



Proceedings of the
Web Archiving and Digital Libraries (WADL)
Workshop 2025

36th ACM Conference on Hypertext and Social Media (HT 2025).

September 15, 2025

Editors:

Mat Kelly

Brenda Reyes Ayala

Location:

Chicago, Illinois, USA
(hybrid event)

Contents

1	Workshop Organizers	2
1.1	Organizing Committee	2
1.2	Program Committee	2
2	WADL proposal	3
3	Call for papers	7
4	Workshop schedule	8
5	Accepted papers	9
5.1	Coming Back Differently: A Case Study of Near Death Experiences of Webpages	10
5.2	Toward Robust URL Extraction for Open Science: A Study of arXiv File Formats and Temporal Trends	14
5.3	Lost but Preserved - A Web Archiving Perspective on the Ephemeral Web	19
5.4	Medieval Citation Networks as Digital Hyperlinks: Transformer-Based Authorship Attribution in Historical Text Collections	22

I. Workshop Organizers

1.1 Organizing Committee

The 2025 iteration of the WADL workshop was organized by:

- Mat Kelly, Drexel University, mkelly at drexel dot edu
- Brenda Reyes Ayala, University of Alberta, reyesaya at ualberta dot ca

1.2 Program Committee

- Sumitra Duncan (Frick Art Reference Library)
- Joshua Finnell (Colgate University)

- Shawn M. Jones (Google)
- Lauren Ko (University of North Texas Libraries)
- Michael L. Nelson (Old Dominion University)
- Alexander C. Nwala (College of William & Mary)
- Nicholas Taylor (Los Alamos National Laboratory)
- Michele C. Weigle (Old Dominion University)
- Laura Wrubel (Stanford University)

II. WADL proposal

Web Archiving and Digital Libraries (WADL) 2025

Mat Kelly
Drexel University
Philadelphia, PA, USA
mkelly@drexel.edu

Brenda Reyes Ayala
University of Alberta
Edmonton, AB, Canada
reyesaya@ualberta.ca

ABSTRACT

The hypertext ecosystem is inherently dynamic, with digital content continuously evolving and disappearing. Web archiving and digital libraries serve as crucial infrastructures for preserving hypertext structures, enabling scholarly research on the evolution of hyperlinked content, and supporting knowledge continuity across various domains.

This workshop will explore cutting-edge approaches in web archiving and their intersection with hypertext research, emphasizing AI-driven preservation, user engagement with digital archives, decentralized archival systems, and ethical considerations. WADL 2025 will feature interactive sessions, hands-on demonstrations, and cross-disciplinary discussions that directly align with ACM Hypertext 2025 themes, including hypertext history, digital storytelling, automation, decentralized hypertext, and ethical concerns in web preservation.

CCS CONCEPTS

• **Information systems** → **Digital libraries and archives.**

KEYWORDS

Web Archiving, Digital Libraries, Community Building, Digital Preservation

ACM Reference Format:

Mat Kelly and Brenda Reyes Ayala. 2025. Web Archiving and Digital Libraries (WADL) 2025. In *Adjunct Proceedings of the 36th ACM Conference on Hypertext and Social Media (HT Adjunct 2025), September 15–18, 2025, Chicago, IL, USA*. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3720533.3756896>

1 WORKSHOP DESCRIPTION

WADL 2025 will provide a platform for researchers, practitioners, and archivists to collaborate on the design and implementation of tools, standards, and frameworks for web archiving and digital libraries, with a specific emphasis on hypertext structures and their preservation. The workshop will cover the full lifecycle of web-based content, from creation and publishing to archiving, access, and analysis.

WADL 2025 will cover all topics of interest, including but not limited to:

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

HT Adjunct 2025, September 15–18, 2025, Chicago, IL, USA

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1533-4/2025/09

<https://doi.org/10.1145/3720533.3756896>

- continue to build a diverse community of people integrating web archiving with digital libraries,
- help attendees learn about useful methods, systems, and software in this area,
- help define and outline future research and practice in this area, to enable more high-quality web archiving,
- produce an archival publication that will help advance technology and practice, and
- promote synergistic efforts including collaborative projects and proposals.

The objectives of this workshop include the following.

- **Enhancing Hypertextual Archives:** Exploring how web archives serve as historical repositories of hypertext content and how link-based data can be analyzed over time.
- **Leveraging AI and NLP in Digital Libraries:** Investigating how machine learning and NLP techniques can improve web archive searchability, classification, and metadata extraction.
- **Decentralized Web Archiving:** Discussing the role of blockchain, peer-to-peer (P2P) storage, and federated models in building resilient web archives.
- **Ethical and Social Implications of Web Archiving:** Addressing misinformation, censorship, diversity, and biases in digital preservation.
- **Interactive and User-Centric Archival Interfaces:** Exploring hyperlinked knowledge graphs, linked open data, and immersive storytelling techniques that leverage web archives.

2 RELEVANCE TO THE HYPERTEXT COMMUNITIES

This workshop directly aligns with ACM Hypertext 2025 themes:

- Hypertext & Web History – Studying the evolution of linked content and its preservation.
- Digital Storytelling & Memory – Leveraging web archives for narratives and historical analysis.
- AI & Automation – Applying machine learning and NLP to improve archiving and retrieval.
- Decentralized and Distributed Hypertext – Exploring blockchain and P2P models for resilient archiving.
- Social & Ethical Implications – Addressing biases, misinformation, and ethical concerns in archiving.

By explicitly integrating hypertextual methodologies into web archiving research, this workshop will strengthen the connections between the Hypertext, Digital Libraries, and Web Science communities.

3 BIOGRAPHICAL DETAILS

Mat Kelly is an assistant professor at Drexel University’s College of Computing and Informatics. He holds a Ph.D. in Computer Science from Old Dominion University. He has been heavily involved in organizational (e.g., program chair [1]) and refereeing roles in communities like those at the ACM/IEEE-CS Joint Conference on Digital Libraries and is a managing editor at the International Journal on Digital Libraries (IJDL). He was a co-organizer of the two previous iterations of the WADL workshop, held at JCDL 2023 [5] and 2022 [8]. He has led as PI on studies funded by the US Institute of Museum and Library Services (IMLS) exploring the dynamics of preserving online advertisements [6] and personalized, ephemeral web content [7]. More information can be found at: <https://matkelly.com>.

Brenda Reyes Ayala holds a PhD in Information Science from the University of North Texas. She is currently an Associate Professor at the School of Library and Information Studies at the University of Alberta in Edmonton, Canada. She is a Program Committee Member of JCDL, was co-chair of WADL 2023, and was Volunteer Chair of JCDL in 2018. She leads the project “Preserving our digital heritage: Improving the detection of quality problems in web archives”, which is funded by the Social Sciences and Humanities Research Council of Canada. In the past she has worked as a web archivist at the University of North Texas Libraries and the National Library of Spain. More information can be found at <https://reyesayala.github.io>.

4 MOTIVATION

In recent years, interest in web archiving and digital libraries has grown significantly. We anticipate an attendance of approximately 30-35 participants, including a strong representation of students.

5 WORKSHOP AND SUBMISSION FORMATS

Submissions should be between 3-5 pages for a 15-minute presentation and one page for posters and lightning talks. Submissions will use the ACM Proceedings template¹.

All submissions will be peer-reviewed by the program committee, and accepted contributions will be compiled into the WADL program. We anticipate the program to include aspects from multiple disciplines such as Computer Science, Library and Information Science, Web Science, Social Sciences, History, Journalism, etc.

6 LENGTH

We propose a full-day workshop in hybrid format. We anticipate invited speakers, presentations of selected papers and posters, as well as demonstrations and panels.

We plan to schedule a mix of pre-recorded and live in-person presentations (10-15 minutes each), ensuring ample time for discussion and interaction. In addition, we plan to have a small poster session where virtual and in-person attendees can present early-stage work.

¹<https://www.acm.org/publications/proceedings-template>

7 PREVIOUS EDITIONS OF THE WORKSHOP SERIES

The most recent related workshop, WADL 2023 [5] was held as a hybrid, in-person/virtual event², in conjunction with ACM/IEEE-CS JCDL 2023. This event consisted of the presentation of 5 peer-reviewed and accepted papers as well as lightning and drop-in talks by attendees both in-person at the event and remotely attending. The agenda remains available at <https://fox.cs.vt.edu/wadl2023.html>. WADL 2022 was held virtually due to the pandemic, but hosted 13 peer-reviewed paper talks and two invited presentations throughout the full-day session. The schedule remains available at <https://fox.cs.vt.edu/wadl2022.html>. Likewise, schedules for previous iterations of the workshop from 2020³, 2019⁴, 2018⁵, 2017⁶, and 2016⁷ are available at the links provided. While the workshop has often coincided with JCDL, the authors seek to expand the community of interested researchers that attend ACM Hypertext.

In the 2023 iteration of the conference, we received very positive feedback from participants and are therefore strongly encouraged to continue the workshop in 2025. Two previous WADL meetings resulted in the publication of a special issue in the IEEE TCDL Bulletin such as in 2016 [4] and 2015 [3] and the workshop organizers are motivated to revitalize this effort with WADL 2025. Other workshop proceedings will be openly accessible from VTechWorks⁸, Virginia Tech’s institutional repository.

A previous workshop, WIRE [9], focused on research of archival holdings and on making use of archives that preserve the web. The first workshop on Web Archiving and Digital Libraries, WADL 2013, led to a summary [2] after a group responded to the call for meeting as part of the JCDL 2013 workshop program.

An earlier similar workshop at a prior JCDL conference took place in Ottawa in 2011⁹, partly as a result of the emergence of a cooperative to explore web archiving¹⁰. Broader in scope but related are the annual General Assembly meetings of the International Internet Preservation Consortium (IIPC)¹¹. In addition, various sponsored programs have connected, like a closely related initiative¹² funded by the Andrew W. Mellon Foundation.

8 ANY PLAN FOR FURTHER PUBLICATION

As was done for previous WADL events, the organizers will evaluate venues to offer an opportunity to publish invited contributions that were presented in their preliminary stages at the workshop. In the past, the workshop has also led to a call for contributions for a special issue of IJDL. We intend to invest this effort again and edit an IJDL special issue on web archiving following WADL 2025.

²<https://fox.cs.vt.edu/wadl2023.html>

³<https://fox.cs.vt.edu/wadl2020.html>

⁴<https://fox.cs.vt.edu/wadl2019.html>

⁵<https://fox.cs.vt.edu/wadl2018.html>

⁶<https://fox.cs.vt.edu/wadl2017.html>

⁷<https://fox.cs.vt.edu/wadl2016.html>

⁸<https://vtechworks.lib.vt.edu/>

⁹<http://infolab.stanford.edu/wac/>

¹⁰<https://cs.harding.edu/wag2011/>

¹¹<https://netpreserve.org/>

¹²<https://library.columbia.edu/collections/web-archives.html>

9 PROGRAM COMMITTEE MEMBERS

Our previous committee members have helped to ensure a quality program for the workshop that is tailored toward its topic. The following individuals have previously expressed a willingness to evaluate submissions to the workshop. We anticipate requesting their service for peer-review of WADL 2025 submissions.

- Sumitra Duncan, Frick Art Reference Library
- Joshua Finnell, Colgate University
- Shawn M. Jones, Google
- Lauren Ko, University of North Texas Libraries
- Michael L. Nelson, Old Dominion University
- Alexander Nwala, College of William & Mary
- Nicholas Taylor, Los Alamos National Laboratory
- Michele C. Weigle, Old Dominion University
- Laura Wrubel, Stanford University

10 SPECIAL REQUIREMENTS

We anticipate our program to be as inclusive as possible to represent the international community of web archiving and digital libraries researchers and practitioners. As with previous iterations of WADL, we anticipate hosting a hybrid conference. To facilitate this, we request technical assistance to ensure seamless virtual participation while the primary event takes place at the Hypertext conference venue in Chicago.

REFERENCES

- [1] Cristina Ceballos (Ed.). 2024. *JCDL '23: Proceedings of the 2023 ACM/IEEE Joint Conference on Digital Libraries* (Santa Fe, New Mexico, USA). IEEE Press.
- [2] Edward A. Fox and Mohamed M. Farag. 2013. Report on the Workshop on Web Archiving and Digital Libraries (WADL 2013). *ACM SIGIR Forum* 47, 2 (2013), 128–133. <https://doi.org/10.1145/2568388.2568840>
- [3] Edward A. Fox, Zhiwu Xie, and Martin Klein. 2015. Web Archiving and Digital Libraries 2015 (WADL 2015) Overview. *Bulletin of IEEE Technical Committee on Digital Libraries* 11, 2 (2015), 1–2. <https://bulletin.jcdl.org/Bulletin/v11n2/papers/intro.pdf>
- [4] Edward A. Fox, Zhiwu Xie, and Martin Klein. 2017. Web Archiving and Digital Libraries (WADL) 2016: Highlights and Introduction to this Special Issue. *Bulletin of IEEE Technical Committee on Digital Libraries* 13, 1 (2017), 1–2. <https://bulletin.jcdl.org/Bulletin/v13n1/papers/intro.pdf>
- [5] Mat Kelly, Brenda Reyes Ayala, Zhiwu Xie, and Edward A. Fox. 2023. Web Archiving and Digital Libraries (WADL) 2023. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*. IEEE, Santa Fe, New Mexico, 314–315. <https://doi.org/10.1109/JCDL57899.2023.00074>
- [6] Mat Kelly, Alex Poole, Michael L. Nelson, and Michele C. Weigle. 2022–2024. Saving Ads: Assessing and Improving Web Archives' Holdings of Online Advertisements. Institute of Museum and Library Services (IMLS) National Leadership Grant.
- [7] Mat Kelly, Alex Poole, Michael L. Nelson, and Michele C. Weigle. 2024–2026. Preserving Personalized Advertisements for More Accurate Web Archives. Institute of Museum and Library Services (IMLS) National Leadership Grant.
- [8] Martin Klein, Mat Kelly, Zhiwu Xie, and Edward A. Fox. 2022. Web archiving and digital libraries (WADL) 2022. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries* (Cologne, Germany) (JCDL '22). Association for Computing Machinery, New York, NY, USA, Article 57, 2 pages. <https://doi.org/10.1145/3529372.3530920>
- [9] Matthew Weber, David Lazer, and Kris Carpenter Negulescu. 2014. WIRE 2014 Workshop - Working with Internet Archives for Research. <https://web.archive.org/web/20241110081706/https://www.lazerlab.net/event/wire-workshop-working-internet-archives-research>

III. Call for papers

To be held in conjunction with 36th ACM Conference on Hypertext and Social Media (HT 2025).

- September 15 - 18, 2025
- Chicago, Illinois, USA

The Web Archiving and Digital Libraries Workshop (WADL) is back! Join us in-person or online at the ACM Hypertext Conference in Chicago this September!

WADL 2025 will continue its tradition of providing a forum and collaboration platform for international leaders from academia, industry, and government to discuss challenges, and share insights, in designing and implementing concepts, tools, and standards in the realm of web archiving. Together, we will explore the integration of web archiving and digital libraries, over the complete digital resource life cycle: creation/authoring, uploading, publishing on the web, crawling/collecting, compressing, formatting, storing, preserving, analyzing, indexing, supporting access, etc.

WADL 2025 will cover all topics of interest and specifically invite contributions from practitioners. Topics include but are not limited to:

- Enhancing Hypertextual Archives: Exploring how web archives serve as historical repositories of hypertext content and how link-based data can be analyzed over time.
- Leveraging AI and NLP in Digital Libraries: Investigating how machine learning and NLP techniques can improve web archive searchability, classification, and metadata extraction.
- Decentralized Web Archiving: Discussing the role of blockchain, peer-to-peer (P2P) storage, and federated models in building resilient web archives.
- Ethical and Social Implications of Web Archiving: Addressing misinformation, censorship, diversity, and biases in digital preservation.
- Interactive and User-Centric Archival Interfaces: Exploring hyperlinked knowledge graphs, linked open data, and immersive storytelling techniques that leverage web archives.

Submission: Submission should be in ACM 2-column format.

Submission Length

- 3-5 pages for a 15 minute presentation

- 1 page for 5 minute lightning talk

A selection of the best papers from WADL will be invited to submit an extended version for publication in a special issue on web archiving in the International Journal on Digital Libraries (IJDL).

- Due date: ~~July 6, 2025~~ **July 14, 2025**
- Notifications: August 3, 2025

Submissions should be made via EasyChair.

IV. Workshop schedule

Central U.S. Time, September 15, 2025

Time	Session
9:00 am - 9:20 am	Opening Remarks and Technical Setup (slides)
9:20 am - 9:40 am	Paper: <i>Coming Back Differently: A Case Study of Near Death Experiences of Webpages</i> by Lesley Frew, Michael Nelson, and Michele Weigle (recording)
9:40 am - 10:00 am	Paper: <i>Toward Robust URL Extraction for Open Science: A Study of arXiv File Formats and Temporal Trends</i> by Rochana R. Obadage, Lamia Salsabil, Sawood Alam, William A. Ingram, Bipasha Banarjee, Edward A. Fox, and Jian Wu
10:00 am - 10:20 am	Paper: <i>Medieval Citation Networks as Digital Hyperlinks: Transformer-Based Authorship Attribution in Historical Text Collections</i> by Jonathan Schler, Nati Ben-Gigi, Binyamin Katzoff, and Maayan Geffet-Tamir
10:20 am - 10:30 am	Transition Buffer
10:30 am - 11:00 am	Coffee Break (<i>Hosted by ACM Hypertext</i>)
11:00 am - 11:05 am	Session Restart & Quick Welcome Back
11:05 am - 11:25 am	Paper: <i>Lost, but Preserved - A Web Archiving Perspective on the Ephemeral Web</i> by Sawood Alam and Mark Graham
11:45 am - 11:55 am (flexible)	Open Lightning Talks (<i>Drop-in, short-format contributions</i>)
11:55 am - 12:10 pm (flexible)	Community Discussion: <i>The Future of WADL</i>
12:10 pm - 12:30 pm (flexible)	Workshop Wrap-Up & Closing Reflections
12:30 pm onward	Lunch (<i>Hosted by ACM Hypertext</i>)

Table 4.1: WADL 2025 Schedule

V. Accepted papers

The following paper submissions were peer reviewed and presented at the WADL Workshop.

- Lesley Frew, Michael Nelson, and Michele Weigle, “Coming Back Differently: An Exploratory Case Study of Near Death Experiences of Webpages”
- Rochana R. Obadage, Lamia Salsabil, Sawood Alam, William A. Ingram, Bipasha Banarjee, Edward A. Fox, and Jian Wu, “Toward Robust URL Extraction for Open Science: A Study of arXiv File Formats and Temporal Trends”
- Sawood Alam and Mark Graham, “Lost, but Preserved - A Web Archiving Perspective on the Ephemeral Web”
- Jonathan Schler, Nati Ben-Gigi, Binyamin Katzoff, and Maayan Geffet-Tamir, “Medieval Citation Networks as Digital Hyperlinks: Transformer-Based Authorship Attribution in Historical Text Collections”

All accepted papers will be deposited into the University of Alberta Education & Research Archive and be invited to submit an extended version of their work to be published in the International Journal on Digital Libraries (IJDL).

Coming Back Differently: An Exploratory Case Study of Near Death Experiences of Webpages

Lesley Frew

Department of Computer Science
Old Dominion University
Norfolk, Virginia, USA
lfrew001@odu.edu

Michael L. Nelson

Department of Computer Science
Old Dominion University
Norfolk, Virginia, USA
mln@cs.odu.edu

Michele C. Weigle

Department of Computer Science
Old Dominion University
Norfolk, Virginia, USA
mweigle@cs.odu.edu

Abstract

In this case study, we use web archives to analyze 8,824 webpages that were taken offline and subsequently put back online, thus experiencing a “near death experience.” We enumerate the stages of a webpage’s near death experience, including the change from a successful HTTP status code to non-successful and back, the intermediate stage with markers such as an under construction banner, and an analysis of how the pages came back differently.

CCS Concepts

• Information systems → Users and interactive retrieval; Digital libraries and archives.

Keywords

versioned document collections, web archives, government documents

This work is licensed under a Creative Commons “Attribution-NonCommercial-ShareAlike 4.0 International” license.



1 Introduction

The web is ephemeral. According to a recent study, the median lifespan of a webpage is 2.3 years [15]. However, referring to a webpage as having a *lifespan* implies that a webpage begins, then eventually and abruptly ends. For many webpages this is true, but there are also webpages that temporarily stop working, but then resume functioning again. Upon resolving, the webpage could be different than before it stopped functioning. This phenomenon highlights a violation of the *ceteris paribus* assumption that is common in online research [7].

In this case study, we analyze 8,824 webpages on the United States Centers for Disease Control (CDC) website (cdc.gov) that were taken offline and subsequently put back online in early 2025, thus experiencing a “near death experience.” For example, Figure 1 shows a CDC webpage taken offline, and Figure 3 shows a page that has returned with a notification banner that it will be undergoing changes.

2 Identifying Webpages with Near Death Experiences

We made a domain match CDX query¹ on February 17, 2025 for cdc.gov webpages captured in 2025, which resulted in 6,505,819 URL matches, canonicalized by SURT. We then filtered these matches

¹<https://github.com/internetarchive/wayback/tree/master/wayback-cdx-server>

for pages with an HTTP 200 status code followed by at least one non-200 status code and then followed by a 200 status code. This resulted in 25,000 matches. Many of these matches were likely embedded resources, such as images, so we further filtered the matches for items ending in `htm`, `html`, or a slash. This resulted in 8,824 candidates for webpages that experienced a near death experience. Pages with internal redirects, such as pages without a slash redirecting to pages with a slash, appear in the CDX as a candidate but did not truly have a near death experience.

For each of these webpages, we used the Memento protocol [14] to request an archived version of the webpage from the Wayback Machine from near January 1, 2025 and March 1, 2025. We also collected a version of the webpage on the live web in June 2025. CDC webpages contain a meta property with their last updated date, and we separated the webpages into two groups: pages with announced updates via the meta tag as of June 2025 and pages without announced updates. There were 276 pages with announced updates. We then further analyzed these pages manually and discovered additional silent updates to the pages in February 2025, discussed further in Section 3.3.

3 Near Death Experience Stages

We find that pages go through three stages in their near death experience. First, they experience a period of time where they have a non-200 HTTP status code, which is the *clinical death stage*. Next, the pages resolve again but contain a marker that they are imminently changing or in danger, which is the *liminal stage*. Finally, the pages come back, possibly with semantically different content.

3.1 Clinical Death Stage

In the clinical death stage, webpages change from resolving with a successful 200 HTTP status code to any non-successful HTTP status code, as shown in Table 1. Web archives can capture pages with non-successful HTTP status codes, compared to timeout or resolution errors, which are not HTTP events. Figure 1 shows an archived version of a webpage with a 404 status code. In this dataset, pages were only offline for a brief period of time due to a judicial order[5]. In other datasets and contexts, the amount of time a webpage is in this stage would vary.

Table 1 shows the status codes of the pages in this dataset. In order to filter out false positive redirects, we filtered for pages that had a prefix of “http” to remove `http` to `https` redirects, and also removed non-200 pages with a different but canonical URL than the 200 match (such as ending in a slash or not), which removed 3,000 false positive redirects from the results. The status codes shown in Table 1 include only those with at least 20 pages in the dataset. In

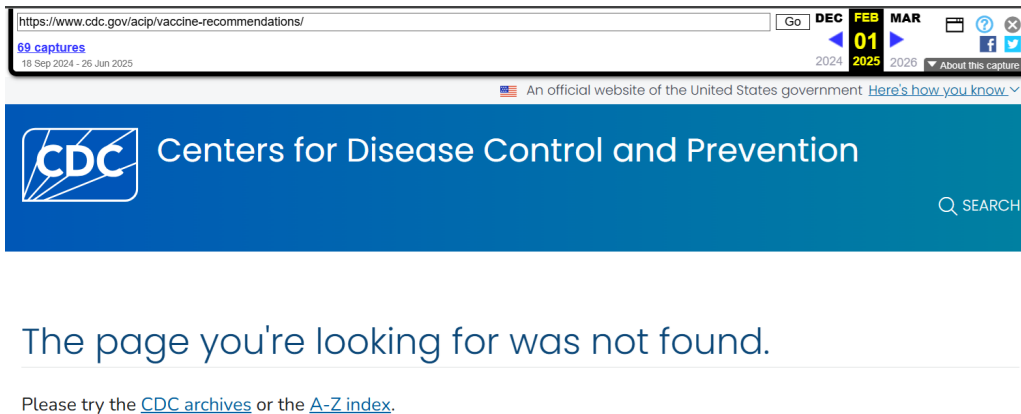


Figure 1: Many CDC webpages have at least one non-successful HTTP status code in February 2025. This webpage’s 404 status code corresponds to the clinical death stage of the webpage’s near death experience.



Figure 2: Data.CDC.gov webpages redirected to this error sink, which corresponds to the clinical death stage of the webpage’s near death experience.

Status Code	All CDX matches	HTML matches only
301 Moved Permanently	9791	4496
302 Found	1153	2
403 Forbidden	8133	793
404 Not Found	3063	1593

Table 1: Status codes of CDC near death webpages, for status codes with at least 20 pages. Redirects 3xx are the most common status code, followed by Forbidden and Not Found.

Mortality Weekly Report site was temporarily moved resulting in 8,021 redirects. These three sites comprise nearly all of the redirects. The error sinks themselves had a status of 200. The presence of so many error sinks highlights the need to explore all status codes when investigating near death webpage experiences, even seemingly transient status codes like redirects.

In this dataset, the clinical death stage corresponds to intentional take down. In other datasets, the clinical death stage may correspond to access prevention for bots or geographical regions [1] or transient server unavailability [8].

In addition to the 404 Not Found status shown in Figure 1, we also found 403 Forbidden and 3xx Redirect statuses as shown in Figure 2.

We further analyzed the redirects. data.cdc.gov had 736 redirects all to the same page as shown in Figure 2, also known as an *error sink* [6]. The National Healthcare Safety Network (NHSN) had 1,730 pages redirecting to another error sink. The Morbidity and

3.2 Liminal Stage

Similar to how humans who experience near death experiences are referred to as being in a liminal state during a coma [9], webpages that experience a near death experience can also exist in a liminal state for a period of time. During this state, these webpages are on the boundary between life and death. The webpages have

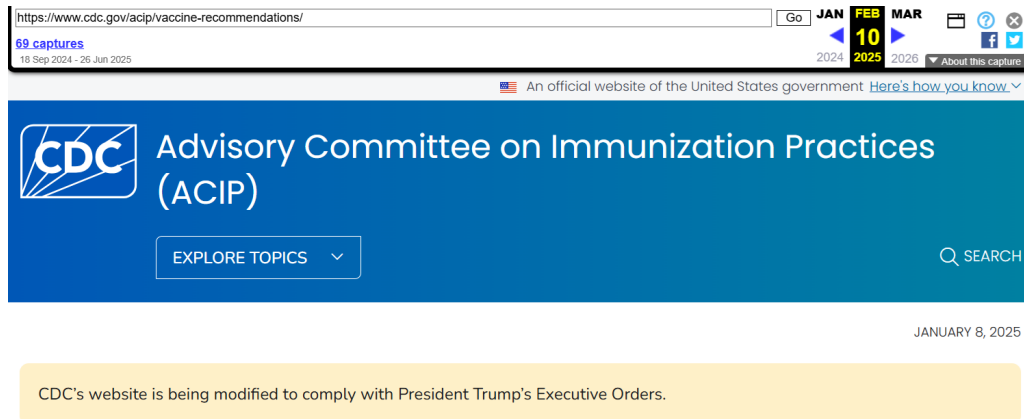


Figure 3: CDC webpages displayed a banner in February 2025 while they were edited to comply with new executive orders. This banner corresponds to the liminal stage of the webpage’s near death experience.

begun resolving with a 200 HTTP status code again during the liminal stage. Figure 3 shows an archived version of the Advisory Committee on Immunization Practices Vaccine Recommendations webpage displaying a banner referencing pending changes. These webpages are highly fragile and are at risk of changing or being entirely permanently deleted. The explicit fragility of this webpage resulted in public efforts to preserve it, as evidenced by captures in February 2025 from Save Page Now and Archive Team. In other datasets and contexts, additional markers of individual webpages or entire websites in this stage could include being labeled “under construction”,² or including a note on the site’s main landing page [12].

3.3 Coming Back Differently

In this dataset, we found three categories of pages: pages that experienced ongoing changes, pages completely unchanged after coming back, and pages that falsely claim to be unchanged. Of the 8,224 pages, 7,257 of them were completely unchanged after coming back. Without additional captures in web archives showing the non-200 status code and the liminal state banner, we would not know these pages had undergone a near death experience at all.

Up to 1,569 pages contained unannounced, silent updates, with 957 of these pages containing the meta tag with the date last updated. Our calculation uses a predefined government website boilerplate removal algorithm [10], though customizing this algorithm would allow us to further differentiate between pages with updated main content and pages with updated sidebar content that was not properly filtered out by the existing algorithm. Figure 4 shows an example of a silent update. The last updated date of this page for both versions is January 8, 2025, but there was a section of the page deleted between January 24 and February 10, 2025. Of these 957 pages with incorrect meta tags, 638 contained text replacements, 243 contained only deletions, and 76 contained only additions. We computed these changes using set calculations on tokens [2]. The median number of words deleted was 14 and the median number of

²<http://www.textfiles.com/underconstruction/>

words added was 4. CDC.gov does not use the HTTP Last-Modified header, so the meta tag is the best way to examine this property in this instance. Only 276 of these pages had announced updates as of June 2025. With one-eighth of pages in this dataset experiencing these silent updates, we conclude that the last updated date is untrustworthy.

There were 276 pages that experienced ongoing changes, with announced changes by June 2025. These pages also experienced a non-200 and non-redirect status in their clinical death stage. Of these, a minimum of 33 pages contained unannounced, silent updates as of March 2025. The reason why this is a lower bound is that webpages that were updated in between January 1 and the presidential inauguration on January 20th would return as announced updates under this setup, as the pair compared was from January 1 and March 1.

We analyzed the additions manually, but found no semantically meaningful additions. Rather, we found these pages experienced changes outside of their defined content, in areas such as the header or footer, or added internal table-of-contents style navigation to the page. We used Nost et al.’s 2020 boilerplate removal for government webpages [10], which demonstrably needs further updates for 2025.

4 Discussion

The applications stemming from web pages with near-death experiences involve two areas: monitoring and presentation. Web archives have seen a rise in many more webpages being captured via citizen archiving and monitoring groups [4] such as EDGI [10]. These monitoring organizations need a way to flag pages in the liminal stage for increased crawling coverage, to make the best use of their finite resources. Already, Save Page Now uses different crawling technology than is used for wide crawls in order to preserve these intentionally saved pages with higher fidelity [11]. While pages that temporarily go offline, perhaps because of a transient server issue, and then return identically will be marked as such in a web archive as a warc/revisit, pages in the liminal stage will have a new hash code, which opens the opportunity for automatic identification.

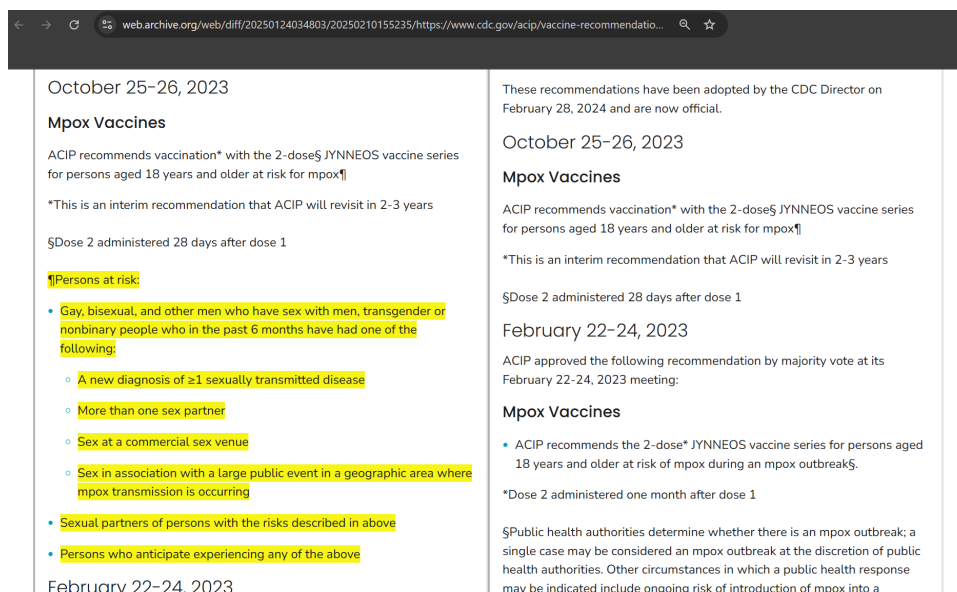


Figure 4: This CDC webpage experienced changes between January 24 and February 10, as shown in the Wayback Machine Changes Tool, but the last updated meta tag date for both pages is January 8.

The second application area of interest is in the presentation of the changes on these pages. Past work focused on presenting pages to best capitalize on cognitive phenomena such as pre-attentive processing [3], but future work will need to identify how to highlight things that should have changed but did not, such as the last modified date area on the page.

5 Future Work

Future work will include analyzing additional markers of stages of webpage near death experiences in larger longitudinal data sets, such as ClueWeb [13], Common Crawl³, or the Not Your Parents' Web dataset[6, 15]. We aim to showcase additional websites that have undergone near death experiences for different amounts of time, and how the stages differ because of that. We also plan to further investigate the motivations for silent updates, both on this CDC dataset and other more heterogeneous datasets.

6 Conclusions

In this work, we presented a case study of CDC webpages that experienced a near death experience: they went offline for a period of time, as captured in web archives, then returned with a banner signaling impending changes. We analyzed the eventual changes to the webpages, and showed that the last modified date is not trustworthy due to a significant amount of silent, unannounced updates on these pages. We use this case study as simply one example of this common phenomenon of near death experiences of webpages, outlining the stages of this experience to enable future study on other datasets.

³<https://commoncrawl.org/>

References

- [1] Anat Ben-David and Adam Amram. 2018. The Internet Archive and the socio-technical construction of historical facts. *Internet Histories* 2, 1-2 (2018), 179–201.
- [2] Lesley Frew. 2024. *Surfacing text changes in archived webpages*. Master's thesis. Old Dominion University.
- [3] Lesley Frew, Michael L Nelson, and Michele C Weigle. 2023. Making changes in webpages discoverable: A change-text search interface for web archives. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 71–81.
- [4] Lesley Frew, Michael L Nelson, and Michele C Weigle. 2025. Temporally Extending Existing Web Archive Collections for Longitudinal Analysis. *arXiv preprint arXiv:2505.24091* (2025).
- [5] Lauren Gardner and Kyle Cheney. 2025. Judge orders Trump admin to restore removed health agency webpages. <https://www.politico.com/news/2025/02/11/health-agency-webpage-removal-lawsuit-00203582>.
- [6] Kritika Garg, Sawood Alam, Michele C. Weigle, Michael L. Nelson, and Dietrich Ayala. 2025. Not Here, Go There: Analyzing Redirection Patterns on the Web. In *Proceedings of the 17th ACM Web Science Conference*. doi:10.1145/3717867.3717925
- [7] David Karpf. 2012. Social science research methods in Internet time. *Information, communication & society* 15, 5 (2012), 639–661.
- [8] Andy Lawrence and Lenny Simon. 2021. Annual outage analysis 2021. *Uptime Institute Intelligence, UII-46 v1*. IPM (2021).
- [9] Limor Meoded Danon. 2016. Between My Body and My "Dead Body" Narratives of Coma. *Qualitative health research* 26, 2 (2016), 227–240.
- [10] Eric Nost, Gretchen Gehrke, Grace Poudrier, Aaron Lemelin, Marcy Beck, Sara Wylie, on behalf of the Environmental Data, and Governance Initiative. 2021. Visualizing changes to US federal environmental agency websites, 2016–2020. *PLOS ONE* 16, 2 (02 2021), 1–27. doi:10.1371/journal.pone.0246450
- [11] Jessica Ogden, Edward Summers, and Shawn Walker. 2024. Know (ing) Infrastructure: The Wayback Machine as object and instrument of digital research. *Convergence* 30, 1 (2024), 167–189.
- [12] Magdalena Olszanowski. 2020. *girl. is. a. four. letter. word The Collective Practices of Amateur Self-Image (in) ing and Personal Website Production 1996 to 2001*. Ph.D. Dissertation. Concordia University Montreal, Quebec, Canada.
- [13] Arnold Overwijk, Chenyan Xiong, and Jamie Callan. 2022. ClueWeb22: 10 Billion Web Documents with Rich Information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3360–3362. doi:10.1145/3477495.3536321
- [14] Herbert Van de Sompel, Michael Nelson, and Robert Sanderson. 2013. RFC 7089 - HTTP framework for time-based access to resource states—Memento. <https://tools.ietf.org/html/rfc7089>
- [15] Michele C. Weigle. 2024. Some URLs Are Immortal, Most Are Ephemeral. <https://ws-dl.blogspot.com/2024/09/2024-09-20-some-urls-are-immortal-most.html>.

Toward Robust URL Extraction for Open Science: A Study of arXiv File Formats and Temporal Trends

Rochana R. Obadage
Old Dominion University
Norfolk, VA, USA
rochana@cs.odu.edu

Lamia Salsabil
Old Dominion University
Norfolk, VA, USA
lsals002@odu.edu

Sawood Alam
Internet Archive
San Francisco, CA, USA
sawood@archive.org

William A. Ingram
Bipasha Banerjee
Virginia Tech
waingram@vt.edu, bipashabanerjee@vt.edu

Edward A. Fox
Virginia Tech
Blacksburg, VA, USA
fox@vt.edu

Jian Wu
Old Dominion University
Norfolk, VA, USA
j1wu@odu.edu

Abstract

In this work, we study how URL extraction results depend on input format. We compiled a pilot dataset by extracting URLs from 10 arXiv papers and used the same heuristic method to extract URLs from four formats derived from the PDF files or the source LaTeX files. We found that accurate and complete URL extraction from any single format or a combination of multiple formats is challenging, with the best F1-score of 0.71. Using the pilot dataset, we evaluate extraction performance across formats and show that structured formats like HTML and XML produce more accurate results than PDFs or Text. Combining multiple formats improves coverage, especially when targeting research-critical resources. We further apply URL extraction on two tasks, namely classifying URLs into open-access datasets and software and the others, and analyzing the trend of URLs usage in arXiv papers from 1992 to 2024. These results suggest that using a combination of multiple formats achieves better performance on URL extraction than a single format, and the number of URLs in arXiv papers has been steadily increasing since 1992 to 2014 and has been drastically increasing from 2014 to 2024. The dataset and the Jupyter notebooks used for the preliminary analysis are publicly available at <https://github.com/lamps-lab/arxiv-urls>.

CCS Concepts

• **Information systems** → **Digital libraries and archives**; *Data analytics*; • **Applied computing** → **Digital libraries and archives**; *Publishing*.

Keywords

Open Access Datasets and Software, Extraction Performance, Reproducibility, Preserving OADS

ACM Reference Format:

Rochana R. Obadage, Lamia Salsabil, Sawood Alam, William A. Ingram, Bipasha Banerjee, Edward A. Fox, and Jian Wu. 2025. Toward Robust URL Extraction for Open Science: A Study of arXiv File Formats and Temporal Trends. In *Proceedings of Web Archiving and Digital Libraries Workshop 2025*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WADL 2025, Chicago, Illinois, USA

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
<https://doi.org/XXXXXXX.XXXXXXX>

(WADL 2025). ACM, New York, NY, USA, 5 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Recent research found that accessibility of datasets and software is crucial for reproducing the research findings of a paper [2]. As a digital library, arXiv [3, 12], hosting over 2.3 million open-access full-text papers, serves as a critical platform for early research dissemination and long-term archiving. However, the digital resources these papers reference face a troubling reality: they disappear at an alarming rate [14, 16]. This link rot undermines the accessibility of shared datasets and software, which can further play a negative impact on reproducibility of research findings.

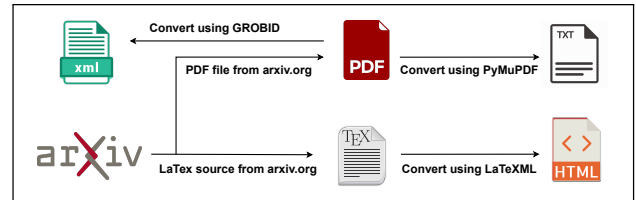


Figure 1: Different file formats for an arXiv paper and how they are obtained for our study.

Extracting URLs from PDF documents is an important task for a variety of downstream tasks such as URL classification, meta-analysis, and web archiving. Most existing methods first extract Text from PDFs [5, 9, 22] and then use heuristic methods to identify URLs. However, this intuitive method has several drawbacks. First, mainstream text extraction tools do not preserve the text flow in the original paper. This leads to truncated lines and thus URLs. Second, URLs may appear in the text and annotation layers of PDFs but most text extractors neglect URLs in annotation layers. Finally, in certain tasks, it is necessary to identify the locations of URLs, such as the footnotes or references, but it is very challenging to use a heuristic method (such as regular expressions) to accurately identify the locations.

arXiv provides papers in multiple formats including PDF and LaTeX source. Recently, arXiv launched an initiative to use HTML to publish research papers. The authors first submit the source files in LaTeX, which are then automatically converted to HTML format. However, the complexities of conversion may introduce

noise or artifacts that prevent URLs from being accurately extracted. Compared with PDFs, HTML documents are more machine readable and thus much easier to extract URLs. In addition, GROBID is an open-source machine learning library designed to extract, parse, and re-structure scholarly papers. Its output is in TEI format, a standard for encoding and representing texts in scholarly papers. The output of GROBID marks rich semantic information such as metadata, references and segmented sections. How the performance of URL extraction depends on the input document format has not been well investigated. Our work aims to fill this gap using arXiv papers.

arXiv is a server that allows authors to submit and publish preprint papers. Since its launch in 1991, this digital service provided by arXiv has been extremely popular across a variety of scientific disciplines, and several sister preprint servers for fields not covered by the original arXiv, such as bioRxiv and medRxiv, have been established afterward. Understanding the dependence on the input format will help us to improve the performance of URL extraction. The comparison results will also guide us to develop data-ensemble methods that aggregate extraction results from different inputs. We systematically evaluate URL extraction performance across four full-text formats (Figure 1) and analyze temporal trends in URL usage patterns from 1992 to 2024.

Our contributions are twofold. First, we provide a comparative evaluation of URL extraction performance across four file formats, identifying which formats yield the best performance. Second, we present a longitudinal analysis of URLs in arXiv papers, revealing how the overall usage of URLs in a yearly sample has evolved over three decades.

2 Related Work

Researchers have explored various methods for extracting URLs from scholarly documents, but most approaches focus on a single file format [5, 9, 22]. Large-scale efforts like S2ORC [17] and Semantic Scholar [11] mainly extract URLs from PDFs or preprocessed Text using regular expressions and heuristic filters to identify valid web addresses. Tools such as GROBID [1] and PyMuPDF [21] are designed for PDF-based extraction, targeting structured information and references. GROBID uses PDFBox to extract Text from PDFs before applying ML models to parse the document.

Some recent studies demonstrated the benefits of using HTML or XML formats, because these preserve more semantic structure [17, 18] in the original document and are more “machine-friendly”, to retain the fidelity of the original content. However, these studies do not focus specifically on URL extraction and do not perform direct comparisons across multiple formats as input.

Studies show that links in academic papers decay over time, with rates varying by domain and hosting platform. Klein et al. documented widespread link rot [15], and Hennessey et al. found that some domains are more vulnerable than others [13]. A 2025 study of over 12,000 GitHub repositories found that more than 10% had no archived version and that even archived pages often suffered from damage or missing source files [6]. Escamilla et al. (2023) showed that a hybrid classifier can identify open-access data and software URIs across both major and niche platforms [10]. However, these works either focus on particular domains, e.g., Git

Hosted Platforms (GitHub, GitLab, etc.) or use off-the-shelf software without thorough evaluation.

The current literature has two key gaps. First, most URL extraction tools perform information extraction based on text converted from PDFs and do not compare performance across different formats. As we will show, this may miss a significant number of URLs due to the broken content flow when text is extracted from PDFs. Second, there are no benchmarks designed specifically for extracting URLs from scholarly papers. “Hence, it is difficult to assess or improve how well a system captures URLs extracted from various formats.

3 Dataset

To evaluate the URL extraction performance, we must build a dataset containing research papers in multiple formats (PDF, Text, XML, and HTML). We constructed our pilot dataset from arXiv (version: January 2024), which contains over 2.3 million open-access full-text research papers since 1991. We used stratified random sampling to select 1,161 papers, drawing three papers from each “year, month” stratum from 9107 to 2401. We applied PyMuPDF (v1.24.13) [21] to extract Text directly from PDFs. Among the sampled papers, 726 had LaTeX source files. Only 17 included HTML versions from arXiv, hence we used the external tool LaTeXML (v0.8.8) [4] to convert LaTeX source files to HTML, which is the same tool used by arXiv to generate HTML files. Out of 726 LaTeX source files, the tool successfully converted 204 into HTML. The remaining files could not be converted due to inherent limitations and conversion errors in the tool.

Table 1: File format coverage and conversion tools used for the randomly selected sample of 1,161 arXiv papers

Format	Conversion Tool	# Papers	# Papers with extracted URLs
PDF	–	1,161	–
Text	PyMuPDF [21]	1,161	260
LaTeX	–	726	252
HTML	LaTeXML [4]	204	134
XML	GROBID [1]	60	60

After collecting and converting the files, we automatically excluded papers that do not contain any URL in any format. Table 1 summarizes the number of papers for each format and the number of papers from which we were able to extract at least one URL. Although all 1,161 papers were available in both PDF and Text form, only 60 papers contained URLs in all three formats: Text, LaTeX, and HTML. We selected this subset and used GROBID (v0.8.1) [1] to convert their PDFs into XML format. Finally, from the 60 papers that contained URLs in all five formats, we randomly selected 10 papers as the pilot dataset for our preliminary evaluation. The pipeline to generate different formats is shown in Figure 1.

4 URL Extraction Performance

4.1 Ground Truth

To build the ground truth, we manually inspected the PDF versions of 10 selected papers (see Section 3) and identified all valid URLs

(ground truth). In our study, we consider a URL to be **valid** if it appears in the PDF version. This process resulted in a ground truth set of 87 valid URLs.

As a preliminary study, we apply a heuristic method for all file formats and try to identify URLs using markups of regular expressions. For the Text format, we applied regular expression-based pattern matching to identify URL candidates. For the LaTeX format, we extracted URLs from .tex and .bbl files using both regular expressions and the \url and \urladdr anchors. For HTML files, we extracted URLs enclosed by the <a> tags, and for XML files, we selected elements containing a target attribute (eg., <ref target="https://github.com/kermit2/grobid"/>).

After the extraction process, we excluded self-referencing URLs (these are the links that point back to the same paper or its hosting page on arXiv). To evaluate the performance of each format, we extended the ground truth by creating a superset of URLs. This superset includes all URL-like candidates extracted from the Text, LaTeX, HTML, and XML formats in addition to the human-verified URLs from the PDF versions. Extending the set was necessary to accurately compute precision, which requires the total number of extracted URL strings, including both valid and **invalid** (URL not found in the PDF) URLs.

Framing it as a set-matching problem, we evaluate the URL extraction performance for each individual format and format combination using precision, recall, and F1 score, based on the total number of extracted URLs and the subset identified as valid URLs. We use the following definitions:

- **Precision** = extracted valid URLs / total extracted URL strings
- **Recall** = extracted valid URLs / total valid URLs (ground truth)

Table 2 presents the URL extraction performance across individual file formats and their combinations, evaluated against a ground truth of 87 valid URLs, manually identified from the PDF versions of 10 selected papers.

Table 2: URL extraction performance across format combinations. P = Precision; R = Recall; V. URLs = Number of valid URLs identified through manual PDF inspection.

Format	V. URLs	P	R	F1
Text	22	0.42	0.25	0.31
LaTeX	24	0.57	0.28	0.38
HTML	56	0.67	0.64	0.65
XML	39	1.00	0.45	0.62
Text + LaTeX	34	0.41	0.39	0.40
Text + HTML	65	0.53	0.75	0.62
Text + XML	51	0.63	0.59	0.61
LaTeX + HTML	69	0.61	0.79	0.69
LaTeX + XML	56	0.76	0.64	0.69
HTML + XML	60	0.69	0.69	0.69
Text + LaTeX + HTML	72	0.49	0.83	0.62
Text + LaTeX + XML	59	0.55	0.68	0.61
Text + HTML + XML	69	0.55	0.79	0.65
LaTeX + HTML + XML	73	0.62	0.84	0.71
Text + LaTeX + HTML + XML	76	0.50	0.87	0.64

First, the results indicate that accurate and complete URL extraction from scholarly papers is a challenging task, and no single format or combination of formats achieves a perfect or even nearly perfect result. Among single formats, HTML achieved the highest F1 score (0.65), with a balanced precision (0.67) and recall (0.64), highlighting its comparative effectiveness for URL extraction. XML exhibited perfect precision (1.00) but only moderate recall (0.45), indicating that while all extracted URLs were valid, it missed many that were present in the PDF files. LaTeX and Text formats showed relatively lower recall (0.28 and 0.25 respectively), with LaTeX outperforming Text slightly in terms of precision (0.57 vs. 0.42).

Combining multiple formats substantially improved recall and the overall F1 scores. For example, the LaTeX + HTML combination achieved an F1 score of 0.69, with a high recall (0.79) and a precision (0.61). Notably, combining three formats (LaTeX + HTML + XML) yielded the highest F1 score (0.71) and recall (0.84). The four-format combination (Text + LaTeX + HTML + XML) achieved the highest recall (0.87) but a slightly reduced F1 score (0.64), primarily due to a lower precision (0.50) caused by an increase in non-valid URL-like strings from the Text format.

The results underscore the importance of leveraging multiple formats to improve URL extraction. While single formats can offer partial coverage, combining other formats, especially HTML and XML, delivers more comprehensive and accurate extraction.

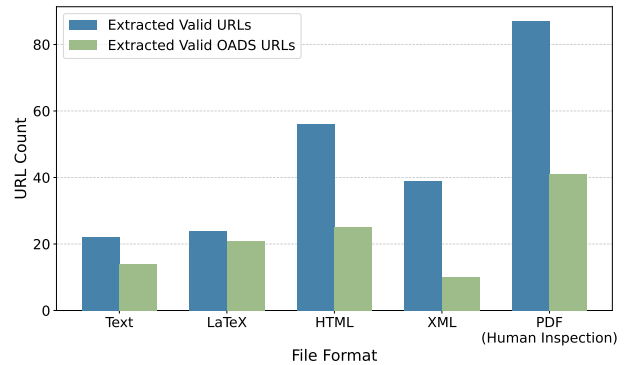


Figure 2: Number of valid and OADS URLs extracted per file format, compared to human inspection of 10 PDF papers.

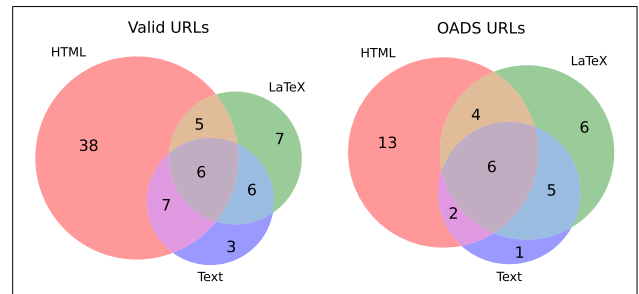


Figure 3: Overlap of valid and OADS URLs across Text, LaTeX, and HTML formats for the 10 manually annotated papers.

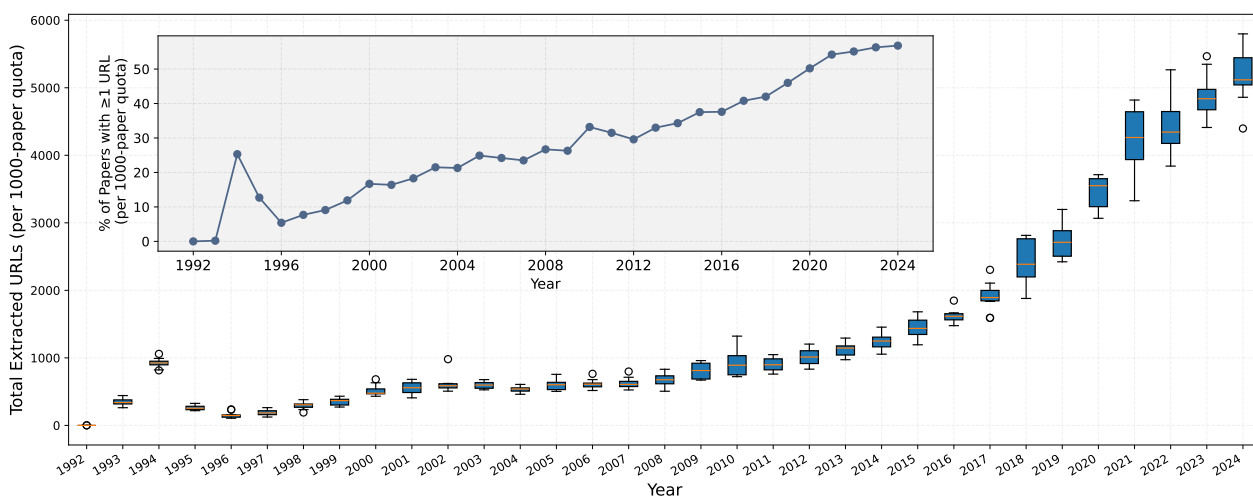


Figure 4: A composite distribution of extracted URLs from arXiv papers. The boxplot for each year is obtained by 10 random draws of 1,000 papers with replacement from all papers published in that year. Inset: Percentage of papers containing at least one URL, based on a single random sample of 1,000 papers per year.

4.2 OADS URLs

The URLs extracted encompass a wide variety of URLs. Analyzing them, we focus on a particular type of URLs that link to open-access datasets and software (OADS) [20]. These URLs have shown to be important for reproducing the findings reported in research papers in AI [2]. To this end, we manually labeled the 87 valid URLs extracted from the PDFs and found that 41 of them are OADS URLs. Figure 2 displays the number of valid OADS URLs and all valid URLs extracted for each format. Among all formats, HTML yielded the highest number of OADS URLs (25) but LaTeX yielded the highest fraction of OADS URLs (87.5%). The Text format yields a low OADS URL extraction number (14) but a decent extraction rate (63.6%). We compare the overlap of valid and OADS URL availability across these three formats in Figure 3.

As shown in Figure 3, several URLs were extracted from all three formats, but each format also included several URLs not extracted in the others. Figure 3 further shows that relying on a single format can miss a significant number of URLs when they are extracted using a regular expression-based method. Using multiple formats increases the chance toward a more complete set of URLs. A similar conclusion holds for OADS URLs. In particular, extraction from HTML format yields the highest number of valid URLs and OADS URLs.

5 Temporal Trends in arXiv URLs

Understanding how scholars reference web-based resources over time is crucial for designing reliable preservation strategies and improving long-term access to research artifacts. As open science increasingly relies on external datasets, software, and repositories linked via URLs, examining historical patterns of URL usage helps identify shifting practices and emerging dependencies. Here we present a longitudinal analysis of URLs in 2.3 million arXiv papers to reveal how linking behaviors have evolved over three decades.

As a pilot study, we built a subset of arXiv papers by randomly sampling 1000 arXiv papers per year from 1992 to 2024, resulting in a total of ~33,000 papers.

To account for sampling variability and ensure robust yearly estimates, we repeated the sampling process ten times. Each time, we obtain the number of URLs extracted for each year. By combining the data collected in 10 times, we produce a composite URL count distribution in Figure 4. We extracted URLs from each paper’s Text format using regular expression-based pattern matching. Although the Text format did not yield the most accurate or complete URL extractions compared to other formats, we included it in our study because it was readily and consistently available across the full arXiv corpus at the time of analysis. It allowed us to assess the trend with minimal preprocessing.

Figure 4 inset presents the percentage of papers containing at least one URL over time from the 1000 selected paper samples per year. The adoption of URLs in arXiv papers has increased significantly over the past three decades. In 1993, fewer than 0.02% of sampled papers included even a single URL, reflecting the limited use of the internet in scholarly communication at the time. Adoption remained below 10% through the mid-1990s (except for 1994), but began rising steadily after 1998, coinciding with the broader availability of web browsers and growing reliance on online resources in academic research [8, 19].

6 Discussion

Our pilot study indicates that no single file format provides accurate and complete coverage of URLs found in the original scholarly papers. Indeed, the format of the input document strongly affects the URL extraction. Structured formats such as HTML and LaTeX usually yield better results because they keep the content clean and machine-readable. HTML includes semantic tags that separate URLs from other text, and LaTeX preserves the original content

without layout issues. In contrast, text extracted from PDFs often deform URLs through formatting complexities such as line breaks, hyphenation, or embedded graphics. We found that no single format captures OADS links consistently. Combining multiple formats, particularly HTML, XML, and LaTeX, improves coverage by ~39% to ~51% compared with relying on only the Text format. However, combining formats can also increase false positives. One solution is to filter or score links based on confidence.

Our temporal analysis reveals an accelerating dependency on external digital resources in scientific research. The increase in papers containing linked resources, from only 0.02% in the early 1990s to approximately 55% in recent years, marks a fundamental shift in scholarly communication and knowledge sharing. The notable spike in 1994, where around 25% of papers contained URLs, stood out from the much lower rates in surrounding years. At first, this seemed like an artifact in the data, but further investigation confirmed it was real and reflected the early growth of the World Wide Web. By late 1993, over 500 web servers were active [7], and several papers from 1994, particularly in Computer Science and Physics, cited these early web resources, contributing to a temporary but genuine increase in URL usage. This finding illustrates how individual researchers or research groups can drive the adoption of new technologies before broader community acceptance occurs. The 1994 anomaly represents early adopters experimenting with web-based resource sharing, indicating the widespread adoption that would follow several years later.

Our study faces several limitations that potentially affect the generalizability of our findings. The first is the small sample size because not all formats are available for most papers. Additionally, in the temporal trend analysis, the results are based on URLs extracted from Text files. Further experiments using other input formats should be conducted to validate and confirm these findings.

7 Conclusion and Future work

In this study, we examined URL extraction performance across multiple arXiv file formats using a pilot dataset, revealing clear differences in extraction accuracy and completeness. Markup formats such as HTML and XML generally provide more accurate and comprehensive URL extraction compared to Text format. Our results highlight that relying on a single file format risks missing important links, especially those related to open-access datasets and software. Combining formats improves coverage but requires attention to the unique challenges each format presents. Our temporal analysis reveals how referencing web-based resources in scholarly papers has steadily increased over the past thirty years.

Several research directions emerge from our findings that could substantially advance scholarly preservation efforts. Scaling human annotation to more papers would provide more robust ground truth data and enable detailed error analysis across different research domains. Expanding our analysis to include S2ORC and PubMed would provide insights across scholarly repositories in a broader range of domains.

References

- [1] 2008–2025. GROBID. <https://github.com/kermitt2/grobid.swh:1.dir:dab86b296e3c216e2241968f0d63b68e8209d3c>

- [2] Kehinde Ajayi, Muntabir Hasan Choudhury, Sarah M. Rajtmajer, and Jian Wu. 2023. A Study on Reproducibility and Replicability of Table Structure Recognition Methods. https://doi.org/10.1007/978-3-031-41679-8_1, 3–19 pages.
- [3] arXiv. 2025. arXiv.org: e-Print archive for Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance, and Statistics. <https://arxiv.org>. Accessed: 2025-07-01.
- [4] Deyan Ginev Bruce R. Miller. 2004. LaTeXML A LaTeX to XML/HTML/MathML Converter. <https://github.com/bruceMiller/LaTeXML>. [Accessed 27-06-2025].
- [5] Duy Duc An Bui, Guilherme Del Fiol, and Siddhartha Jonnalagadda. 2016. PDF text classification to leverage information extraction from publication reports. *Journal of Biomedical Informatics* 61 (2016), 141–148. doi:10.1016/j.jbi.2016.03.026
- [6] David Calano, Michael Nelson, and Michele Weigle. 2025. GitHub Repository Complexity Leads to Diminished Web Archive Availability. In *Proceedings of the 17th ACM Web Science Conference 2025 (WebSci '25)*. Association for Computing Machinery, New York, NY, USA, 449–459. doi:10.1145/3717867.3717920
- [7] CERN. 2024. A short history of the Web. <https://home.cern/science/computing/birth-web/short-history-web>. Accessed: 2025-07-01.
- [8] Sandra L De Groot, Mary Shultz, and Deborah D Bleic. 2014. Information-seeking behavior and the use of online resources: a snapshot of current health sciences faculty. *J Med Libr Assoc* 102, 3 (July 2014), 169–176.
- [9] Jingcheng Du, Dong Wang, Bin Lin, Long He, Liang-Chin Huang, Jingqi Wang, Frank J Manion, Yeran Li, Nicole Cossrow, and Lixia Yao. 2025. Use of deep learning-based NLP models for full-text data elements extraction for systematic literature review tasks. *Scientific Reports* 15, 1 (June 2025), 19379.
- [10] Emily Escamilla, Lamia Salsabil, Martin Klein, Jian Wu, Michele C. Weigle, and Michael L. Nelson. 2023. It's Not Just GitHub: Identifying Data and Software Sources Included in Publications. In *Linking Theory and Practice of Digital Libraries*, Omar Alonso, Helena Cousijn, Gianmaria Silvello, Mónica Marrero, Carla Teixeira Lopes, and Stefano Marchesin (Eds.). Springer Nature Switzerland, Cham, 195–206.
- [11] Rodney Kinney et al. 2025. The Semantic Scholar Open Data Platform. arXiv:2301.10140 [cs.DL] <https://arxiv.org/abs/2301.10140>
- [12] Paul Ginsparg. 2011. It was twenty years ago today ... arXiv:1108.2700 [cs.DL] <https://arxiv.org/abs/1108.2700>
- [13] Jason Hennessey and Steven Xijin Ge. 2013. A cross disciplinary study of link decay and the effectiveness of mitigation techniques. *BMC Bioinformatics* 14, 14 (Oct. 2013), S5.
- [14] Susan Howell and Amber Burtis. 2022. The continued problem of URL decay: an updated analysis of health care management journal citations. *J Med Libr Assoc* 110, 4 (Oct. 2022), 463–470.
- [15] Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. 2014. Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLOS ONE* 9, 12 (12 2014), 1–39. doi:10.1371/journal.pone.0115253
- [16] Viktor Lakic, Luca Rossetto, and Abraham Bernstein. 2023. Link-Rot in Web-Sourced Multimedia Datasets. In *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I* (Bergen, Norway). Springer-Verlag, Berlin, Heidelberg, 476–488. doi:10.1007/978-3-031-27077-2_37
- [17] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4969–4983. doi:10.18653/v1/2020.acl-main.447
- [18] Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2018. Information extraction from scientific articles: a survey. *Scientometrics* 117, 3 (Dec. 2018), 1931–1990. doi:10.1007/s11192-018-2921-5
- [19] Pavel Panchevka and Chris Harrelson. 2025. History of the Web. In *Web Browser Engineering*. Oxford University Press. doi:10.1093/9780198913887.003.0003 arXiv:<https://academic.oup.com/book/0/chapter/498097450/chapter-pdf/61200141/isbn-9780198913887-book-part-2.pdf>
- [20] Lamia Salsabil, Jian Wu, Muntabir Hasan Choudhury, William A. Ingram, Edward A. Fox, Sarah M. Rajtmajer, and C. Lee Giles. 2022. A Study of Computational Reproducibility using URLs Linking to Open Access Datasets and Software. In *Companion Proceedings of the Web Conference 2022 (Virtual Event, Lyon, France) (WWW '22)*. Association for Computing Machinery, New York, NY, USA, 784–788. doi:10.1145/3487553.3524658
- [21] Julian Smith. 2016. PyMuPDF. <https://pypi.org/project/PyMuPDF/>. [Accessed 27-06-2025].
- [22] Ke Zhou, Richard Tobin, and Claire Grover. 2014. Extraction and analysis of referenced web links in large-scale scholarly articles. In *IEEE/ACM Joint Conference on Digital Libraries*. 451–452. doi:10.1109/JCDL.2014.6970220

Lost, but Preserved – A Web Archiving Perspective on the Ephemeral Web

Sawood Alam
Internet Archive
San Francisco, CA, USA
sawood@archive.org

Mark Graham
Internet Archive
San Francisco, CA, USA
mark@archive.org

ABSTRACT

The World Wide Web, our era’s most dynamic information ecosystem, is characterized by its transient nature. Recent studies have highlighted the alarming rate at which web content disappears or changes, a phenomenon known as “link-rot”. A 2024 Pew Research Center study revealed that 38% of webpages from 2013 were inaccessible a decade later and a quarter of the URLs from their entire dataset spanning across a decade were found dead. Even more striking, Ahrefs, an SEO company, reported that at least 66.5% of links to sites created in the last nine years are now dead. These findings echo earlier research by Zittrain et al., which uncovered significant link-rot in journalistic references from New York Times articles.

While these statistics paint a grim picture of digital impermanence, they often overlook a crucial factor: the role of web archives. This work aims to reframe the link-rot discussion by considering the preservation efforts of various web archiving institutions. Our research revisiting the Pew dataset yielded a surprising discovery: only one in ten URLs from the original study were truly missing as opposed to one in four, the remaining bulk had at least one capture in the Wayback Machine. This finding suggests that the digital landscape, when viewed through the lens of web archiving, may be less ephemeral than commonly perceived.

KEYWORDS

Link-Rot, Web Archiving, Ephemeral Web, PEW Research, Wayback Machine

1 INTRODUCTION

Last year, the Pew Research Center published a link-rot study, “When Online Content Disappears” [1]. They stated, “38% of webpages that existed in 2013 are no longer accessible a decade later”. They further noted, “a quarter of all webpages that existed at one point between 2013 and 2023 are no longer accessible”. This is not an isolated report that quantified the rate of loss of the online information. Numerous other link-rot studies in the last two decades have reported similar numbers or worse, depending on the context and samples. For example, Ahrefs, an SEO company, earlier this year reported, “At Least 66.5% of Links to Sites in the Last 9 Years Are Dead” [2]. In 2021, Jonathan Zittrain published an article in the Atlantic, “The Internet Is Rotting” [3], in which their team analyzed about 2 million external links from the New York Times (NYTimes)¹ articles and reported that 25% of deep links have rotted. They further noted that 72% of the older links from 1998 were dead. A recent longitudinal study on link-rot from the Old Dominion University (ODU), “Some URLs Are Immortal, Most Are Ephemeral” [4],

¹<https://www.nytimes.com/>

analyzed 27.3 million URL samples from the Wayback Machine² since 1996 and reported that about 65% of the sampled URLs were found dead on the live web, when checked in 2023. Brewster Kahle, the founder of the Internet Archive, has been citing numbers from the early days of the web and stating the average life of web pages to be anywhere from 40 to 100 days. Different studies have looked at the problem from different perspectives and contexts, hence it is often difficult to compare them side-by-side, but they all agree on the fact that an increasing number of links are rotting with the passage of time. However, some of these studies (not all) have failed to acknowledge the existence of web archives, such as the Wayback Machine, where a portion of the dead web might be preserved and can be used as a fallback when a reference leads to a broken link.

In this work we go through some of the link-rot studies and look at them from the perspective of the Wayback Machine to see how much of the dead web can be rescued. Table 1 shows the status of the dead and rescued web at a glance as sampled by a few different studies. This work was presented at the IIPC WAC 2025 conference³ and the recording of the talk is available on YouTube⁴.

2 METHODOLOGY

Below are brief descriptions of some terminologies that we use in this work:

- *Alive*: URLs that return 200 OK HTTP status code when resolved
- *Dead*: URLs that return an HTTP error status codes, TCP connection errors, or DNS failures when resolved
- *Preserved*: URLs that are *Alive* on the live web as well as present in a web archive
- *Rescued*: URLs that are *Dead* on the live web, but are present in a web archive
- *Endangered*: URLs that are *Alive* on the live web, but are not present in any web archive
- *Vanished*: URLs that are *Dead* on the live web and also not present in any web archive
- *Archived*: *Preserved* + *Rescued*
- *Accessible*: *Preserved* + *Rescued* + *Endangered*

2.1 Datasets

Pew Research Center has generously shared their dataset with us. Their dataset contains 5.4 million unique URLs in general, news, government, and Wikipedia references categories sampled from the CommonCrawl archive⁵ and Wikipedia pages. They also reported

²<https://web.archive.org/>

³<https://netpreserve.org/ga2025/>

⁴<https://www.youtube.com/watch?v=gmU3vFbs2GM>

⁵<https://commoncrawl.org/>

Table 1: Dead links from various link-rot studies rescued by the Wayback Machine.

Study	Year	Sample Period	Sample Size (URLs)	Dead	Rescued
Pew (All)	2024	2013-2023	5.4M	26%	16%
Pew (General)	2024	2013-2023	1M	27%	13%
Zittrain NYT	2021	2013-2013	88K	40%	38%
ODU NYPW	2024	1996-2021	27.3M	65%	65%

on Tweets in their post, but that dataset was not shared with us due to the restrictions posed by the usage policies. Each URL had its categories, live status in the form of HTTP status code (including TCP or DNS failures), terminal URL and status code in case of redirects, and the year it was sampled from. The original dataset was stored in Parquet files, so performed some extractions and transformations to to it to suit our process. Then we checked URLs against the Wayback Machine to see if and when each of those were archived the first time and recorded this information in the dataset. In our study we analyzed this dataset in two forms: 1) *All* 5.4 million URLs together and 2) one million of *General* sample only.

We requested access to the dataset of about 2 million URLs from the Zittrain’s NYTimes outlinks study, but could not get it. Hence, we created our own dataset by downloading all the NYTimes pages published in 2013 that are present in the Wayback Machine, extracting all the outlinks from them, and excluding all the links to pages from NYTimes itself. We were able to collect about 88 thousand such URLs this way. Then we checked the live web status of each of the URLs (after following up to 5 redirects, if any). Then we checked for their presence in the Wayback Machine. One noticeable difference from the original dataset in that we only sampled data from one year while the original dataset included URLs from a span of a decade.

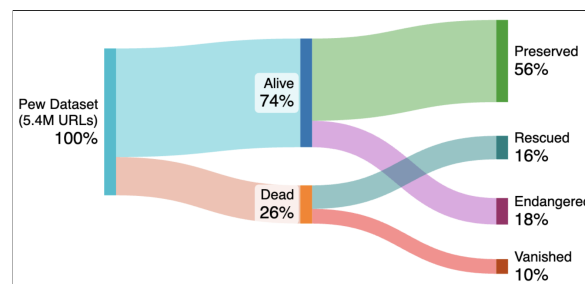
We reported findings of ODU’s “Not Your Parents’ Web” research here directly, without any further analysis, as the work already covered what would be useful here and we were collaborators in that work.

3 EVALUATION

Below, we look at four different samples/studies to see how much of those samples are present in the Wayback Machine.

3.1 Pew (All)

When we do not take any web archives into account, about a quarter of all the 5.4 million sampled URLs would be considered inaccessible or *Dead* as illustrated in Figure 1. However, **when we leverage the Wayback Machine to access otherwise dead URLs, the fraction of inaccessible or vanished URLs drops from one in every four down to only one in every ten.** The Wayback Machine has about 72% of the entire dataset *archived*, of which 56% are *preserved* from the URLs that are still *alive* on the live web and 16% are *rescued* from the *dead*. There are 18% of the URLs from the sample that are still *alive*, but have not been *archived* in the Wayback Machine yet, which we call *endangered*, as they may become *vanished* if they cease to exist on the live web ever. It is worth noting that we did not account for any captures of these URLs that might be present in any of the many smaller web

**Figure 1: Archiving status of all the URLs from the Pew dataset in the Wayback Machine.**

archives beyond the Wayback Machine, which if accounted for, might increase the percentage of the *accessible* URLs a little more. Moreover, we relied on HTTP status codes and did not look into the contents of the pages to check for any *soft-404s* [5] or other irrelevant content, which might change the numbers further.

3.2 Pew (General)

A subset of about 1 million URLs from the Pew dataset is a sample of general web pages from the last decade, spanning across 11 years from 2013 to 2023. They noted that about a quarter of the URLs from this subset were *dead* in 2023, with older URLs having a greater percentage of loss, all the way to 38% for links from 2013. We recreated their yearly graph in Figure 2 in orange color with an overlay of *rescued* URLs by the Wayback Machine in green color. We found that **about 38% of the 38% dead URLs from 2013 (i.e., about 15% of the total) are rescued by the Wayback Machine.** Moreover, about a quarter of the accumulative URLs of the general sample which were considered *dead*, about half of them were *rescued* by the Wayback Machine. It is worth noting that the last three years in Figure 2 seem to be rescued almost completely, but it is due to some data contamination as we have started ingesting CommonCrawl data from the recent years into the Wayback Machine, which happens to be the source of the sample of the Pew dataset.

3.3 Zittrain NYT

We then looked at the dataset we created from the archived pages of NYTimes. We found that **40% of the external links from NYTimes pages from 2013 were found dead on the live web, but 96% of the URLs are archived in the Wayback Machine.** This means, only about 2% URLs from this sample have *vanished*. However, this impressive number needs to be taken with a grain of salt because we do not have the original URL sample and our

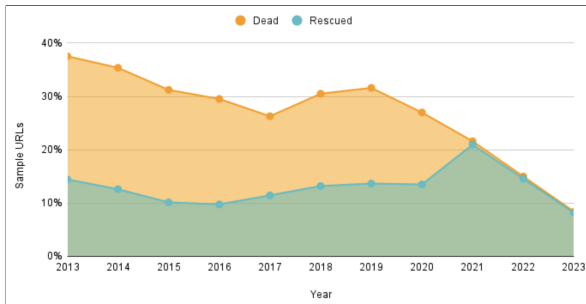


Figure 2: Yearly archiving status of URLs from the general sample of the Pew dataset in the Wayback Machine.

own sample is derived from pages present in the Wayback Machine, which has an inherent bias of outlinks from those pages being more likely to be archived than the outlinks of the pages that are not present in the Wayback Machine. That said, we will be keen to revisit these numbers if and when we get access to the original sample of URLs used in Zittrain’s study.

3.4 ODU NYPW

A recent, and perhaps the most comprehensive, longitudinal link-rot study from ODU, to which we are a collaborator, analyzed 27.3 million URLs sampled from the index of the Wayback Machine spanning over more than two and a half decades. They reported **about 65% of the sampled URLs from 1996 to 2021 were found dead in 2023**. A significant number of these samples were not even resolving the DNS, indicating that many of those domain names were not registered anymore. They found that most of the URLs die rapidly in the first few years of their existence, but some of the longest living sites are not *dead* yet. Luckily, **all of the dead URLs in this sample are rescued by the Wayback Machine** by the virtue of it being the source of the sample in the first place. This also means, the ODU study would not be able to tell the percentage of *vanished* URLs.

4 CONCLUSIONS AND FUTURE WORK

In summary, all of the link-rot studies, with varying numbers, indicate that the web is brittle and an increasing number of web resources die with the passage of time. However, we found that web archives like **the Wayback Machine play an increasingly important role in rescuing the dead web and minimizing the fracture of the knowledge graph on the web**, but there is a lot more to do. For example, **the Turn All References Blue (TARB) project has fixed more than 23 million broken links (and counting) on hundreds of wikis** with the help of the InternetArchiveBot⁶ and the Wayback Machine.

While there is not a lot that can be done to resurrect the *vanished* web other than attempting to find alternate locations where the content might have moved to (via projects like FABLE⁷), we are determined to minimize the percentage of the endangered URLs. However, there are some internal and external factors that limit our

⁶<https://meta.wikimedia.org/wiki/InternetArchiveBot>

⁷<https://webresearch.eecs.umich.edu/fable/>

ability to make it ZERO, such as, resource limitations, JavaScript-heavy pages, bot blocking, loginwalls, paywalls, deepweb, lack of timely discovery, etc. We strive to narrow down the potential loss of our cultural heritage via different means such as ingesting feeds from MediaCloud⁸, GDELT⁹, Wikipedia EventStream¹⁰, and more recently, becoming part of the IndexNow¹¹ initiative for link discovery soon after corresponding page creation or update on the web.

5 ACKNOWLEDGMENTS

We thank our friends at the Pew Research Center and the Old Dominion University, our colleagues Jake LaFountain and Stephen Balbach, and our intern Rachel Auslander for their help and support in this work.

REFERENCES

- [1] A. Chapekis, S. Bestvater, E. Remy, and G. Rivero, “When Online Content Disappears,” <https://www.pewresearch.org/data-labs/2024/05/17/when-online-content-disappears/>, 2024.
- [2] P. Stox, M. Pecánek, and J. Hardwick, “At Least 66.5% of Links to Sites in the Last 9 Years Are Dead,” <https://ahrefs.com/blog/link-rot-study/>, 2024.
- [3] J. L. Zittrain, “The Internet Is Rotting,” <https://www.theatlantic.com/technology/archive/2021/06/the-internet-is-a-collective-hallucination/619320/>, 2024.
- [4] M. C. Weigle, K. Garg, S. Alam, D. Ayala, and M. L. Nelson, “Some URLs Are Immortal, Most Are Ephemeral,” <https://ws-dl.blogspot.com/2024/09/2024-09-20-some-urls-are-immortal-most.html>, 2024.
- [5] L. Meneses, R. Furuta, and F. Shipman, “Identifying ‘Soft 404’ Error Pages: Analyzing the Lexical Signatures of Documents in Distributed Collections,” in *Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries*, ser. TPDL ’12, vol. 7489, 2012, pp. 197–208.

⁸<https://www.mediacloud.org/>

⁹<https://www.gdeltproject.org/>

¹⁰https://wikitech.wikimedia.org/wiki/Event_Platform/EventStreams_HTTP_Service

¹¹<https://www.indexnow.org/>

Medieval Citation Networks as Digital Hyperlinks: Transformer-Based Authorship Attribution in Historical Text Collections

Jonathan Schler*
Holon Institute of Technology (HIT)
Holon, Israel

Nati Ben-Gigi*
Binyamin Katzoff*
Maayan Geffet-Tamir*
Bar-Ilan University
Ramat-Gan, Israel

Abstract

Digital libraries containing historical manuscripts face persistent challenges in authorship attribution, particularly for anonymous or misattributed texts where traditional bibliographic metadata is incomplete or disputed. We demonstrate that citation networks in medieval texts function as primitive hyperlink structures, creating navigable knowledge graphs that encode stable authorial signatures across centuries-old document collections. Our transformer-based framework leverages three complementary components: (i) a BERT-CRF deep learning pipeline achieving accuracy of $F1 \approx 0.90$ in automatically extracting references from medieval Hebrew and Aramaic texts, (ii) cosine similarity analysis of citation frequency vectors that capture each author's unique "citation fingerprint," and (iii) network-based indicators quantifying cross-community influence patterns in historical corpora. Applied to a corpus of 62.5 million tokens spanning the rabbinic literature of the 10th-15th century, our system successfully extracted more than 230,000 references and constructed comprehensive citation networks. We validate the approach through a contested attribution case: commentary on Tractate Bava Metzia attributed to the medieval scholar "Ritva." Our analysis reveals distinct citation profiles between the attributed text and verified Ritva works (cosine similarity: 0.32), confirming scholarly suspicions of multiple authorship. The methodology identifies Rabbi Shem Tov ibn Gaon as the likely author of disputed sections (similarity: 0.959), corroborated by historical evidence. This work positions medieval citation practices as precursors to modern web hyperlink structures, demonstrating how transformer-based NLP can unlock authorship information embedded in historical reference networks. The language-agnostic methodology offers digital libraries scalable tools for automated manuscript attribution, applicable beyond medieval texts to any citation-rich historical corpus. Our approach bridges computational hypertext analysis with traditional humanities scholarship, providing new pathways for AI-enhanced organization and discovery in digital manuscript collections.

*All authors contributed equally to this research.

HT, Chicago, IL

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
<https://doi.org/XXXXXXX.XXXXXXX>

Keywords

digital hyperlinks, network analysis, authorship attribution, citation network, machine learning, text processing, historical networks

ACM Reference Format:

Jonathan Schler, Nati Ben-Gigi, Binyamin Katzoff, and Maayan Geffet-Tamir. 2025. Medieval Citation Networks as Digital Hyperlinks: Transformer-Based Authorship Attribution in Historical Text Collections. In *Proceedings of Web Archiving and Digital Libraries (WADL) Workshop 2025 (HT)*. ACM, New York, NY, USA, 5 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Long before Berners-Lee conceived the World Wide Web in 1989, medieval scholars had already established sophisticated hyperlinked knowledge networks through systematic citation practices [9]. These historical reference patterns, embedded in manuscripts spanning centuries, represent early examples of linked information structures—networks that connected ideas, authorities, and texts across vast temporal and geographical distances, prefiguring modern hypertext systems.

Digital libraries today confront significant challenges in organizing and attributing historical manuscript collections, particularly in cases involving anonymous or disputed authorship where traditional bibliographic metadata proves insufficient [3]. The problem becomes especially acute in specialized corpora such as medieval religious literature, where centuries of copying, compilation, and mis-attribution have obscured original authorship patterns. Recent scholarship has demonstrated that computational approaches can address these attribution challenges through novel methodologies that leverage citation patterns as stable authorial signatures [16, 23].

Contemporary advances in natural language processing and network analysis offer new pathways for addressing these attribution challenges. Citation networks in historical texts function as computational hyperlink structures, encoding stable authorial signatures that persist across an author's corpus [2]. Unlike traditional stylistic approaches that rely on linguistic features [11, 21], citation-based methods leverage the intellectual fingerprint embedded in how authors reference earlier authorities—patterns that remain remarkably consistent across different works and topics.

The methodological foundation for this approach builds upon established research in authorship attribution. Machine learning algorithms have been successfully applied to create writing profiles based on stylistic features such as word frequency, grammatical characteristics, and syntactic patterns [1, 10]. However, these methods face limitations when applied to historical texts, especially

in morphologically rich languages like Hebrew, where standard NLP resources are limited [22]. Citation-based approaches offer a complementary methodology that can overcome these linguistic barriers while providing domain-agnostic solutions.

Building on three complementary studies from our research program, this paper presents a transformer-based framework that treats medieval citation practices as precursors to modern web hyperlinks. Our approach integrates: (i) automatic reference extraction using a BERT-CRF pipeline trained on Rabbinic Hebrew achieving a F1 accuracy of 0.90 [3], (ii) cosine similarity analysis of citation frequency vectors for authorship verification, and (iii) network-based diversity indicators that quantify influence patterns across historical scholarly communities [2].

The methodology addresses a fundamental challenge in digital humanities: developing scalable computational approaches for authorship analysis in large historical text collections where traditional attribution methods fail. Applied to contested medieval commentaries, our citation fingerprint approach successfully identifies distinct authorial signatures, confirming scholarly hypotheses about disputed texts and proposing specific alternative attributions with quantifiable confidence measures.

This work contributes to the intersection of hypertext research and digital libraries by demonstrating how historical citation networks prefigure modern web structures. The language-agnostic methodology offers scalable tools for automated manuscript attribution, extending beyond medieval texts to any citation-rich historical corpus. By positioning citation analysis as a form of ancient hypertext navigation, we bridge computational approaches with traditional humanities scholarship, providing new frameworks for AI-enhanced organization and discovery in digital manuscript collections.

2 Related Work

Citation network analysis has emerged as a prominent methodology in digital humanities, enabling researchers to map intellectual relationships and trace knowledge transmission across historical corpora. Early work by [15] demonstrated the application of citation analysis to classical texts, utilizing natural language processing techniques to extract canonical citations and study intertextuality, while [4] constructed citation networks among recent monographs on Venetian history, revealing disciplinary clusters and identifying influential works. Recent advances by [8] survey comprehensive approaches to citation network analysis in digital humanities, highlighting the growing sophistication of computational methods for historical text analysis. In the context of Rabbinic literature specifically, [23] developed semi-automatic approaches to generate networks mapping relationships of Jewish sages across generations, utilizing lexical and syntactic patterns to identify names and relationships within Halachic debates. [16] applied quantitative social network analysis to the Babylonian Talmud, revealing dense core networks formed through relationships among influential rabbis and uncovering insights about historical connections and transmission pathways. These studies demonstrate the particular richness of citation networks in religious scholarly traditions, where systematic referencing practices create dense intellectual webs spanning centuries.

Computational authorship attribution represents a well-established field with methodologies ranging from traditional stylometric analysis to modern machine learning approaches. [11] provide a comprehensive survey of computational methods, highlighting the effectiveness of features such as word frequency, character n-grams, and syntactic patterns in distinguishing authorial signatures, while [6] extends this analysis to modern attribution methods, emphasizing the importance of function words and stylistic markers. However, traditional approaches face significant limitations when applied to historical texts, especially in morphologically rich languages like Hebrew, where standard NLP resources often prove insufficient [22]. Citation-based approaches offer complementary methodologies that can overcome these linguistic barriers. [7] proposed methods for analyzing citation patterns in academic papers to determine authorship, achieving success rates between 30-50% through citation profile matching and self-citation pattern identification, while [19] combined citation analysis with traditional stylometric methods, achieving approximately 60% success rates. In Rabbinic literature specifically, [5] developed supervised machine learning models to identify citations in Hebrew-Aramaic documents, and [12] leveraged these systems to analyze influence networks in Jewish Responsa literature. Recent transformer-based advances [18, 20] have enhanced performance for historical document processing, with [17] introducing BEREL for Rabbinic Hebrew analysis, demonstrating that citation patterns remain remarkably stable across an author’s corpus [2].

3 Methodology

Our approach to citation-based authorship attribution consists of three integrated components: (i) automated reference extraction from historical texts using transformer-based deep learning, (ii) construction of citation fingerprint vectors that capture authorial reference patterns, and (iii) similarity analysis using network-based measures to identify authorial signatures. This section details each component and demonstrates how they combine to create a scalable framework for manuscript attribution in digital libraries.

3.1 Automated Citation Extraction Pipeline

The foundation of our methodology relies on accurate extraction of citations from medieval Hebrew and Aramaic texts, a task complicated by the morphological richness of these languages and the non-standardized citation practices of historical authors [3]. We developed a multi-layered system that decomposes the complex reference extraction task into manageable subtasks.

Our corpus consists of medieval Rabbinic literature spanning the 10th-15th centuries, comprising over 62.5 million tokens from approximately 120 authors across six geographic regions, with pre-processing addressing orthographic inconsistencies, abbreviations, and morphological variations through a comprehensive thesaurus containing 240 authors’ names and 280 book titles [2].

3.1.1 BERT-CRF Architecture for Reference Identification. Following recent advances in transformer-based language models for historical texts, we employ a BERT-CRF architecture specifically adapted for Rabbinic Hebrew [17]. The model performs two sequential classification tasks that address the structural complexity of references in medieval texts.

The first task identifies reference boundaries within continuous text, addressing the challenge that references often appear in sequences without clear punctuation markers. This boundary detection model tags words that separate consecutive or recursive references, utilizing dedicated labels for different separation types. For recursive references—where one citation contains another citation—the model employs specialized tags to distinguish nested citation structures.

The second task performs component identification within detected reference boundaries, classifying words as author names, book names, reference terms like "chapter" or "page", and other citation elements. Training data consists of manually annotated references from a representative subset of the corpus: 3,301 references with 20,477 named entities for component identification, and 4,744 references with 23,184 entities for boundary detection [3].

The BERT-CRF models achieve strong performance with component identification attaining F1 score of 0.896 and boundary detection achieving F1 score of 0.856. When integrated into the complete pipeline including name normalization and validation steps, the system achieves overall precision of 0.896, recall of 0.905, and F1 score of 0.901 [3].

3.2 Citation Fingerprint Construction

The core innovation of our approach lies in treating citation patterns as stable authorial signatures that can be quantitatively compared across texts and authors. We formalize this concept through citation fingerprint vectors that capture the frequency distribution of an author's references to earlier authorities.

For each author in the corpus, we construct a citation vector \mathbf{v}_a where each dimension corresponds to a cited authority from the medieval period. Formally, if $A = \{a_1, a_2, \dots, a_n\}$ represents the set of all cited authorities in the corpus, then:

$$\mathbf{v}_a = [c_{a,1}, c_{a,2}, \dots, c_{a,n}]$$

where $c_{a,i}$ represents the number of times author a cites authority a_i . This representation captures both the diversity of sources an author draws upon and the relative frequency with which they reference different authorities.

A critical challenge involves accounting for significant variations in corpus size across authors. To address this issue, we employ cosine similarity for vector comparison, which measures the angle between vectors rather than their magnitude:

$$\text{similarity}(\mathbf{v}_a, \mathbf{v}_b) = \frac{\mathbf{v}_a \cdot \mathbf{v}_b}{|\mathbf{v}_a| \cdot |\mathbf{v}_b|}$$

This metric ranges from 0 to 1, where values close to 1 indicate similar citation patterns regardless of the absolute number of citations.

3.3 Network-Based Similarity Analysis

Building upon established methods in bibliometric analysis, we extend citation fingerprint comparison to network-level analysis that considers individual authorial patterns and community-wide citation structures [2].

Before applying citation fingerprints to attribution problems, we validated the stability of citation patterns within known authors'

corpora. For established authors with multiple works, we divided their complete works into random sections and computed similarity between the resulting citation profiles. Results demonstrate remarkable consistency: the Ramban's commentary sections achieve 0.99 similarity, while other major authors (Rashba, Ritva, Maharam Chalava) consistently score above 0.98 [2]. This validation confirms that citation patterns represent stable authorial characteristics that persist across different works and topics.

3.4 Authorship Attribution Protocol

The complete attribution methodology integrates citation extraction, fingerprint construction, and similarity analysis into a systematic protocol for evaluating disputed attributions. For a text of unknown or disputed authorship, we: (1) extract all citations using the BERT-CRF pipeline, (2) construct a citation fingerprint vector for the disputed text, (3) compare this vector against established citation profiles of candidate authors using cosine similarity, (4) rank potential attributions by similarity scores, and (5) validate results through content analysis and historical plausibility assessment. We require a minimum of 530 citations in disputed texts to ensure statistical reliability (10 times as many data-points as vector dimensions), with similarity scores above 0.85 indicating possible authorial correspondence and scores below 0.4 suggesting different authorship.

4 Case Study: The Ritva Attribution Problem

To demonstrate the effectiveness of our citation-based authorship attribution methodology, we apply it to a well-known disputed attribution in medieval Rabbinic literature: the commentary on Tractate Bava Metzia attributed to Rabbi Yom Tov of Seville (known by the acronym "Ritva"). This case study illustrates how computational analysis can resolve centuries-old scholarly debates while providing quantifiable evidence for attribution decisions.

4.1 Historical Context and Scholarly Debate

The commentary on Bava Metzia attributed to the Ritva presents a classic authorship attribution problem. Two distinct commentaries circulate under the Ritva's name: the *Hiddushei HaRitva* (accepted as authentic) and a commentary printed in Amsterdam in 1729 whose attribution has been questioned by scholars since the 18th century.

The disputed commentary was first partially printed in the *responsa* of Maharam Galanti (Venice, 1608) covering folios up to 12b, then published in its entirety in Amsterdam in 1729. However, prominent 18th-century scholars including Maharit Algazi and the Hida challenged this attribution based on: (i) discrepancies with Ritva quotations in the 16th-century compilation *Shitah Mekubetzet*, and (ii) differences in writing style and citation patterns compared to authenticated Ritva works.

Rabbi Halpern's analysis [14] proposed that the Amsterdam commentary comprises two distinct parts written by different authors: the first part (folios 1-11) by a student of the Rashba, possibly an earlier version of the Ritva's own commentary, and the second part (folios 12-end) by unknown author(s), possibly scholars from Provence. Rabbi Lichtenstein [13] subsequently argued that the

first part was authored by Rabbi Kreskes Vidal, also a student of the Rashba.

4.2 Computational Analysis

We applied our citation fingerprint methodology to test these scholarly hypotheses and propose alternative attributions based on quantitative evidence.

4.2.1 Baseline Citation Profile Construction. We first established a baseline citation profile for the authentic Ritva by analyzing all accepted works in our corpus: his commentaries on multiple Talmudic tractates, *Hilkhot Berakhot*, and responsa. This comprehensive profile, constructed from 847 distinct citations, represents the Ritva’s characteristic pattern of referencing earlier authorities. To validate the stability of this profile, we compared citation patterns across different authentic Ritva works. The average similarity between individual tractate commentaries and the complete Ritva profile was $\mu = 0.82$ ($\sigma = 0.11$), confirming consistent citation behavior across his authenticated corpus.

4.2.2 Attribution Testing of the Disputed Commentary. Initial analysis of the complete disputed commentary revealed a striking divergence from the authentic Ritva profile. The similarity score of 0.32 falls well below our threshold for positive attribution (0.85), strongly suggesting different authorship. This quantitative result supports the scholarly consensus that the Amsterdam commentary was not written by the Ritva.

4.2.3 Two-Part Analysis. Following Halpern’s hypothesis of composite authorship, we divided the disputed commentary at folio 12 and analyzed each section independently.

First Part Analysis (Folios 1-12): The citation profile of the first section yielded a similarity score of 0.87 when compared to the authentic Ritva works, indicating substantial correspondence. However, comparison with other contemporary authors revealed an even stronger match: Rabbi Kreskes Vidal achieved a similarity score of 0.91. While both scores fall within the range of positive attribution, the slightly stronger correspondence with Vidal supports Lichtenstein’s attribution hypothesis.

Second Part Analysis (Folios 12-end): The second section showed minimal similarity to the authentic Ritva (0.17), confirming that this portion was definitely not authored by him. Systematic comparison against all 120 authors in our corpus identified the most likely alternative attribution.

4.3 Novel Attribution Discovery

Our computational analysis identified Rabbi Shem Tov ibn Gaon as the most probable author of the second part of the disputed commentary, with a remarkable similarity score of 0.959. Rabbi Shem Tov ben Abraham ibn Gaon (c. 1250-1330) was born in Soria, northern Castile, studied under the Rashba, and authored *Migdal Oz*, one of the earliest commentaries on Maimonides’ *Mishneh Torah*. Significantly, he explicitly references his own lost Talmudic commentary called "*Shita*" in *Migdal Oz*, writing: "Many commentators have written on this, the first chapter of Bava Metzia, and we expanded upon it in its place in our *Shita*." This attribution had not been previously proposed in traditional scholarship and represents a novel contribution enabled by large-scale computational analysis.

Table 1: Citation Similarity Analysis for Disputed Commentary Section

Author	Similarity Score
R. Shem Tov ibn Gaon	0.959
Ra’avad the third	0.835
Rabbi Shmuel Hasardi	0.728
Authenticated Ritva works	0.170

To validate the computational attribution, we conducted comparative content analysis between the disputed commentary section and *Migdal Oz*, identifying several instances of parallel reasoning and similar interpretative approaches. One striking example appears in both works’ treatment of *Hilkhot Gezeila Va’Aveida* 17:11, where both texts present identical arguments, citing the same Talmudic sources and reaching identical conclusions about the given dispute. The attributed section contains approximately 800 citations, well above our minimum threshold of 530 for statistical reliability, with the extremely high similarity score (0.959) representing the strongest match in our entire corpus analysis.

5 Results and Discussion

5.1 Methodological Validation and Performance Assessment

Our citation-based authorship attribution framework demonstrates strong performance across multiple evaluation criteria, establishing its viability as a scalable tool for digital library applications. The technical components achieve consistently high accuracy with reference extraction attaining $F1 = 0.901$, while citation fingerprint analysis successfully resolves the contested Ritva attribution with quantifiable confidence measures. The stability of citation patterns across known authors’ corpora provides crucial validation, with consistency scores exceeding 0.98 for established authors (Ramban, Rashba, Ritva, Maharam Chalava) confirming that citation fingerprints represent stable authorial characteristics rather than random textual variation [2]. Cross-validation through content analysis strengthens confidence in computational results, as demonstrated by the identification of parallel reasoning between the attributed commentary section and Rabbi Shem Tov’s *Migdal Oz*, showing convergence between quantitative analysis and qualitative examination.

5.2 Comparison with Traditional Attribution Methods

Traditional approaches to manuscript attribution rely primarily on stylistic analysis, content examination, and historical documentation, but face significant scalability limitations when applied to large digital collections. Our computational approach offers several key advantages: citation fingerprint analysis provides quantifiable similarity measures enabling systematic comparison across large author sets, maintains consistency in evaluation criteria avoiding subjective variations, and enables analysis at scales that reveal subtle authorial signatures invisible to manual examination. However, computational methods complement rather than replace traditional scholarship, as domain expertise remains essential for interpreting

results, assessing historical plausibility, and understanding intellectual contexts that shape citation practices. The most effective attribution analysis combines computational efficiency with scholarly interpretation, as demonstrated in our validation of the Rabbi Shem Tov attribution through content analysis.

5.3 Applications to Digital Libraries and Hypertext Research

The methodology’s language-agnostic design and reliance on citation patterns rather than linguistic features enable application beyond medieval Hebrew literature to any citation-rich historical corpus, with potential applications including academic paper attribution for resolving authorship disputes, automated manuscript cataloging, plagiarism detection through unusual citation patterns, and web archive analysis to track intellectual influence patterns in digital scholarship. Our positioning of medieval citation networks as ancient hyperlink structures opens new avenues for hypertext research by demonstrating how both systems create navigable knowledge structures where references serve as pathways between related information sources—medieval scholars relied on citation networks to discover relevant authorities and trace intellectual lineages just as modern web users navigate hyperlinked content. Digital libraries can leverage these insights by treating citation networks as primitive knowledge graphs that encode semantic relationships between texts, authors, and ideas, with visualization tools mapping citation networks geographically and temporally to reveal intellectual transmission patterns and scholarly influence networks previously hidden in traditional library catalogs, thus bridging historical scholarship practices with modern information organization methods.

6 Conclusion

This paper demonstrates how medieval citation networks function as ancient hyperlink structures, encoding stable authorial signatures that enable computational authorship attribution in historical digital libraries. Our transformer-based methodology successfully resolves a centuries-old attribution debate, identifying Rabbi Shem Tov ibn Gaon as the previously unknown author of disputed commentary sections with 95.9% similarity confidence, illustrating how AI and NLP techniques can unlock authorship information embedded in historical reference networks.

The technical contributions establish a scalable framework for citation-based authorship analysis, with the BERT-CRF pipeline achieving $F1 \approx 0.90$ accuracy in medieval Hebrew reference extraction and citation fingerprint methodology offering digital libraries language-agnostic tools for manuscript attribution. The methodology’s applicability extends beyond medieval texts to any citation-rich corpus, including modern academic literature, web archives, and collaborative scholarly environments, with the proposed attribution satisfying historical plausibility criteria including temporal alignment, intellectual tradition, and geographic feasibility.

By positioning citation networks as fundamental hypertext structures, this work reveals how scholarly communities created navigable knowledge graphs long before digital technologies emerged, suggesting that linked information systems represent persistent human approaches to knowledge organization. The convergence

of historical scholarship and computational analysis exemplified here demonstrates the value of interdisciplinary collaboration in addressing complex problems in digital cultural heritage, providing new pathways for AI-enhanced organization and discovery in digital manuscript collections.

References

- [1] Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM* 52, 2 (2009), 119–123.
- [2] Nati Ben-Gigi, Maayan Zhitomirsky-Geffet, Binyamin Katzoff, and Jonathan Schler. 2024. Citation network analysis for viewpoint plurality assessment of historical corpora: The case of the medieval rabbinic literature. *PLOS ONE* 19, 7 (2024), e0307115.
- [3] Nati Ben-Gigi, Maayan Zhitomirsky-Geffet, Jonathan Schler, and Binyamin Katzoff. 2024. Automatic Construction of the Citation Network from the Medieval Jewish Responsa Literature. *ACM Journal on Computing and Cultural Heritage* 18, 2 (2024), 1–18.
- [4] Giovanni Colavizza. 2017. The core literature of the historians of Venice. *Frontiers in Digital Humanities* 4 (2017), 14.
- [5] Yaakov HaCohen-Kerner, Nadav Schweitzer, and Dror Mughaz. 2011. Automatically identifying citations in Hebrew-Aramaic documents. *Cybernetics and Systems: An International Journal* 42, 3 (2011), 180–197.
- [6] Xie He, Arash Habibi Lashkari, Nikhill Vombatkere, and Dilli Prasad Sharma. 2024. Authorship attribution methods, challenges, and future research directions: A comprehensive survey. *Information* 15, 3 (2024), 131.
- [7] Shawndra Hill and Foster Provost. 2003. The myth of the double-blind review? Author identification using only citations. *ACM SigKDD Explorations Newsletter* 5, 2 (2003), 179–184.
- [8] Sehrish Iqbal, Saeed-UI Hassan, Naif Radi Aljohani, Salem Alelyani, Raheel Nawaz, and Lutz Bornmann. 2021. A decade of in-text citation analysis based on natural language processing and machine learning techniques: an overview of empirical studies. *Scientometrics* 126, 8 (2021), 6551–6599.
- [9] Michael Kelly and K Patrick Fazioli. 2023. *social and intellectual networking in the early middle ages*. punctum books.
- [10] Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*. ACM, 489–495.
- [11] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology* 60, 1 (2009), 9–26.
- [12] Moshe Koppel and Nadav Schweitzer. 2014. Measuring direct and indirect authorial influence in historical corpora. *Journal of the Association for Information Science and Technology* 65, 10 (2014), 2138–2144.
- [13] Eliyahu Lichtenstein. 1980. *Mavo le-idushei HaRitva al Masechet Gittin*. Mossad Harav Kook, Jerusalem. Introduction to the Ritva’s Novellae on Tractate Gittin.
- [14] Ritva. 1962. *Hiddushei Ritva, Baba Metzia*. Private publication, London. Edited from manuscripts and published for the first time.
- [15] Matteo Romanello. 2016. Exploring citation networks to study intertextuality in classics. *DHQ: Digital Humanities Quarterly* 10, 2 (2016).
- [16] Michael L Satlow and Michael Sperling. 2022. The Rabbinic citation network. *AJS Review: The Journal of the Association for Jewish Studies* 46, 2 (2022), 291–319.
- [17] Avi Shmidman, Joshua Guedalia, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Eli Handel, and Moshe Koppel. 2022. Introducing BEREL: BERT embeddings for Rabbinic-encoded language. *arXiv preprint arXiv:2208.01875* (2022).
- [18] Pranaydeep Singh, Gorik Rutten, and Els Lefever. 2021. A pilot study for BERT language modelling and morphological analysis for ancient and medieval Greek. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. 128–137.
- [19] Richard Snodgrass. 2006. Single-versus double-blind reviewing: An analysis of the literature. *ACM Sigmod Record* 35, 3 (2006), 8–21.
- [20] Thea Sommerschild, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando De Freitas. 2023. Machine learning for ancient languages: A survey. *Computational Linguistics* 49, 3 (2023), 703–747.
- [21] Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60, 3 (2009), 538–556.
- [22] Omri Suissa, Avshalom Elmalech, and Maayan Zhitomirsky-Geffet. 2022. Text analysis using deep neural networks in digital humanities and information science. *Journal of the Association for Information Science and Technology* 73, 2 (2022), 268–287.
- [23] Maayan Zhitomirsky-Geffet and Gila Prebor. 2019. SageBook: Toward a cross-generational social network for the Jewish sages’ prosopography. *Digital Scholarship in the Humanities* 34, 3 (2019), 676–695.