TrendMachine:
A Temporal Webpage Resilience Portal

Sawood Alam	*Internet Archive*
Mark Graham	*Internet Archive*
Kritika Garg	*Old Dominion University*
Michele C. Weigle	*Old Dominion University*
Michael L. Nelson	*Old Dominion University*
Dietrich Ayala	*Protocol Labs*

@WaybackMachine @WebSciDL @ProtocolLabs

ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), June 27, 2023, Santa Fe, NM

Supported in part by Protocol Labs and Filecoin Foundation

Research Question

How healthy has a web page been throughout its lifetime?

Temporal and Spatial Landscape of Archival Analysis

Long Duration

Single Webpage

Webpage Collection

- TMVis
- Wayback Machine Changes
- **TrendMachine**

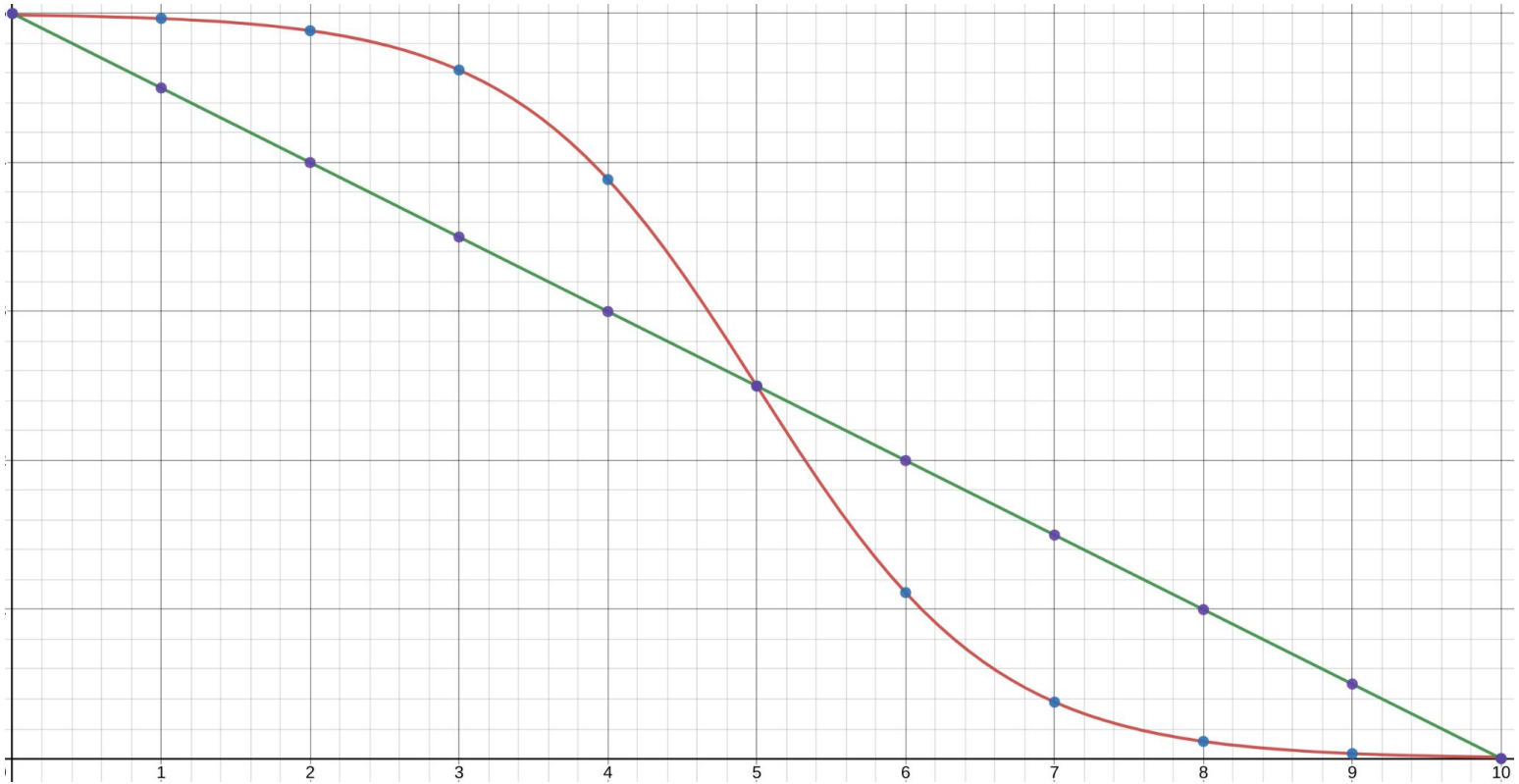
- MementoMap
- CDX Summary
- Archives Unleashed Toolkit

- Memento Damage
- Archival ACID Test
- Reconstructive

- Warrick
- Wayback Machine Downloader
- Video Archiving Insights

Short Duration

Modeling Web Page Health: Linear vs. S-Curve



Sigmoid Function for Web Page Resilience

$$Resilience_t = \frac{Spread}{1 + e^{Shift - \frac{t}{Slope}}}$$

Spread: How far up or down the value can go from its starting position?

Shift: How soon any significant change in the value can begin?

Slope: How quickly the value reaches close to the maximum change?

TrendMachine: Composite Sigmoid Parameters of Resilience

TrendMachine: Overview

URL

https://wikipedia.org/

Fill Missing Days

10 − +

Filling Policy

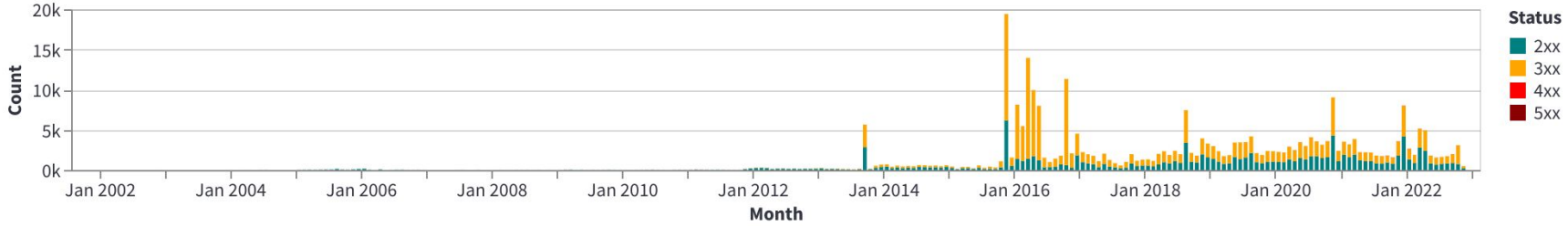
Closest

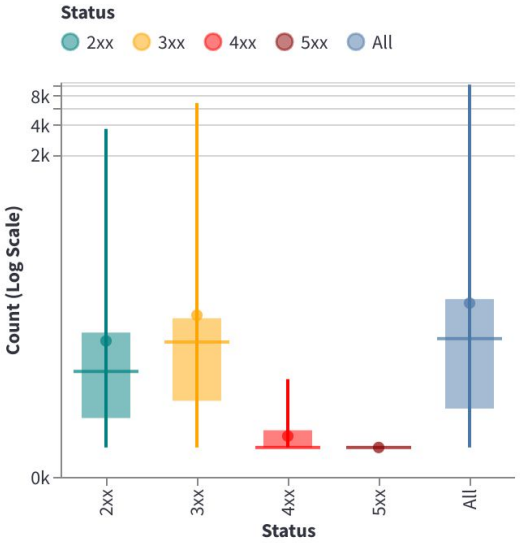
Sigmoid Parameters

Captures	Span	Gaps	Resilience	Fixity	Chaos
299,264	21y7m	2,807	0.99665	0.69436	0.56415
↑ 39.241% (OK)	↑ 3m10d (Last)	↑ 1,155 (Filled)	↓ -0.00016	↑ 24.611% (Changed)	↑ 0.63000

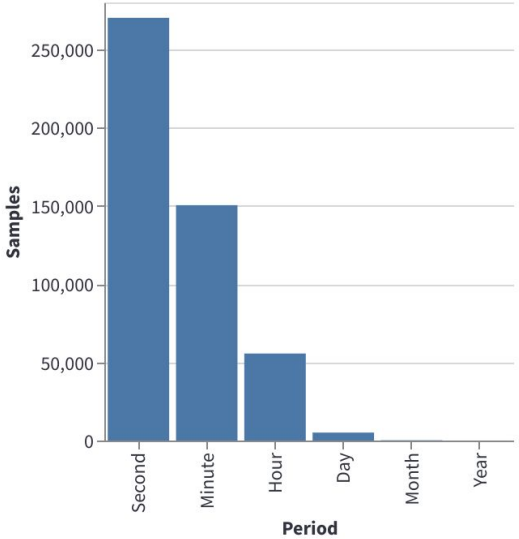
Code: https://github.com/internetarchive/trendmachine
Demo: https://trendmachine.sawood-dev.us.archive.org/

TrendMachine: Temporal Distribution of Archiving Activities





The page is archived as few as one or zero times and as many as tens of thousands of times in a single day.



Specimen Selection Algorithm

```
PRIORITY = ["2xx", "4xx", "5xx", "3xx"]

FOREACH st OF PRIORITY
  IF st IN statuses(day)
    specimen = statuses(day).match(st)[0]
    BREAK
```

A **3xx** specimen usually suggests that the URL is redirecting to somewhere other than a variation of the same URL.

DAY1	DAY2	DAY3	DAY4
4xx	3xx	5xx	3xx
3xx	3xx	3xx	5xx
2xx	3xx	5xx	3xx
5xx		4xx	5xx
2xx		4xx	

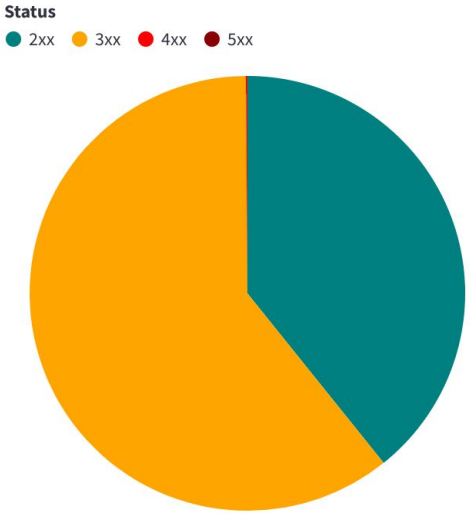
Filling Missing Observations

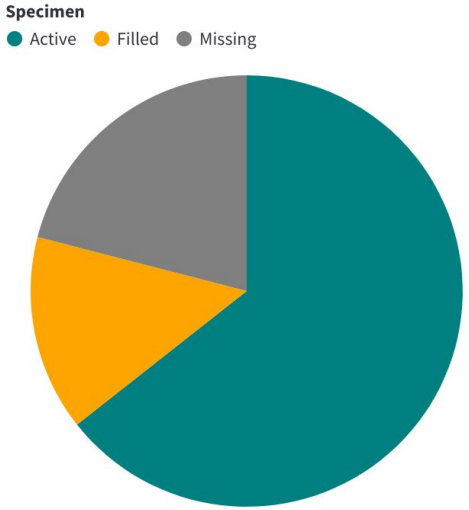
Policy	DAY1	DAY2	DAY3	DAY4	DAY5	DAY6
Identical	2xx	**2xx**	2xx	4xx		2xx
Closest	2xx	2xx	**2xx**	**4xx**	4xx	2xx
Forward	2xx	2xx	**2xx**	**2xx**	4xx	2xx
Backward	2xx	2xx	**4xx**	**4xx**	4xx	2xx
ANY	2xx					2xx

Do not fill the gap if the status codes before and after are not identical.

Do not fill the gap if it is larger than a configured threshold.

TrendMachine: TimeMap Status Codes vs. Daily Specimens





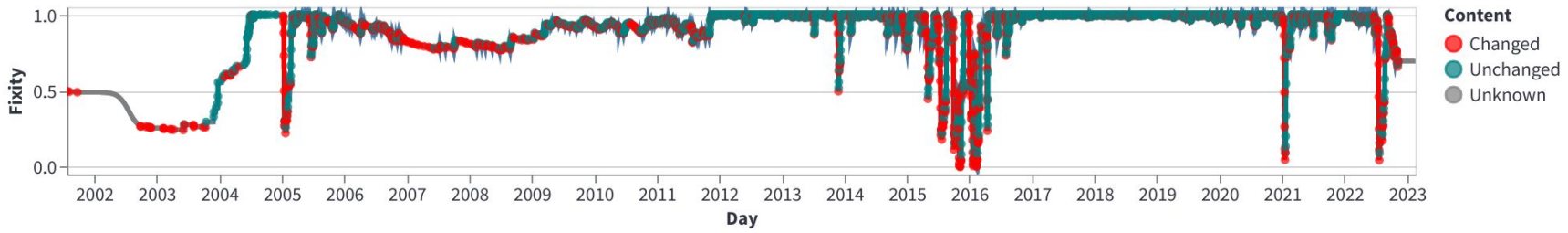
Most of the self-redirect 3xx observations (HTTP/HTTPS or WWW/Apex domain) are eliminated in daily specimens.

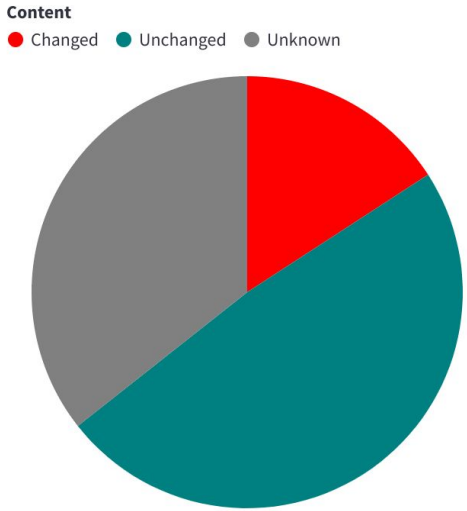
About one third of the days since the first observation have no captures, of which some are filled using a filling policy.

TrendMachine: Resilience

- Resilience score is calculated using Sigmoid function on status codes of daily specimens
- Initial value of 0.5 and normalized between 0 and 1
- After the first few observations, Wayback Machine did not archive it for several months in 2002
- Towards the end of 2002, Resilience score went up slowly due to infrequent archiving
- In 2003 “wikipedia.org” started to redirect to “en.wikipedia.org”
- After 2005, Resilience of the Wikipedia home page has mostly been stable and high

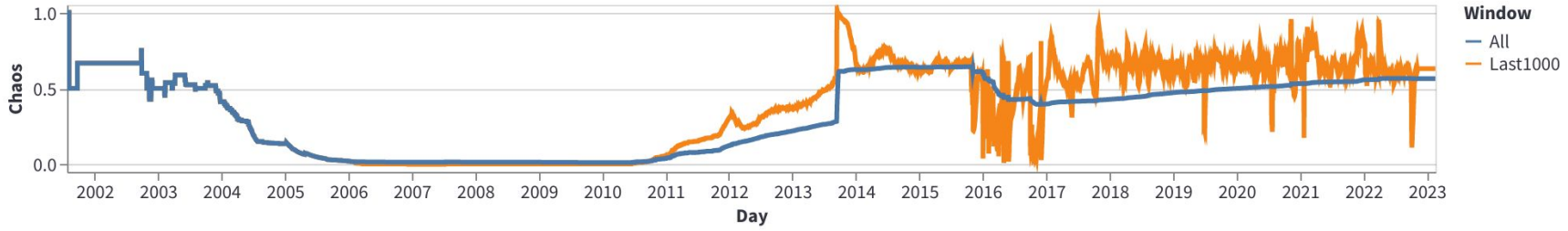
TrendMachine: Fixity





- Fixity score (normalized) is calculated using Sigmoid function on content digests of daily specimens
- Content digest reported in CDX can be sensitive to Content-Encoding, resulting in false alarms, even when the underlying content remains unchanged

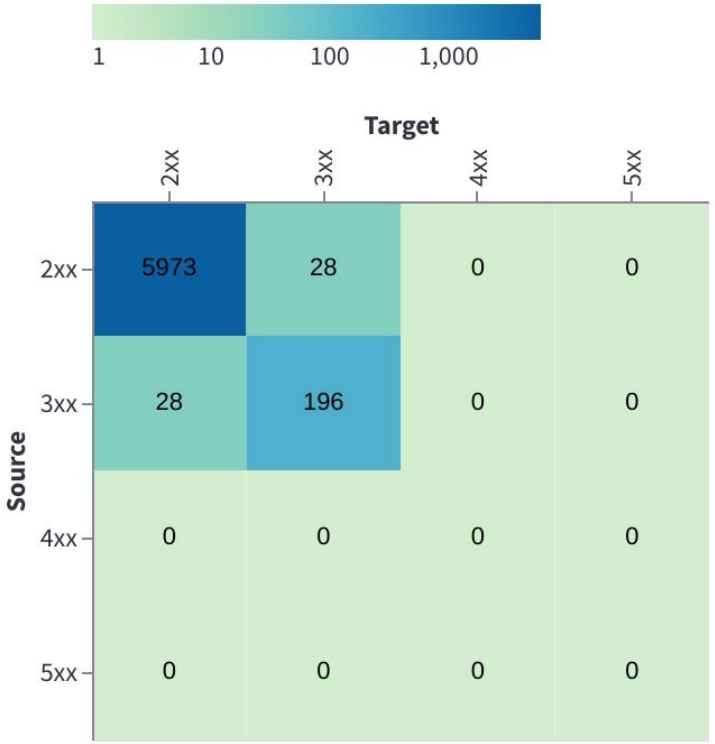
TrendMachine: Chaos



$$\text{Chaos} = \frac{|\ 2xx,\ \cancel{2xx},\ \cancel{2xx},\ 3xx,\ \cancel{3xx},\ 2xx\ |}{|\ 2xx,\ 2xx,\ 2xx,\ 3xx,\ 3xx,\ 2xx\ |} = \frac{3}{6} = 0.5$$

- Chaos score (normalized) is calculated using a Run-Length Encoding inspired technique on all status codes of the CDX data in which consecutive duplicates are removed in the numerator
- An alternate sliding-window calculation is performed on the last N observations as the score becomes insensitive to recent changes on large TimeMaps
- A high Chaos along with a high Resilience is often an indication of canonical redirects (e.g., adoption of HTTPS and/or consolidation of WWW and Apex domain)

TrendMachine: Status Code Transitions



- Large numbers along the major diagonal indicate status code stability for extended periods of time
- Large numbers in non-diagonal cells suggest frequent changes in Resilience curve
- Web pages with high Resilience score for extended periods usually exhibit large numbers in the top-left cell (2xx -> 2xx)
- A large number in the 3xx -> 3xx cell usually indicates extended periods of redirection to other URLs (e.g., URL restructuring, login wall, domain change, and parked domain)

TrendMachine: Compare First and Last Mementos

First Capture: 2001-07-27

HomePage [Home]

HomePage | RecentChanges | Preferences | Random Page
You can edit this page right now! It's a free, community project

Welcome to Wikipedia, a collaborative project to produce a complete encyclopedia from scratch. We started in January 2001 and already have **over 6,000 articles.** We want to make over 100,000, so let's get to work--*anyone* can edit any page--copyedit, write a little, write a lot. See the Wikipedia FAQ for information on how to edit pages and other questions. If you're visiting Wikipedia for the first time, welcome! *The content of Wikipedia is covered by the GNU Free Documentation License.*

Philosophy, Mathematics, and Natural Science
Astronomy and Astrophysics -- Biology -- Chemistry -- Earth Sciences -- Mathematics -- Philosophy -- Physics -- Science -- Statistics

Social Sciences
Anomalous Phenomena -- Anthropology -- Archaeology -- Countries of the world -- Economics -- Geography -- History -- History of Science and Technology -- Language -- Linguistics -- Politics -- Psychology -- Sociology

Applied Arts and Sciences
Agriculture -- Architecture -- Business and Industry -- Communication -- Computing -- Education -- Engineering -- Family and Consumer Science -- Health Sciences -- Law -- Library and Information Science -- Public Affairs -- Technology -- Transport

Culture
Classics -- Critical Theory -- Dance -- Entertainment -- Film -- Games -- Hobbies -- Literature -- Music -- Opera -- Painting -- Performing Arts -- Recreation -- Religion -- Sculpture -- Sports -- Theater and Drama -- Tourism -- Visual Arts and Design

Other Category Schemes
About Wikipedia category schemes -- Library of Congress catalog scheme -- Dewey Decimal System -- Wikipedia arranged by topic -- Year in Review -- Historical anniversaries -- Reference tables -- Biographical Listing

International Wikipedias
About the International Wikipedias -- [Catalan (Català)] -- [Chinese (Hanyu)] -- [German (Deutsch)] -- [Esperanto] -- [French (Français)] -- [Hebrew (Ivrit)] -- [Italian (Italiano)] -- [Japanese (Nihongo)] -- [Portuguese (Português)] -- [Russian (Russkiy)] -- [Spanish (Castellano)] -- [Swedish (Svensk)]

Wikipedia
FAQ -- Policy -- Article news -- Announcements -- Requested articles -- Help desk -- New topics -- Brilliant prose -- Mailing list -- Feature requests -- Bug reports -- Commentary -- Public Domain Resources -- Wikipedia and Nupedia -- Sandbox -- License -- Wikipedians

Today is Friday, [July 27]?, 2001, servertime (Pacific Standard Time).

/Talk

Last Capture: 2023-01-18

WIKIPEDIA
The Free Encyclopedia

English
6 585 000+ articles

日本語
1 354 000+ 記事

Русский
1 875 000+ статей

Français
2 477 000+ articles

Deutsch
2 751 000+ Artikel

Español
1 823 000+ artículos

Italiano
1 785 000+ voci

中文
1 323 000+ 条目 / 條目

فارسی
941 000+ مقاله

Português
1 096 000+ artigos

EN

Read Wikipedia in your language

Wikipedia is hosted by the Wikimedia Foundation, a non-profit organization that also hosts a range of other projects.
You can support our work with a donation.

Commons
Freely usable photos & more

Wikivoyage
Free travel guide

Wiktionary
Free dictionary

Wikibooks
Free textbooks

Wikinews
Free news source

Wikidata
Free knowledge base

Download Wikipedia for Android or iOS

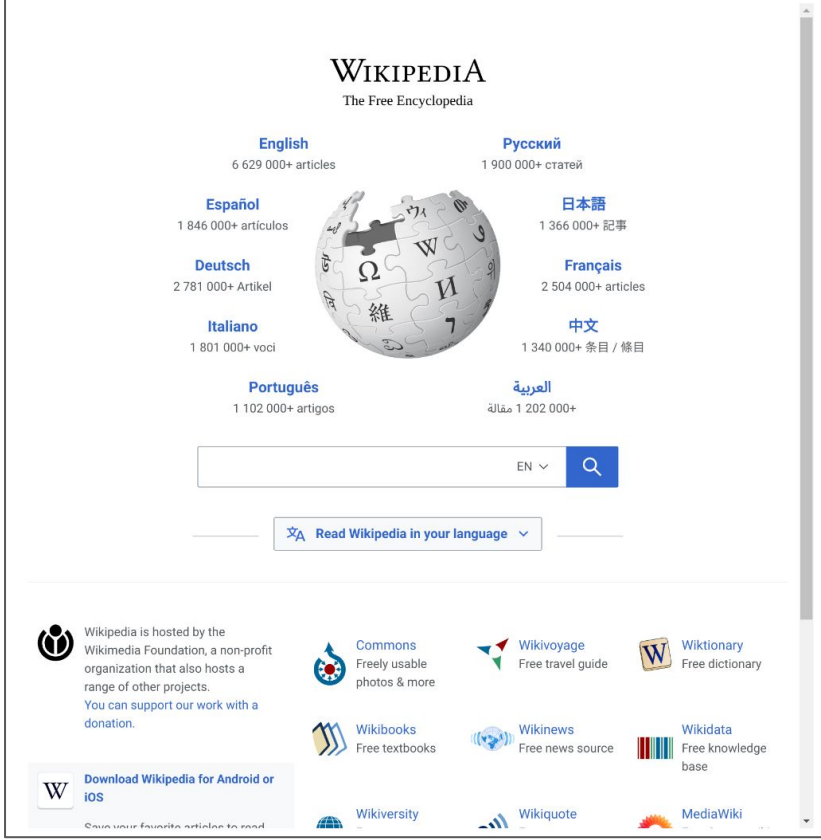
Wikiversity

Wikiquote

MediaWiki

TrendMachine: Live Web Page With Headers

Live Page



HTTP Headers

```
HTTP/1.1 301 Moved Permanently
date: Thu, 23 Mar 2023 11:16:59 GMT
server: mw1411.eqiad.wmnet
location: https://www.wikipedia.org/
content-length: 234
content-type: text/html; charset=iso-8859-1
vary: X-Forwarded-Proto
age: 42926
x-cache: cp1079 miss, cp1081 hit/131906
x-cache-status: hit-front
server-timing: cache;desc="hit-front", host;desc="cp1081"
strict-transport-security: max-age=106384710; includeSubDomains; preload
report-to: { "group": "wm_nel", "max_age": 604800, "endpoints": [{ "url": "https://intake-logging.w
nel: { "report_to": "wm_nel", "max_age": 604800, "failure_fraction": 0.05, "success_fraction": 0.0}
set-cookie: WMF-Last-Access=23-Mar-2023;Path=/;HttpOnly;secure;Expires=Mon, 24 Apr 2023 12:00:00 GMT
x-client-ip: 207.241.234.233
```

```
HTTP/1.1 200 OK
date: Thu, 23 Mar 2023 09:45:26 GMT
cache-control: s-maxage=86400, must-revalidate, max-age=3600
server: ATS/9.1.4
etag: W/"126fa-5f6ca1234f42a"
last-modified: Mon, 13 Mar 2023 15:50:32 GMT
content-type: text/html
content-encoding: gzip
vary: Accept-Encoding
age: 48419
x-cache: cp1079 hit, cp1081 hit/803714
x-cache-status: hit-front
server-timing: cache;desc="hit-front", host;desc="cp1081"
strict-transport-security: max-age=106384710; includeSubDomains; preload
report-to: { "group": "wm_nel", "max_age": 604800, "endpoints": [{ "url": "https://intake-logging.w
nel: { "report_to": "wm_nel", "max_age": 604800, "failure_fraction": 0.05, "success_fraction": 0.0}
x-client-ip: 207.241.234.233
accept-ranges: bytes
content-length: 18106
```

Potential Use Cases

- Detect points of interest in a large TimeMap
- Sample captures/mementos from TimeMaps for visual summarization
- Detect archival sinks (like login pages, paywalls, and misconfigured redirects)
- Detect poor-quality pages like Soft-404 and parked domains
- Detect potential link-rot (and fix them when possible, like in a wiki page)
- Optimize crawl jobs by minimizing wasteful downloads and maximizing coverage
- Archival quality assurance
- Cluster pages of a large archival collection in different categories

Future Work

- Report heuristics-based archival summary by combining various scores
- Report/embed captures/mementos that can be points of interest
- Calculate Fixity using less-sensitive digests (e.g., SimHash)
- Calculate Chaos after applying convolutions to smooth out alternate changes
- Allow alternate web page health models (not just Sigmoid functions)
- Deploy in production by integrating with Wayback Machine

Summary

A mathematical model to quantify temporal health of a web page

An interactive portal with configuration options for experiments

Resilience, Fixity, Chaos, Distributions, Transitions, etc. reports

An evolving open-source codebase and demo deployment

Code: https://github.com/internetarchive/trendmachine
Demo: https://trendmachine.sawood-dev.us.archive.org/