HDSM
Humanities Data Science & Methodology, TU Darmstadt

Digital History 2024 Switzerland

TECHNISCHE UNIVERSITÄT DARMSTADT
Institut für Geschichte

Data-Driven Approaches to Studying the History of Museums on the Web: Challenges and Opportunities for New Discoveries

Dr. Nadezhda Povroznik
Wissenschaftliche Mitarbeiterin
FG - Humanities Data Science and Methodology
Institut für Geschichte
Technische Universität Darmstadt

September 12–13

Digital History 2024 Switzerland

University of Basel

DigiHistCH24 conference theme:
‹Historical Research, Digital Literacy and Algorithmic Criticism›

September 12–13

Digital History 2024 Switzerland

University of Basel

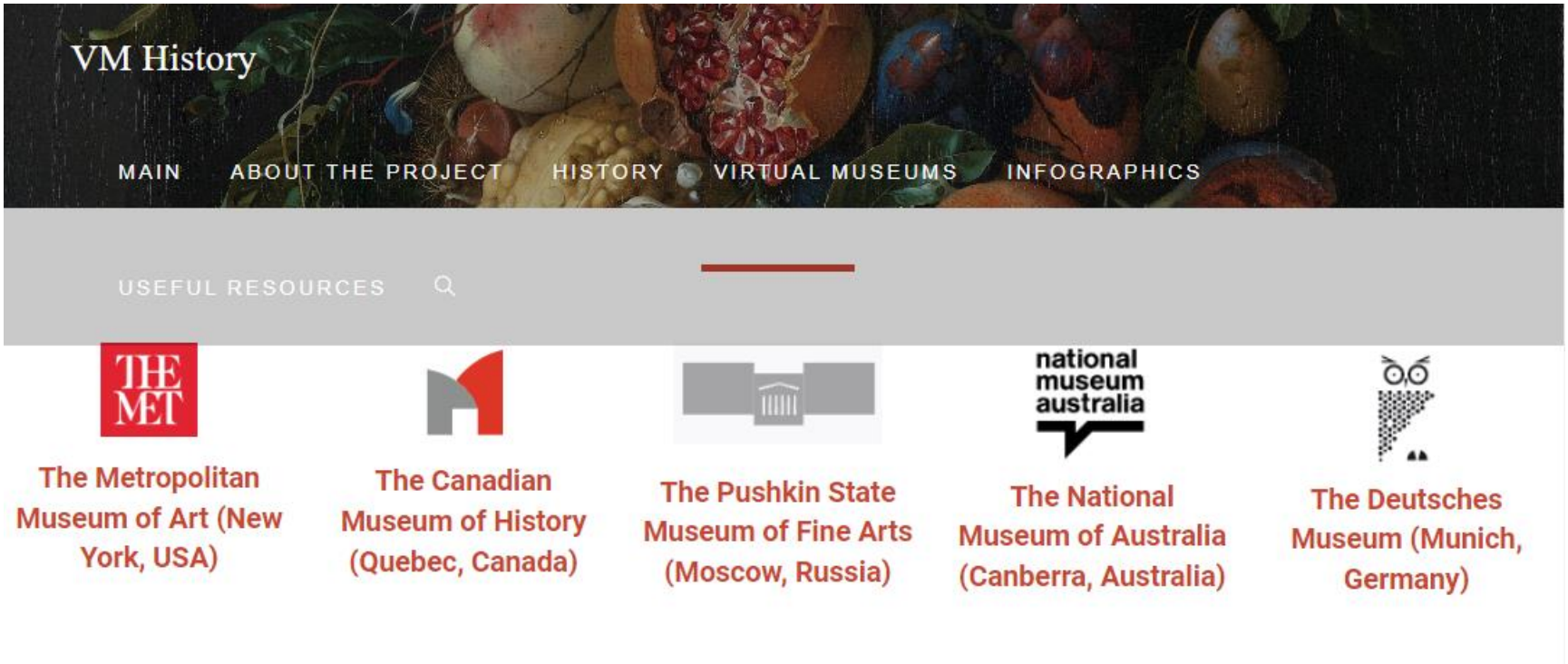
DigiHistCH24 conference theme:
‹Historical Research, Digital Literacy and Algorithmic Criticism›

Digital History of Virtual Museums

https://virtualmuseumhistory.com/

Digital History of Virtual Museums: Case studies

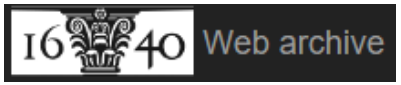
https://virtualmuseumhistory.com/virtual-museums/



The image is generated by ChatGPT 4o.

The prompt: illustrate data-driven approach to the history of the museums on the web

INTERNET ARCHIVE


1640 Web archive

Trove

digitálne pramene

WAS
WEB
ARCHIVE
SINGAPORE

YORK
UNIVERSITE
UNIVERSITY

BnF
Bibliothèque
nationale de France

Stanford LIBRARIES

EU Web archive

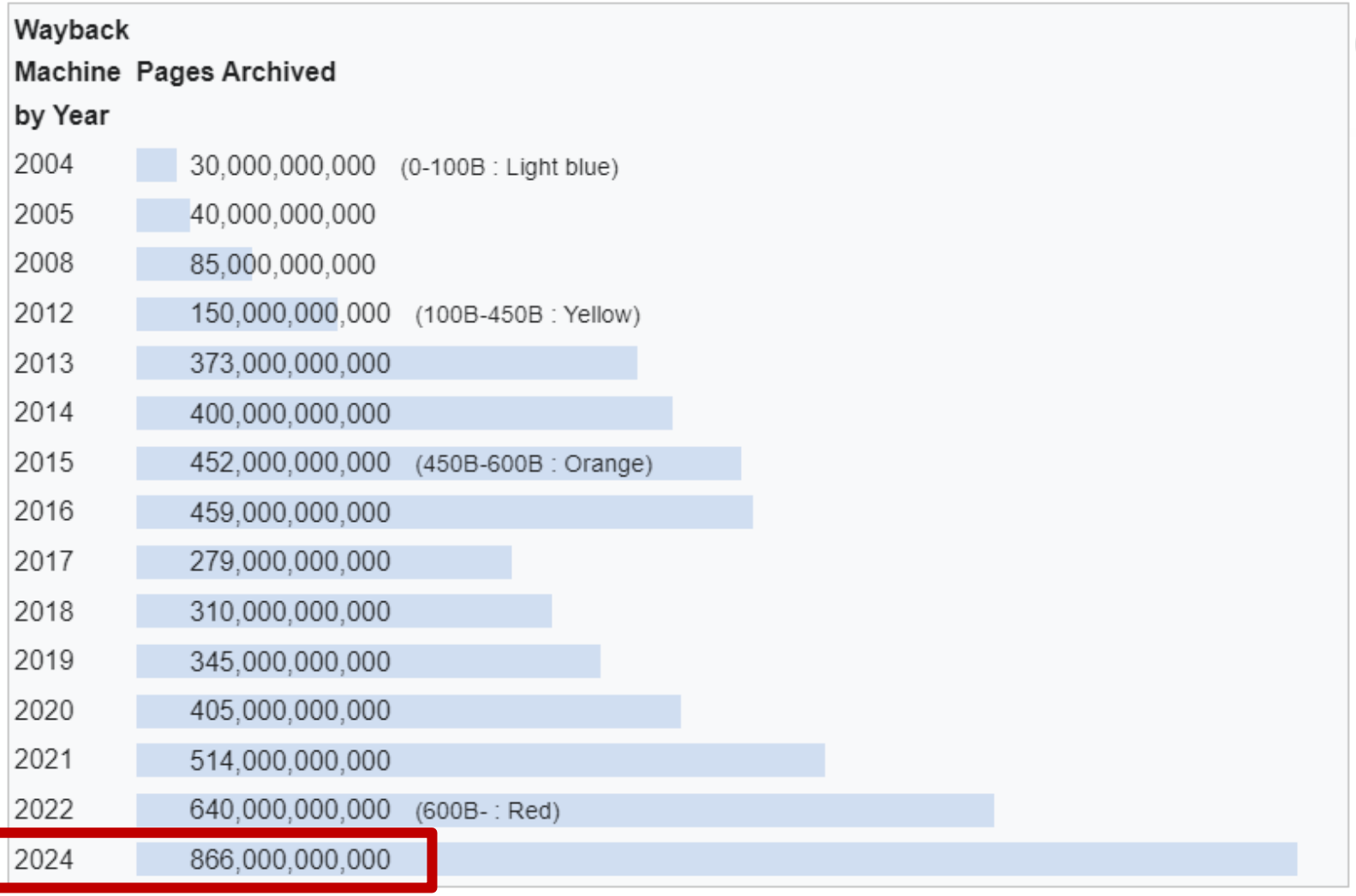
Bibliothèque nationale du Luxembourg
Nationalbibliothéik

国立国会図書館
National Diet Library, Japan

DEUTSCHE
NATIONAL
BIBLIOTHEK

UKWA
UK WEB ARCHIVE

Wayback Machine Growth



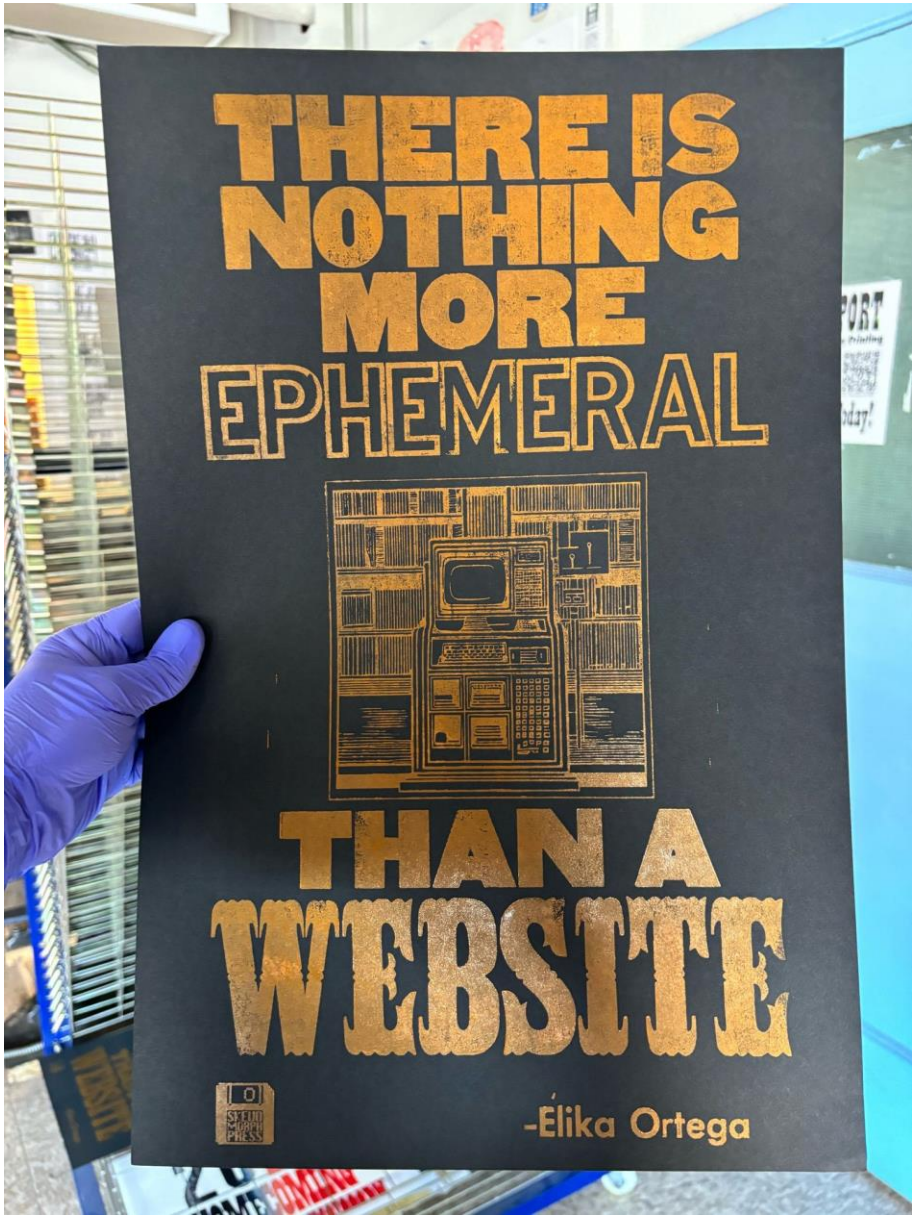
Wayback Machine. Wikipedia, https://en.wikipedia.org/wiki/Wayback_Machine

Big-Data Issues

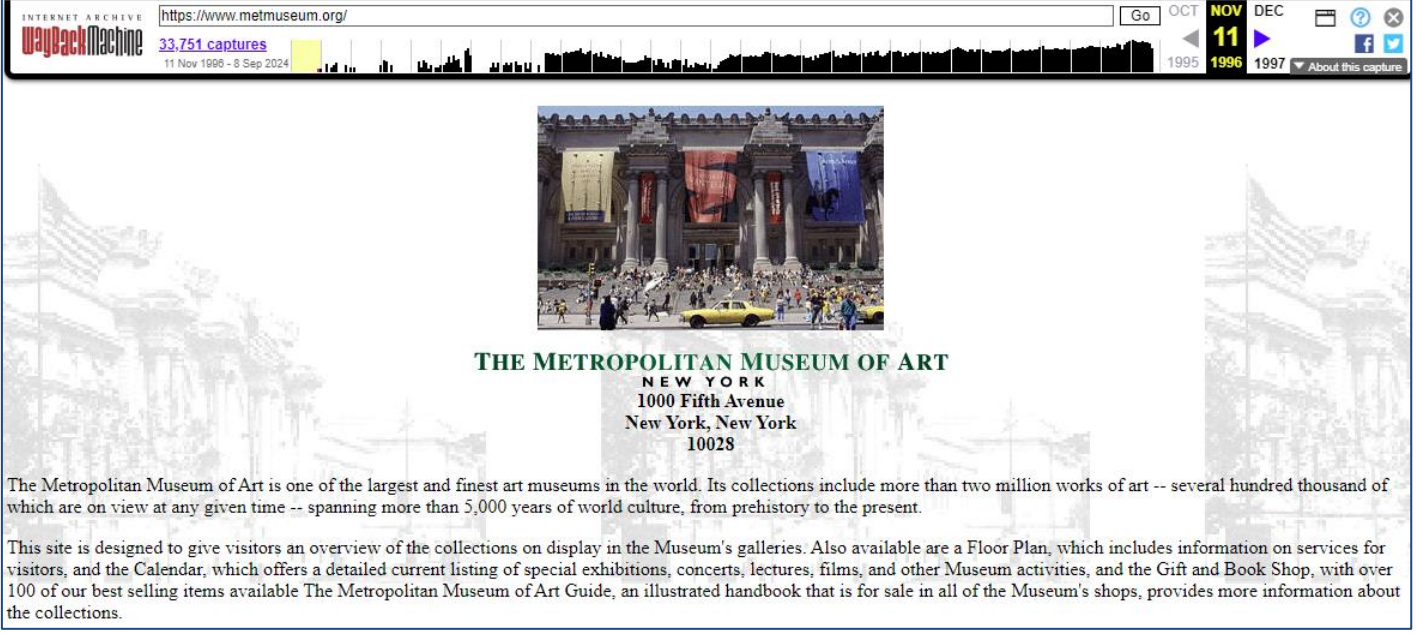
- **huge in volume**, consisting of terabytes or petabytes of data;
- **high in velocity**, being created in or near real-time;
- **diverse in variety**, being structured and unstructured in nature;
- **exhaustive in scope**, striving to capture entire populations or systems (n = all);
- **fine-grained in resolution** and uniquely indexical in identification;
- **relational in nature**, containing common fields that enable the conjoining of different data sets;
- **flexible**, holding the traits of extensionality (can add new fields easily) and **scaleability** (can expand in size rapidly).

Kitchin, Rob. 2014. "Big Data, New Epistemologies and Paradigm Shifts." *Big Data & Society* 1 (1). https://doi.org/10.1177/2053951714528481

What do we study?

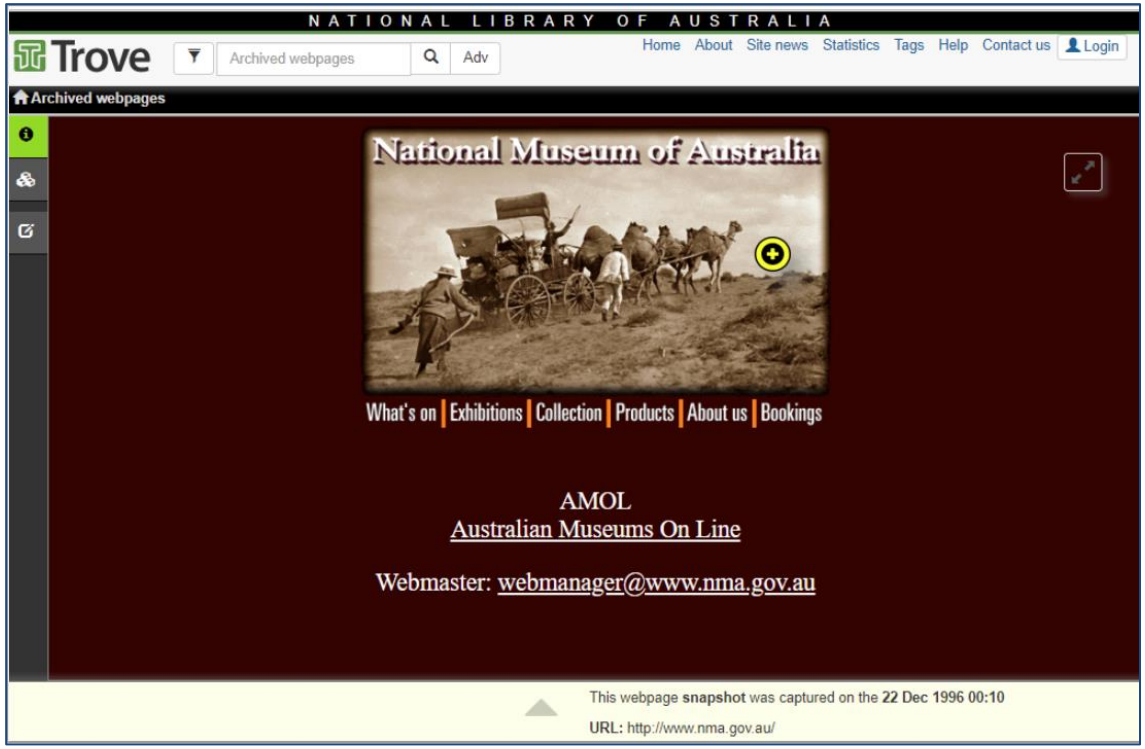


Skeuomorph Press & BookLab at UIUC
https://bsky.app/profile/ryancordell.bsky.social/post/3l3esvnozbh26



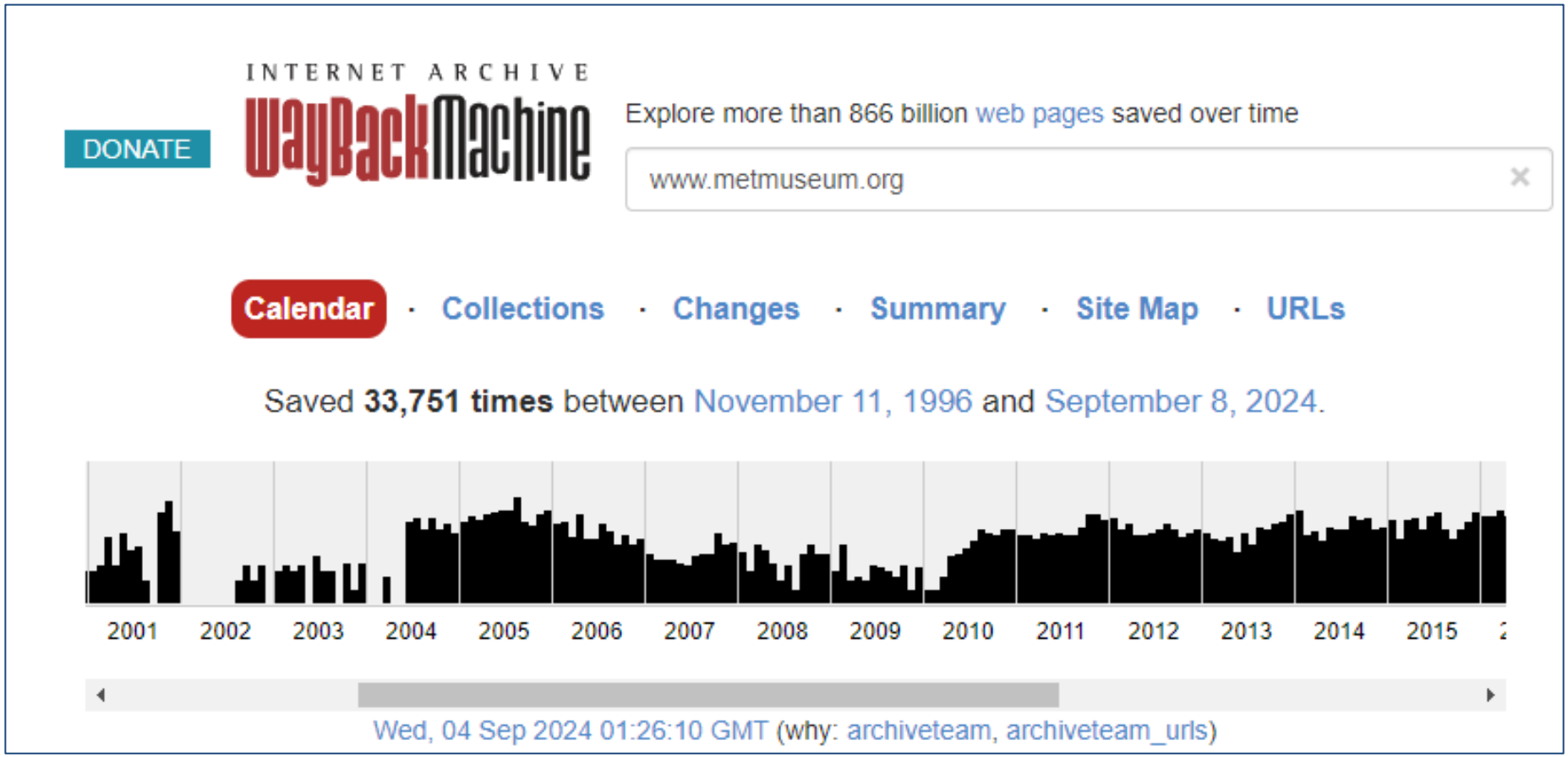
The Metropolitan Museum of Modern Art, Wayback Machine, Internet Archive. Snapshot on 11.11.1996, https://web.archive.org/web/19961111003737/http://www.metmuseum.org

Case Studies



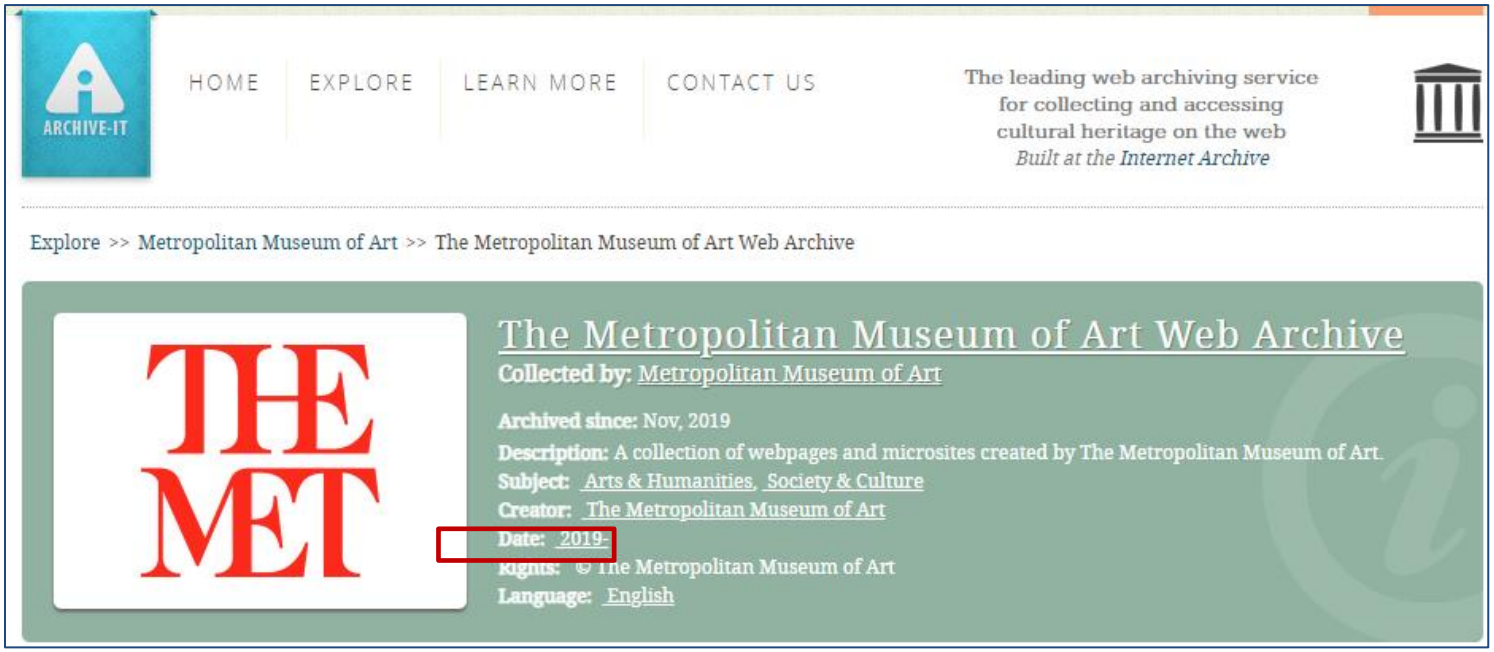
National Museum of Australia, TROVE. Snapshot on 22.12.1996, https://webarchive.nla.gov.au/awa/19961221131034/http://www.nma.gov.au

Obtaining Data: The MET



metmuseum.org, Calendar, Wayback Machine, Screenshot on 08.09.2024,
https://web.archive.org/web/20240000000000*/metmuseum.org

Obtaining Data: The MET



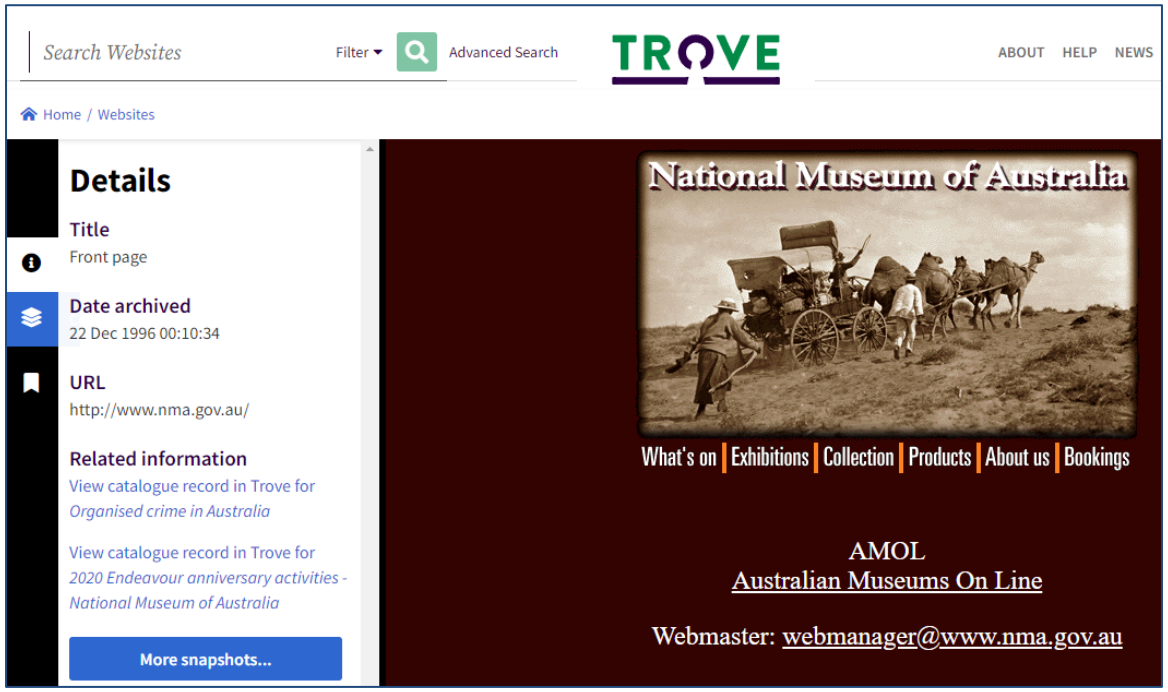
The Metropolitan Museum of Art Web Archive, Archive-IT, https://www.archive-it.org/collections/13135

Subject

Art (348)
Artists (252)
Interviews (244)
Art History (241)
Publication (238)
Exhibitions (47)
Publications (24)
History (22)
Programs (17)
Archives (16)
Libraries (16)
Blogs (14)
Fashion (8)
Education (7)
Politics and culture (4)

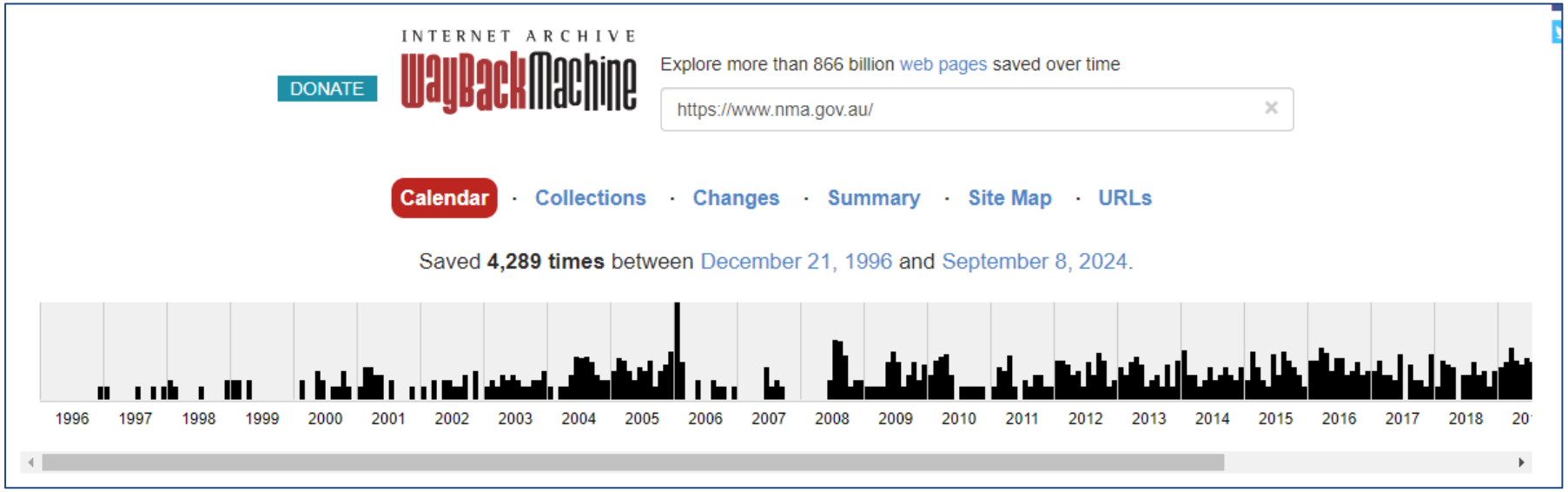
Date

2015- (241)
2020- (45)
2021- (17)
2000- (15)
2019- (6)
2023- (6)
2020 (5)
2011- (3)
2018- (3)
2021 (3)
Various (3)
2022 (2)
1938- (1)
2000 (1)
2007- (1)



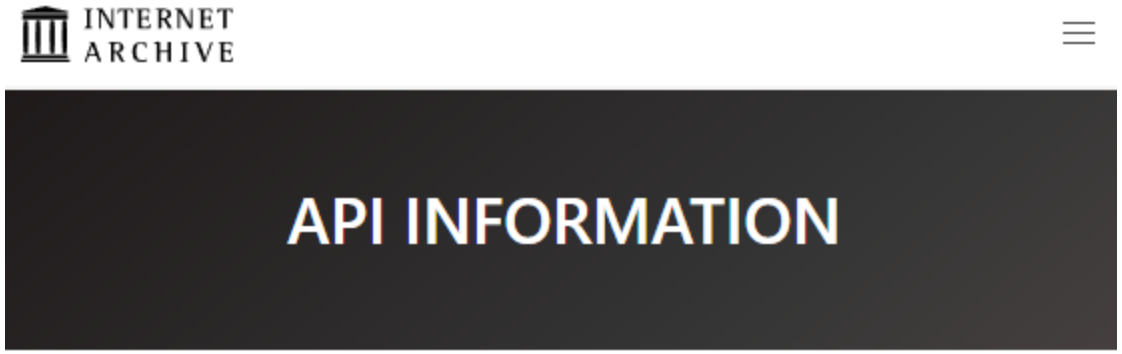
National Museum of Australia, TROVE. Snapshot on 22.12.1996, https://webarchive.nla.gov.au/awa/19961221131034/http://www.nma.gov.au

NMA



Nma.gov.au, Calendar, Wayback Machine, Screenshot on 08.09.2024, https://web.archive.org/web/20240000000000*/Nma.gov.au

How do we obtain data?



INTERNET ARCHIVE
WayBackMachine

host www.metmuseum.org

	Captures	Unique URLs
text/html	11,137,508	5,693,445
image/jpeg	4,431,051	305,778

Wayback Machine Downloader

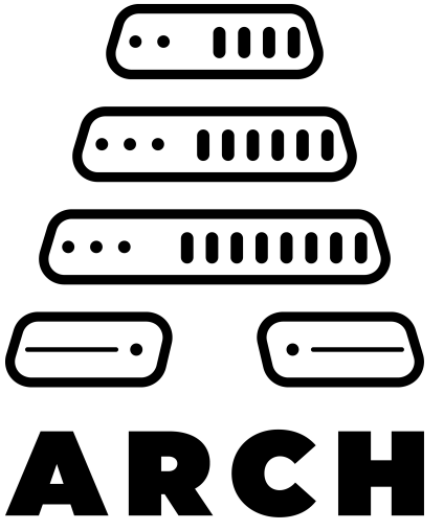
https://github.com/hartator/wayback-machine-downloader

```
Downloading https://www.metmuseum.org/ to websites/www.metmuseum.org/ from Wayback Machine archives.

Getting snapshot pages.....................................................................................
... found  1372457 snapshots to consider.

file_list_curated: 1342067
1342067 files to download:
http://www.metmuseum.org:80/ -> websites/www.metmuseum.org/19961111003737/index.html (1/1342067)
http://www.metmuseum.org:80/ -> websites/www.metmuseum.org/19961219004113/index.html (2/1342067)
http://www.metmuseum.org:80/ -> websites/www.metmuseum.org/19970217203149/index.html (3/1342067)
http://www.metmuseum.org:80/ -> websites/www.metmuseum.org/19970416052424/index.html (4/1342067)
http://www.metmuseum.org:80/ -> websites/www.metmuseum.org/19970511181850/index.html (5/1342067)
http://www.metmuseum.org:80/ -> websites/www.metmuseum.org/19970619043010/index.html (6/1342067)
http://www.metmuseum.org:80/ -> websites/www.metmuseum.org/19971024144359/index.html (7/1342067)
http://www.metmuseum.org:80/ -> websites/www.metmuseum.org/19971210061150/index.html (8/1342067)
http://www.metmuseum.org:80/ -> websites/www.metmuseum.org/19980118222400/index.html (9/1342067)
http://www5.metmuseum.org:80/ -> websites/www.metmuseum.org/19981201090220/index.html (10/1342067)
http://metmuseum.org:80/ -> websites/www.metmuseum.org/19990117021812/index.html (11/1342067)
http://metmuseum.org:80/ -> websites/www.metmuseum.org/19990208005322/index.html (12/1342067)
```

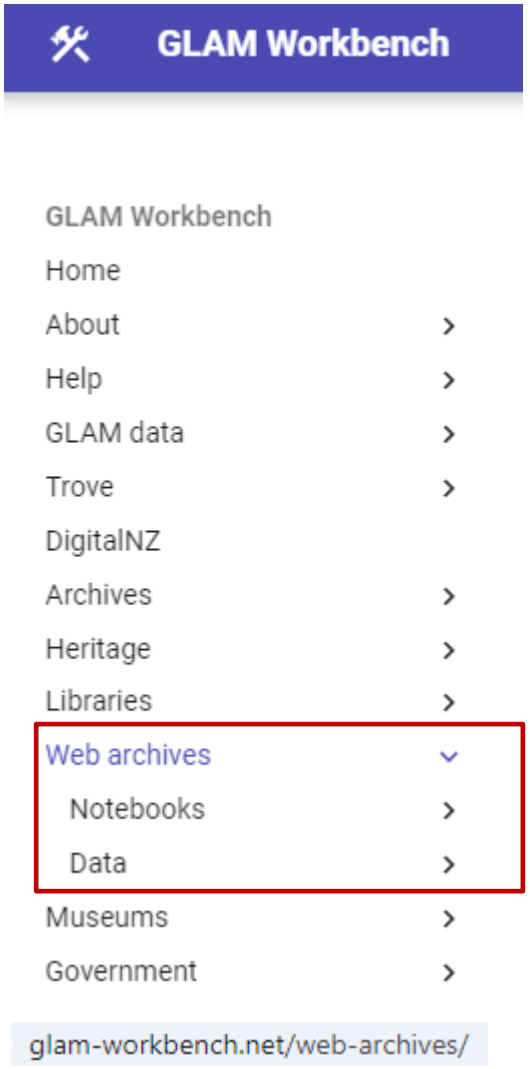
Research Infrastructures

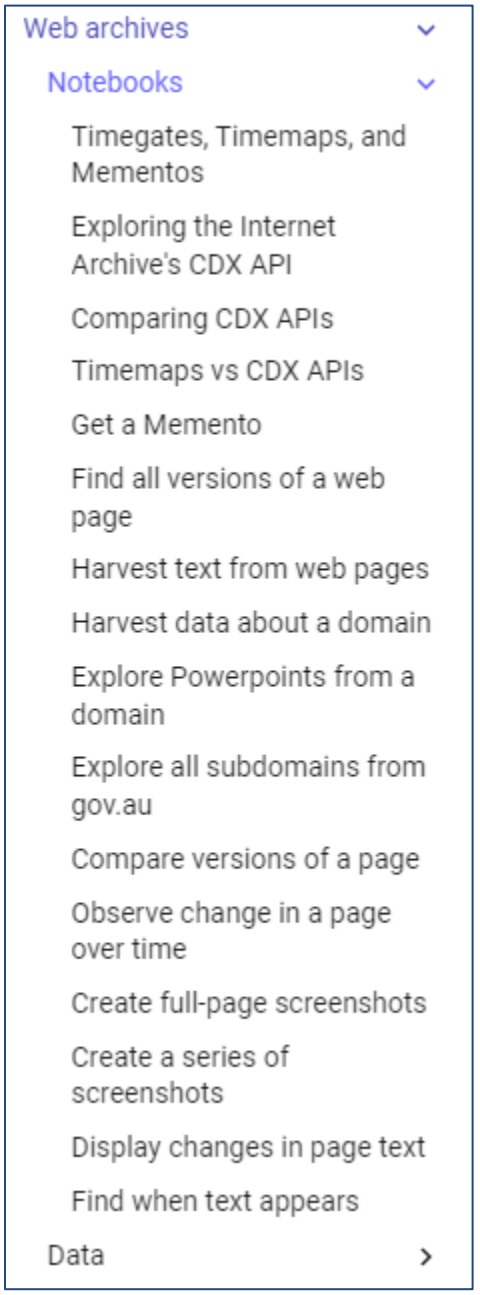


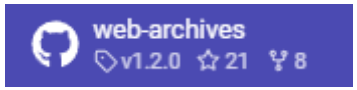
ARCH

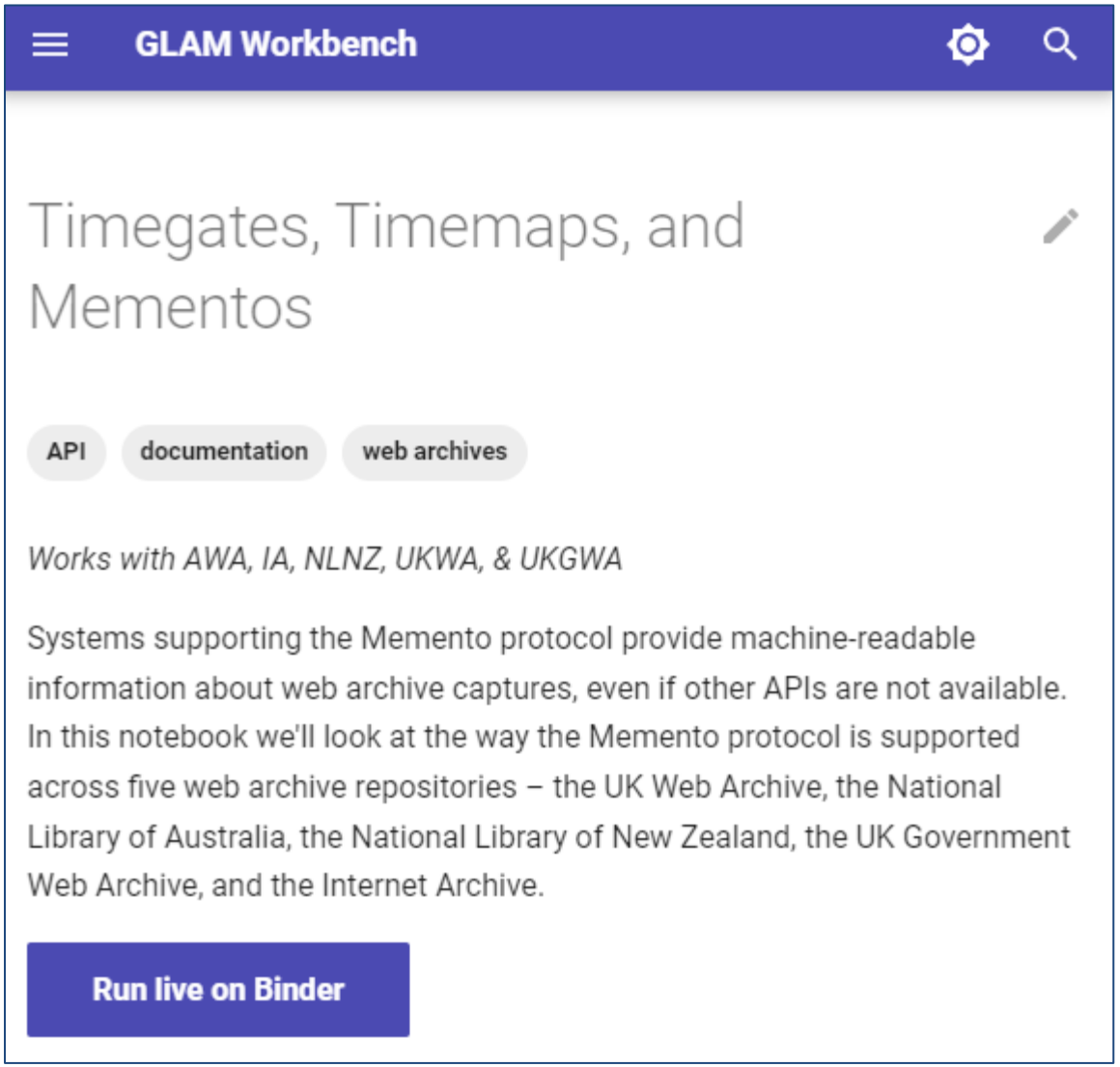
Archives Research Compute Hub

github.com/internetarchive/arch









GLAM Workbench, https://glam-workbench.net/

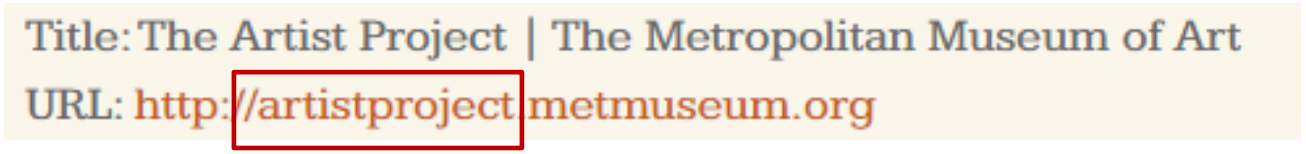
What do we obtain? **URL-analysis**

https://web.archive.org/web/20080515145825/http://www.metmuseum.org/explore/van_gogh/vangogh_alternate.html

https://web.archive.org/
web/
20080515145825/
http://www.metmuseum.org/
explore/
van_gogh/
vangogh_alternate.html

What do we obtain? **Subdomains issue**

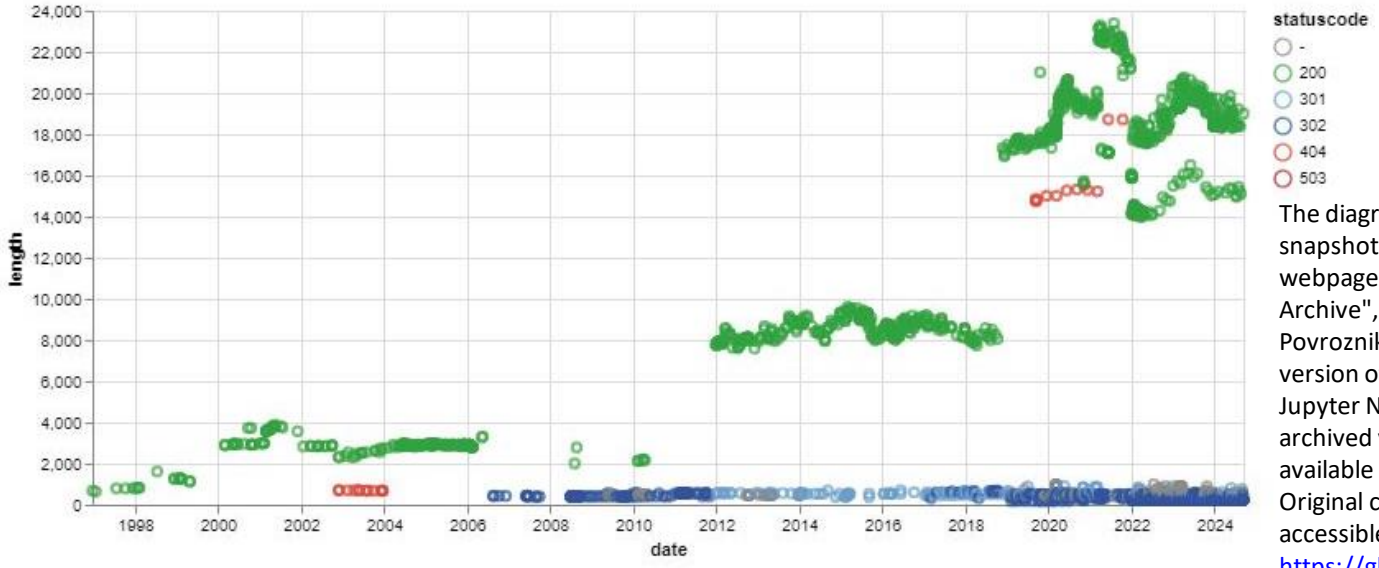
https://web.archive.org/
web/
20080515145825/
http://www.metmuseum.org/
explore/
van_gogh/
vangogh_alternate.html



The Metropolitan Museum of Art Web Archive, Archive-IT, https://www.archive-it.org/collections/13135

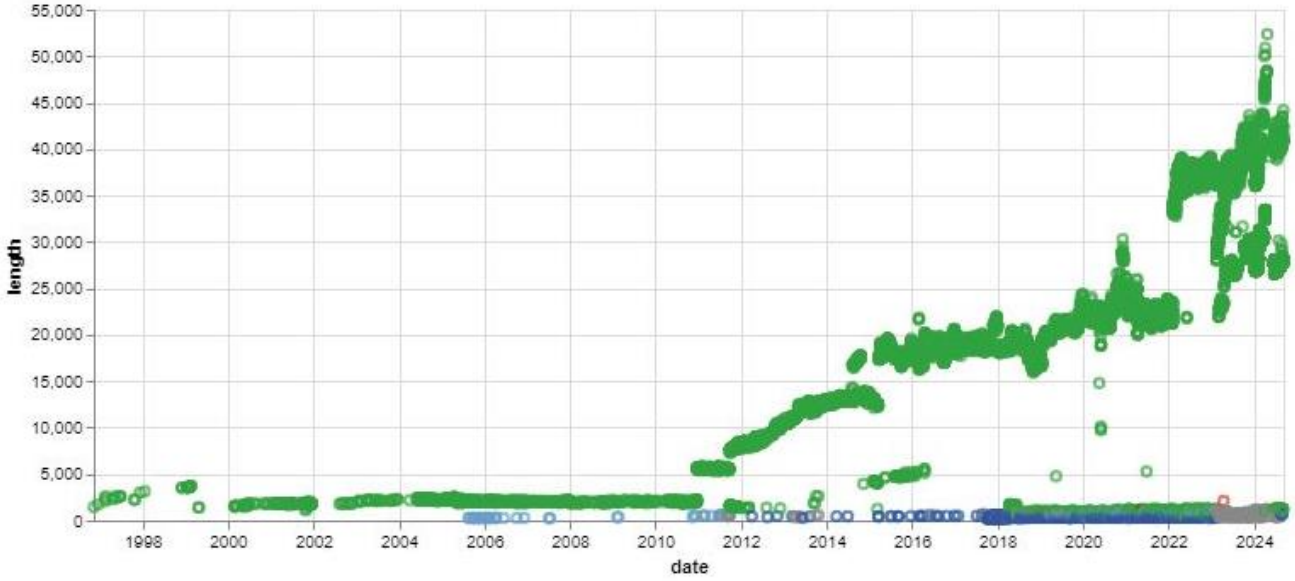
What do we obtain? **A set of the webpages**

NMA



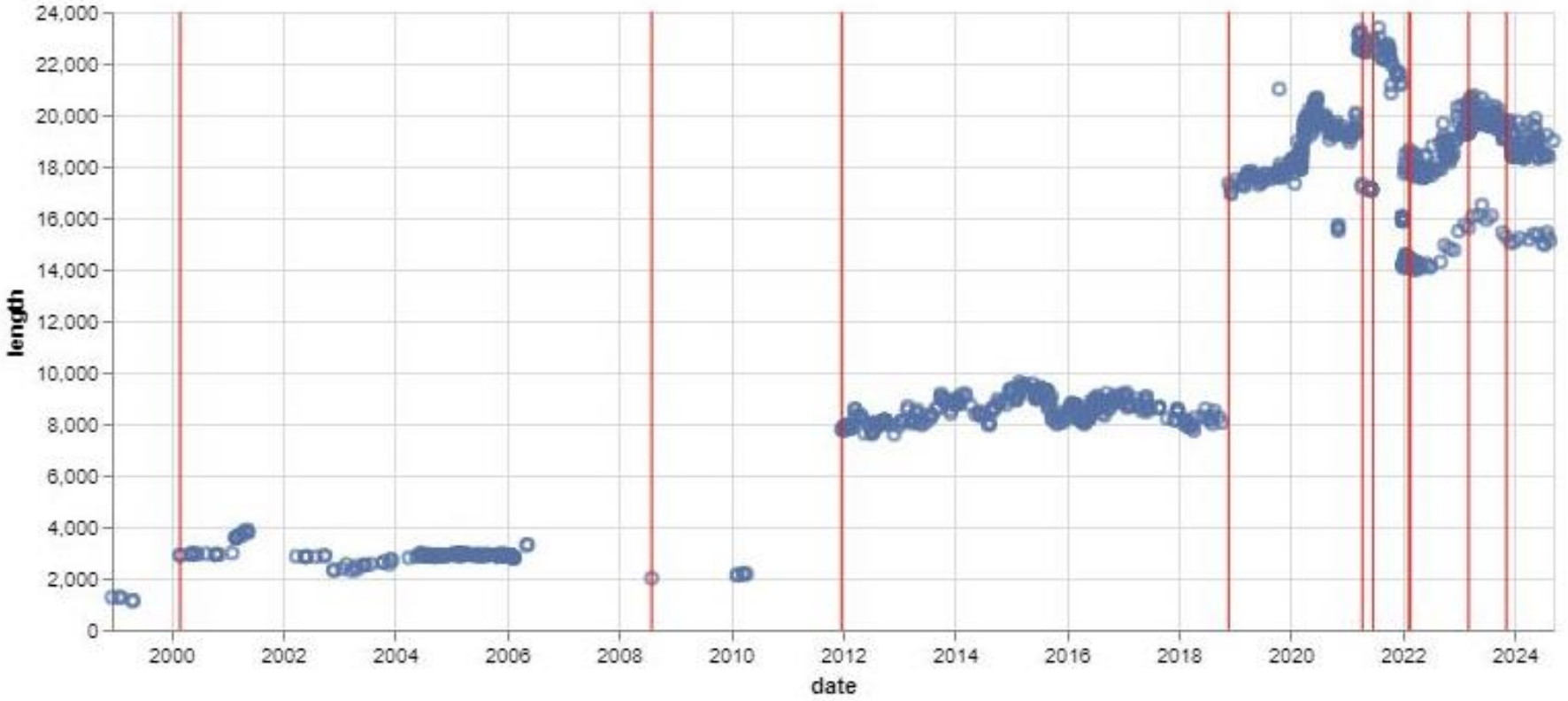
The diagram "Distribution of the snapshots of the nma.gov.au webpage preserved on the Internet Archive", created by Nadezhda Povroznik, is based on a modified version of the code from the Jupyter Notebook "Find all the archived versions of a web page," available on GLAM Workbench. Original code by Tim Sherratt, accessible at GLAM Workbench, https://glam-workbench.net/web-archives/get-all-versions

MET



The diagram "Distribution of the snapshots of the metmuseum.org webpage preserved on the Internet Archive“, created by Nadezhda Povroznik, is based on a modified version of the code from the Jupyter Notebook "Find all the archived versions of a web page," available on GLAM Workbench. Original code by Tim Sherratt, accessible at GLAM Workbench, https://glam-workbench.net/web-archives/get-all-versions

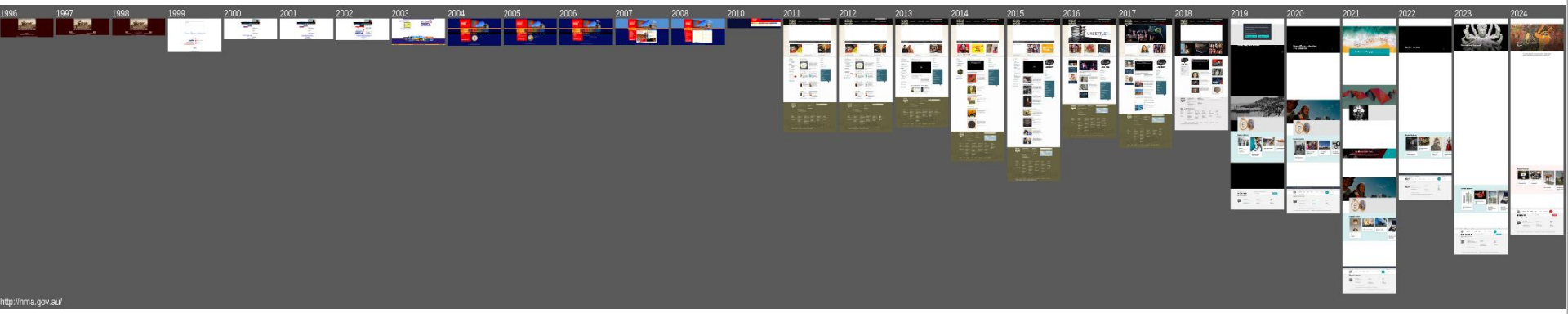
What do we obtain? **A set of the webpages**



NMA: Periodisation

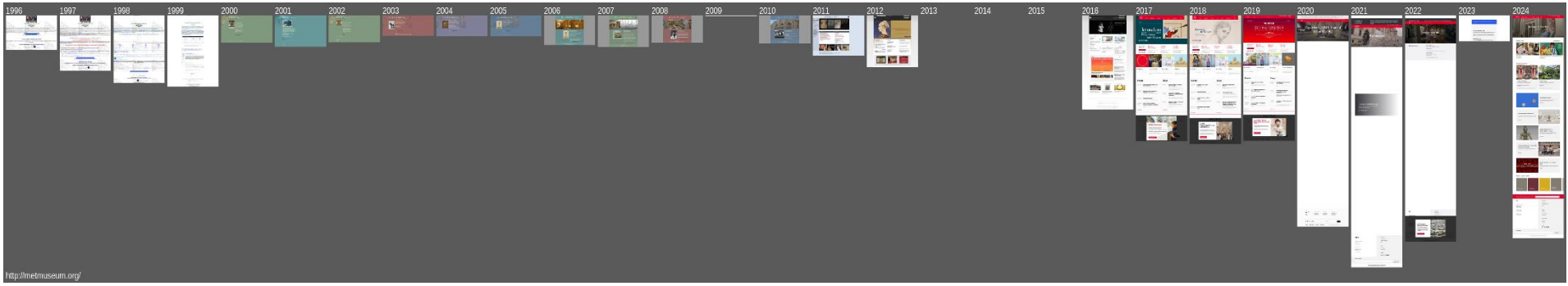
The diagram "Distribution of the snapshots of the nma.gov.au webpage preserved on the Internet Archive", created by Nadezhda Povroznik, is based on a modified version of the code from the Jupyter Notebook "Find all the archived versions of a web page," available on GLAM Workbench. Original code by Tim Sherratt, accessible at GLAM Workbench, https://glam-workbench.net/web-archives/get-all-versions

What do we obtain?
A Complex of formats



NMA

The diagram “Screenshots 1996-2024 of the nma.gov.au webpage", created by Nadezhda Povroznik, is based on a modified version of the code from the Jupyter Notebook "Using screenshots to visualise change in a page over time," available on GLAM Workbench. Original code by Tim Sherratt, accessible at GLAM Workbench, https://glam-workbench.net/web-archives/create-screenshots-over-time/



MET

The diagram “Screenshots 1996-2024 of the metmuseum.org webpage", created by Nadezhda Povroznik, is based on a modified version of the code from the Jupyter Notebook "Using screenshots to visualise change in a page over time," available on GLAM Workbench. Original code by Tim Sherratt, accessible at GLAM Workbench, https://glam-workbench.net/web-archives/create-screenshots-over-time/

What do we obtain?

Complex of formats

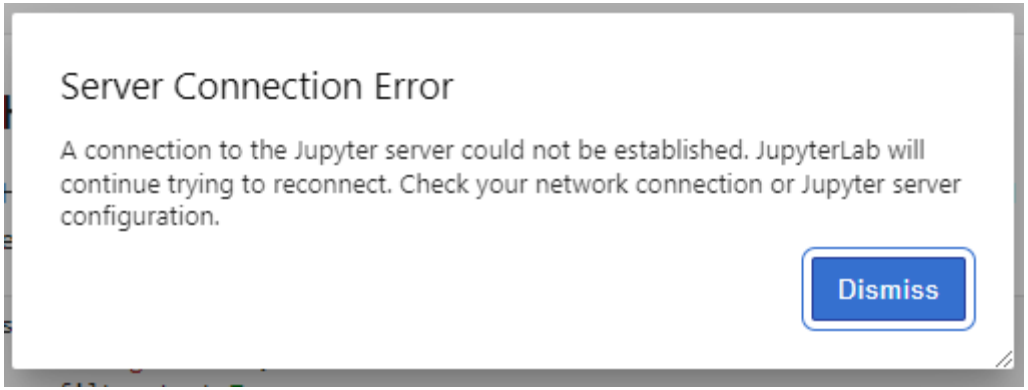
Building a Corpora

Harvesting a single url or list of urls

Get all human-visible text from all captures of a single url in the Australian Web Archive.

```
save_texts_from_url("ia", "http://nma.gov.au/", filter_text=False)
```

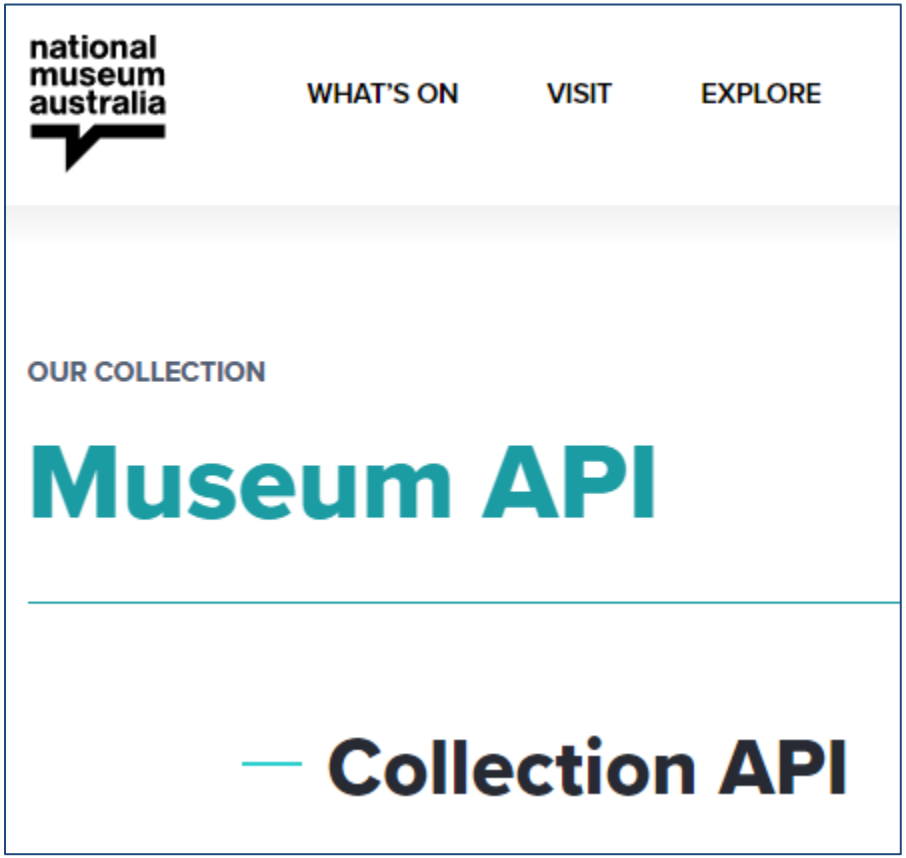
Captures: 0% 19/4299 [02:15<17:56:14, 15.09s/it]



The Metropolitan Museum of Art Collection API

The Metropolitan Museum of Art Collection API, https://metmuseum.github.io

Museums‘ APIs: Object Images



NMA, Collection API, https://www.nma.gov.au/about/our-collection/museum-api

Complexity

host www.metmuseum.org

Indexed on April 11, 2024.

See also: · Parent Domain Summary: metmuseum.org · TLD Summary: org

Saved **33,798 times** between November 11, 1996 and September 10, 2024.

MIME-types: Any MIME-types | Year Start: 1996 | Year End: 2024

All | text | image | application | font | audio | video

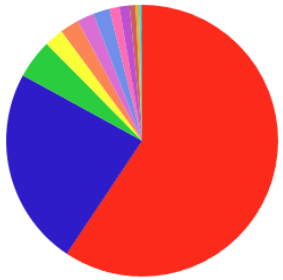
Summary on MIME-types Count

Quick search on MIME-types..

<< < 1 2 ... > >>

	Captures	Unique URLs
text/html	11,137,508	5,693,445
image/jpeg	4,431,051	305,778
text/css	869,592	50,560
image/png	428,525	3,861
image/svg+xml	424,359	492
font/woff2	374,467	1,049
image/gif	368,511	17,229
application/json	225,383	16,729
application/javascript	224,760	178,209
application/octet-stream	134,124	420

Captures



metmuseum.org, Summary, Wayback Machine, Screenshot on 08.09.2024, https://web.archive.org/web/20240000000000*/metmuseum.org

September 12–13

Digital History 2024 Switzerland

University of Basel

DigiHistCH24 conference theme:
‹Historical Research, Digital Literacy and Algorithmic Criticism›

DigiHistCH24 conference theme:
‹Historical Research, Digital Literacy and Algorithmic Criticism›

- Valuable data
- New insights
- Source criticism
- Web preservation

- Complexity in obtaining data
- Lowering the entry barrier

- Completeness
- Reliability
- Biases
- Limits
- Challenges

Thank you!