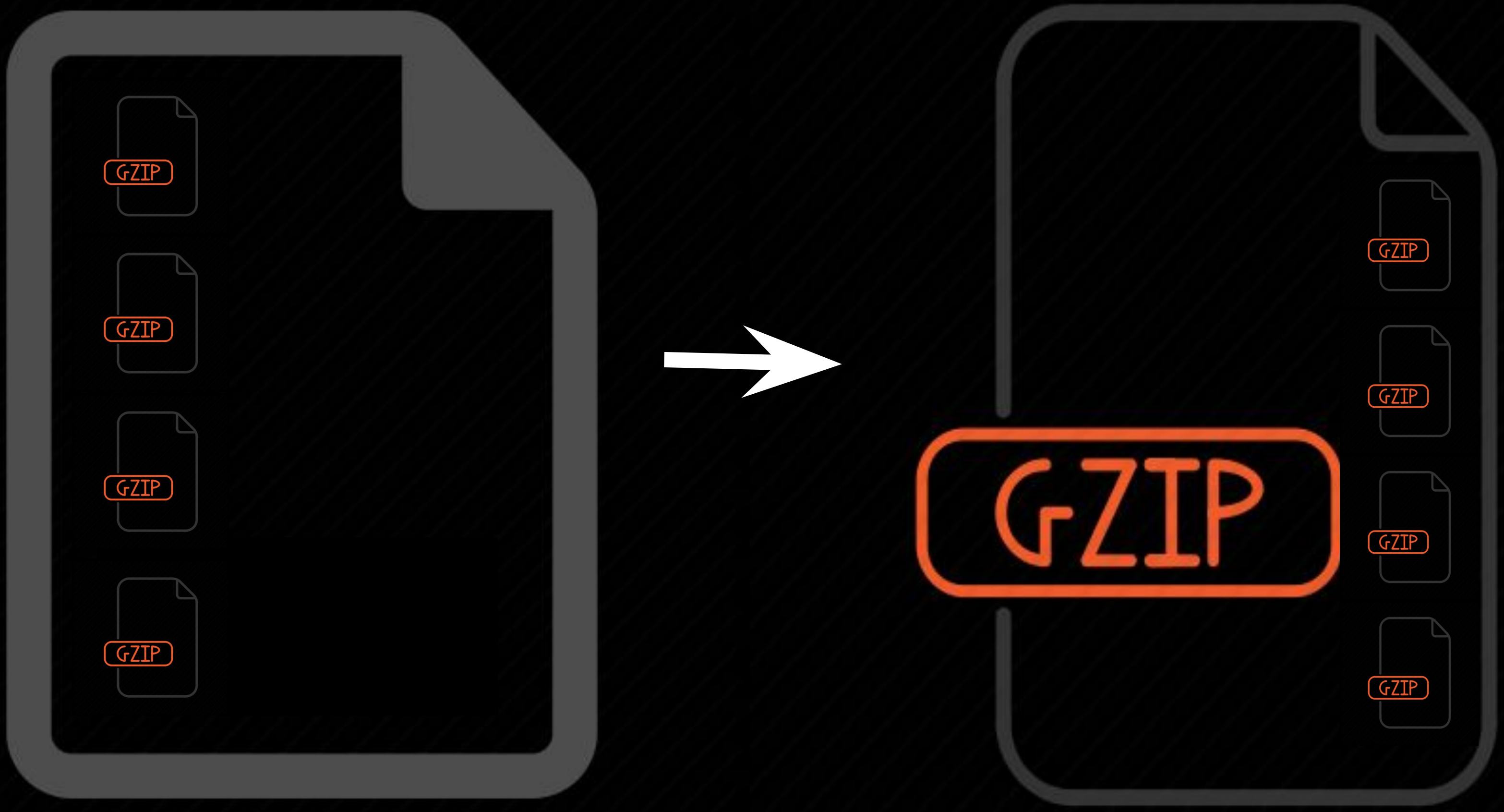
|< &lt; PREV RANDOM NEXT &gt; &gt;|

www.xkcd.com, accessed 09/04/2019


HTML
Created by AomAm from Noun Project
JS
Created by AomAm from Noun Project
CSS
Created by AomAm from Noun Project
JPG
Created by AomAm from Noun Project
GZIP
GZIP
GZIP
GZIP

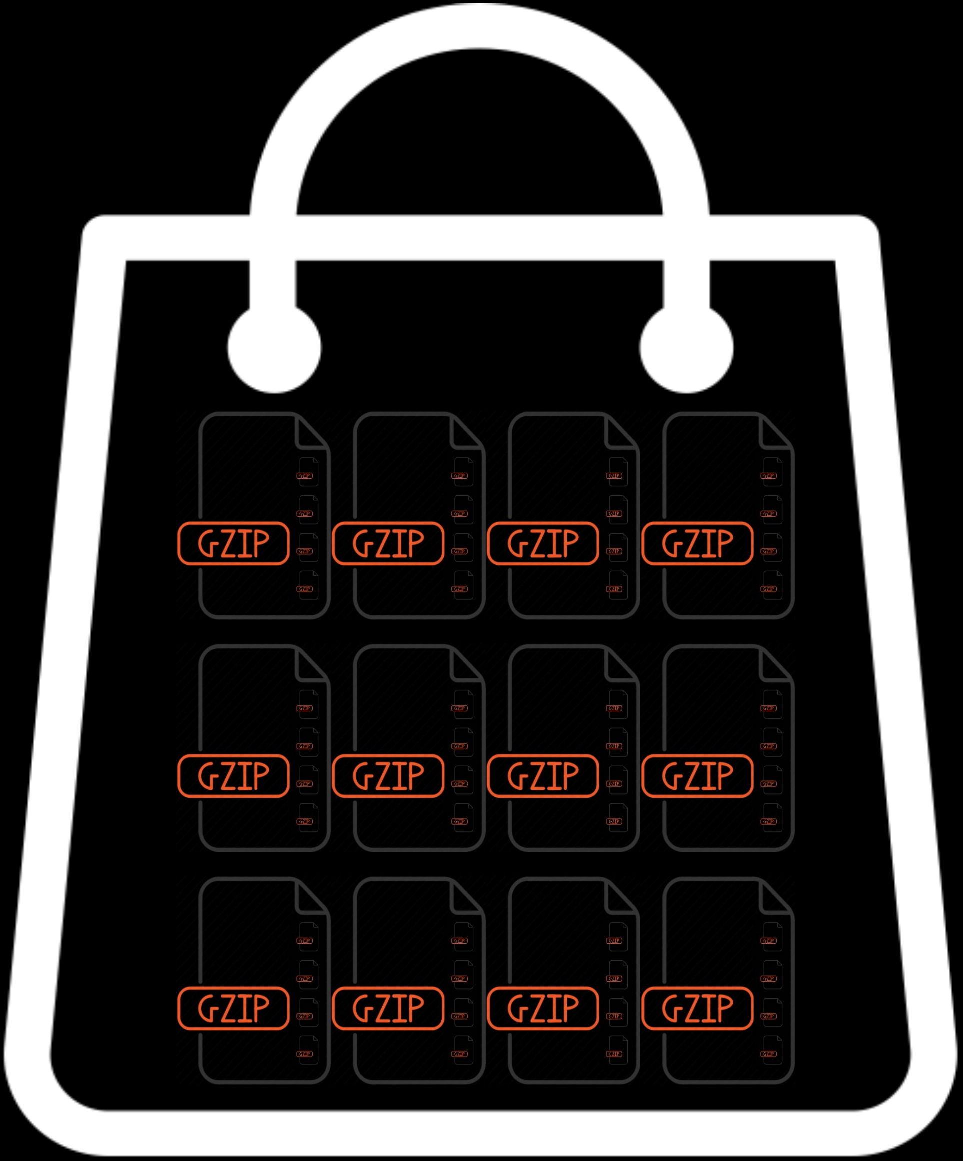
.warc

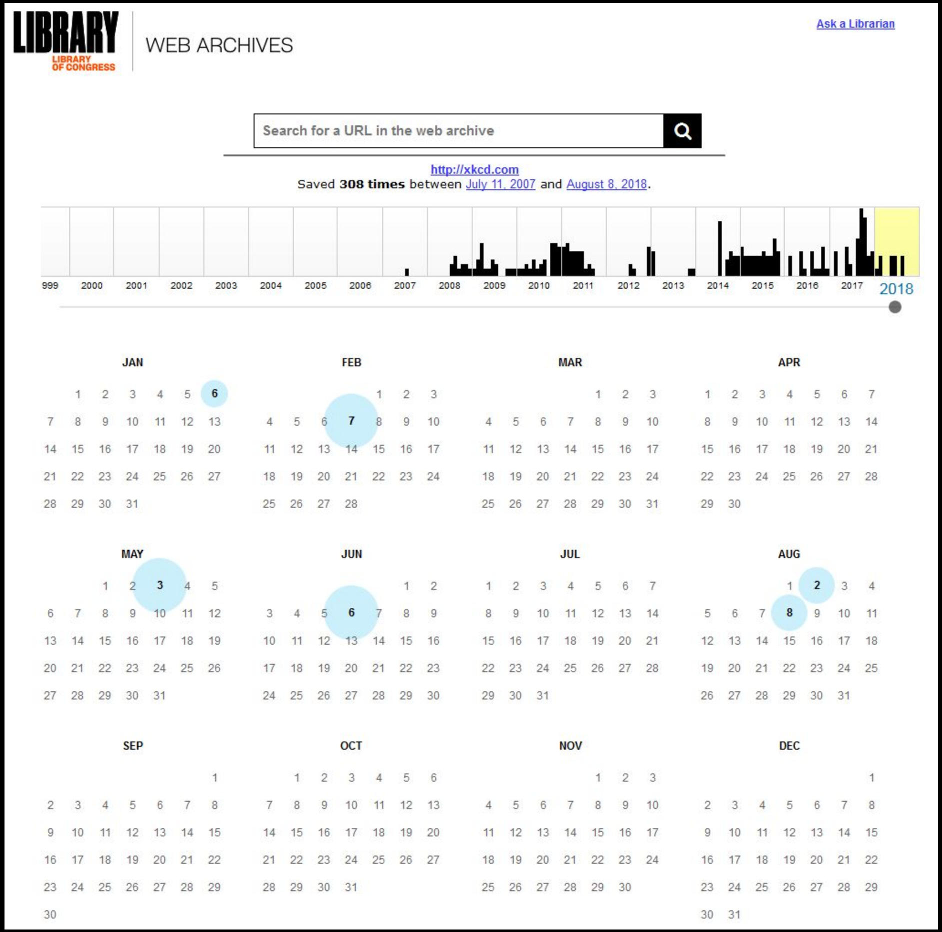
.warc.gz - 1GB



BagIt Bags - 1TB

- Copy on long-term storage
- Copy on presentation storage





http://xkcd.com
Saved **308 times** between July 11, 2007 and August 8, 2018.

JAN

	1	2	3	4	5	**6**
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

FEB

				1	2	3
4	5	6	**7**	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28			

MAR

				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

APR

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

MAY

		1	2	**3**	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

JUN

					1	2
3	4	5	**6**	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

JUL

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

AUG

			1	**2**	3	4
5	6	7	**8**	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

SEP

						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

OCT

	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

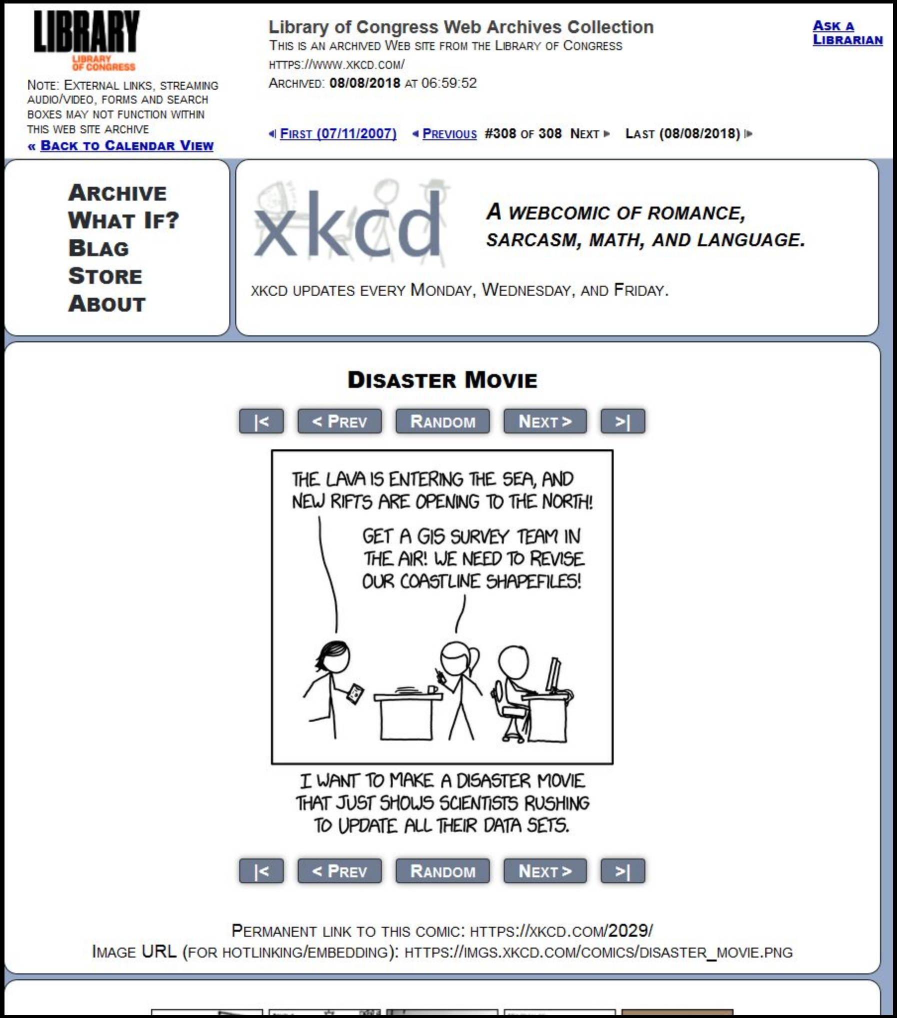
NOV

				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

DEC

						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

webarchive.loc.gov/all/*/http://xkcd.com


LIBRARY
LIBRARY OF CONGRESS
Ask a Librarian
Archive
What If?
Blag
Store
About
xkcd
|<
< Prev
Random
Next >
>|
THE LAVA IS ENTERING THE SEA, AND NEW RIFTS ARE OPENING TO THE NORTH!
GET A GIS SURVEY TEAM IN THE AIR! WE NEED TO REVISE OUR COASTLINE SHAPEFILES!
|<
< Prev
Random
Next >
>|

Library of Congress Web Archives Collection
This is an archived Web site from the Library of Congress
https://www.xkcd.com/
Archived: **08/08/2018** at 06:59:52

Note: External links, streaming audio/video, forms and search boxes may not function within this web site archive
« **Back to Calendar View**

First (07/11/2007) Previous **#308** of **308** Next Last (08/08/2018)

A webcomic of romance, sarcasm, math, and language.

xkcd updates every Monday, Wednesday, and Friday.

Disaster Movie

I WANT TO MAKE A DISASTER MOVIE THAT JUST SHOWS SCIENTISTS RUSHING TO UPDATE ALL THEIR DATA SETS.

Permanent link to this comic: https://xkcd.com/2029/
Image URL (for hotlinking/embedding): https://imgs.xkcd.com/comics/disaster_movie.png

webarchive.loc.gov/all/20180808065952/https://www.xkcd.com/

Data Lifecycle - Access

Wayback Indexes

- Sit between the content (W/ARCs) on presentation storage and the Wayback Machine software to replay content

- Each line represents a single archived web object with related metadata

Fields:

- canonized URL
- capture timestamp
- full URL
- reported MIME type of original document
- response code
- checksum
- document length (size)
- W/ARC offset
- W/ARC filename

Using Wayback Indexes Creatively

- Counting total & unique objects
- Qualifying those numbers based on:
  - HTTP response (do 404s count as objects since their records take up space?)
  - Renderable content (200s; uncorrupted)

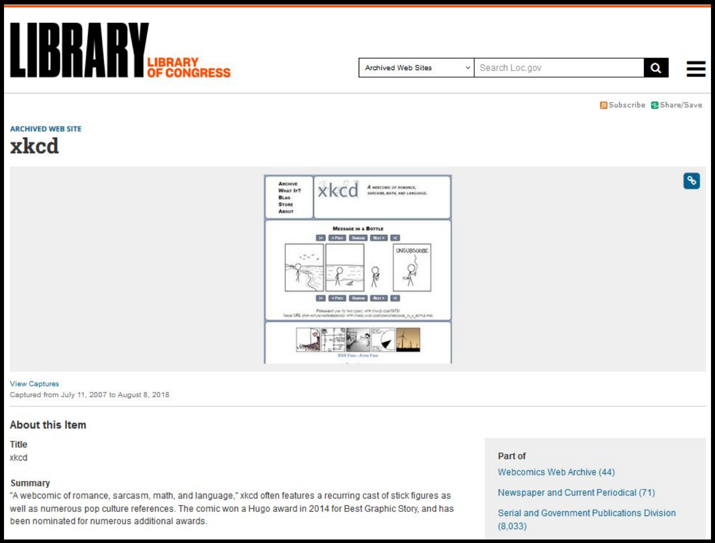
- Number of objects per "collection" or "web archive" as defined by LC

- Grouping unique objects by MIME type
- Counting objects of certain MIME type per "collection" or "web archive"

HTML
Created by AomAm
from Noun Project

CSS
Created by AomAm
from Noun Project

JPG
Created by AomAm
from Noun Project



LIBRARY OF CONGRESS

Archived Web Sites | Search Loc.gov

Subscribe Share/Save

ARCHIVED WEB SITE

xkcd

View Captures
Captured from July 11, 2007 to August 8, 2018

About this Item

Title
xkcd

Summary
"A webcomic of romance, sarcasm, math, and language," xkcd often features a recurring cast of stick figures as well as numerous pop culture references. The comic won a Hugo award in 2014 for Best Graphic Story, and has been nominated for numerous additional awards.

Part of
Webcomics Web Archive (44)
Newspaper and Current Periodical (71)
Serial and Government Publications Division (8,033)

www.loc.gov/item/lcwaN0009950/

Thank you!

Grace Thomas

@gracehthom

grth@loc.gov
webcapture@loc.gov