Ancient GeoCities: A Dataset of Temporally Annotated Web Pages

Ira Kokoshko
Berlin School of Library and Information Science
Humboldt-Universität zu Berlin
Berlin, Germany
ira.kokoshko@hu-berlin.de

Robert Jäschke
Berlin School of Library and Information Science
Humboldt-Universität zu Berlin
Berlin, Germany
robert.jaeschke@hu-berlin.de

Abstract

Estimating the creation time of web pages is a lasting problem with direct implications for web archiving, temporal information retrieval, and the study of web evolution. While existing approaches often rely on link-based propagation or external timestamping services, they are not suited for older web content, where such signals are sparse or absent. In this work, we present a large-scale, systematically annotated dataset of historical web pages derived from the GeoCities corpus. The dataset aggregates and normalizes multiple temporal signals extracted from HTTP headers and HTML content, including first posted dates, last updated statements, and copyright years, and analyzes their coverage and consistency. As a first step toward content-based temporal inference, we additionally explore the alignment of these signals with zero-shot year estimates produced by a large language model. By releasing this dataset, we simplify temporal analysis of the GeoCities corpus and provide a foundational resource for training and evaluating models that infer page creation times from content alone.

CCS Concepts

• **Information systems** → *Web crawling*; *Data cleaning*.

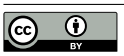
Keywords

GeoCities, Dataset, Temporal Annotation, Creation Date

ACM Reference Format:
Ira Kokoshko and Robert Jäschke. 2026. Ancient GeoCities: A Dataset of Temporally Annotated Web Pages. In *18th ACM Web Science Conference (WebSci '26), May 26–29, 2026, Braunschweig, Germany*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3795766.3799783

1 Introduction

Determining the temporal origin of web content is an essential but challenging task, for instance, in studies of the early Web. Historians, digital humanities scholars, and sociologists often use web archives to reconstruct the cultural, social, and technological practices of the past Web [10]. However, large web archives rarely provide information about the creation time of a web page. Archives such as the Internet Archive usually only preserve the time of crawling that can differ significantly from the time of creation. For the GeoCities end-of-life crawl [4] this difference can be more than a decade, which erases the chronological order and complicates analysis.



WebSci '26, Braunschweig, Germany

ACM ISBN 979-8-4007-2504-3/26/05
https://doi.org/10.1145/3795766.3799783

Estimating the creation time of archived web pages thus is an important research task. Existing approaches mostly rely on external signals such as URL shortening services, social media, or search engine timestamps [8]. Link-based methods are applied by Nunes et al. [6] based on the assumption that connected web resources demonstrate similar update patterns. A sophisticated link-based approach motivated by a probabilistic model of the Web is proposed by Ostroumova Prokhorenkova et al. [7]. While these methods are effective for 'modern' web content, they are ineffective for older web content, like GeoCities web pages, that is pre-dating social networks. Other work focuses on constructing and exploring citation networks using standardized genres such as news articles [9], which is not directly transferable to the diverse, DIY content typical for 1990s and early 2000s personal web pages.

To address this problem, we create a temporarily enriched dataset obtained by scanning the Internet Archive's GeoCities end-of-life crawl.[1] GeoCities operated from 1994 to 2009 and became a platform serving approximately seven million users distributed across tens of millions of HTML pages in various neighborhoods and individually [5]. Its shutdown by Yahoo! in 2009 was a significant loss, but rescue crawls conducted in the last months before shutdown preserved a unique snapshot of the early personal web [3].

Our goal is to recover temporal signals embedded in the archived pages themselves and aggregate them into a structured dataset. Using a combination of HTML parsing, text extraction, and regular expression-based detection methods, we collect a range of potential indicators of the date of a page's initial creation or last update. These include explicit expressions such as "last updated" or "first posted", copyright statements, and dates recorded in HTTP headers. The result is a dataset of web pages with temporal annotations that can serve as a basis for training and evaluating machine learning models aimed at estimating the creation dates of web pages directly from their content.

In addition to extracting metadata, the dataset enables the systematic analysis of the consistency and reliability of temporal signals at different technical levels: HTTP headers vs. HTML content. The contributions of this paper are as follows:

- a simple yet robust approach to estimate creation times of web pages in the GeoCities corpus,
- an understanding of the temporal distribution of the GeoCities corpus, and
- a dataset of temporal annotations for a subset of the GeoCities corpus.[2]

We publish the extraction pipeline, summary statistics, and dataset structure to ensure reproducibility and future extensions.

[1] https://archive.org/web/geocities.php
[2] Available at https://doi.org/10.5281/zenodo.17970511

This paper is organized as follows: In Section 2 we present related work, in Section 3 we describe our approach to create and analyze the dataset, and in Section 4 we provide exploratory results. We conclude the paper in Section 5 with a discussion.

2 Related Work

Existing approaches to estimate the creation time of web pages often rely on signals that are typically not available for historical crawls, for instance, links from social social media or other external sources. One of the first such approaches was undertaken by Nunes et al. [6]. Their work supplements conventional methods relying on HTTP headers by additionally examining a document's vicinity. Firstly, through parsing the HTML structure of a document, its neighbors are derived. Subsequently, they are analyzed to compute an average update date, which serves as a dating mechanism for web documents lacking HTTP headers. The examination of neighboring resources increased the retrieval rate of valid HTTP headers from 53% to 86% in a dataset of 10 000 URLs from the Yahoo! directory.

SalahEldeen and Nelson [8] estimate creation dates for URLs by leveraging multiple sources, including the URL-shortening service Bitly (the first time someone shortened the URL), Topsy (the first time someone tweeted the URL), Memento aggregator (the first web archive appearance), Google's last crawl time, and the LastModified HTTP header, while also examining backlinks and applying the same techniques to linked resources. The method was evaluated using a manually verified gold standard dataset consisting of 1 200 URLs, successfully estimating creation dates for 75.90% of resources.

Ostroumova Prokhorenkova et al. [7] propose a link-based approach using a probabilistic model of web graph structure evolution, which utilizes the publication dates of web pages as its parameters. This approach generalizes and extends one-step link-based methods suggested in [6] and [8] in two directions. Firstly, Ostroumova Prokhorenkova et al. [7] suggest and test different propagation functions. Secondly, they compare the one-step and the multi-step propagation. For the evaluation two distinct datasets are used: a sample of 4 million web pages crawled by Yandex, 93 000 of them provide credible time, and the publicly accessible MemeTracker dataset that encompasses about 12 million web pages with known dates. All improvements over the baseline approaches achieved on both datasets demonstrate statistical significance.

Link-based methods for estimating the publication dates of web pages emerged in response to the well-known limitations of content-based methods. The researchers noted that HTML documents frequently contain multiple candidate dates with no obvious way to determine which one reflects the page's actual creation time; that these dates appear in a wide variety of formats, often requiring extensive normalization; and that many pages include no explicit date information at all in their visible content. These challenges made link-based approaches a practical alternative. However, the recent development of large language models offers a compelling approach to estimate creation time based on the textual content. Such models are capable of interpreting ambiguous and inconsistent textual cues in ways that earlier approaches could not. Yet training and evaluating these models requires a collection of web pages annotated with reliable temporal signals, which to the best of our knowledge does not exist. Thus, we describe the construction of such a dataset designed to fill this gap.

3 Method

3.1 Dataset

The Internet Archive's GeoCities end-of-life crawl comprises 8 897 compressed WARC files with an overall size of 5.3 TB. To process the WARC data at the terabyte-scale we use the fastwarc library,[3] which provides a high-performance interface for accessing WARC records [1]. The WARC files contain WARC records – requests, responses, and metadata. The WARC response records contain the HTTP response headers and the HTTP response body. For each WARC record we extract the HTTP Last-Modified header field. For each WARC file a corresponding CDX file serves as a lightweight index of URLs, timestamps, MIME types, and WARC payload digests. The WARC payload digest is a cryptographic hash (SHA-1) of the payload content of a WARC record, which means that pages with the same payload digest have (very likely) identical content, even if their URLs or crawl times differ. It allows us to detect duplicate or template pages efficiently without reprocessing full content and avoid repeated computation. Thus, the CDX files enable selective search without the need to scan the actual WARC files.

3.2 Filtering and Extraction

To create a usable subset of WARC records that contain HTML pages whose creation time can be estimated, we employ the following set of filtering criteria for each WARC response record (i. e., its WARC type is "response") using the CDX index:

HTTP 200: The HTTP status code is 200 ("OK").
MIME text/html: The MIME type (as reported by the server) is text/html.
Host GeoCities: The host name of the URL ends with geocities.com or geocities.yahoo.com.
Payload digest unique: The WARC payload digest is unique within the whole dataset.
Text extraction successful: Plain text could be extracted from the HTML.

The first two steps (HTTP 200 and MIME text/html) ensure that only successfully crawled HTML pages are included. The third step (Host GeoCities) is necessary, since the dataset also contains content from other hosts. This could be embedded content, for example, images or HTML pages included in frames using the HTML `<iframe>` tag. The fourth step (Payload digest unique) is an approach to remove error pages that appeared shortly before GeoCities was shut down. Initial exploration of the dataset revealed that many (especially dynamic) pages had already been inactive by the time of the 2009 end-of-life crawl and were replaced by different standardized Yahoo! shutdown error pages (see Figure 1 for an example). These error pages contain identical HTML bodies and thus have identical WARC payload digests, providing a reliable mechanism to detect them. The final filtering (Text extraction successful) step employs resiliparse,[4] a robust toolkit that allows us to convert each HTML page into

[3] https://pypi.org/project/FastWARC/
[4] https://pypi.org/project/Resiliparse/

```
1 (?i)(?P<label>first\s+(posted|created)|(site|page)\s+was\s+created)[\s:;,.]*(?P<date>[A-Za-z0-9 ,./:-]{4,60})
2 (?i)©[^0-9]{0,20}(?P<start>\d{4})(?:\s*[--]\s*(?P<end>\d{4}))?
3 (?i)(?P<label>last\s+(updated|modified|revised))\s*[:;]?\s*(?P<date>[A-Za-z0-9 ,./:-]{4,60})
```

Listing 1: Regular expressions used to identify temporal signals.

Figure 1: The most common type of error page that appears in 294 006 WARC records.

normalized plain text. In Section 4.1 we show how many records pass each filtering step.

3.3 Rule-based Date Estimation

Once the plain text is obtained, we search for temporal metadata indicating the creation date of a web page using regular expressions. These expressions target three categories of textual cues noticed during the initial exploration of the dataset:

- explicit statements about the initial publication of a page (e. g., "First posted") – line 1 in Listing 1,
- copyright years, which frequently appear either as single year or as year ranges – line 2 in Listing 1, and
- statements about later modifications (e. g., "Last updated") – line 3 in Listing 1.

We parse each extracted date string to a Date object using Python's dateutil module. For partial date expressions, where only the year is found, we interpret four-digit year values as referring to the mid of the year (i. e., first of July). That ensures a neutral estimate that avoids biasing all dates toward the start or end of the year. Because pages may contain multiple temporal indicators, we define an order for selecting a single best guess. Priority is given first to explicit creation dates, then to the earliest copyright year mentioned, followed by last updated dates. This procedure provides an estimate of the creation date obtained from the text of a web page.

3.4 LLM-based Date Estimation

We conducted an experiment to examine the extent to which an LLM's assessment of a web page's creation date aligns with the dates obtained from the textual content. We randomly sampled 100 000 records from those records for which we could estimate a creation date. From the text of the pages we removed the segments that were matched by the corresponding regular expressions, e. g., "Last updated on November 11, 1996", and queried an LLM (Qwen3-30B-A3B-Instruct-2507-FP8) with the following prompt:

> Read the following web-page text and estimate the year when the page was originally created. Return only one 4-digit year.

One challenge for text-based approaches is to distinguish between different times relevant for or represented by a document. For example, apart from the time of a document's creation (which we are interested in), there is also the document's *focus time* [2]. With our regular expressions and the prompt for the LLM we explicitly aim at the creation time. Still, there could be ambiguities, for example, between the creation time of the textual *content* and the actual *web page* which contains that content. Distinguishing these two cases is quite challenging. Technically, the web pages on GeoCities could have only been created (or uploaded) between 1994 and 2009. Thus, we remove records where our estimation approaches yield a year outside of that range. These might be estimation errors or (in case of years before 1994) hint at content that was created before the launch of GeoCities.

3.5 Dataset Generation

To store the results, we keep the crawl structure consisting of 149 segments, in which each folder is named after the corresponding segment identifier,[5] for example, `GEOCITIES-20090829030404-00020-ia400131-c`. For each input WARC file in a segment we create a corresponding CSV file which lists for every successfully processed WARC record the following information: (1) its URL, (2) its payload digest, (3) the capture timestamp, (4) the HTTP Last-Modified header, (5) the raw and (6) normalized temporal signals from the plain text, and (7) the computed best guess. Together, these files represent a structured catalog of temporal metadata across the entire GeoCities collection.

4 Results

Our aim is to gain insights into the filtering and extraction process and to provide an overview on the temporal signals that we could find. We also compare the different signals to better understand their consistency.

4.1 Filtering and Estimation

Table 1 summarizes the successive filtering steps applied to the original GeoCities dataset in order to obtain a subset of web pages suitable for temporal signal extraction. Starting from a total of 317 million WARC records, approximately 60% correspond to HTTP responses with status code 200. Further restricting the data to successful HTML responses reduces it to 63.7 million records (20% of the total). Applying an additional constraint that URLs must belong to a GeoCities domain yields 59 million records. We removed template pages and error messages by de-duplicating records based on their WARC payload digests. This resulted in 56.7 million unique

[5] https://archive.org/details/geocities

Table 1: The number of records after each filtering step in the GeoCities dataset.

subset after filtering	records	percent
149 WARC Segments	317 151 386	100.00
+HTTP 200	192 216 902	60.61
+MIME text/html	63 684 009	20.08
+Host GeoCities	58 957 720	18.59
+Payload digest unique	56 731 511	17.89
+Text extraction successful	55 593 383	17.53

Table 2: The number of records for subsets of different temporal signals (percentage relative to the 55 593 383 records for which text could be successfully extracted).

type of signal	records	percent
HTTP Last-Modified	48 169 123	86.6
HTML	2 930 834	5.3
Copyright years	1 983 746	3.6
Last updated	1 094 791	2.0
First posted	134 178	0.24
HTTP or HTML	48 230 859	86.8
HTTP and HTML	2 869 098	5.2

payloads, showing that several million GeoCities URLs share identical content. Such de-duplication is especially relevant for temporal analysis, because repeated generic pages (often automatically generated in 2009 due to the site's shutdown) can cause temporal bias. The vast majority of unique pages could be parsed and converted into plain text.

As Table 2 shows, 48 million pages contain at least one temporal signal, either from the HTTP header or from the HTML content. The distribution highlights a strong imbalance between HTTP-based and HTML-based temporal signal types. HTTP Last-Modified headers are present for almost all pages with signals. In contrast, the HTML-based dates are comparatively rare: only about 134 000 pages contain a detectable first posted signal, about one million pages contain a last updated expression, and two million pages contain copyright year information. On the other hand, for more than 2.8 million pages we have signals from two different sources, which we will compare in Section 4.3. Only a small fraction of pages contain multiple HTML-based temporal cues, which suggests that redundancy in content-derived dates is rare. For example, last updated and copyright is found in only 89 301 pages.

4.2 Temporal Distribution

Figure 2 illustrates the chronological progression of pages for which HTML-based temporal signals are available, using the creation best guess as a proxy for page creation year. The values are approximately normal distributed and centered in the early 2000s, consistent with the historical growth and decline of GeoCities. The peaks around 1997 and 2000 could correspond to periods of rapid user adoption of the platform and increased content creation.

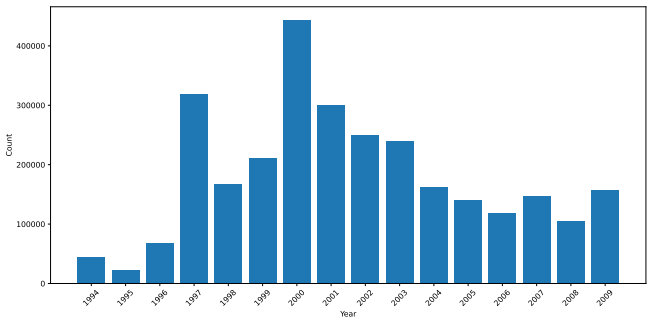


Figure 2: Temporal distribution of best guesses for the creation year of web pages based on HTML content.

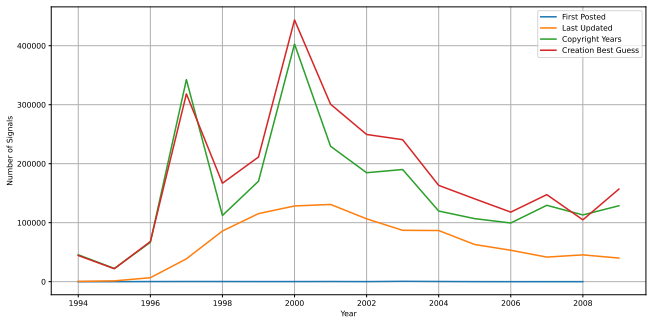


Figure 3: Temporal distribution of the different HTML-based signals.

Figure 3 breaks down this temporal distribution by signal source, plotting four curves corresponding to creation best guess, first posted, last updated, and copyright years. The figure highlights differences in both frequency and temporal spread across signal types. Copyright years are relatively common and broadly distributed, while first posted dates are quite sparse. Last updated signals are in the intermediate position. The copyright years and last updated graphs follow broadly similar trajectories and differ mainly in scale rather than temporal placement, except for two peaks in 1997 and 2000. The overall consistency suggests that the signals capture temporal properties of the same content rather than arbitrary or noisy dates.

4.3 Consistency of Signals

To get a better understanding of the consistency between different signals, we examine their differences. Each bar in Figures 4 and 5 shows for how many pages the LLM estimated a year that was the corresponding number of years earlier/later than the HTML-based guess (Fig. 4) or the HTTP Last-Modified date (Fig. 5). Zero means perfect agreement; negative (positive) values indicate that the LLM predicts the page is newer (older). The mean absolute error (MAE) between the HTML-based guess and the LLM is 3.21 years and 3.7 years when comparing the LLM estimate to the HTTP-based guess. This larger error could indicate that last modified dates reflect later updates rather than the original creation of a page.

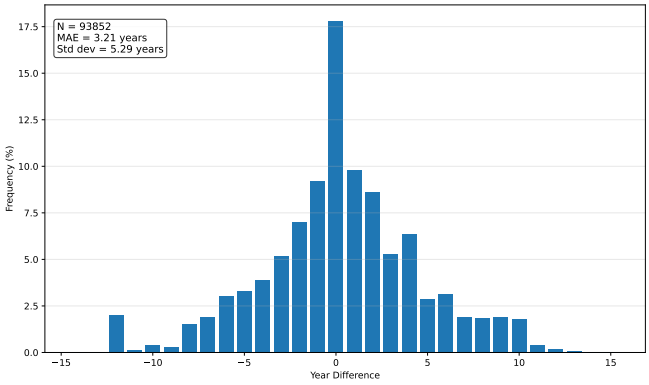


Figure 4: Distribution of year differences (HTML best guess − LLM estimated year).

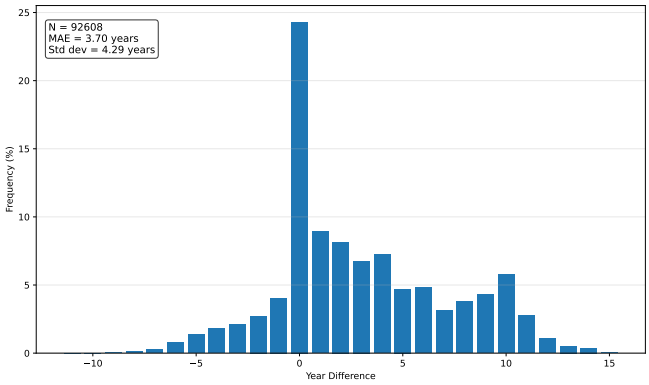


Figure 5: Distribution of year differences (HTTP Last-Modified − LLM estimated year).

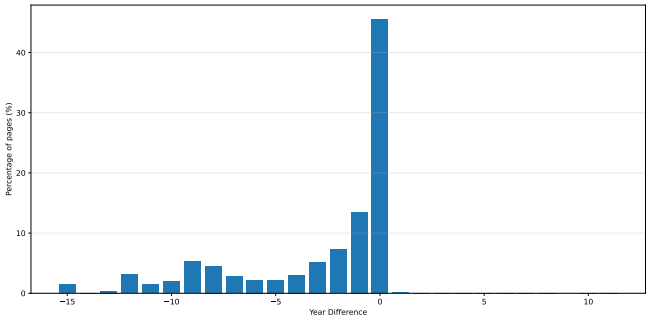


Figure 6: Distribution of differences between HTTP and HTML estimates (HTML best guess − HTTP Last-Modified).

We see this confirmed in Figure 6 which shows, for the 2 869 098 pages for which both HTTP-based and HTML-based temporal signals are available, the difference between the two signals. We observe a distinctly asymmetric distribution with values ranging from approximately −15 to 10 years. It is noticeable that the distribution is heavily concentrated near zero: for 59.2% of the pages the two signals match with an accuracy of ± one year. Negative differences indicate that HTML-based estimates are earlier than the corresponding HTTP Last-Modified headers. This pattern is consistent with the interpretation that HTTP headers often reflect later maintenance, whereas HTML signals capture original creation or early publication times. Only 0.145% of records reveal positive differences.

5 Discussion

With our study we address the estimation of web page creation times in historical web corpora, using the GeoCities archive to demonstrate both promises and limitations of content-based temporal signals. While the HTTP Last-Modified header provides temporal information for a large fraction of pages, HTML-based signals are comparatively sparse, yet, when available they provide valuable additional evidence. Specifically, they can hint at the creation time of a web page which can be different from its time of last change, which is typically reflected by the HTTP Last-Modified header. We systematically applied the filtering, extraction, and normalization procedures across the entire corpus using a transparent and reproducible pipeline. However, we recognize that the absence of a ground truth or gold standard for page creation dates is a central limitation. The dates we have extracted from the HTML content are user-generated, often inconsistently formatted, and were sometimes retrospectively edited. Although this limits their suitability as a gold standard, they are still valuable for analyzing and understanding the GeoCities corpus and for training models. Manual annotation at scale would be a huge effort but still remain uncertain in many cases, especially when no explicit temporal information is available.

Zero-shot prompting an LLM to estimate page creation years from raw page text yields limited accuracy, which raises important questions about whether more specialized prompting strategies, model selection, or fine-tuning approaches could improve performance. The dataset we introduce provides a foundation for training and evaluating models for temporal inference. Apart from the text, other clues from other content types (e. g., timestamps within image metadata) could also be used. Without ground truth data the robustness of the proposed approach cannot be evaluated directly. Instead, we rely on internal consistency checks and cross-signal agreement (HTTP vs. HTML signals) as indirect validation measures. This highlights both the methodological challenge of estimating the creation time for historical web pages and the need for community efforts to gradually improve annotation quality over time.

Since the GeoCities dataset is, in a sense, 'final', it is well suited for research and iterative enrichment. By combining temporal signals with content features (e. g., GIF images) our dataset can support studies of early meme and visual culture. Researchers could trace when specific graphics or textual catchphrases appeared and how they spread across GeoCities neighborhoods. One could also analyze when different thematic communities (e. g., fandoms, hobby groups) emerged, peaked, and declined. For the future, we envision that the GeoCities dataset is continuously enriched with other signals and annotations, for example, about named entities, topics, events, sentiment, social references etc., thereby transforming

it into a 'living' research resource for the community. Such enrichment can be performed in a distributed manner by publishing additional data that can easily be linked to the individual WARC files and records, as we have done with our dataset. We therefore would like to encourage other researchers to join us in exploring and annotating the GeoCities corpus.

References

[1] Janek Bevendorff, Martin Potthast, and Benno Stein. 2021. FastWARC: Optimizing Large-Scale Web Archive Analytics. arXiv:2112.03103 [cs.IR] doi:10.48550/arXiv.2112.03103

[2] Adam Jatowt, Ching-Man Au Yeung, and Katsumi Tanaka. 2013. Estimating Document Focus Time. In *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management* (San Francisco, California, USA) *(CIKM '13)*. ACM, New York, NY, USA, 2273–2278. doi:10.1145/2505515.2505655

[3] Katie Mackinnon. 2022. The death of GeoCities: seeking destruction and platform eulogies in Web archives. *Internet Histories* 6, 1-2 (2022), 237–252. doi:10.1080/24701475.2022.2051331

[4] Ian Milligan. 2017. Welcome to the web: The online community of GeoCities during the early years of the World Wide Web. In *The Web as History*, Niels Brügger and Ralph Schroeder (Eds.). UCL Press, London. https://uwspace.uwaterloo.ca/handle/10012/11859

[5] Ian Milligan. 2019. Exploring Web Archives in the Age of Abundance: A Social History Case Study of GeoCities. In *The SAGE Handbook of Web History*, Niels Brügger and Ian Milligan (Eds.). SAGE Publications Ltd, London, Chapter 23, 344–358. doi:10.4135/9781526470546

[6] Sérgio Nunes, Cristina Ribeiro, and Gabriel David. 2007. Using Neighbors to Date Web Documents. In *Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management* (Lisbon, Portugal) *(WIDM '07)*. ACM, New York, NY, USA, 129–136. doi:10.1145/1316902.1316924

[7] Liudmila Ostroumova Prokhorenkova, Petr Prokhorenkov, Egor Samosvat, and Pavel Serdyukov. 2016. Publication Date Prediction Through Reverse Engineering of the Web. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (San Francisco, California, USA) *(WSDM '16)*. ACM, New York, NY, USA, 123–132. doi:10.1145/2835776.2835796

[8] Hany M. SalahEldeen and Michael L. Nelson. 2013. Carbon Dating the Web: Estimating the Age of Web Resources. In *Proceedings of the 22nd International Conference on World Wide Web* (Rio de Janeiro, Brazil) *(WWW '13 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1075–1082. doi:10.1145/2487788.2488121

[9] Andreas Spitz, Jannik Strötgen, and Michael Gertz. 2018. Predicting Document Creation Times in News Citation Networks. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1731–1736. doi:10.1145/3184558.3191633

[10] Eveline Vlassenroot, Sally Chambers, Emmanuel Di Pretoro, Friedel Geeraert, Gerald Haesendonck, Alejandra Michel, and Peter Mechant. 2019. Web archives as a data resource for digital scholars. *International Journal of Digital Humanities* 1, 1 (01 April 2019), 85–111. doi:10.1007/s42803-019-00007-7