Summer 2024

Surfacing Text Changes in Archived Webpages

Lesley Frew
Old Dominion University, lfrew001@odu.edu

SURFACING TEXT CHANGES IN ARCHIVED WEBPAGES

by

Lesley Frew
B.A. May 2011, University of Virginia
M.T. May 2011, University of Virginia

A Thesis Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

COMPUTER SCIENCE

OLD DOMINION UNIVERSITY
August 2024

Approved by:

Michele C. Weigle (Director)

Michael L. Nelson (Member)

Sampath Jayarathna (Member)

ABSTRACT

SURFACING TEXT CHANGES IN ARCHIVED WEBPAGES

Lesley Frew
Old Dominion University, 2024
Director: Dr. Michele C. Weigle

Webpages change over time, and web archives hold copies of historical versions of webpages. Users of web archives, such as journalists, want to find and view changes on webpages over time. However, the current search interfaces for web archives do not adequately support this task. For the web archives that include a full-text search feature, multiple versions of the same webpage that match the search query are shown individually without enumerating changes, or are grouped together in a way that hides changes. We present a change text search engine that allows users to find changes in webpages. We describe the implementation of the search engine backend and frontend, including a tool that allows users to view the changes between two webpage versions in context as an animation. We also propose changes to the Internet Archive's Wayback Machine replay navigation banner to further support users viewing change over time. We evaluate the search engine with U.S. federal environmental webpages that changed between 2016 and 2020. The change text search results page can clearly show when terms and phrases were added or removed from webpages. The inverted index can also be queried to identify salient and frequently deleted terms in a corpus. We align the dataset to with a real-world click dataset, showing that users were searching for the same environmental terms that were ultimately deleted.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Michele C. Weigle, for believing in me and for seeing my potential as a researcher. I also want to thank her for being my coach and collaborator by continuously pushing me to my intellectual limits. I also want to thank her for always thinking about me and helping me, as well as for her invaluable intellectual contributions to this research. I'm looking forward to our continued collaboration. I would also like to thank my co-advisor, Dr. Michael L. Nelson, for always being honest with me, for his support during challenging situations, and for welcoming me into the Web Science and Digital Libraries Research Group. I would also like to thank him for our repartee and bouncing ideas off of each other. With all three of us together, our ideas are woven together throughout this thesis, and the sum is thus greater than our individual parts. I would also like to thank Dr. Sampath Jayarathna for his continuous support throughout my master's degree and for serving as a committee member.

I would also like to thank the Web Science and Digital Libraries Research Group for their support and friendship. Of course I thank my family for their support. I would also like to thank the McLean High School Math Department as well as my principal, the math and computer science Fairfax County Public Schools central office staff, and the Northern Virginia Community College dual enrollment office for their support and encouragement. Finally, I'm grateful to the funding agencies that supported me during my master's degree for tuition (Fairfax County Public Schools and the Virginia Department of Education's program for dual enrollment teachers), conference attendance (Old Dominion University and the Association for Computing Machinery) and conference awards (University of Illinois Urbana-Champaign School of Information Sciences).

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

The current ways in which we interact with the Web, including our browsers and popular search engines, predispose us to only see the most recent versions of webpages. Some webpages update their content frequently, while others remain relatively static. Web archives preserve these historical versions of webpages. Different web archives have different holdings, and aggregating the holdings increases the number of webpage versions available for viewing.

Journalists frequently use web archives as evidence.[1] These journalists reference archived copies of webpages not only when the pages are unavailable, but also when the pages have changed over time [51]. Journalists are just one type of user aiming to view change over time via web archives. This thesis explores improvements to web archive information retrieval user interfaces that better enable users to find and view changes over time in web archives.

1.1 PROBLEM

As pointed out by Teevan [132], simply saving past versions of webpages without providing proper means for discovery is insufficient to help people find the historical content changes that they are seeking.

One example of real webpages that have changed over time is US federal environmental webpages between presidential terms. The public relies on the government to provide access to unbiased and trustworthy information. However, between 2016 and 2020, substantial edits and dele-

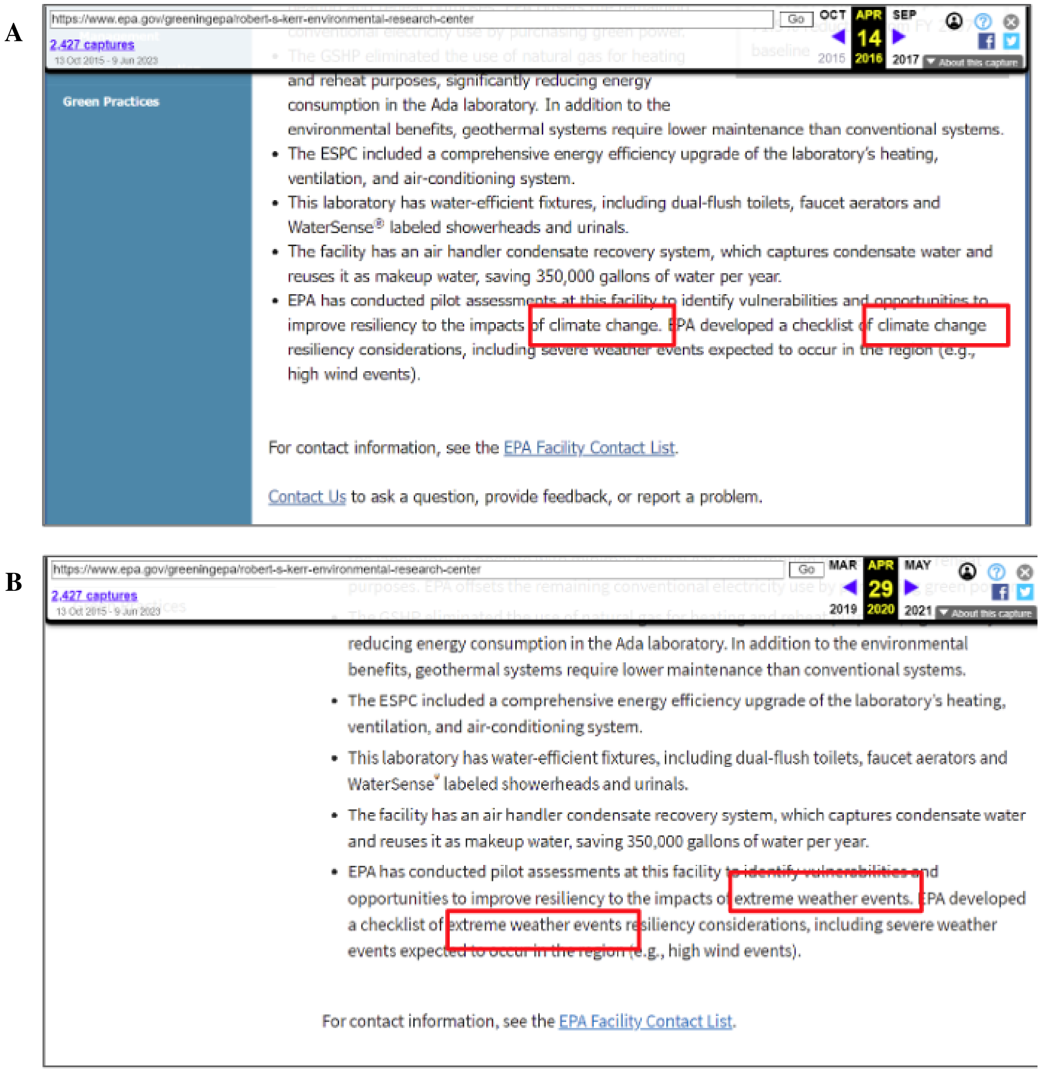
[1] `https://archive.org/about/news-stories/search?mentions-search=Wayback+Machine`

tions were made to US federal environmental website content and entire pages were deleted. For example, on the EPA Kerr Research Center page, every instance of "climate change" was replaced with "extreme weather events".[2] Figure 1 shows these changes. The Environmental Data and Governance Initiative (EDGI) tracked these changes to increase public awareness and to strengthen government accountability due to the lack of digital legal deposit laws in the United States [110].

However, none of the current web archive search interfaces support searching for these known deleted terms. Current search functionality in web archives is limited, ranging from no search capabilities at all, to metadata search only, to full text search over only some collections [57]. None of the current interfaces allow the user to query for webpage changes. Yet, we know that webpages change over time, and we present evidence from two formative investigations showing that real users want to use web archives to view change over time.

Another open problem in web archives information retrieval is how to present multiple captures of the same webpage that match the query term. Some web archives only include one version, which reduces clutter, while others include all versions, increasing clutter but giving users access to all matching versions [70]. Figure 2 shows grouped and non-grouped web archive search results. The non-grouped page is cluttered by identical results, while the grouped version does not indicate how the page has changed over time with respect to the query term. Grouping search results by change terms shows multiple versions of a page as one result in a meaningful way.

[2] `https://web.archive.org/web/diff/20160414222520/20200429003612/https://www.epa.gov/greeningepa/robert-s-kerr-environmental-research-center`



A

and reheat purposes, significantly reducing energy consumption in the Ada laboratory. In addition to the environmental benefits, geothermal systems require lower maintenance than conventional systems.

- The ESPC included a comprehensive energy efficiency upgrade of the laboratory's heating, ventilation, and air-conditioning system.
- This laboratory has water-efficient fixtures, including dual-flush toilets, faucet aerators and WaterSense® labeled showerheads and urinals.
- The facility has an air handler condensate recovery system, which captures condensate water and reuses it as makeup water, saving 350,000 gallons of water per year.
- EPA has conducted pilot assessments at this facility to identify vulnerabilities and opportunities to improve resiliency to the impacts of climate change. EPA developed a checklist of climate change resiliency considerations, including severe weather events expected to occur in the region (e.g., high wind events).

For contact information, see the EPA Facility Contact List.

Contact Us to ask a question, provide feedback, or report a problem.

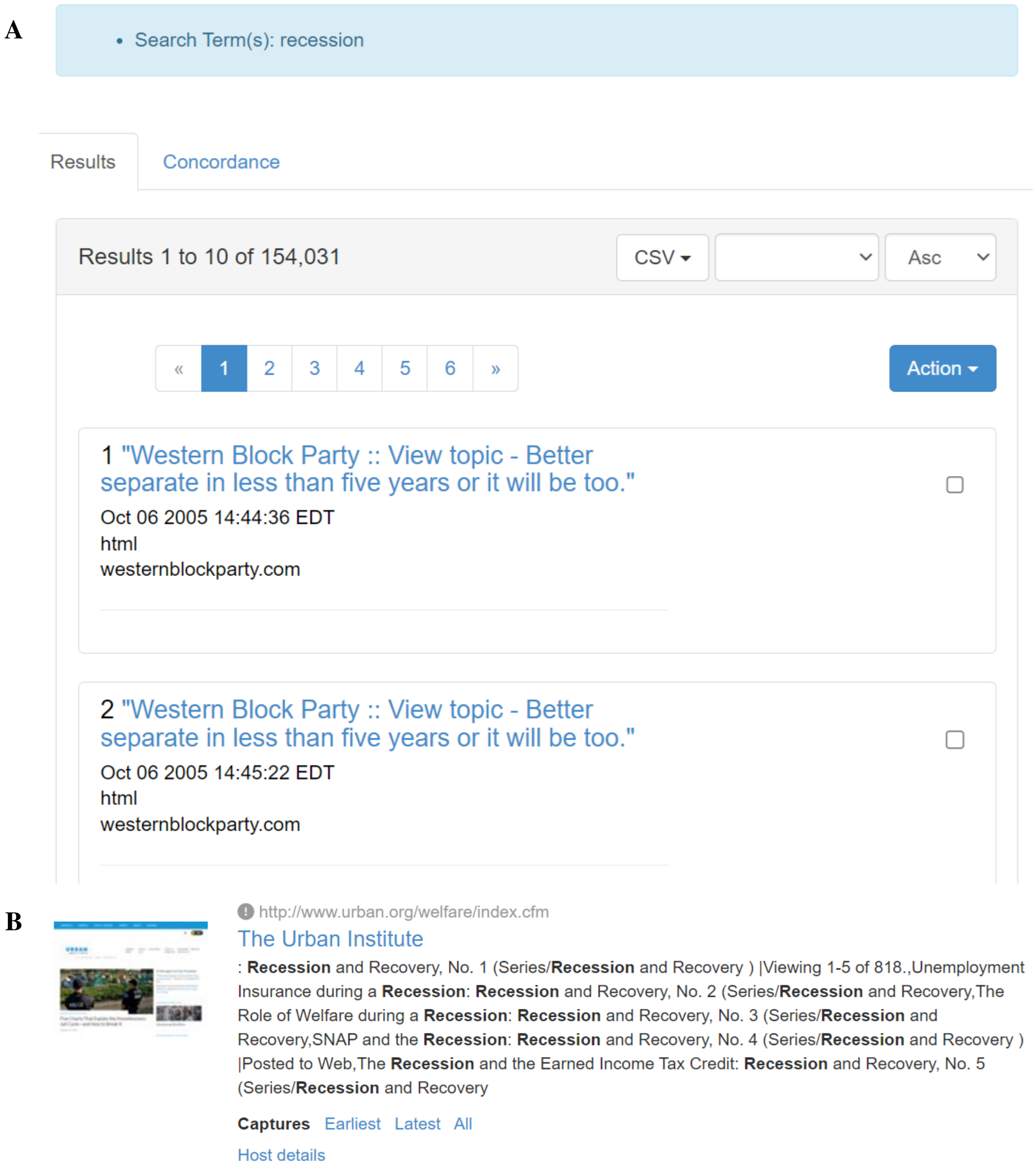
B

reducing energy consumption in the Ada laboratory. In addition to the environmental benefits, geothermal systems require lower maintenance than conventional systems.

- The ESPC included a comprehensive energy efficiency upgrade of the laboratory's heating, ventilation, and air-conditioning system.
- This laboratory has water-efficient fixtures, including dual-flush toilets, faucet aerators and WaterSense® labeled showerheads and urinals.
- The facility has an air handler condensate recovery system, which captures condensate water and reuses it as makeup water, saving 350,000 gallons of water per year.
- EPA has conducted pilot assessments at this facility to identify vulnerabilities and opportunities to improve resiliency to the impacts of extreme weather events. EPA developed a checklist of extreme weather events resiliency considerations, including severe weather events expected to occur in the region (e.g., high wind events).

For contact information, see the EPA Facility Contact List.

Figure 1. Changes on US federal sites between 2016 and 2020 showing removal of the phrase "climate change." *(A)* The 2016 version of the Kerr Research Center page uses the phrase "climate change." *(B)* The 2020 version of the Kerr Research Center page uses the phrase "extreme weather events."



- Search Term(s): recession

Results 1 to 10 of 154,031

1 "Western Block Party :: View topic - Better separate in less than five years or it will be too." □
Oct 06 2005 14:44:36 EDT
html
westernblockparty.com

2 "Western Block Party :: View topic - Better separate in less than five years or it will be too." □
Oct 06 2005 14:45:22 EDT
html
westernblockparty.com

http://www.urban.org/welfare/index.cfm
The Urban Institute
: **Recession** and Recovery, No. 1 (Series/**Recession** and Recovery) |Viewing 1-5 of 818.,Unemployment Insurance during a **Recession**: **Recession** and Recovery, No. 2 (Series/**Recession** and Recovery,The Role of Welfare during a **Recession**: **Recession** and Recovery, No. 3 (Series/**Recession** and Recovery,SNAP and the **Recession**: **Recession** and Recovery, No. 4 (Series/**Recession** and Recovery) |Posted to Web,The **Recession** and the Earned Income Tax Credit: **Recession** and Recovery, No. 5 (Series/**Recession** and Recovery

Figure 2. Examples of web archive search results with and without grouping for the term "recession." *(A)* Webarchives.ca ungrouped search results, where the top two results are the same page with captures differing by one minute. *(B)* Internet Archive Wayback Machine collection search, with only one result per page. The main version of the page linked includes the query term 28 times. The *earliest* capture linked does not include the query term.

To summarize the problem, users want to view how webpages have changed over time using web archives. However, the current search interfaces do not allow for users to search for terms that have changed over time, and the presentations of the results is ineffective for this purpose.

In this thesis, we address three research questions:

1. How can we make changes in webpages discoverable and understandable?
2. How can we increase efficiency in web archive user navigation for viewing change over time?
3. How can aggregated webpage changes of a corpus be used computationally to provide compelling evidence for edit intentions?

Research Question 1 and Research Question 2 are addressed in Chapter 5 through the creation of a change-text search interface for web archives. Research Question 3 is addressed in Chapter 6 through evaluating changes on US federal environmental sites between 2016 and 2020.

1.2 CONTRIBUTIONS

This thesis makes the following important contributions to the area of information seeking behavior in web archiving:

- This thesis shows that users, such as journalists, use web archives to view change over time.
- Next, this thesis presents a backend to support a change text search index using Lucene.
- Third, this thesis presents a frontend for a change text search interface including a search engine results page that grouped multiple versions of archived webpages by their changed

terms, a major contribution towards the open problem of how to present multiple versions of the same archived webpage in search results.

- Fourth, this thesis presents two difference visualizations in the change text search interface, the sliding difference viewer and the change text animation, to assist the user in analyzing changes on webpages.
- Fifth, this thesis shows that users who view more than one version of a webpage in web archives need additional navigational tools to help them accomplish the task of viewing changes over time.
- Sixth, this thesis presents a new navigational banner for replay in web archives that supports users in viewing changes on webpages over time.
- Lastly, this thesis offers evidence in support of the effectiveness of the change text search index by evaluating a dataset of federal environmental webpages, showing environmental terms deleted over time and matching these terms with user queries.

1.3 THESIS ORGANIZATION

This thesis consists of seven chapters. This first chapter presents the overarching motivation for the thesis work and approaches taken to address the problem. Chapter 2 surveys how humans seek information, how the live web and past web function, how search engines function, and what datasets exist that contain changes on webpages over time. Chapter 3 focuses on why users want to make and view webpage changes, how users behave in web archives and searches, and how this influences web archive search implementations, how users view change and how this influences tools to view change, and then discusses ethical implications for creating a change-text search interface. Chapter 4 presents two formative investigations about why and how users try to view change in web archives. These analyses lead to proposed interfaces that improve information behavior flow and access in web archives, which are presented in Chapter 5. The first proposed system to address Research Question 1 is a backend search index that enables users to search for changes, along with a frontend search interface and two tools to view change over time: a sliding diff tool and a change animation tool. The second proposed interface to address Research Question 2 improves the functionality of the replay navigation banner for users trying to view changes over time, and also improves how archived redirect pages are presented to these users. Chapter 6 addresses Research Question 3 by presenting two evaluations of the change text search index using federal environmental webpages. The first evaluation validates existing research about deleted environmental terms, and introduces new common deleted terms throughout the corpus. The second evaluation aligns the dataset with real user queries for those pages in the same timespan, and shows that user were searching for the same the environmental terms that were ultimately deleted from the pages. Chapter 7 outlines opportunities for future work and summarizes the thesis.

CHAPTER 2

BACKGROUND

This chapter is devoted to explaining essential information needed to understand the contributions of the thesis. First, we examine how human cognition influences information seeking in order to inform our change text search interface design. Next, we describe how pages change on the Web in order to determine what we need to measure and calculate. Next, we examine how web archives function in order to utilize them properly and efficiently in our change text search system, since they contain the past versions of websites with changes. Next, we describe how search architecture functions, and special considerations for web archive search systems, in order to implement these special considerations for our system. Finally, we examine what datasets of webpages with changes would be appropriate to use to evaluate the change text search system.

2.1 HUMAN COGNITION AND ELECTRONIC INFORMATION SEEKING

Facilitating search, or information seeking, in web archives is a primary contribution of this thesis. Before we cover details of how search is, or is not, enabled in web archives, we must understand how human cognition, or understanding, both affects and supports information seeking behaviors. This will inform our search interface design.

In this section, we summarize the insights from Marchionini [103] on human cognition and electronic information seeking. Users seek information in order to learn. This process may be modeled by identifying the unknown, formulating questions to search for the answer to this unknown, and then the answer to the original question leads to new unknowns. Human cognition has a significant effect on how information seeking systems should be designed, so that users can accomplish their search tasks.

A user's personal information infrastructure influences how they behave during search. One aspect of this infrastructure is any prior mental models they have of information retrieval systems. Another aspect is how humans model knowledge. Many cognitive skills contribute to the personal information infrastructure, including inference skills as well as organizational skills. Additional meta-cognitive skills also contribute, such as planning. Finally, a user's attitudes towards information seeking affect their personal information infrastructure as well. All of these factors will lead a user to develop a model for the information seeking system. The model may resemble a physical place where the user conducts transactions, or it may resemble another electronic system.

Human memory capabilities have a strong influence on search. Humans can remember a smaller amount of information in their working memory than their long term memory. Information systems that employ too much novelty will exhaust their users' working memory with concerns about how to operate the system rather than keeping the memory available for focusing on the knowledge in the search task.

Another hindrance to electronic information seeking is aligning how humans and computers intercommunicate. Humans can communicate implicitly and can make assumptions about intent, while computers require explicit commands. Humans also wish to complete open search tasks, but these open search tasks must be solved by communicating a series of closed tasks to the computer.

Users with different information seeking needs behave differently when searching. These users' needs necessitate different tasks, which have different cognitive requirements. How well an information retrieval system implements supports for each type of task necessarily dictates how successful a user will be in succeeding in their task. An ideal information seeking system will be developed iteratively to support users in tasks they do not even know they wish to solve yet.

When accomplishing their tasks, users wish to stay on task. Therefore, they want to be presented with results that match their tasks. This idea is called *precision*. Conversely, users want to feel that they have not missed any possible results that could inform their knowledge. This idea is called *recall*. In search systems, precision and recall may not both be optimally possible, and one may increase at the expense of the other.

Taking advantage of human cognition functionality like subconscious pre-attentive processing and familiarity will help users accomplish their goals. Search systems can aid users with preprocessing by highlighting in search results. Enabling native commands to the operating system will allow users to focus their cognitive load on the knowledge of their search task. Another way to support users with functioning in an information seeking environment is to provide easily accessible and prominent help tutorials.

When querying, systems should allow the user to query but also limit them to what is possible to query for. The user should also be able to change the granularity of the collection to refine their search. Users need to be presented with a system that supports them in making a model of across document browsing. When presenting the search results, the main display should help the user easily determine the aboutness of the document.

One additional important cognitive concept that influences information seeking is information forgaging. Information Foraging [118] is a theory for how people find information. Because people have a finite amount of time to accomplish tasks, such as information seeking, they subconsciously or consciously evaluate the utility of given actions compared to their effect on their ultimate goal. People assess relevance via information scent, including metadata of documents. In addition to deciding the relevance of any given document, information first needs to be sorted into clusters, and general information gathering strategies about where to look in the first place are also core tenets of the information foraging framework

2.2 THE EPHEMERAL LIVE WEB

Facilitating finding changes on webpages is a primary contribution of this thesis. In order to inform our change text search interface design, we must understand the types of changes that occur on the Web.

Web pages change over time. The phenomenon of a page slowly changing its aboutness over time is called content drift [81]. Pages are also sometimes entirely deleted from the web. When other pages link to these deleted webpages, it is known as link rot [81].

Web pages are identified by their Uniform Resource Identifier (URI) [22] and are located by their Uniform Resource Locator (URL) [23]. URLs can change over time. Thus, it is possible for both a page's content and its URL to change over time. When a page's URL changes but the original server maintains a directive with the new location, it is called a redirect.

Users access pages on the live web via a web browser using Hypertext Transfer Protocol (HTTP) In addition to transmitting the page content if present, HTTP also transmits a status code to reflect the completion status of the request. These status codes are defined in RFC 9110 [48]. Successful requests have a 200 OK HTTP status. Pages that redirect have a 3xx status code. In particular, 301 means that the resource has moved permanently and the old URL should no longer be used, while a 302 indicates the resource is found and moved temporarily, and the old URL could still be used. A third 3xx status is 300 Multiple Choice. This status would indicate that the user could choose manually, or the browser could choose automatically, one of multiple representations for the requested resource. When the HTTP request cannot be fulfilled, a 4xx status code for a client-side error is returned. Pages that are not found, for instance, if they have been deleted, typically have a 404 status code. Incorrect authentication results in a 401 or 403 status code. Another class of status codes is 5xx for a server side error. This indicates that the request to the server was valid, but there is something preventing the server from completing the request. A 500 Internal Server Error code is the most generic, while more specific server problems have their own codes. Sometimes, there is a mismatch between the codes and the content. For example, a soft 404 [20] indicates that the server is returning a 200 status, but the semantics encoded in the human-readable HTML indicate a 404 would have been appropriate.

2.3 THE PAST WEB

This thesis's primary contribution is surfacing text changes on webpages. Web archives are the primary repository of past versions of webpages. In order to successfully implement a change-text search backend, we must understand how web archives function.

While the navigable collection of all webpages' most recent version is referred to as the Live Web, past versions of webpages make up the Past Web. These past versions of webpages may be saved in web archives. Web archives contain a vast amount of untapped potential for analyzing change over time.

2.3.1 Web Archive Framework

There are multiple web archives [2], [18], [46], including national web archives like the Portuguese Web Archive,[1] subscription-based web archives for organizations like Archive-It,[2] library

[1] https://arquivo.pt/

[2] https://archive-it.org/

web archives like the Library of Congress Web Archive,[3] and more comprehensive web archives like the Internet Archive's Wayback Machine.[4] Web archiving initiatives go back at least twenty years [39].

The Memento Protocol [82], [106], [136] is the standard HTTP content negotiation protocol for web archives. It links the live web and the past web. The original resource as it would have been requested on the live web is identified by a URI-R. In order to request a page in a web archive, users must include both the URI-R and the desired past datetime in their request. Memento-compliant web archive servers will query for and return the archived version of the webpage closest to the requested datetime. Listing 2.1 shows an HTTP request for the version of epa.gov/acidrain using a TimeGate at the Wayback Machine closest to March 23, 2016 using the Accept-Datetime header along with the HTTP response for this request, with the closest memento being on March 22.

Listing 2.1. HTTP request and response for epa.gov/acidrain when server is Memento-compliant

```
curl -I -v -H "Accept-Datetime:  Tue, 23 Mar 2016 00:00:00 GMT"
http://web.archive.org/web/http://epa.gov/acidrain/


HTTP/1.1 302 FOUND
Date: Sun, 11 Aug 2024 17:22:48 GMT
x-archive-redirect-reason: found capture at 20160322000840
location: http://web.archive.org/web/20160322000840/https://www3.epa.gov/acidrain/


HTTP/1.1 200 OK
```

[3] https://webarchive.loc.gov/

[4] https://web.archive.org/

```
Date: Sun, 11 Aug 2024 17:22:48 GMT
memento-datetime:  Tue, 22 Mar 2016 00:08:40 GMT
link: <https://www3.epa.gov/acidrain/>; rel="original",
<http://web.archive.org/web/timemap/link/https://www3.epa.gov/acidrain/>;
rel="timemap"; type="application/link-format",
<http://web.archive.org/web/https://www3.epa.gov/acidrain/>;
rel="timegate",
<http://web.archive.org/web/19970420085456/http://www.epa.gov:80/acidrain/>;
rel="first memento"; datetime="Sun, 20 Apr 1997 08:54:56 GMT",
<http://web.archive.org/web/20160321235530/https://www3.epa.gov/acidrain>;
rel="prev memento"; datetime="Mon, 21 Mar 2016 23:55:30 GMT",
<http://web.archive.org/web/20160322000840/https://www3.epa.gov/acidrain/>;
rel="memento"; datetime="Tue, 22 Mar 2016 00:08:40 GMT",
<http://web.archive.org/web/20160329221858/http://www.epa.gov/acidrain>;
rel="next memento"; datetime="Tue, 29 Mar 2016 22:18:58 GMT",
<http://web.archive.org/web/20240809051617/https://www.epa.gov/acidrain>;
rel="last memento"; datetime="Fri, 09 Aug 2024 05:16:17 GMT"
```

Each archived version of a webpage can be accessed directly with its own URI, known as a "memento," "capture," or "snapshot." It is identified by its URI-M. For example, the URI-R for the Environmental Protection Agency Laws & Regulations page is `https://www.epa.gov/laws-regulations`. Each web archive determines the structure of URI-Ms. Three URI-Ms for the EPA Laws & Regulations page are shown in Table 1. At some web archives, like the Internet Archive and Archive-It, the URI-R and the datetime are transparent in the URI-M. At other web archives, like Archive Today, the URI-M is more opaque and shorter.

Table 1. URI-M structure for https://www.epa.gov/laws-regulations varies across multiple web archives. In some web archives the URI-R and datetime are transparent in the URI-M while in others they are opaque.

Archive	URI-M	Date
Internet Archive	https://web.archive.org/web/20160619022359/ https://www.epa.gov/laws-regulations	2016-06-09
Archive-It	https://wayback.archive-it.org/4638/20160611152955/ https://www.epa.gov/laws-regulations	2016-06-11
Archive Today	https://archive.is/xSZ7X	2017-01-24

A listing of all archived versions of a webpage on a specific server can be queried. This listing is called a *TimeMap*, and is identified by a URI-T. An example TimeMap at the Internet Archive Wayback Machine for the EPA webpage is `http://web.archive.org/web/timemap/link/http://www.epa.gov`. TimeMaps can be represented in various formats, such as Link, JSON, or HTML.

In addition to the Memento Protocol which provides an API specifying how to get a listing of captures for a URI, the Internet Archive also provides an API for its crawl indices (CDX).[5] Memento does not depend on CDX deployment, but the CDX has additional information not present in a TimeMap. The CDX API takes a URL as input, and outputs a listing that includes all captures,

[5] `https://github.com/internetarchive/wayback/tree/master/wayback-cdx-server`

their crawl timestamps, and the status code of the page at the time of capture. It is important to note that crawl time is different from page edit date [105], [107]. If a page has a low rate of change, it might very well be crawled many times in between changes. Alternatively, a page that is being archived at a slower rate than its change rate will not have all of its versions archived.

Multiple similar URLs can resolve to the same webpage. Since the URL is the key used for lookup in a web archive, a technique called Sort-friendly URI Reordering Transform [126], or SURT, is used to match these multiple versions of a URL. SURT can be used independent of web archives as well, for example, to standardize datasets of URLs for comparing to each other. For example, *https://www2.epa.gov/*, *http://www.epa.gov*, *https://epa.gov/*, and *http://www3.epa.gov:80/* all refer to the same *epa.gov* site and have same SURT key, *gov,epa)/*. [6] Figure 3 shows these variations in epa.gov URLs.

2.3.2 Web Archive Replay Functionality

Web archives are stored in the Web ARChive (WARC) file format [66]. Because replay is a goal of digital preservation, WARC files contain full HTTP response headers, HTML documents, and other resources such as images and JavaScript files. WARC files accomplish this functionality by aggregating WARC records. A WARC record includes metadata needed to identify the content within the WARC file, followed by HTTP headers, then finally the content of the resource, such as the HTML or the binary image data. If the WARC record is created from a resource retrieved using a system supporting the Memento protocol, the HTTP headers in the WARC record would include information about the crawl date of the resource.

The state of the art open source tool for replaying WARC files is PyWB [91]. In order to

[6] `http://web.archive.org/cdx/search/cdx?url=epa.gov`

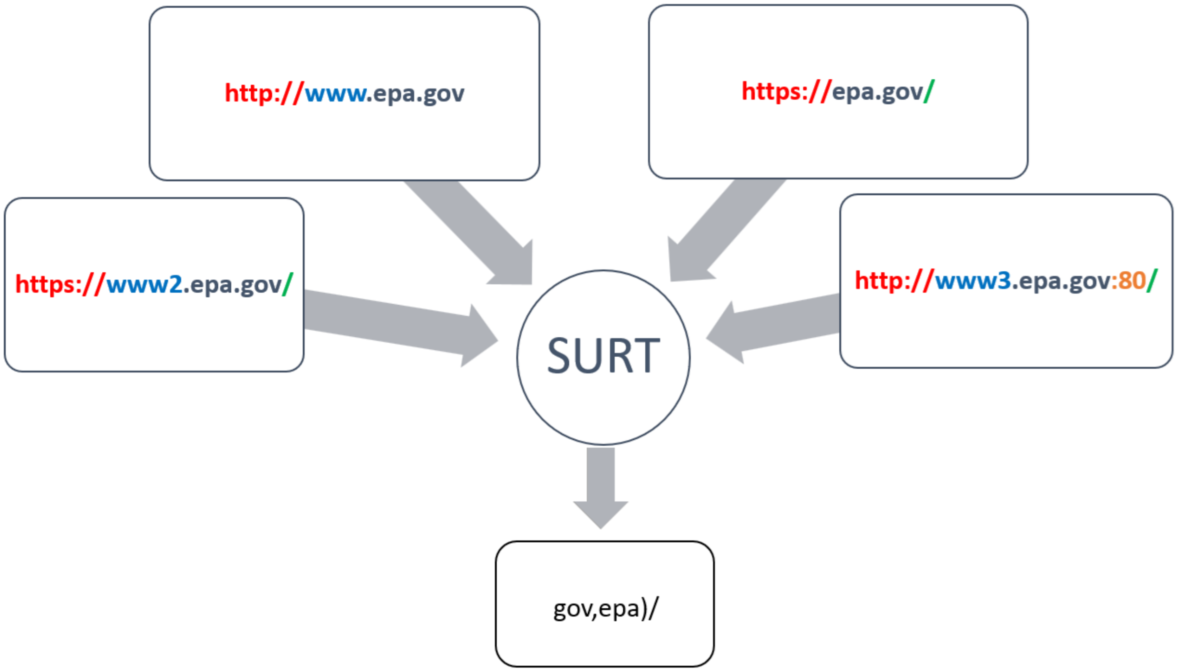


Figure 3. SURT can be used to Match Multiple Versions of a URL. Scheme, canonical subdomain, canonical ports, and slashes are all accounted for.

emulate the experience of viewing the archived webpage as it was when it was archived from the live web, the replay system must display all of the HTML for the page along with all of its resources, including images, CSS, JavaScript, and so on. The straightforward way to accomplish loading the embedded resources is to rewrite all of the links on the page from their URI-Rs to a URI-M with the same datetime as the page being replayed. This will trigger a lookup in the web archive for the resource archived at the closest datetime to the page being replayed.

Sometimes not all of the resources from a page are archived when the page is archived [15]. If a referenced resource cis never archived at that URI-R, it will not be able to be replayed in the web archive. This phenomenon of missing resources is called memento damage [26]. When resources are missing, the user cannot experience the page as it was at the time it was archived.

2.3.3 Viewing Changes in Web Archives

Temporal search engines for web archives do not incorporate the changes between versions of webpages into their results, but some other tools do exist to help users find and view changes on known pages. Sherratt et al. [124] created a tool to help users find changes in a specific web page's version history as a part of the GLAM Workbench web archives Jupyter notebook collection. Because this tool does not index any web page content, each URL must be searched individually. The tool also only includes a linear search option, which hinders its speed. The results highlight the query term in context, but not in context of the previous versions of the page. Summers [130] created a command line tool that allows users to input a webpage and terms of interest and the output is the version when the terms were deleted. Because this tool does not index any web page content, each URL must be searched individually. The tool uses binary search, which increases its speed compared to the GLAM tool. Another tool, WikiBlame,[7] allows for users to search for changes in a specific Wikipedia page. While this tool does allow for binary search, it does not index any page content, so only one page can be searched at a time rather than an entire group of related pages. Wikipedia includes a differences tool as part of its version history viewer,[8] allowing users to view changes in a static context.

The Internet Archive Wayback Machine allows users to compare two different versions of a webpage using its Changes tool.[9] The Changes tool, shown in Figure 4, shows users the differences between two captures at this one particular web archive. Users must know the two dates they wish to compare, as there is not any search functionality integrated with the tool. The Changes tool

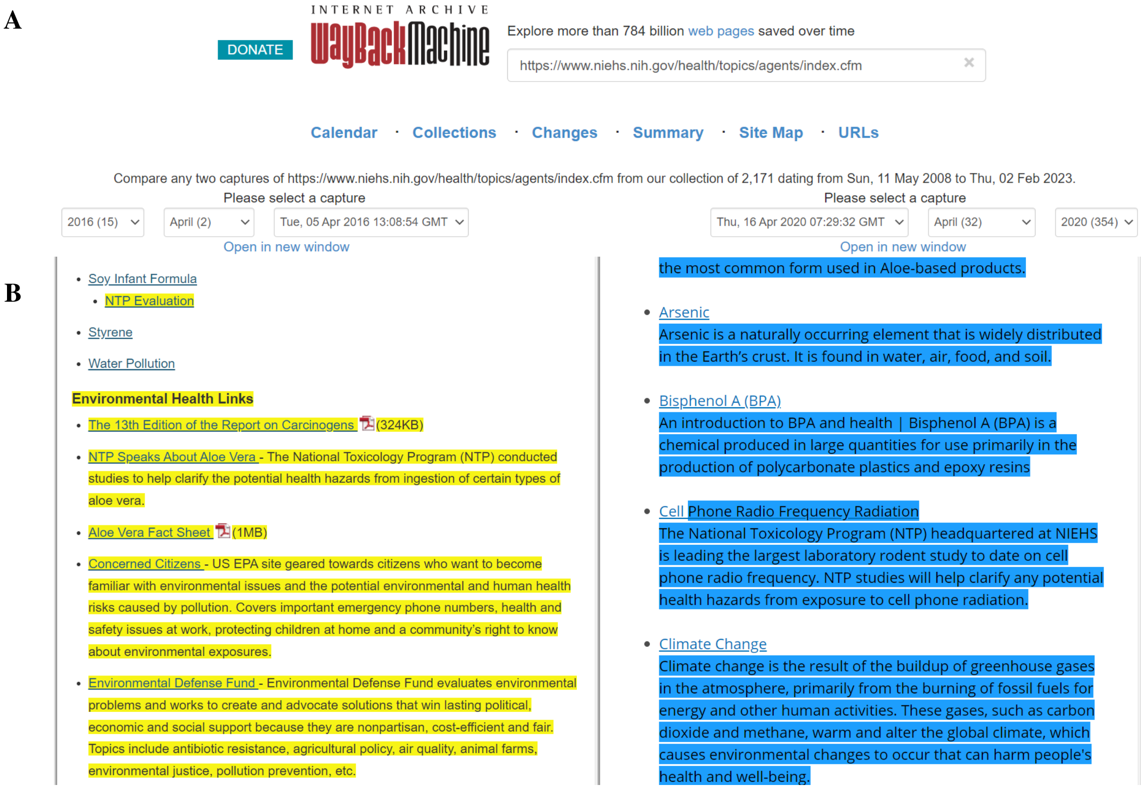
[7] http://wikipedia.ramselehof.de/wikiblame.php

[8] https://en.wikipedia.org/wiki/Help:Diff

[9] https://web.archive.org/web/changes/

was built upon the Environmental Data and Governance Initiative's Web-Monitoring-Diff suite.[10] This service is not currently connected with any tool that allows users to search for changes across versions, and only lets users compare versions from the Internet Archive. There is currently no version of the Changes tool that allows users to compare versions of webpages that exist across multiple archives. The Changes tool is only able to compare two versions of a webpage at one time, and the comparison is currently presented in a side by side static context.

[10] https://github.com/edgi-govdata-archiving/web-monitoring-diff



Explore more than 784 billion web pages saved over time

https://www.niehs.nih.gov/health/topics/agents/index.cfm

Compare any two captures of https://www.niehs.nih.gov/health/topics/agents/index.cfm from our collection of 2,171 dating from Sun, 11 May 2008 to Thu, 02 Feb 2023.

Please select a capture: 2016 (15) | April (2) | Tue, 05 Apr 2016 13:08:54 GMT

Please select a capture: Thu, 16 Apr 2020 07:29:32 GMT | April (32) | 2020 (354)

- Soy Infant Formula
  - NTP Evaluation
- Styrene
- Water Pollution

Environmental Health Links

- The 13th Edition of the Report on Carcinogens (324KB)
- NTP Speaks About Aloe Vera - The National Toxicology Program (NTP) conducted studies to help clarify the potential health hazards from ingestion of certain types of aloe vera.
- Aloe Vera Fact Sheet (1MB)
- Concerned Citizens - US EPA site geared towards citizens who want to become familiar with environmental issues and the potential environmental and human health risks caused by pollution. Covers important emergency phone numbers, health and safety issues at work, protecting children at home and a community's right to know about environmental exposures.
- Environmental Defense Fund - Environmental Defense Fund evaluates environmental problems and works to create and advocate solutions that win lasting political, economic and social support because they are nonpartisan, cost-efficient and fair. Topics include antibiotic resistance, agricultural policy, air quality, animal farms, environmental justice, pollution prevention, etc.

the most common form used in Aloe-based products.

- Arsenic
  Arsenic is a naturally occurring element that is widely distributed in the Earth's crust. It is found in water, air, food, and soil.
- Bisphenol A (BPA)
  An introduction to BPA and health | Bisphenol A (BPA) is a chemical produced in large quantities for use primarily in the production of polycarbonate plastics and epoxy resins
- Cell Phone Radio Frequency Radiation
  The National Toxicology Program (NTP) headquartered at NIEHS is leading the largest laboratory rodent study to date on cell phone radio frequency. NTP studies will help clarify any potential health hazards from exposure to cell phone radiation.
- Climate Change
  Climate change is the result of the buildup of greenhouse gases in the atmosphere, primarily from the burning of fossil fuels for energy and other human activities. These gases, such as carbon dioxide and methane, warm and alter the global climate, which causes environmental changes to occur that can harm people's health and well-being.

Figure 4. The Internet Archive's Wayback Machine Changes tool shows users the differences between two captures. It only works for captures at this one web archive, and there is no way to search the changes to find the dates and times. *(A)* The Wayback Machine Changes tool requires the user to know the date and time of both versions of the page in order to create a comparison. *(B)* The Wayback Machine Changes tool helps users examine additions (blue) and deletions (yellow). The term pollution, on the left in yellow, was removed. ©2023 IEEE

2.4 SEARCH ARCHITECTURE

Facilitating search in web archives is a primary contribution of this thesis. Understanding how search systems function is necessary to implement a web archive search backend. Identifying needs specific to web archives is essential to creating a successful implementation for change-text search.

In order for users to efficiently query for documents, the information in the documents must be extracted and transformed into some kind of database. The preferred database for search is an inverted index, which allows for lookup by term rather than individually searching each document. Lookup by term implies that the information in each document must be tokenized. Users also need a way to submit their query to a search platform that communicates with the inverted index and returns the results in a deliberate order. This involves a frontend search interface as well as the backend platform that connects to the index to execute the queries. In this section, we detail the capabilities of the Apache Search Suite, explain some of the intricacies of tokenization, and then discuss existing full text search workflows for web archives.

2.4.1 Apache Search Suite

Apache Lucene [127] is an open-source, high-performance core search system with a skip list implementation for its term index.[11] Lucene is the foundation for the Apache Solr search platform.[12] Apache Tika extracts text and metadata from files for the Lucene index.[13] Lucene, Solr, and Tika are written in Java. Solarium is the premier PHP client library for Solr.[14] Solarium can be used to create a user interface that translates user-friendly queries into formal Lucene query syntax.

[11] `https://lucene.apache.org/`

[12] `https://solr.apache.org/`

[13] `https://tika.apache.org/`

[14] `https://github.com/solariumphp/solarium`

2.4.2 Indexing Workflow

When documents are indexed, their text contents are tokenized into terms. Tokenization in Lucene follows the Unicode Standard Annex #29 (UAX #29) [38]. Phrase queries are supported on individual fields by also storing term positions with a document. Tokenization also allows for emphasizing search terms and phrases in text snippets shown to the user in the search results, which is called highlighting. Besides the Java Lucene implementation, UAX #29 has an implementation in PHP,[15] and also in Python [125].

2.4.3 Search Architecture for the Archived Web

Solrwayback is a bundle of tools for web archive search, including Lucene, Solr, Apache Tika for metadata extraction, the Apache Tomcat server, the UK Web Archive Discovery indexer, and the Solrwayback user interface [45]. This workflow is shown in Figure 5. The first part of the workflow is indexing the WARCs. While there is no native support for WARC file indexing in Apache Tika, the UK Web Archive Discovery indexer [69] can extract text from WARC records and transmit that text to the Apache suite indexing workflow. The Web Archive Discovery indexer also includes a boilerplate removal option, which can be turned on to avoid indexing JavaScript files. The rest of the workflow regards how users query for information. Users access the user interface and make queries, which are interpreted by Solr. Solr connects directly to Lucene to read from the index. The results are returned and presented to the user.

Web archives are inherently temporal, so it is important for a web archive search backend to support temporal queries. Lucene supports document date ranges and queries across date ranges

[15] `https://emptyheap2019.github.io/posts/parse-words-php/`

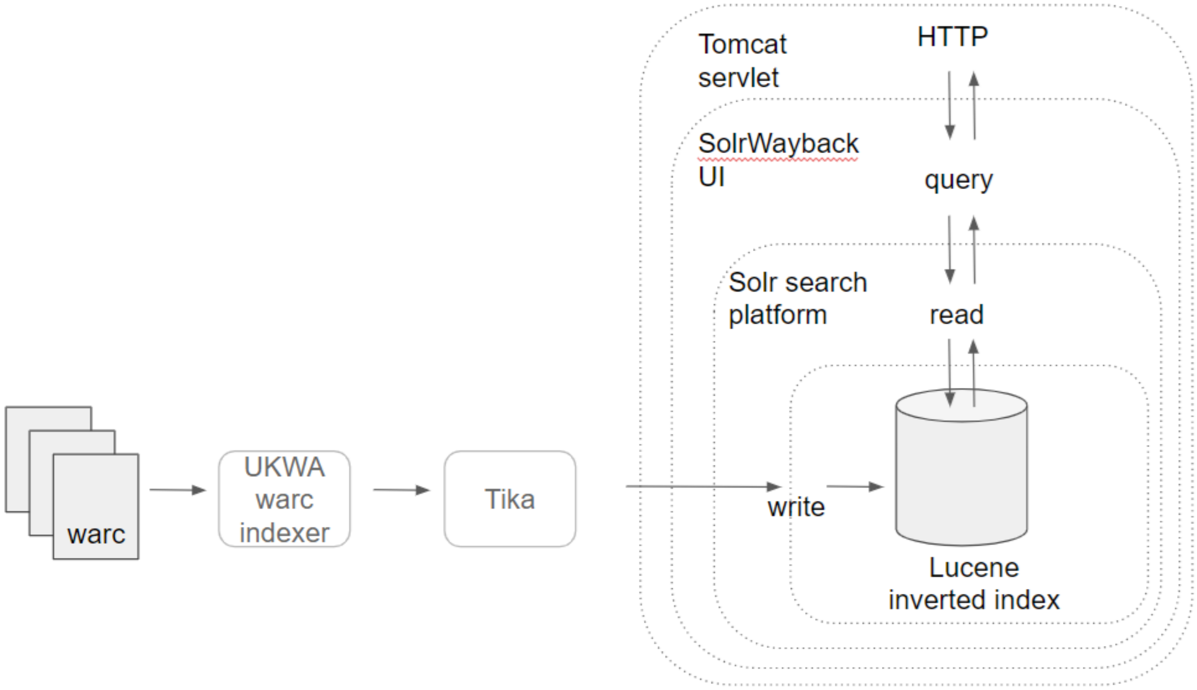


Figure 5. Solrwayback workflow. Indexing uses the UKWA WARC indexer to populate data for Tika to post to the Lucene index. HTTP requests and responses go through Tomcat. The Solrwayback user interface provides a way for users to query. The queries are sent to Solr which reads from the Lucene index.

with variable granularities, and Solr also supports a date range field to complement Lucene's functionality [127].

2.4.4 Web Archiving Lookup User Considerations

> *Access severely constrains the "save everything" mentality since without proper indexing, finding information becomes impossible.* –Gary Marchionini

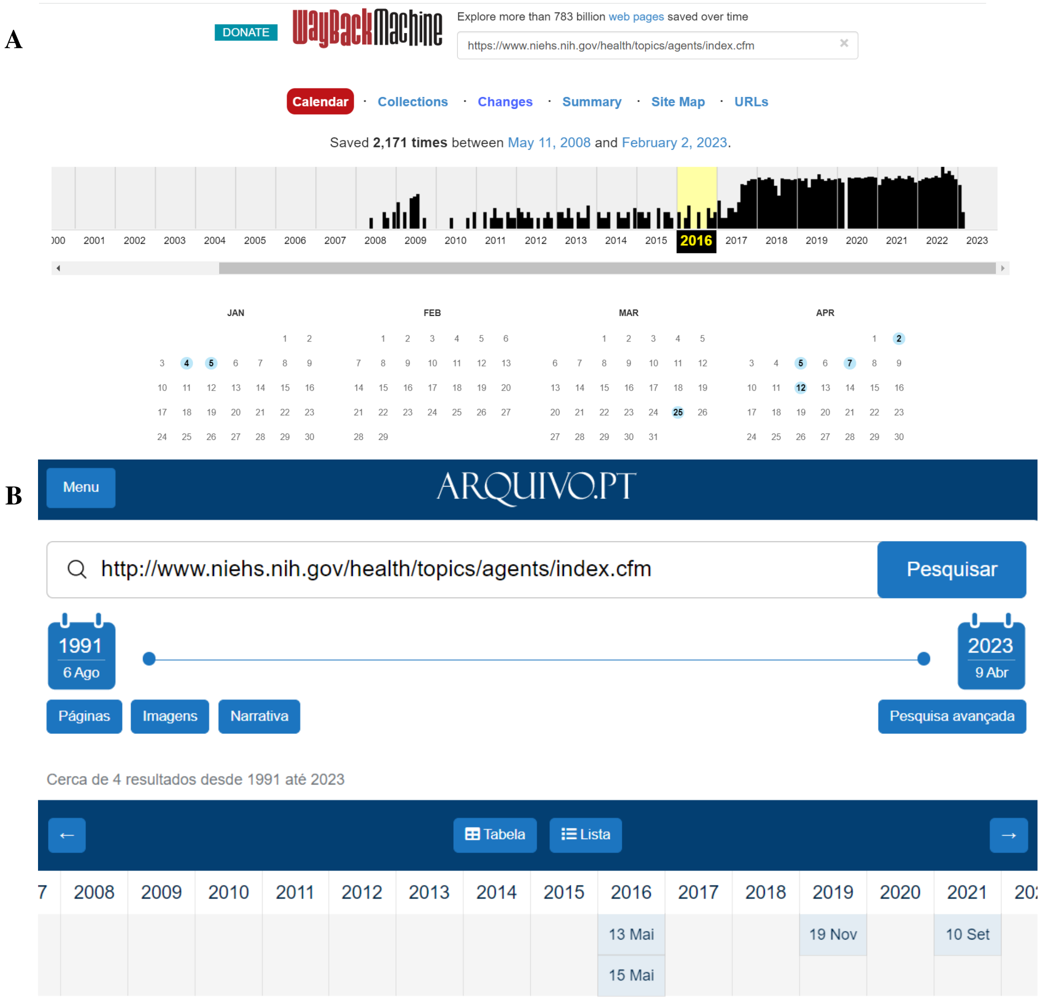
Users of web archives want to find and view changes on webpages over time. However, the current search interfaces for web archives do not support this task. For the web archives that include a full-text search feature, multiple versions of the same webpage that match the search query are shown individually without enumerating changes, or are grouped together in a way that hides changes.

As of June 2024, the Internet Archive Wayback Machine contains 866 billion webpages available for lookup and replay.[16] The Internet Archive clearly employs a “save everything” mentality. Most webpage captures are not indexed by full text; rather, users must know the URL for lookup. The choice to provide a less usable index for a larger amount of material lowers user access to information [103]. Providing access improves users’ ability to find information, but also brings up ethical ramifications for user privacy rights, which we discuss in Chapter 3.

2.4.5 Existing Web Archive Search Interfaces

Full-text search has been identified [119] as a highly requested feature for web archives. Many web archives only include URI lookup of holdings rather than full-text search. Figure 6 shows two examples of web archive URI lookup. Users can only look up webpages with known addresses in this type of search, whereas with full-text search users can type words to find their page instead of being required to know its URI.

[16] `https://web.archive.org/web/20240618001150/https://web.archive.org/`



Saved 2,171 times between May 11, 2008 and February 2, 2023.

Cerca de 4 resultados desde 1991 até 2023

Figure 6. Two examples of web archive URI lookup of https://www.niehs.nih.gov/health/topics/agents/index.cfm. URI lookup does not allow the user to search by page text. *(A)* Internet Archive Wayback Machine URI lookup *(B)* Portuguese Web Archive URI lookup ©2023 IEEE

This figure shows HTML TimeMaps, though other TimeMap formats like JSON and Link do not support full-text search either, or show any information about change of page text over time. One web archive lookup interface that does provide information about major page changes is Archive Today, shown in Figure 7. Because the lookup interface includes thumbnails, the user can see the major changes to the page over time. However, there is still no information presented about how the text on the page has changed.

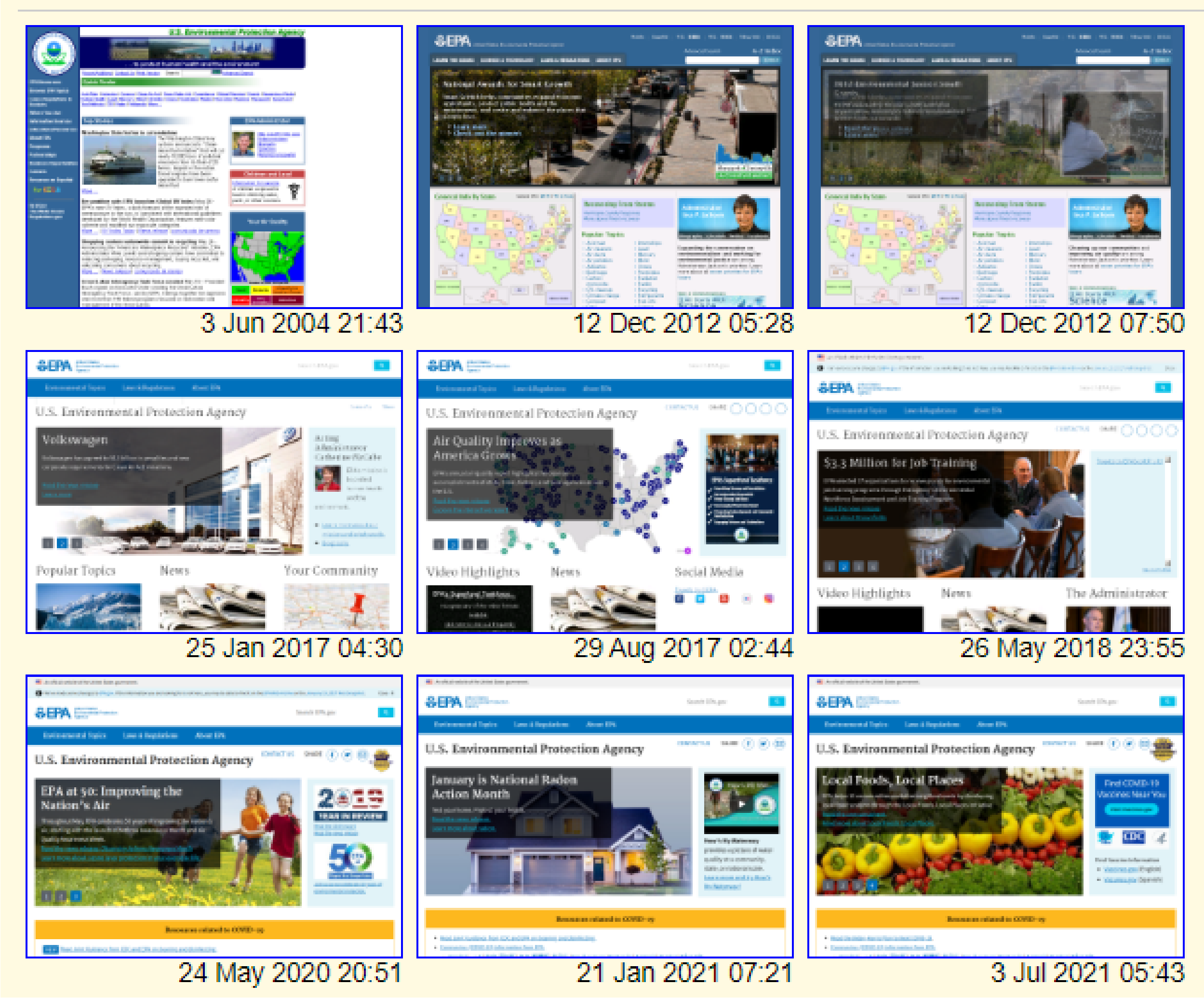


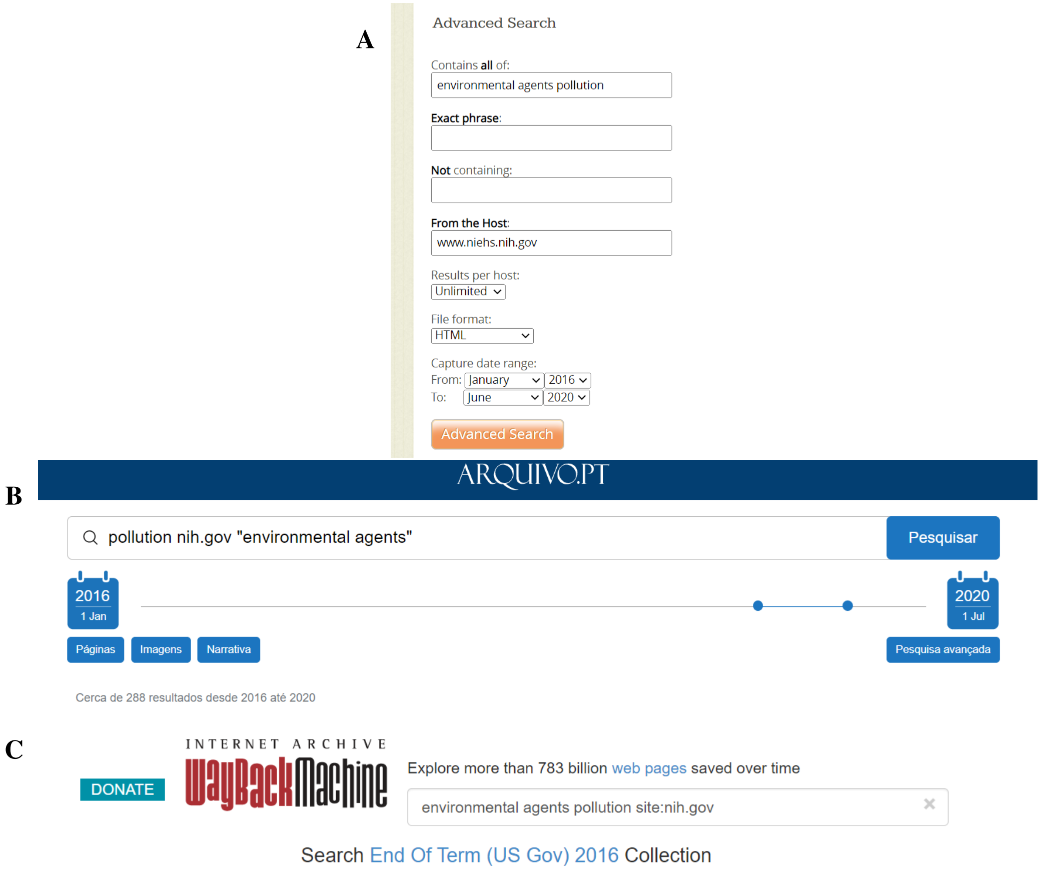
Figure 7. Archive Today URI lookup of epa.gov with thumbnails showing the major changes of the page over time

The Internet Archive's Wayback Machine does provide a search engine, but it is based on metadata, such as page title and domain, rather than full-text indexing. There are some web archives, such as the Portuguese Web Archive [104], the UK Web Archive,[17] and collections in Archive-It, that do support full-text search. Figure 8 shows the query fields currently available to users in some of these search engines. All three query forms allow for filtering by fields such as domain or date. However, none of the query interfaces have the ability for users to specify a query for a deleted term.

While date filtering and domain filtering are helpful, they do not address how to display multiple versions of the same web page that match the search query. These pages may or may not have differences in content, and these interfaces fail to inform the user how the content in the matching versions has changed over time. None of the current interfaces allow the user to query for webpage changes.

Figure 9 shows the search engine results pages for some of these search engines. Archive-It shows each matching version of the page in the results, which introduces clutter. Arquivo.pt has a setting to show a maximum amount of page versions to reduce clutter, but this leads to relevant versions not being returned. The Wayback Machine collection search only shows one page version per query. This reduces clutter, but the reason why that one version was chosen is unclear and there is no easy way to view the other matching versions or the changes between them.

[17] `https://www.webarchive.org.uk/ukwa/`



Advanced Search

Contains **all** of: environmental agents pollution

From the Host: www.niehs.nih.gov

Results per host: Unlimited

File format: HTML

Capture date range:
From: January 2016
To: June 2020

pollution nih.gov "environmental agents"

2016 1 Jan — 2020 1 Jul

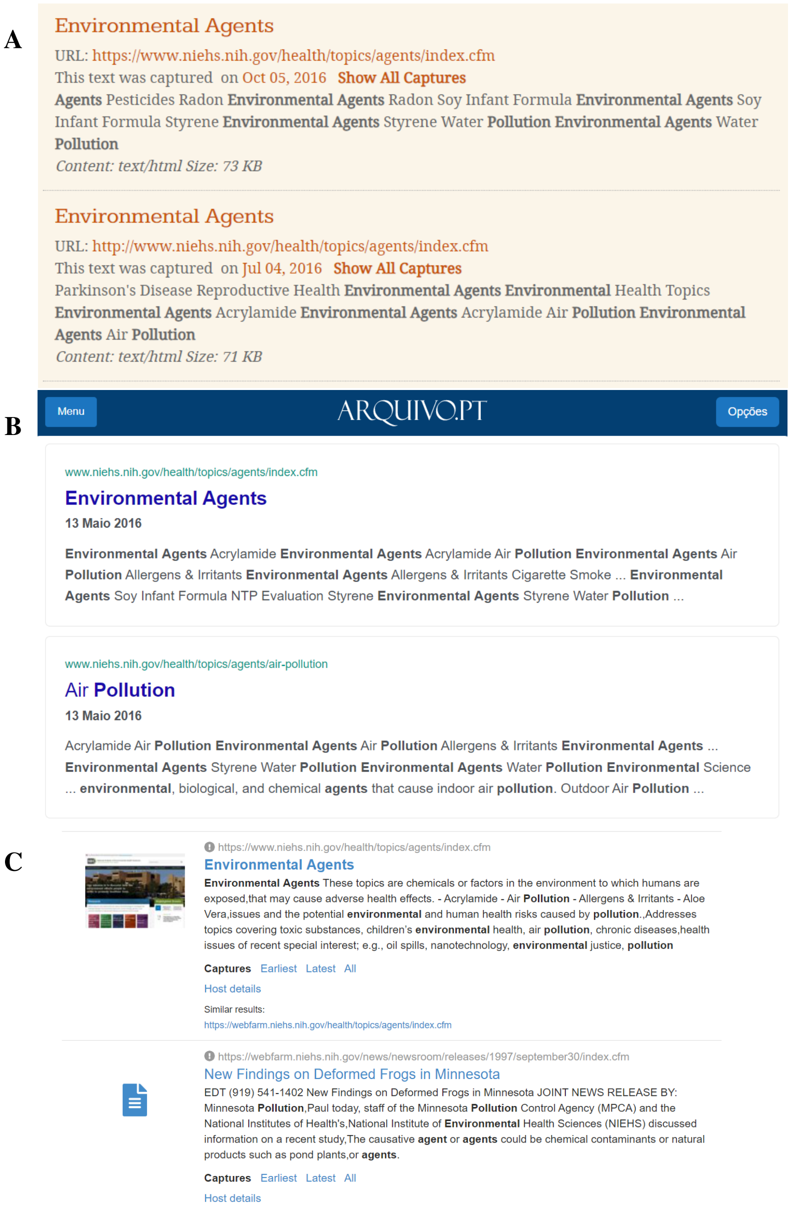
Cerca de 288 resultados desde 2016 até 2020

Explore more than 783 billion web pages saved over time

environmental agents pollution site:nih.gov

Search End Of Term (US Gov) 2016 Collection

Figure 8. Some web archives have full-text search with various filtering features over small collections or over the entire web archive. None of the search interfaces allow for users to search for terms that have been removed from webpages. *(A)* The general search box for Archive-It *(B)* The search box over the entire Portuguese Web archive allows for full-text search and filtering by date. *(C)* The search box at the Internet Archive's Wayback Machine for the 2016 End of Term Archive collection allows for full-text search and filtering by site, but only over one end of term collection at a time. ©2023 IEEE



Environmental Agents
URL: https://www.niehs.nih.gov/health/topics/agents/index.cfm
This text was captured on Oct 05, 2016 **Show All Captures**
Agents Pesticides Radon **Environmental Agents** Radon Soy Infant Formula **Environmental Agents** Soy Infant Formula Styrene **Environmental Agents** Styrene Water **Pollution Environmental Agents** Water **Pollution**
Content: text/html Size: 73 KB

Environmental Agents
URL: http://www.niehs.nih.gov/health/topics/agents/index.cfm
This text was captured on Jul 04, 2016 **Show All Captures**
Parkinson's Disease Reproductive Health **Environmental Agents Environmental** Health Topics **Environmental Agents** Acrylamide **Environmental Agents** Acrylamide Air **Pollution Environmental Agents** Air **Pollution**
Content: text/html Size: 71 KB

www.niehs.nih.gov/health/topics/agents/index.cfm
Environmental Agents
13 Maio 2016
Environmental Agents Acrylamide **Environmental Agents** Acrylamide Air **Pollution Environmental Agents** Air **Pollution** Allergens & Irritants **Environmental Agents** Allergens & Irritants Cigarette Smoke ... **Environmental Agents** Soy Infant Formula NTP Evaluation Styrene **Environmental Agents** Styrene Water **Pollution** ...

www.niehs.nih.gov/health/topics/agents/air-pollution
Air **Pollution**
13 Maio 2016
Acrylamide Air **Pollution Environmental Agents** Air **Pollution** Allergens & Irritants **Environmental Agents** ... **Environmental Agents** Styrene Water **Pollution Environmental Agents** Water **Pollution Environmental** Science ... **environmental**, biological, and chemical **agents** that cause indoor air **pollution**. Outdoor Air **Pollution** ...

https://www.niehs.nih.gov/health/topics/agents/index.cfm
Environmental Agents
Environmental Agents These topics are chemicals or factors in the environment to which humans are exposed,that may cause adverse health effects. - Acrylamide - Air **Pollution** - Allergens & Irritants - Aloe Vera,issues and the potential **environmental** and human health risks caused by **pollution**.,Addresses topics covering toxic substances, children's **environmental** health, air **pollution**, chronic diseases,health issues of recent special interest; e.g., oil spills, nanotechnology, **environmental** justice, **pollution**
Captures Earliest Latest All
Host details
Similar results:
https://webfarm.niehs.nih.gov/health/topics/agents/index.cfm

https://webfarm.niehs.nih.gov/news/newsroom/releases/1997/september30/index.cfm
New Findings on Deformed Frogs in Minnesota
EDT (919) 541-1402 New Findings on Deformed Frogs in Minnesota JOINT NEWS RELEASE BY: Minnesota **Pollution**,Paul today, staff of the Minnesota **Pollution** Control Agency (MPCA) and the National Institutes of Health's,National Institute of **Environmental** Health Sciences (NIEHS) discussed information on a recent study,The causative **agent** or **agents** could be chemical contaminants or natural products such as pond plants,or **agents**.
Captures Earliest Latest All
Host details

Figure 9. None of these three web archive search interfaces group the versions to show the change in the page over time. *(A)* Archive-It collection SERP for Figure 8A's query showing title, URI, date, text snippet, metadata, replay link, additional captures link. Two page versions are shown individually. *(B)* Arquivo.pt SERP for Figure 8B's query showing URI, title, date, text snippet, and replay link. *(C)* Wayback Machine collection SERP for Figure 8C's query showing URI, title, text snippet, screenshot, replay link, and additional captures link. ©2023 IEEE [52]

These three web archive full-text search interfaces show the difficulty in presenting information about a web page with multiple versions. Versions matching the search query are either grouped together without explicitly enumerating all matching versions, or included separately as multiple results. None of the interfaces group the versions to show the change in the page over time.

2.5 WEBPAGE CHANGE AND QUERY DATASETS

This thesis presents a change-text search backend and frontend, which requires a dataset of webpages with changes to validate the implementation. Most existing archived webpage datasets include only one version of the page, which does not make it possible to analyze change over time. Other datasets that do include more than one version of the page were created with a mindset of monitoring. In this section, we present datasets that can be used to evaluate the change-text search system.

2.5.1 General Web Crawl Datasets

Existing data sets of webpage crawls commonly used for academic purposes, such as ClueWeb [112] and Common Crawl,[18] aim to collect snapshots of a large amount of unique URLs. The three ClueWeb datasets cover 2009, 2012, and 2022, and contain one to two billion webpages each. Common Crawl datasets started in 2008, and are collected monthly at the present time. The most recent Common Crawl from July 2024 contains 2.5 billion webpages.

Large datasets, like ClueWeb and CommonCrawl, crawl at different points in time based on a seed rather than always archiving the same pages. The benefits of this approach are to archive new pages found via new links, and to avoid pages that have been deleted. However, since these

[18] https://commoncrawl.org/

data sets did not aim to collect multiple versions of the same page, there is no guarantee of any particular page being crawled regularly, which would hinder the ability to analyze change over time. Additionally, in order for an analysis of webpage change to be of consequence, the changes on those webpages need to be meaningful.

2.5.2 End of Term Web Archive

The archival of federal websites, especially at the end of a president's term, is an important task undertaken by multiple organizations. The End of Term Web Archive is created through a partnership between five organizations, including the Internet Archive and the Library of Congress [116], [123]. This web archive includes a full-text search feature,[19] but each end of term crawl includes only one capture of each web page. Phillips et al. [115] compared the 2008 and 2012 end of term collections to identify changes in crawl dates and webpage addresses, but individual terms were not analyzed.

2.5.3 EDGI Dataset

Nost et al. [110], on behalf of the Environmental Data and Governance Initiative (EDGI), monitored changes on 30 US federal environmental agency websites between 2016 and 2020. They compared the change in 56 pre-chosen environmental terms and phrases on 40,000 webpages using the web archive holdings at the Internet Archive. Their dataset includes a file counted_urls.csv which has URLs for pages and their 2016 and 2020 mementos at the Internet Archive, a sample of which is shown in Listing 2.2. Another file, obama_count.csv counts the 56 terms and phrases on each 2016 memento, as shown in Listing 2.3. Another file, trump_count.csv, contains the term

[19] `http://eot.us.archive.org/search/`

counts in 2020.

Listing 2.2. counted_urls.csv shows paired mementos, NA for capture not found

```
url - o,
final captured url - t
https://www3.epa.gov/enviro/facts/multisystem.html,
http://web.archive.org/web/20160612091334id_
/https://www3.epa.gov/enviro/facts/multisystem.html,
http://web.archive.org/web/20200101042321id_/
https://www3.epa.gov/enviro/facts/multisystem.html
https://www.osha.gov/pls/imis/inspectionNr.html,
http://web.archive.org/web/20160322140439id_/
https://www.osha.gov/pls/imis/inspectionNr.html,NA
```

Listing 2.3. obama_count.csv shows term counts in 2016, 999 for non-200 HTTP status

```
adaptation,agency mission,air quality,anthropogenic,benefits,brownfield...
17,0,1,0,5,0...
999,999,999,999,999,999...
```

In order to determine if a term was added or removed between 2016 and 2020, each of the term count files can be loaded and the corresponding line in each file can be compared value by value. In order to know which page the term counts belong to, the corresponding line in the URL file can be matched up.

In their analysis of the data, Nost et al. showed that approximately 20% of the EPA's website was removed between 2016 and 2020, inhibiting public access to this environmental information.

They found differences in terminology changes by agency type, and by depth of page. They found that certain key terms, like “climate change,” were removed from most federal environmental websites in this time period.

Because this dataset contains known changes, it is particularly well suited for evaluation of a change text search system. The 56 terms and phrases identified by EDGI can be used to verify the accuracy of the algorithms used to implement the change text calculations. It will also be feasible to identify additional terms and phrases deleted on the websites beyond the ones EDGI tracked in order to evaluate the comprehensiveness of the change text index as well.

2.5.4 ORCAS Dataset

The Open Resource for Click Analysis in Search (ORCAS) [35] is a dataset with Microsoft Bing queries and associated clicks from 2017 to 2020. The dataset employs k-anonymity, so each query in the dataset was made by a large number of different users. This protects user privacy, but also ensures that the queries in the dataset are popular. ORCAS is a part of the Microsoft Machine Reading Comprehension (MS MARCO) dataset collection, including a passage relevance dataset [19]. We use ORCAS to link query terms and deleted terms for pages in the evaluation the change text search index (Chapter 6).

2.5.5 Archive Query Log Dataset

The Archive Query Log (AQL) [120] is a dataset created from archived search engine results pages (SERPs), allowing researchers to examine how queries and SERPs change over time. The dataset allows researchers to access 356 million queries, but only seven percent of the corresponding SERPs are included in the dataset. This limits analysis to known queries. For example, since all of the queries are available in the dataset, researchers can determine the answer to research questions about only queries. They can also download the associated SERPs to analyze the results for these queries. However, if a researcher would like to determine the queries for a given result page, that is not possible with the current dataset. There is also no guaranteed associated click for a page either.

2.6 SUMMARY

In this chapter, we presented basic information necessary to understand the work in the thesis. First, we explained how users' mental models, memory, and tasks influence their information seeking capabilities and behavior. We need to define users' mental models in web archives, the tasks they use web archives to accomplish, and memory constraints related to understanding change in order to build a successful search system. Next, we examined how pages' URIs and content change over time, and how HTTP status codes function. We incorporate these ideas into our methodology for finding and viewing changes on webpages.

Next, we explained how the Memento Protocol functions, including URI-Ms and TimeMaps. We explained how archived HTTP status codes can be found in CDX files. We showed how SURT can be used to canonicalize URLs. We presented WARC files as files that hold archived webpages. We introduced PyWB as a tool to replay archived webpages. We introduced the idea of memento damage, which affects the ability to view a past version of a webpage. We also examined the Wayback Machine Changes Tool. We will use the Memento Protocol to incorporate archived webpages into our search system. We will use CDX files to efficiently determine archived status codes for indexing purposes. We will use SURT to align multiple datasets. We must find a tool that can index WARC files. We will use PyWB to create an animated difference viewer. We will show that some users view multiple versions of webpages because of memento damage to one of the versions. We will use the technology that enables the Changes Tool to function to create the animated difference viewer.

In the search architecture section, we introduced Lucene and Solr as established search systems. We explained how indexing requires tokenization. We introduced SolrWayback as an established web archive search interface. We showed that not all web archives have full text search, and we showed that the existing search engine results pages do not have a good solution for how to group webpage versions. We will use Lucene, Solr, and SolrWayback to build our search system. We will incorporate tokenization into our search engine results page for highlighting and for our animated difference tool. We will create a search system that supports querying for changes and grouping the results in a meaningful way.

Finally, we introduce the EDGI dataset which contains known changes and the ORCAS dataset which contains query/click pairs. We will use the EDGI dataset to evaluate the change text search system. We will align the EDGI and ORCAS datasets to show how term changes and queries are related for these webpages.

CHAPTER 3

RELATED WORK

This chapter is devoted to explaining the work that other researchers have completed that is similar to the problem of searching for changes in webpages. First, we detail why users make changes to webpages and why other users want to view these changes. Next, we describe how users behave in both search systems and in web archive systems. We examine approaches to search in web archives. Next, we describe how users view changes, and then examine approaches to implement tools that help users view change. Finally, we survey ethical issues on the Web and detail how these issues relate to searching for changes in webpages.

3.1 WEBPAGE CHANGES: MOTIVATIONS OF WEB ACTORS

The primary contribution of this thesis is a system for searching for changes on webpages. In order to build such a system, we must understand why webpages change, and what exactly users are looking for when they want to view these changes. Many webpages undergo changes. Human editors can make the changes, and human viewers can view and understand the changes. This section discusses the motivations of these actors, the editors and the viewers, for making and viewing changes on webpages.

3.1.1 Measuring Changes on the Web

Adar [4] monitored 55,000 websites for changes at a weekly rate between August and September, 2006. Adar et al. [4], [6] examined changes on the Web in 2009 and found that webpages change at different rates depending on different properties like top level domain and path depth. Webpages also vary in the amount of change they undergo during each edit. Different parts of webpages also change at different rates and in different amounts.

Cho et al. [32] and Fetterly et al. [47] examined changes on the Web in 2003, and also found differences in change rates based on top level domain, although Fetterly et al. note that this phenomenon is more strongly correlated to frequency of change rather than degree of change. Fetterly et al. also found that the size of the page is correlated with both the frequency and degree of change.

Change on webpages is measured in a variety of ways. Similarity measures, such as Levenshtein [95] and Jaccard [67], compare two versions of a document's text and return a percentage similarity. Levenshtein is based on edit distance while Jaccard is based on the difference in ratio between the union and intersection of the terms. When comparing change over multiple versions, Klein et al. [88], [90] found that one consideration is comparing incremental, or sliding, change versus comparing change to the first version, also known as rooted or anchored change. These change measures just described represent whole page change measures rather than term-by-term change measures.

Adar [4] also investigated how individual terms change on pages over time. He found that terms that are removed from a page can reflect natural temporal environmental factors, such as when the seasons change. He found that terms that are not removed from a page generally represent the aboutness of the page, or otherwise represent standard page items such as navigational terms. The way to tell these types of static terms apart is to consider which terms are common in the corpus. Term Frequency - Inverse Document Frequency (TF-IDF) [129] is a measure that can be used to determine terms that are uncommon in a corpus and also can be used to weight query terms and documents to improve the relevance of results.

One limitation of change measurement algorithms is when the page has changed beyond a certain threshold. Dalal et al. [37] found that when the degree of change on a page was large, the relevance of the page to a topical collection was no longer discernable.

Another limitation of change measurement algorithms is estimating rates of change when the data is irregular. Crawling can be too frequent, capturing many duplicates and this wasting crawling resources, or not frequent enough, which results in missed intermediary changes. Cho et al. [32] and Coffman et al. [43] developed change estimation measures to address this issue. In a live web search engine, the goal of using these algorithms would be to maximize page freshness. This is because of the possibility of content drift, including deleted terms, affecting the relevance of the page for the query. In a web archive, the goal of using these algorithms would be to maximize the amount of changes captured.

3.1.2 User Motivations to View Webpage Changes Over Time

One of the motivations that users have for viewing webpage changes is to view new information or updates that have been added to the page, which is referred to as revisiting the page. This type of access is also referred to as monitoring when specifically watching for change. [4]

The International Internet Preservation Consortium Access Working Group identified multiple use cases where users would need to view web page changes over time [65]. One use case involved a professional studying how the presentation of information has changed over time. This user needs access to a tool that can display the differences between versions of pages.

Ogden et al. synthesized web archive use patterns [111]. One pattern is called monitoring, where users wish to see additions, deletions, and changes in content and links. Another pattern is attribution, where journalists and other users use web archives to perform fact checking. Users investigate the current page's content compared to the content of a historical version of the page in the web archive.

According to a study conducted by Teevan et al. in 2007, 39% of search queries represent users trying to re-find previously viewed pages [133]. Teevan also investigated locating previous versions of pages [132]. Users were either unable to retrace their path to find the page they were seeking, or would click on a link to a page that had been removed or a page whose content had been altered. These changes involved content being deleted from a page, such as a post being deleted from a message board thread, pages being moved on websites by site administrators, and content being unintentionally removed because of software failures. Because there are so many reasons why a page might have changed over time, it is important that a user is able to locate and view past versions of pages so that they can re-find information they have previously seen.

Jackson et al. created an exploratory search interface for web archives [70], as shown in Figure 2A (Chapter 1). When designing the search engine results page, they state that they chose to list every version of a webpage as a separate document because their users were interested in viewing how web pages changed over time, such as deleted content.

3.1.3 Editor Motivations to Make Webpage Changes Over Time

There are neutral reasons for human editors to make changes to webpages. One reason is if the information on the page becomes stale. The human editor may notice, even perhaps from another webpage, that the information on their webpage is no longer up to date. They will make changes so that the website content is brought back up to date [4].

On Wikipedia, Yang et al. [140] identified common intentions that motivated edits. Some are neutral, such as updating information as stated above, copy editing and refactoring, and adding citations. Other types of edits actually serve to make the article more neutral, such as editing the point of view or removing information that is untrue. Editors will also add information that supports the content already on the page, clarify existing information, or simplify existing verbiage.

There are also multiple reasons that are not neutral that cause human editors to make changes on webpages. One reason is to improve the ranking of the webpage on search engine results pages. This is referred to as rank incentivized document manipulation, or search engine optimization [92]. Another reason that human editors make non-neutral changes to webpages is to make the webpage less neutral, which is called tendentious editing. Groups of people who are known to benefit from furthering a specific view point on a controversial subject include politicians promoting their party's viewpoints, businesses promoting their products, and others.

Tam et al. [131] developed a framework for evaluating changes in artifacts. The framework can be viewed from multiple perspectives, such as focusing on the artifact or the editor. The framework addresses when, where in the artifact, and how the artifact has changed, as well as why the changes were made.

Rank incentivized edits

While the exact ranking methods used in Bing are not public knowledge [92], it is possible to manipulate a document's ranking without necessarily change the meaning or quality of the page, such as by increasing the frequency of query terms [137].

One way to gauge access is retrievability. A page's retrievability [16] is a measure independent of ranking that quantifies a user's ability to find the page in an information retrieval system. Typically, webmasters aim to raise their webpages' retrievability through means such as search engine optimization [92]. Webmasters, including webmasters of government webpages, are also able to analyze server logs to determine the specific queries from search engines that lead users to navigate to their webpages; Ibáñez et al. [64] showed how this strategy was employed for the European Data Portal. With access to this data, webmasters can determine which query terms lead visitors to their webpages and increase the frequency of those query terms on the page.

Vasilisky et al. [138] used web archives to compare the change on 13 versions of each page in ClueWeb09. They found that the terms in pages' queries became more prevalent on the page over time, and called this measure the QueryTermsRatio, denoted $rqtr(d_i{}^q)$. In equation 1, $d_i{}^q$ refers to the query term ratio of the *i*th document, which is calculated as $|q_i|/|t_i|$, or the count of the query terms in the document divided by the term count in the document. Document 0 is defined to be the most recent version of the page and *i* is negative. Vasilisky et al. found that the average QueryTermsRatio for the corpus remained negative over 12 months; more query terms were added to the page over time in pursuit of search engine optimization. We use the QueryTermsRatio measure to evaluate the change text search index (Chapter 6).

$$rqtr(d_i{}^q) = d_i{}^q - d_0{}^q = |q_i|/|t_i| - |q_0|/|t_0| \tag{1}$$

Due to rank incentivized behavior, the ratio of query terms to page terms of a corpus increases over time [138], and a corpus where the ratio of query terms to page terms decreases over time would be irregular. For example, The American Marketing Association of New York created a sample optimization of the article "Reflections on The Birth of the NYAMA Mentoring Program" for the query phrase "marketing mentorship program."[1] The old version of the article had 399

[1] `amanewyork.org/wp-content/uploads/2017/03/Blog-Writing-Standards-AMA-NY-SEO-Basics.pdf`

words, 2 “marketing,” 0 “mentorship” and 15 “program,” for a QueryTermsRatio of $17/399 =$ 0.0426. The new version of the article has 416 words, 6 “marketing,” 1 “mentorship” and 15 “program,” for a QueryTermsRatio of 0.0529. The relative QueryTermsRatio (difference of new from old) is -0.0103, reflecting search engine optimization.

3.2 USER BEHAVIOR IN SEARCH

Because the major contribution of this thesis is a specialized search system, we must understand how users behave when seeking information. Since search in web archives is not yet ubiquitous, studies on user behavior when searching the past web are limited. Therefore, it is worthwhile to examine how users behave when searching the live web, while also acknowledging that the differences in the motivations of users searching the live and past web will inevitably lead to differences in behavior.

3.2.1 Motivations for Searching

According to a study conducted by Teevan et al. in 2007, 39% of search queries represent users trying to re-find previously viewed pages [133]. For cases when the webpage has changed and no longer contains the prior information, this would readily translate to a motivation for searching a web archive. Teevan conducted a prior study analyzing why general users want to interact with lost webpages [132]. She found that for one third of the queries the information had been removed. If web archive search were readily available and easy to use, these users might have next gone to a web archive to locate their missing information.

Broder [25] broadly defined search tasks with their motivations as navigational, informational, or transactional. Past versions of websites cannot be used to make transactions. Navigational tasks would represent a user wanting to view information about a known item. If the URL of the webpage for that item is the same as the URL for the live web, the user could issue a URL query. Otherwise, the user would most likely need to issue a full text query to find the prior URL. Informational tasks would include aggregating information about a subject across a variety of sites, comparing information across sites to determine a conclusion, or learning about the factors to compare found information. Jiang et al. [76] labeled these tasks as known subject, interpretive, and exploratory. Informational tasks in web archives would undoubtedly require full text search.

3.2.2 Measuring User Behavior in Search

When using a search engine, users contribute two types of measurable behavior: submitting queries and clicking on matching pages [76]. Some of the tools researchers have available to measure these user behaviors include query logs and click logs. These logs are both examples of transaction logs [71]. Query logs and click logs may be combined [35] or separate [19]. A query log is a log generated on the search engine server that has a record of user queries and could also contain the query results as well [71]. A weakly anonymized query log will contain an identifier for each user and the time at which they made the query [3], while k-anonymization will group user queries together to more robustly protect users [35]. Query logs can provide information about more users and their actions than a lab study, but can only be used to approximate user motivations [133]. One type of query log is a click log, which includes the query and the user's associated clicks from the search engine results page. Click logs are used to estimate the relevance of a document to a user's query [141].

It is also worthwhile to determine how the document rankings correlate to the clicked pages. For some tasks, users will read search results sequentially, while for other tasks, they do not access search results in a linear fashion [76]. Different user tasks are also correlated to different numbers of clicked results [76]. Finally, it is well known that user clicks follow a power law, or otherwise stated, users rarely click on results beyond the first search result page [17].

Costa et al. [34] measured additional user actions when searching the Portuguese Web Archive, which supports full-text search. They found that users tend to submit short queries (5 or fewer terms), searched for less than one minute, and that many users actually issued the same queries. They also verified that users rarely go beyond the first page of search results.

3.2.3 User Difficulties While Searching

Users can be hindered by some core issues when searching and also browsing on the internet in general. One issue is referred to as "lost in hyperspace" [44]. This issue stems from the difficulties that users have in creating a mental model of networks as opposed to linear books. Users need to be able to orient and navigate themselves from one page to another, whether the next page is familiar or unfamiliar. When users have trouble accomplishing these tasks, it can be because they do not know where they are going, they know where they are going but not how to get there, or they do not know where they currently are in the system. Web archives introduce an even greater chance of becoming lost in hyperspace, if users cannot discern whether they are on the live web or the past web for example. The Portuguese Web Archive introduced a large static banner to combat the disorientation between the live and past web [36].

Another difficulty that users can encounter when searching occurs when the topic is controversial or argumentative. Because the pages matching the query contain controversial topics, the users searching for these pages may be searching for information on opposing sides of the debate, and search engines may return pages purposefully supporting that user's prior beliefs matching their previous browsing history [41] rather than returning results on both sides of the debate. Live web search engines support and implement this filter bubble [113], and past web search engines will need to evaluate the benefits and drawbacks of implementing such a system for its search users.

3.3 USER BEHAVIOR IN WEB ARCHIVES

A major contribution of this thesis is the creation of a system that allows for users to search for and view changes to webpages. Users are able to view webpage changes in web archives. However, users behave differently on the live web and past web because of the different mental models that govern their actions. In order to understand how users behave in web archives, background information about the users must be taken into account. Web archive users have a variety of professions, experience using web archives, and motivations for using web archives. These prior experiences shape users' mental models, give rise to varied difficulties, and result in different measurable access patterns.

3.3.1 Types of Web Archive Users

Users of web archives have a variety of professions. The IIPC Access Working Group [65] identified additional professional users including researchers studying change over time in computational and humanities contexts, professionals investigating data evolution such as in tourism or real estate, and lawyers in civil trademark cases and patent cases. Journalists are another group of web archive users, however, there has been no prior analysis of journalists' use of web archives [111]. Ras et al. [119] identified journalists, academic professionals in the humanities, public institutions, website owners, and the general public as users of web archives. They also identified users by their motivations, such as professional use, academic use, and private use. Additional use cases for web archives include computational and historical use by researchers, technicians repairing legacy technology by accessing deleted manuals, accessing historical text, data, and media, legal use, multilingual use, and summarizing collections through storytelling [56]. Each of these users brings a different model of access, so the users' expectations of the system should match their task.

Beyond profession, web archive users range from novices to experts. Novice users may or may not have even visited a traditional web archive before [119]. Expert users use web archives for both personal and professional purposes [119].

3.3.2 Mental Models of Web Archives

When users encounter a new system for the first time, they make a mental model of the system based on their prior mental models of other real world or computer based systems. One mental model a user might use for web archive search is a library catalog. A library catalog does not contain full text for searching, but rather hand curated or machine curated metadata. One implication of this mental model would be that a user might expect web pages to be categorized by topic similar to the Dewey Decimal system. A negative of this mental model is that catalog search is limited compared to full text search.

A second mental model that a user might use for web archive search is that of a library archive. Library archive contents are typically put into context by humans and then summarized. This mental model would match Jackson et al.'s [70] users who want to view every archived web page that matches their query in chronological order and in context to determine if it is relevant. The biggest drawback of such a system is the time it would take humans to summarize the information.

Finally, a mental model that many users might choose for web archive search is that of live web search. Live web search engine results pages summarize the page or highlight relevant snippets.

They also only include each URL one time in the results. The benefit of this mental model is that live web search is extremely familiar to users. The drawback of this mental model is that live web search does not include the same URL more than once. Web archives have many similar or relevant copies of the page at a URL that may evolve over time and change their relevance. Figuring out if or how to present these duplicates in a meaningful way is an open problem in web archive search.

3.3.3 User Difficulties in Web Archives

Developing a mental model of web archives and learning how to navigate the past web are not trivial tasks for many users. As recently as 2019, Abrams et al. [1] found that users had difficulty distinguishing between whether they were on the live web or past web. They also found that users' lack of understanding of web archives hindered their success as much as an ineffective user interface. Full-text search is therefore an advanced feature that will only benefit users with a strong understanding of the past web.

Abrams et al. [1] conducted a user study on the current search implementation for Archive-It. They found that users have trouble distinguishing when they have navigated from the past web to the live web. The users wanted to be able to integrate the Archive-It search interface with other search interfaces, such as using the browser page search, using a live web search engine like Google, and using another past web gateway like the Internet Archive Wayback Machine. Users want to search with date filters, but this capability was not possible due to how dates were indexed in the system at the time. Users also want a way to search for top level web pages.

Cruz et al. [36] conducted iterative user studies while constructing the search interface for the Portuguese Web Archive. Users had trouble distinguishing between the live web and the past web. Adding a distinctive frame around the replay of past web pages helped to mitigate this problem.

Users also had trouble distinguishing between URL search and full text search, so there is one search box that detects which task the user is attempting. Users wanted to filter by date, but had trouble when the datepicker required them to specify a small granularity, such as a day, when they were trying to search for a larger granularity, such as month or year. Additionally, not all users even noticed the temporal information associated with the archived pages in the search engine results page, even when this information was displayed directly underneath the title of the page.

Ras et al. [119] conducted a user study using the web archives of the National Library of the Netherlands. They found that users want full text search capabilities in web archives. The users wanted to be able to integrate the KB search interface with other search interfaces, such as using the browser page search, using the browser search bar, using a live web search engine like Google, and using another past web gateway like the Internet Archive Wayback Machine. Users also want a way to search for home pages, and to always be able to navigate to a home page at a similar datetime when a subpage is returned as a search result. Users want to search with date filters, but this capability was not possible with the current search implementation. When using a past web replay engine with navigation buttons to display earlier and later versions of the same page, users preferred using these navigation buttons to using the TimeMap with the list of page versions. Also, users were confused about how these buttons operated: they expected that only different versions would be shown, rather than every crawl with possibly identical content to the previous crawl counting as a separate version. Users also do not always type the full URL of a site when searching, so the capability of the search engine to add a scheme to a URL is necessary.

Finally, Alam et al. [8] surveyed web archive navigation banners, highlighted usability issues, and suggested changes. A common change suggested by both Cruz et al. and Alam et al. is to keep the context of the page through the banner and branding without compromising the quality of the replay, for example by blocking part of the page.

3.3.4 Types of User Behavior in Web Archives

AlNoamany et al. [13] found distinct access patterns for human users in web archives. About one-third of users access only one memento. These users are referred to as dip users. AlNoamany et al. also found that these sessions originated from links to mementos on popular websites such as Wikipedia, Reddit, and Snopes. Wikipedia and the Internet Archive have a partnership to change broken links to reference into their archived versions, which benefits both organizations[58]. Another one-third of users access multiple mementos from different pages at the same time in the past. These users are referred to as dive users. Jatowt refers to this behavior as vertical browsing. This browsing style is shown in Figure 11. The remaining users access web archives mostly in mixed patterns, while about 4% of users access the same page at different times in the past. These users are referred to as slide users. Jatowt refers to this behavior as horizontal browsing, for example along a horizontal time line. This browsing style is shown in Figure 10. Users who access the same page at different times may have the same motivations as users who revisit pages, such as trying to view information that has been removed from the page [132].

Jayanetti et al. [74], [75] built upon AlNoamany's work and found that half of users are accessing only one memento. Only about 15% are browsing vertically, and most of the remaining users are accessing archives in mixed patterns. The amount of users browsing horizontally remained constant at about 4%. Jayanetti et al. also found that users usually view mementos from the current year.

Users have a hard time browsing horizontally. If the user is looking for a specific change, they must view many versions of the page in order to find the version they are seeking. Most users

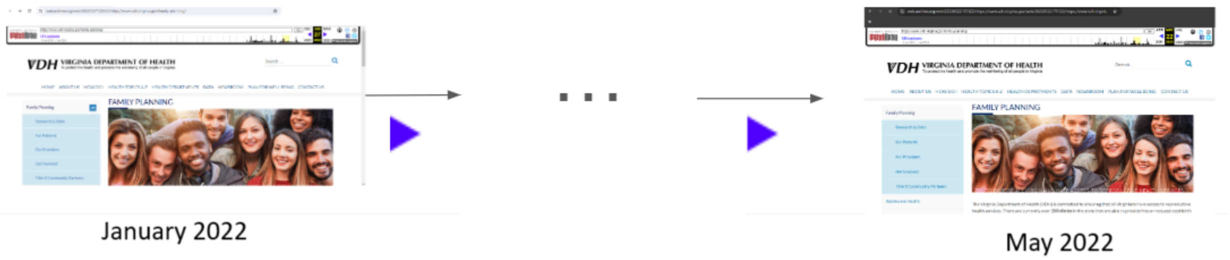
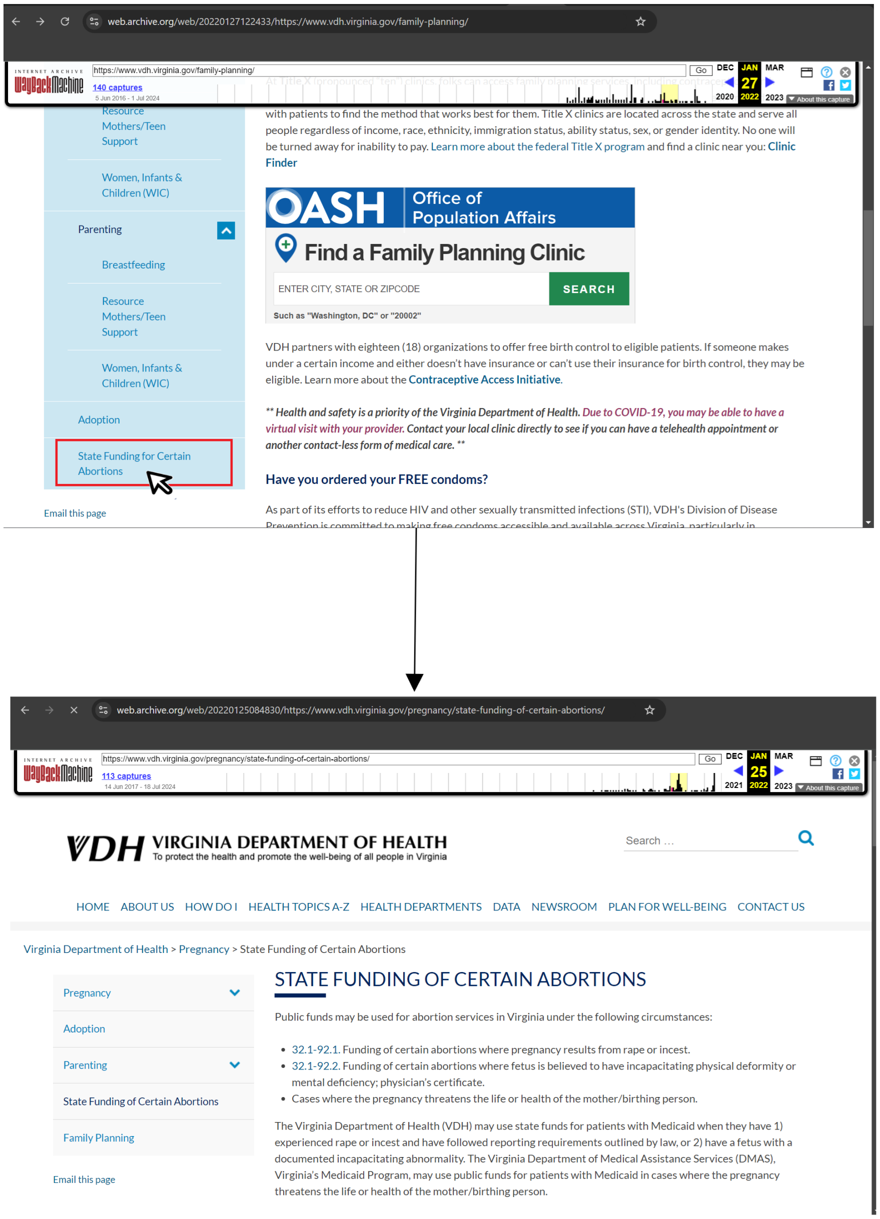

VDH VIRGINIA DEPARTMENT OF HEALTH
FAMILY PLANNING
January 2022
VDH VIRGINIA DEPARTMENT OF HEALTH
FAMILY PLANNING
May 2022

Figure 10. Slide or horizontal browsing of the Virginia Department of Health Family Planning webpage between January and May of 2022.

would likely go sequentially through the versions. A user with a knowledge of computer science algorithms may view pages in a binary search fashion. Either of these options is inefficient and requires trial and error because there is no way currently to search for when a term appeared or disappeared from a page.

Costa et al. [34] analyzed query logs in the Portuguese Web Archive, which includes a full-text search implementation. They found that users usually view mementos from the first year the site was archived. It is possible that Jayanetti and Costa found different results due to the distribution of mementos in different archives by year, or because different categories of users view mementos from different times. The search implementation includes a date filter. About a quarter of users filtered the date of the search results using this interface, while about 4% of users typed an implicit temporal query instead. Users can search by either terms or URL from the same search box, and two-third of users search by terms while one-third search with a known URL.


web.archive.org/web/20220127122433/https://www.vdh.virginia.gov/family-planning/
INTERNET ARCHIVE WaybackMachine
https://www.vdh.virginia.gov/family-planning/
Go
140 captures
5 Jun 2016 - 1 Jul 2024
DEC JAN MAR
27
2020 2022 2023
About this capture
Resource Mothers/Teen Support
Women, Infants & Children (WIC)
Parenting
Breastfeeding
Resource Mothers/Teen Support
Women, Infants & Children (WIC)
Adoption
State Funding for Certain Abortions
Email this page
OASH Office of Population Affairs
Find a Family Planning Clinic
ENTER CITY, STATE OR ZIPCODE
SEARCH
Such as "Washington, DC" or "20002"

with patients to find the method that works best for them. Title X clinics are located across the state and serve all people regardless of income, race, ethnicity, immigration status, ability status, sex, or gender identity. No one will be turned away for inability to pay. Learn more about the federal Title X program and find a clinic near you: Clinic Finder

VDH partners with eighteen (18) organizations to offer free birth control to eligible patients. If someone makes under a certain income and either doesn't have insurance or can't use their insurance for birth control, they may be eligible. Learn more about the Contraceptive Access Initiative.

*** Health and safety is a priority of the Virginia Department of Health. Due to COVID-19, you may be able to have a virtual visit with your provider. Contact your local clinic directly to see if you can have a telehealth appointment or another contact-less form of medical care. ***

Have you ordered your FREE condoms?

As part of its efforts to reduce HIV and other sexually transmitted infections (STI), VDH's Division of Disease Prevention is committed to making free condoms accessible and available across Virginia, particularly in

web.archive.org/web/20220125084830/https://www.vdh.virginia.gov/pregnancy/state-funding-of-certain-abortions/
INTERNET ARCHIVE WaybackMachine
https://www.vdh.virginia.gov/pregnancy/state-funding-of-certain-abortions/
Go
113 captures
14 Jun 2017 - 18 Jul 2024
DEC JAN MAR
25
2021 2022 2023
About this capture
VDH VIRGINIA DEPARTMENT OF HEALTH
To protect the health and promote the well-being of all people in Virginia
Search …
HOME ABOUT US HOW DO I HEALTH TOPICS A-Z HEALTH DEPARTMENTS DATA NEWSROOM PLAN FOR WELL-BEING CONTACT US
Virginia Department of Health > Pregnancy > State Funding of Certain Abortions
Pregnancy
Adoption
Parenting
State Funding of Certain Abortions
Family Planning
Email this page

STATE FUNDING OF CERTAIN ABORTIONS

Public funds may be used for abortion services in Virginia under the following circumstances:

- 32.1-92.1. Funding of certain abortions where pregnancy results from rape or incest.
- 32.1-92.2. Funding of certain abortions where fetus is believed to have incapacitating physical deformity or mental deficiency; physician's certificate.
- Cases where the pregnancy threatens the life or health of the mother/birthing person.

The Virginia Department of Health (VDH) may use state funds for patients with Medicaid when they have 1) experienced rape or incest and have followed reporting requirements outlined by law, or 2) have a fetus with a documented incapacitating abnormality. The Virginia Department of Medical Assistance Services (DMAS), Virginia's Medicaid Program, may use public funds for patients with Medicaid in cases where the pregnancy threatens the life or health of the mother/birthing person.

January 2022

Figure 11. Dive or vertical browsing around January 2022 of two pages on the Virginia Department of Health website.

3.4 APPROACHES FOR IMPLEMENTING WEB ARCHIVE SEARCH

In this thesis, we implement a search system that utilizes web archives in order to support users who want to view changes on webpages over time. Web archive search is markedly different from search on the live web. In this section, we examine the various approaches to web archive search. We show that none of these approaches can solve the problem of searching for changes in webpages.

3.4.1 The Need for Specialized Indexing

Some researchers have worked towards effective implementations for finding information held in web archives without indexing. Kanhabua et al. [86] chose to leverage live web search engines and align the results with web archive holdings. The shortcoming to this approach is that live web search engines prioritize freshness, so older and deeper web archive holdings cannot be found with this method. Another non-indexing approach was taken by Alkwai et al. [11] by prioritizing users' information seeking intentions to find relevant content rather than any one specific page. Their method functions by finding a similar page in a web archive to pages that have been deleted on the live web. This approach is highly relevant to improving users' ability to navigate between versions of webpages when the URI changes, as discussed in Chapter 4. However, because this system only uses the URI rather than full-text search, it will not be able to make pages with poorly named URIs discoverable.

Other researchers have acknowledged the need for indexing to truly make a majority of web archive holdings discoverable. Beyond the standard indexing workflow, additional considerations are necessary for effectively indexing web archive contents. In particular, because web archives are a type of versioned document collection, the specific way in which each version is indexed is vital to discovery of changes between versions. Berberich et al. [21] developed a temporal coalescing framework that orders all versions of a document over time and then assigns each document a validity range rather than a single datetime. For example, the Virginia Department of Health (VDH) Family Planning webpage was crawled by the Internet Archive on January 27, 2022. In the model using the second of the crawl, this timestamp is only valid for one second on January 27, 2022. The next crawl occurred on January 31. Using the validity range model, the first crawl is valid from the second of crawl on January 27 to the second before the crawl on January 31. The validity allows for the amount of document change to be quantified between versions. Considering again the VDH Family Planning page, all of the versions crawled between January 2022 and May 2022 have the same content. The coalesced validity range for the entire page could then be considered as from January 27 to May 22. This coalescing idea can be applied at the page level or at the term level, and at the page level the idea can be applied with a similarity threshold which results in various amounts of index efficiency. While documents with a certain similarity threshold were combined in Berberich's index to save space, the separate document versions were not combined in the search engine results page, and searching by change was not possible. While temporal date ranges were not implemented when Berberich developed the validity range framework, date range fields are now a standard part of Apache Lucene.

3.4.2 Approaches to Search Engine Results Pages

A few temporal search engines have investigated how to present multiple versions of a page in the search results. Jackson et al. [70] chose to include every relevant version of a web page in their search results page ordered by time. While there is a benefit to showing each individual page, in that grouping web page versions would hide change over time, it also has drawbacks. Kiesel et al. [87] performed a qualitative evaluation on their personal web archiving localized search system, and found that including every version of a web page introduced clutter into the search results pages that affected usability. An example of cluttering is shown in Figure 2A (Chapter 1). Major [102] also identified repeated captures of the same URI in search results as problematic.

Since displaying every matching version of the page clutters the search results, it is worth examining systems that group all versions of a page into one result. Melo et al. [104] added a backend parameter to the Portuguese Web Archive search system to limit the number of versions displayed on a search results page. Jones et al. [80] advanced the presentation of social cards for archived web pages, which are similar to grouped search results. The social cards include explicit information about the date of the crawl for the memento shown, along with links to additional captures for that page. However, an open question when applying this method to web archives would be how that particular version of the page is chosen from many other versions that match the query. While avoiding clutter is the main benefit to grouping, the drawback is that changes between versions are not visible. Neither of these solutions, grouping or individual results, will help users view change when searching web archive holdings.

In all of these systems, searching by change is not possible, and the only way to view the exact changes between versions is by manual inspection. In order to capitalize on using web archives to examine change over time, temporal search interfaces must be extended to support searching for term and phrase changes. These search interfaces will also need to effectively show the changes to the users. Finally, the search tool will need to be able to find meaningful changes in a data set of webpages relevant to existing digital humanities research.

3.4.3 User Queries and Web Archives

In addition to the user interface, the queries a user makes in web archives are also important to analyze in order to make sure the system can support relevant queries.

ORCAS, a click-query dataset described in Chapter 2, is a part of the Microsoft Machine Reading Comprehension (MS MARCO) dataset collection [19]. Researchers have used MS MARCO in conjunction with web archives to examine how changes on the web affect information retrieval systems' accuracy. Fröbe et al. [54] produced a more accurate retrieval model for MS MARCO by using web archives. MacAvaney et al. [99] used web archives to create a dataset similar to MS MARCO for webpages from 2006. Soboroff [128] used web archives to study how document change impacts retrieval models, specifically studying the change of the 2004 GOV2 collection. Soboroff assumed that any change in a document would render it not relevant for the query associated with the previous version of the page, and that a collection's relevance tends to decay as the documents change.

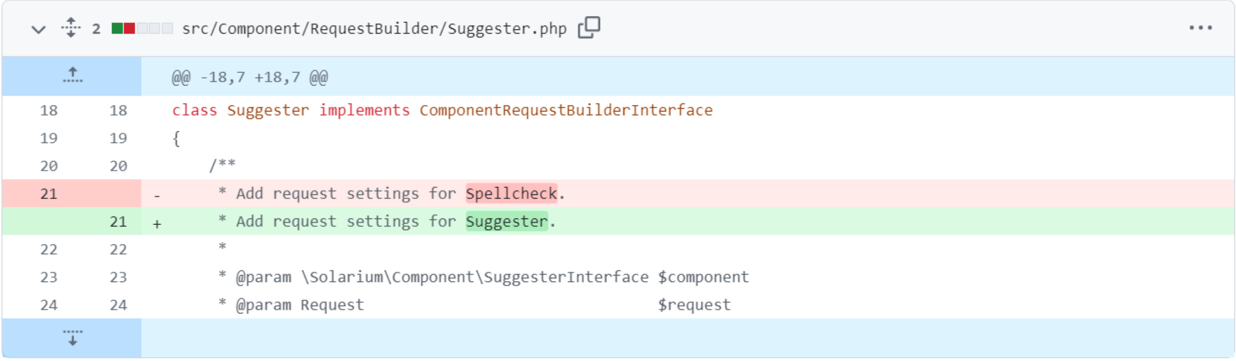
The ephemeral nature of the web affects information retrieval model accuracy. Bajaj et al. [19] acknowledge that many pages in the MS MARCO dataset have changed or are no longer available on the live web. Nguyen et al. [107] found that only considering a single crawl date for each page led to inaccurate retrieval models. Another problem that affects the accuracy of web archive information retrieval models is the presence of duplicate captures in web archives [107].

3.5 USER BEHAVIOR WHEN USING TOOLS TO VIEW CHANGE

Some of the major contributions of this thesis are a search engine results page that shows differences as an integral part of the result, along with two tools to help users view change. In order to successfully create these artifacts, we need to understand how users behave when using tools to view change. In this section, we summarize user studies that differentiate between design choices for viewing and finding changes in various digital settings.

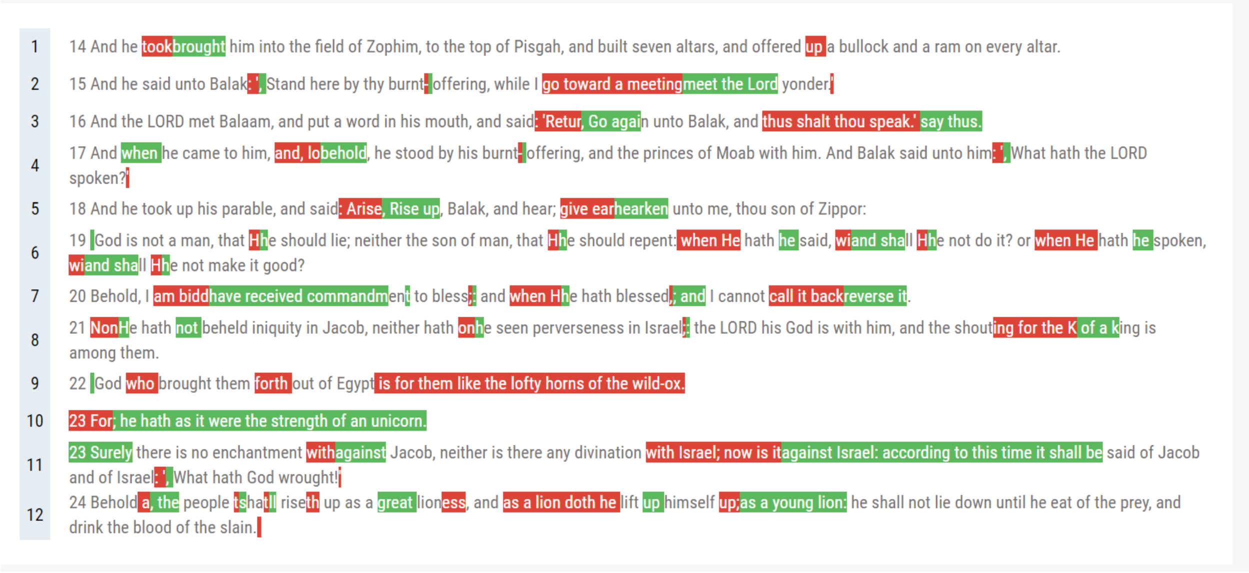
3.5.1 Data Comparison Presentation

There are three main ways that difference calculation outputs can be presented to the user. First, only the differences in a line-by-line format can be shown, as shown in Figure 12. This format is successful when there is a lot of commonality between the two versions, and the user wants to see only the differences. Next, a side-by-side comparison can be shown, as shown in Figure 14. This allows the user to see what has changed and what has stayed the same, but it takes more time for the user to process the larger amount of information that they are being shown. The third way to view a difference is with a combined view, where deletions are shown in one style, and the subsequent insertion is shown right next to the deletion in another style. This is shown in Figure 13. While this view is more compact, if there are too many changes, it can be hard to read [42].



```
src/Component/RequestBuilder/Suggester.php
@@ -18,7 +18,7 @@
class Suggester implements ComponentRequestBuilderInterface
{
    /**
-     * Add request settings for Spellcheck.
+     * Add request settings for Suggester.
     *
     * @param \Solarium\Component\SuggesterInterface $component
     * @param Request                                $request
```

Figure 12. Line-by-Line diff view of Solarium code on Github. This view shows the user each change on lines placed sequentially.



14 And he tookbrought him into the field of Zophim, to the top of Pisgah, and built seven altars, and offered up a bullock and a ram on every altar.

15 And he said unto Balak: ',Stand here by thy burnt-offering, while I go toward a meetingmeet the Lord yonder.'

16 And the LORD met Balaam, and put a word in his mouth, and said: 'Retur, Go again unto Balak, and thus shalt thou speak.' say thus.

17 And when he came to him, and, lobehold, he stood by his burnt-offering, and the princes of Moab with him. And Balak said unto him: ',What hath the LORD spoken?'

18 And he took up his parable, and said: Arise, Rise up, Balak, and hear; give earhearken unto me, thou son of Zippor:

19 God is not a man, that Hhe should lie; neither the son of man, that Hhe should repent: when He hath he said, wiand shall Hhe not do it? or when He hath he spoken, wiand shall Hhe not make it good?

20 Behold, I am biddhave received commandment to bless;: and when Hhe hath blessed; and I cannot call it backreverse it.

21 NonHe hath not beheld iniquity in Jacob, neither hath onhe seen perverseness in Israel;: the LORD his God is with him, and the shouting for the K of a king is among them.

22 God who brought them forth out of Egypt is for them like the lofty horns of the wild-ox.

23 For; he hath as it were the strength of an unicorn.

23 Surely there is no enchantment withagainst Jacob, neither is there any divination with Israel; now is itagainst Israel: according to this time it shall be said of Jacob and of Israel: ',What hath God wrought!'

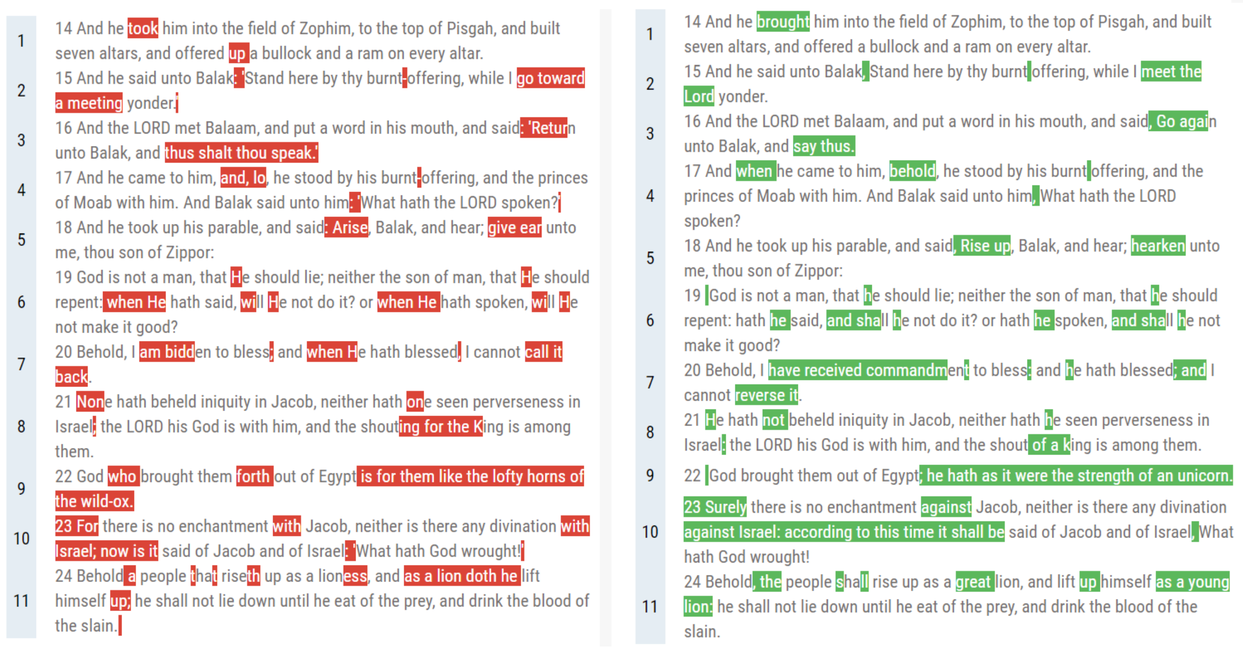
24 Behold a, the people tshatll riseth up as a great lioness, and as a lion doth he lift up himself up;as a young lion: he shall not lie down until he eat of the prey, and drink the blood of the slain.'

Figure 13. Combined diff view of The Book of Numbers Chapter 23, Jewish Publication Society vs King James Version, on countwordsfree.com/comparetexts. This view includes all lines together, but each individual change shown in context in the line.

3.5.2 Visualizing Changes

Gleicher et al. [55] assert that there are three ways to visualize differences in data: juxtaposition in small multiples, superimposition, and explicit derivation of the change. For visualizing differences in text, explicit derivation is not an option, since words cannot be subtracted from each other meaningfully. Thus, the static visualizations are either juxtapositions or superimpositions. Line-by-line and combined diffs are superimposition formats, while side-by-side is a juxtaposition format.

Rensink [121] explains that detecting change involves identifying that there is a change, along with identifying what has changed and where. Careful design choices are necessary to help users



14 And he took him into the field of Zophim, to the top of Pisgah, and built seven altars, and offered up a bullock and a ram on every altar.
15 And he said unto Balak: 'Stand here by thy burnt-offering, while I go toward a meeting yonder.'
16 And the LORD met Balaam, and put a word in his mouth, and said: 'Return unto Balak, and thus shalt thou speak.'
17 And he came to him, and, lo, he stood by his burnt-offering, and the princes of Moab with him. And Balak said unto him: 'What hath the LORD spoken?'
18 And he took up his parable, and said: Arise, Balak, and hear; give ear unto me, thou son of Zippor:
19 God is not a man, that He should lie; neither the son of man, that He should repent: when He hath said, will He not do it? or when He hath spoken, will He not make it good?
20 Behold, I am bidden to bless; and when He hath blessed, I cannot call it back.
21 None hath beheld iniquity in Jacob, neither hath one seen perverseness in Israel; the LORD his God is with him, and the shouting for the King is among them.
22 God who brought them forth out of Egypt is for them like the lofty horns of the wild-ox.
23 For there is no enchantment with Jacob, neither is there any divination with Israel; now is it said of Jacob and of Israel: 'What hath God wrought!'
24 Behold a people that riseth up as a lioness, and as a lion doth he lift himself up; he shall not lie down until he eat of the prey, and drink the blood of the slain.'

14 And he brought him into the field of Zophim, to the top of Pisgah, and built seven altars, and offered a bullock and a ram on every altar.
15 And he said unto Balak, Stand here by thy burnt offering, while I meet the Lord yonder.
16 And the LORD met Balaam, and put a word in his mouth, and said, Go again unto Balak, and say thus.
17 And when he came to him, behold, he stood by his burnt offering, and the princes of Moab with him. And Balak said unto him, What hath the LORD spoken?
18 And he took up his parable, and said, Rise up, Balak, and hear; hearken unto me, thou son of Zippor:
19 God is not a man, that he should lie; neither the son of man, that he should repent: hath he said, and shall he not do it? or hath he spoken, and shall he not make it good?
20 Behold, I have received commandment to bless: and he hath blessed; and I cannot reverse it.
21 He hath not beheld iniquity in Jacob, neither hath he seen perverseness in Israel: the LORD his God is with him, and the shout of a king is among them.
22 God brought them out of Egypt; he hath as it were the strength of an unicorn.
23 Surely there is no enchantment against Jacob, neither is there any divination against Israel: according to this time it shall be said of Jacob and of Israel, What hath God wrought!
24 Behold, the people shall rise up as a great lion, and lift up himself as a young lion: he shall not lie down until he eat of the prey, and drink the blood of the slain.

Figure 14. Side-by-Side diff view of The Book of Numbers Chapter 23, Jewish Publication Society vs King James Version, on countwordsfree.com/comparetexts. This view shows one version of the text with the deletions and another version of the text with the additions.

spot changes. Change detection is processed in visual short term memory. Healey et al. [60] state that preattentive processing is a core technique employed during design to aid in rapid detection of changes, such as in estimation situations.

Researchers are able to measure change detection by recording the reaction time for different interstimulus intervals between change states. Users are able to recognize that a change has happened with a short interval. However, users need a longer interval to recognize the difference between the previous state and the next state [117]. Rensink [121] names this distinction as dynamic change, the act of changing in progress, versus completed change, identifying that a change has occurred.

Rensink [121] identifies multiple animation strategies to use to present changes to users. Three strategies, gap-contingent, saccade-contingent, and blink-contingent, use time to separate the versions with changes. Gap contingent uses a larger set amount of time, while saccade and blink contingent capitalize on the natural timing of eye movements. Two strategies use movement: shift-contingent and cut-contingent. Shift contingent shifts the scene while cut contingent presents a completely different perspective of the scene. Two strategies use distractors to highlight the changes: splat-contingent introduces distractors on purpose, while occlusion-contingent blocks the changing item. Finally, gradual change creates a transition over a few seconds between the initial and changed states. Each of these strategies has been studied to have benefits and drawbacks. Another animation strategy that can be used in conjunction with any of these strategies is repetition, where the animation is repeated a set number of times over a set time period, or until the user recognizes the change.

For viewing differences in the combined view, an alternative to static visualization is to animate each difference one-by-one. The benefit of this modification to the static combined view is that animations can help the user understand the context of the changes on the page, including where, what, and how much has changed [31].

Gur et al. [59] conducted a user study where they evaluated static presentation of changes versus animations. Users were divided into two groups based on their overall ability to identify change. In the upper group, there was no difference in results for static versus animated presentation of changes. However, in the lower group, static side-by-side presentation had a much faster change identification time than animations.

3.5.3 Juxtaposition Versus Superimposition

Lanna [94] created a tool for visualizing code differences. Lanna found that only showing differences, such as in a line-by-line format, was unpopular with users because they wanted to see the full context of the document before seeing the changes. Lanna also found that displaying changes side-by-side was less effective than a combined view when the combined view had a pop-out effect. The users were faster and more accurate in the combined view than in traditional software with side-by-side views. However, the results in this system were negatively affected when the changes were so extensive that horizontal scrolling was needed or vertical spacing was mismatched between versions. Lanna also created a hotkey that allowed the user to rapidly switch between versions, in essence creating a user-controlled animation of the changes. They classified changes as additions, deletions, or modifications.

Niederer et al. [108] created a system called TACO for visualizing changes in tabular data. The system allowed for visualizing changes over the lifetime of the table as well as changes between two specific versions. They used a color scale to indicate the amount of change, similar to the Internet Archive's Wayback Changes calendar view. They also classified changes as addition/deletion, merge/split, reorder, and content. An addition/deletion refers to new cells added, while content refers to changes in existing cells. Reorder means the cells were reordered without changing value. Finally, a merge/split indicates the table grew or shrunk in size. Webpages also are merged or split when they grow or shrink in size. To visualize the changes, they used the juxtaposition and explicit techniques identified by Gleicher et al. [55], but did not use the superimposition technique.

3.6 APPROACHES FOR IMPLEMENTING TOOLS TO VIEW CHANGE

In this thesis, we implement a search system that helps users find and view changes in webpages. This section examines the various approaches for tools to view change. We show that none of these approaches can solve the problem of viewing changes in webpages in the context of a search system.

3.6.1 Data Comparison Algorithms

Users have been capable of using a computer to calculate and view the differences between two text files since the Unix diff program was developed by AT&T labs in the 1970s [63]. The diff is calculated using algorithms developed for the longest common subsequence problem [62]. PHP can be used to calculate and view a diff [30], using a port of the Python difflib implementation of the Gestalt Pattern Matching algorithm.[2]

While these implementations are moderately successful at calculating the differences between the extracted text of a web page, a different algorithm is needed to calculate the differences between pages with markup. AT&T Labs created a prototype HtmlDiff program in 1996 [42]. Zoetrope used the Document Object Model (DOM) to extract element paths and identify a best matching element in a different version of a page [5].

3.6.2 Approaches for Static Visualization Tools

Other types of versioned document collections besides web archives have led to tools that allow for comparison for multiple document versions. For example, Henley et al. [61] developed a tool called Yestercode that allows programmers to use a slider to navigate between different

[2] `https://docs.python.org/3/library/difflib.html`

versions of their code and display the differences between consecutive versions. The collaborative document writing tool DocuViz [139] aims to help users visualize how documents evolved, and Perez-Messina et al. [114] developed a tool to visualize the origin of text segments in collaborative documents.

3.6.3 Approaches to Animation Tools

Thumbnail summarization can show visual, large scale changes on webpages [14]. A tool using this technique is TMVis [98]. This thumbnail approach utilizes web archives, and can be shown as a small multiples chart or played in an animation.

One system that animates changes in text as a complement to displaying the changes side-by-side is Diffamation [31]. This system shows all text changes between two versions of a document in parallel animation with navigation in order to help the user understand where, what, and how much text has changed. Another system that uses text animation to show differences between document versions is SlideDiff [40]. SlideDiff shows changes to text and media in slide presentation versions, even showing a simulated mouse cursor in order to draw the user's attention to the position of each of the changes and help the user infer the intent of the editor due to the animation appearing more life-like.

These goals are also applicable to viewing and understanding edits on webpages. Jatowt et al. [73] created the Journey to the Past framework that included a browser that allowed users to search for changes in a web page across web archives and animate the differences that matched the query terms. The interface for this browser is similar to video replay, with buttons to help the user navigate between versions and control the animation replay. Because this framework does not include indexing, searching for a changed term or phrase must be done one page at a time, rather than at the collection level. Adar et al. [5] developed the Zoetrope system to allow users to search individual web page versions for terms and show the differences over time in a stop-motion style of animation. This system relied on local crawling, and the queries relied on closest matches in the DOM rather than indexing; integrating full-text search with Zoetrope's visualization capabilities was identified as future work. Teevan et al. [134] developed DiffIE, a tool that allows users to view changes since they last visited a page highlighted. The changes were categorized into four types: addition, deletion, movement, and change. A change is when the parent HTML node has the same number of children but the text of one child is different, while additions and deletions are detected entirely by child count differences between versions.

3.7 THE ETHICS OF SEARCHING DELETIONS

In this thesis we propose a new search system. An essential part of proposing a new computational tool is to consider how it can be misused and what the consequences of misuse are. Below, we provide context for ethical concerns in web research. Then, we summarize the ethical concerns of web archiving and use them to create a framework that web archive collection curators can use when determining if full text search is appropriate for their collection.

3.7.1 Ethics and the Web

Three major areas for ethics concerning the Web are data collection, privacy, and representation. These issues are relevant to both the live web and the past web. A lack of ethical data collection results in mistrust between users and the organization. A lack of data privacy results in users opting their data out of web archives, which hinders public access to historical information. A lack of representation presents a skewed version of history, which leads to inaccurate conclusions. In the context of search in web archives, mistrust and missing content make the service unusable. Upholding ethical data practices create a stronger product that will be of higher benefit to the users.

Ethical concerns for data collection on the web center around the legality of web scraping versus the rights of users to own their own data and consent to their data being scraped for fuzzily defined purposes. Proponents of web scraping follow laws like the Digital Millennium Copyright Act [33]. These proponents believe that any content that is copyright will be identified by owners and removed in time. Many, but not all, web archives follow this type of crawling policy [2], [18], [46]. Proponents of owner rights to data argue that scraping the web is laissez faire and organizations that scrape the web should take more care to observe data property rights implicitly held by the content creators [77]. These proponents prefer to work with datasets that are created to be open source, which ensures that the content creators have an equal partnership in contributing to the datasets.

Regarding privacy, the right to be forgotten is a privacy right that has led to legislation in many places [97]. This idea pertains to after data has been collected, and has been made available for access such as in search engine results. Minor children are of particular concern in jurisdictions that have implemented this type of law. The implementation of this type of law has users opt out, rather than opt in. In Europe, search engines in particular make use of this opt-out implementation scenario. While search engines are commonplace in most users' lives, many users are unaware that their historical public web interactions exist in web archives, and are available for public viewing [100].

There is a debate over whether privacy or access should be more important when laws are non-existent. Proponents of privacy believe that users should opt in to having their data included, which protects the privacy of the owners of the data over public access. Proponents of access believe users should opt out of having their data included site by site, increasing the amount of data available for public access at the expense of data owners who do not know their data has been included. [93] Different web archives follow different privacy versus access policies. The Internet Archive follows an opt-out policy, while other web archives, like the UK Web Archive, follow an opt-in policy.

Representation is not equal on the live web. Groups with less representation are identified by gender, race, and socioeconomic status. Representation on the web roughly correlates to the groups impacted by the digital divide, which describes the phenomenon of Internet access differentials by group. It makes sense that groups with less access to the Internet would have less content on the Web. Web archives also have underrepresented groups, such as by ethnicity or by language, in their collection that line up with content representation on the live Web [10], [77], [85].

3.7.2 Summary of Ethical Concerns About Searching Deletions

Both the web and web archiving bring up ethical concerns about data privacy, data ownership rights, collection and representation. Full text search of web archives is not immune to these issues. Fiesler [49] argues that in the case of deleted content, archiving without consent is unethical. Participants believe that using their protected or deleted content for research is unethical, thus Fiesler et al. [50] further recommend not using data sets that contain deleted content. This viewpoint strongly supports the privacy of individual users, but it does not necessarily translate properly for how content in web archives from government entities, politicians, and corporations should be treated.

Many researchers have grappled with the ethics of web archiving itself, and full-text search adds another layer of ethical concerns on top of the problems already identified. Jo et al. [77] identify that consent, privacy, transparency, and representation all contribute to whether or not a dataset is ethical. Crawling the web is considered “laissez-faire data collection” because the origins of the data with respect to privacy and consent are not considered. Many larger web archives operate on an opt-out model, rather than an opt-in model. Content owners may withdraw their consent to have their websites removed from an archive, but they never actively gave consent in the first place. Additional concerns about consent arise when the website is an aggregate of multiple content creators, such as a web forum. The forum owner, along with the users who wrote the content, should all have withdrawal privileges. Privacy concerns especially apply to marginalized groups, such as minors. Some web archives do provide transparency through provenance data on captures. Finally, ensuring that marginalized groups have adequate representation is especially important in aggregate settings, such as machine learning as well as full-text search.

Jules et al. [85] examined ethical concerns in web archiving and identified consent as a major hurdle in web archiving. They also identified misinformation as problematic, since users examining collections with misinformation would need context in order to fully understand and verify the authenticity of the information. Representation of marginalized groups was also identified as an important issue. There is potential for commercial entities to exploit this data. Also, when archivists are members of the community they are archiving, it brings a different perspective to both consent and representation. The authors are clear that these practices apply to all aspects of web archiving, from collecting data to providing access to that data. Since full-text search is a very powerful way to access web archives, these ethical concerns are amplified if left unaddressed.

Lin et al. [96] also looked into the ethics of full-text search for youth websites like GeoCities. They stated that many people are comfortable with web archives because they provide “privacy by obscurity." Adding full-text search to web archives would eliminate this privacy. Because many of the content creators on GeoCities were minors, being able to find text and images from their youth would be exploitative. A second concern is related to consent: is it unlikely that users who used aliases would have consented to their content being archived, and it is also likely that they would withdraw consent because they could lose their anonymity with the technology available today.

Mackinnon [100] laid out a framework for conducting research in web archives, especially with marginalized groups. She also conducted user studies on how participants felt about re-experiencing their personal data in web archives. Her web archive research framework is centered around the idea that data belongs to people, and those people need to have discovery paths for their data along with sovereignty. Mackinnon's user studies confirm Lin et al.'s hypothesis that users would be uncomfortable with those images and text being available today. In her user studies, participants expressed anxiety about not knowing if their personal data was archived because it was hard to discover. Full-text search would provide a remedy to this problem but would amplify privacy concerns. Users were uncomfortable with the data from their youth being preserved in web archives. One user stated, "No one wants this online! No one wants these things in an archive!" These people had never consented to their data being archived many years ago and should have more control over it now. As we saw in the introduction to this blog post, women are often marginalized on the Internet. Mackinnon stresses that web archives simply reflect this phenomenon and advocates that stronger privacy measures are needed for women, trans, and non-binary people in web archives.

3.7.3 Ethical Framework Recommendations

Each author above gave recommendations for ethical guidelines when web archiving, and the recommendations are complementary.

Representation: Jo et al. [77], like Jules et al. [85], recommended that marginalized groups have agency and be a part of the archiving process. Jo et al., as well as Lin et al . [96], suggested that each project create a code of ethics. Lin et al. specified this further by suggesting that the original purpose and discovery goals of the website be considered when determining if full-text search is appropriate for the collection.

Consent and Transparency: Jules et al. [85] state that web crawls are not collected with explicit permission, and collections need to employ additional procedures when collecting data that involve clear-cut consent. Mackinnon [100] recommends informed consent crawls as well as soliciting donation of data. Jo et al. [77] advised that data sets should include provenance information to help identify consent concerns. The Internet Archive's Wayback Machine does include provenance information about its crawls and its Save Page Now feature, and other web archives should also adopt this practice. Documenting the Now has created a Python library to detect provenance information from Wayback captures called waybackprov.[3] Jackson also advocated for transparency through provenance [68]. Mackinnon notes that consent is necessary but not sufficient, and that participants need to be involved in the digital afterlife of their data, which ties back into representation. Regarding full-text search, only collections with vetted consent should have full-text search enabled. There are some smaller web archives that do operate on an opt-in model, so those web archives would have fewer barriers to providing ethical access via full-text search.

Privacy: Both Mackinnon [100] and Lin et al. [96] suggested anonymization as a form of privacy control. Mackinnon refers to this process as data orphaning, which allows the content to remain in the collection with permission from the author in exchange for anonymity. Jo et al. [77],

[3] `https://github.com/DocNow/waybackprov`

along with Mackinnon, advocate for active screening for private data during the curation process. Mackinnon also advocates for data sovereignty, especially related to youth data and the Right to be Forgotten. In order to incorporate these ideas into full-text search, the same right to erasure policies that apply to live web search engines like Google should apply to web archive search engines. Web archives should also develop a protocol for screening for data that should be kept private from search.

Mackinnon also developed care ethics scaffolding. For representation, she emphasizes that the research or product need to benefit the communities that produced the data, and that active measures must be taken to make sure that the communities are not taken advantage of. For consent, the people who created the data must explicitly give permission for their data to be archived and kept. Implementing these policies for web archives with full-text search will mean having the communities who created the data as active stakeholders in the development of the search tool.

3.7.4 Ethics Are Not Absolute

For some web archive collections, it is straightforward to resolve ethical concerns. Webpages like GeoCities, which include sensitive content created by minors, are not appropriate for full-text indexing. US federal webpages are appropriate for full-text indexing because the publications of the US government are public documents that are not entitled to any privacy. What about everything in between these extremes? When do privacy rights prevail, and when do other factors hold more influence?

Caplan-Bricker [28] highlighted the ethical dilemmas that web archive researchers face. The article recounts how Bergis Jules chose to preserve and publish a dataset of deleted tweets because of their perceived historical value, rather than remove the tweets as required by Twitter's developer terms of service at the time. Maddock et al. [101] chose to remove deleted tweets from their dataset, even though it directly hindered their research. They also likened deleting a tweet to withdrawing consent. Web archives contain a large amount of deleted content, which adds to the ethical complexity of making this content discoverable via full-text search. It is possible to index the changes on web pages, but it will not be appropriate to do this for every collection. The collection curators will need to weigh the historical importance of the collection versus the privacy rights of the content creators.

Politicians are afforded less privacy than ordinary citizens. One reason for this is that public figures have a lower expectation of privacy by their own career choice. A second reason for this is that the enfranchised public has a right to know how politicians' publicized political beliefs have changed over time. Another group of people with a lower expectation of privacy are people who have committed crimes. These people delete publicly posted content to hide evidence of illegal activity and motive, which is significantly different than a typical citizen updating their webpage. Curators of these types of web archive collections will need to delineate between content relevant to public knowledge and personal content with a typical expectation of privacy.

Sometimes activists will even archive evidence of a crime in progress, which is what happened with the January 6 United States Capitol attack. Rioters did not consent to having their social media posts about the attack archived, but these archived posts were later used for suspect identification as well as motive for sentencing. Proactive archiving preserves evidence that can be used by law enforcement for justice, but Chapman [29] showed how it can also lead to vigilantism such as doxxing. Doxxing is particularly problematic when it leads to misidentification. One solution could be to give law enforcement more access to these kind of archives than the general public to prevent vigilantism, but this is too idealistic: would they have prosecuted anyone if the riot had gone differently? The racial and economic biases in the US justice system also erode trust in law enforcement. Chapman recommends developing an ethical framework to address this issue for future archival crowdsourcing efforts. Related to this is hesitance to archive things associated with protests [85].

Archived web content from different sources may necessitate different levels of access. Some different levels of access might include internal/employee access, on-site access for researchers, remote access for researchers, public access without full-text search, and public access with full-text search. For example, Jules et al. [85] suggested that on-site access requirements could increase privacy by preventing data scraping. Many archives already implement different access levels. For example, the Library of Congress web archive provides many items available on-site only, while Archive-It has no such constraints. In other countries, these considerations are legal rather than ethical. The Non-Print Legal Deposit regulations in the UK specify that creators must opt-in to make their captures available off-site. The curators of each web archive collection should assess their collection contents to decide what level of access is appropriate. Users should be informed of the reasons for the chosen access level.

3.7.5 Ethics Outlook

Full-text search in web archives is becoming more common. It is important for web archiving organizations to create ethical frameworks before implementing this technology on a wide scale. If the underlying datasets are not created ethically, the search results will perpetuate algorithmic bias and further marginalize underrepresented communities. Using an ethical framework will strengthen users' trust. Curators will need to consider the ethical framework in conjunction with other important factors, such as historical significance, to decide on the appropriate privacy and access measures for their specific collections.

3.8 SUMMARY

In this chapter, we showed that none of the current approaches to web archive search can solve the problem of finding and viewing changes in webpages. First, we examined why users edit webpages, including neutral edits like keeping the page fresh, as well as non-neutral edits such as biased edits and edits that increase a page's standing with search engines. We also examined why users view changes, and learned that users want to examine deleted content and see how pages have changed over time in context. We use these ideas to guide our research in further specifying users tasks for viewing change using web archives (Chapter 4), and when evaluating changes in context (Chapter 6).

Next, we examined user behavior in search and web archives. We identified users of web archives, including journalists, and conduct a more detailed investigation of this group of users in Chapter 4. We described Teevan's methodology to examine why users want to view unavailable content, and will apply this methodology to determine more specific user tasks in Chapter 4. We identified that query logs can be used to estimate the relevance of a page, and use query logs to evaluate the retrievability of pages in Chapter 6. We presented AlNoamany's categorization of web archive users, and further investigate slide user tasks in Chapter 4. We detailed prior work by Ras et al. highlighting users' navigational difficulties, and validate these observations with our own investigation in Chapter 4. We identified the need for full-text search in web archives. We detailed why non-indexing solutions cannot take full advantage of web archives' holdings. We explain why the current ways of grouping search engine results pages in web archives are not feasible for users, and will contribute a new grouping method in Chapter 5.

We detailed the various approaches researchers have developed to view change, and examined systems that implement these approaches. We introduced Henley's concept of a sliding difference viewer, which we implement for web archives in Chapter 5. We also introduced various tools that animate webpages, but none are linked to a tool that can search across an entire corpus. We link our animated difference tool to our search tool in Chapter 5.

Finally, we detailed ethical concerns of searching deletions. We caution about full scale implementation of such a tool due to user data privacy rights.

CHAPTER 4

INVESTIGATING USER TASKS IN WEB ARCHIVES

Before developing improvements to current web archive search and navigation interfaces, the unmet needs of the users must be established. In this chapter, we conduct two formative investigations. First, we analyzed news articles containing references to web archives to identify the user tasks of journalists [51], one of which was viewing change over time. Second, we analyzed a web archive server access log to identify the tasks of users who view more than one archived version of a webpage over time.

4.1 USER TASKS OF JOURNALISTS

Different groups of users collectively have different levels of understanding about web archives. Ainsworth [7] demonstrated the emergence of web archives as evidence in journalism. The list of news articles that he presented as examples was a novel contribution at that time. Users with a strong mental model for the past web could benefit from advanced web archives features such as full-text search. What is the current mental model for web archives held by journalists, and how do web archives help journalists?

4.1.1 Collecting Articles That Reference Web Archives

Teevan [132] analyzed a set of web pages that were collected from a phrase search for "Where'd it go?" with the goal of understanding user behavior about re-finding. By searching for a salient phrase with a news aggregator, we used this same methodology to analyze how journalists currently perceive and use web archives.

We chose the search phrase “Wayback Machine” because of the Internet Archive’s prominence. We conducted one manual search for this phrase on May 11, 2022 using Google News.[1] Since this search yielded appropriate results, we automated the search with the GNews Python API.[2] In addition to specifying keywords, GNews queries can also restrict results to a customizable time span. Based on the article dates and distribution seen in the manual search, we chose a time span of 14 days. Each GNews query returns 100 results in JSON format. We executed this GNews query every two weeks starting May 25, 2022 and ending July 6, 2022.

When Teevan searched for “Where’d it go,” about 25% of the results contained usable information about re-finding behavior. Similarly, not all of the articles collected by searching for the phrase “Wayback Machine” included information about how journalists use web archives. Some of the articles included “wayback machine” as a colloquial phrase, while others were articles about web archives. Ultimately, 106 articles contained evidence showing how journalists use web archives.

We analyzed the articles after each search iteration and categorized them by user task, based on how the journalist was using the memento from the web archive. There are a variety of ways journalists use web archives when investigating stories, but not all user tasks were present in each result set. The distribution of articles coded with each task was also different between result sets; considering the count of articles coded with each task cumulatively resulted in a more accurate task distribution. The last set of articles collected on July 6 showed that the list of tasks and their distribution was stable.

[1] https://news.google.com/

[2] https://github.com/ranahaani/GNews/

4.1.2 User Tasks of Journalists

As shown in Figure 15, journalists use web archives to view unavailable pages, but they also frequently use web archives to view change over time. These users wanted to view term and phrase additions, deletions, and the associated content lifespan.

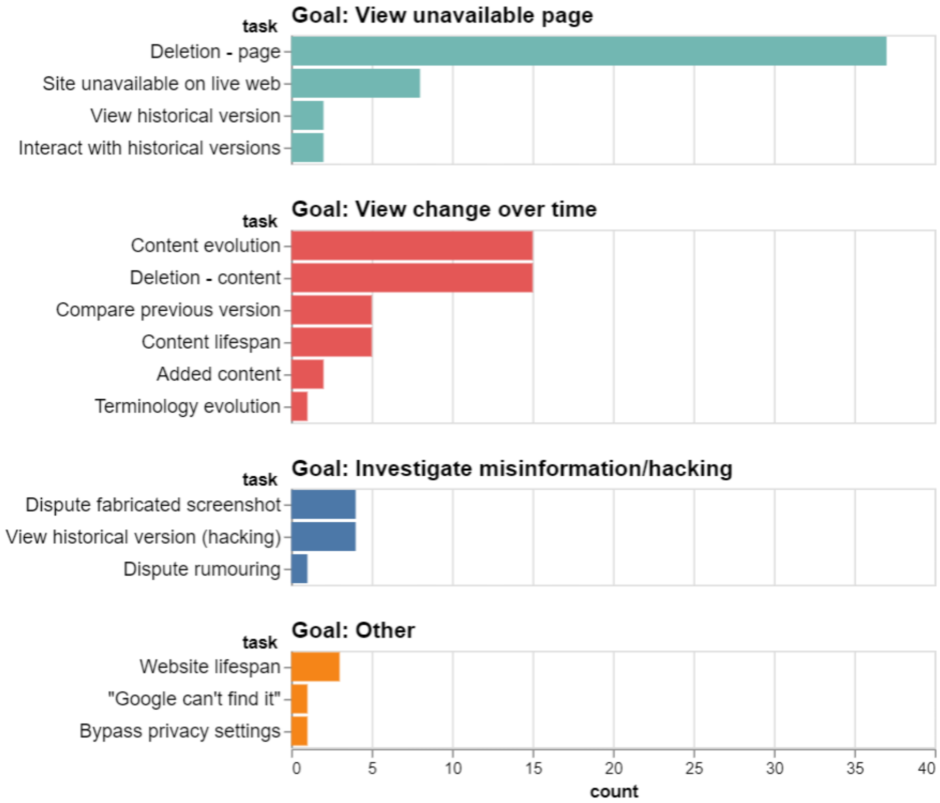


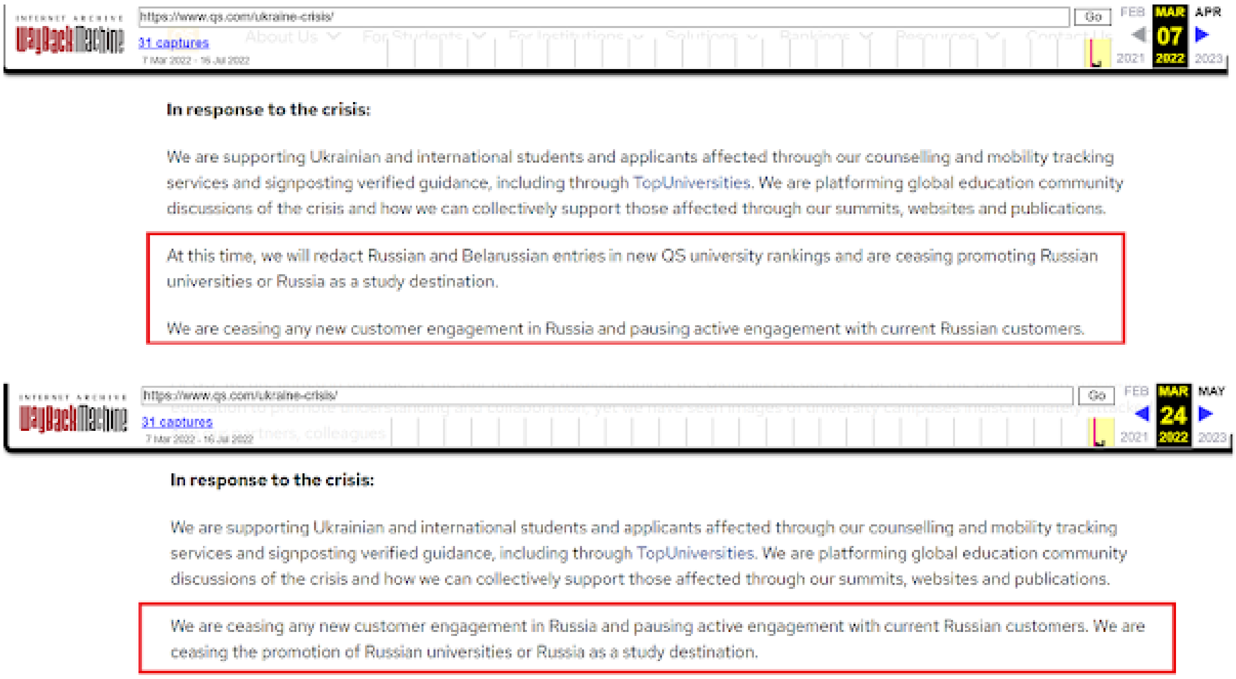
Figure 15. The two most common goals for journalists who use web archives as evidence in their articles is to view unavailable pages and to view page content change over time. ©2023 IEEE [52]

One way that journalists use web archives is to view web pages that are no longer available on the live web. When trying to view unavailable content, the most common task was viewing a single page that had been deleted. For example, these pages could have a 404 HTTP status code, while the parent site is still available on the live web. A more advanced use of web archives by journalists is to investigate how content changes on web pages over time. The top two tasks matching this goal are viewing how content has evolved on a page over time and viewing content (such as a sentence) that has been deleted from a page that is still available on the live web. Other tasks in this category include calculating the lifespan of certain content, determining when content was added to a page, comparing a previous version to the current version, and examining terminology evolution on a page.

Figure 16 shows one example of how journalists use web archives to view deleted content. InsideHigherEd reported on Quacquarelli Symonds changing their policy about excluding Russian universities from rankings in March 2022 [72].

4.1.3 The Need for Robust Links

While journalists have a strong mental model of web archives and use them to support a variety of tasks, they were less successful at linking to the mementos they found. Only two-thirds of the journalists successfully linked to a memento. The most common problem that journalists encountered when trying to link to mementos was that they included no links at all, either to the memento or to the resource on the live web. This problem was present in 20% of the articles. A third of these articles appeared on media sites that only allow internal links or don't have any links in articles, but the other articles contained links to external sites, so the majority of these articles do demonstrate users' difficulty in linking to mementos. Journalists also linked to pages on the live web rather than the archived versions that they used when writing their articles. Some journalists linked to time maps instead of individual mementos, and another journalist had a typo in their memento link.



In response to the crisis:

We are supporting Ukrainian and international students and applicants affected through our counselling and mobility tracking services and signposting verified guidance, including through TopUniversities. We are platforming global education community discussions of the crisis and how we can collectively support those affected through our summits, websites and publications.

At this time, we will redact Russian and Belarussian entries in new QS university rankings and are ceasing promoting Russian universities or Russia as a study destination.

We are ceasing any new customer engagement in Russia and pausing active engagement with current Russian customers.

In response to the crisis:

We are supporting Ukrainian and international students and applicants affected through our counselling and mobility tracking services and signposting verified guidance, including through TopUniversities. We are platforming global education community discussions of the crisis and how we can collectively support those affected through our summits, websites and publications.

We are ceasing any new customer engagement in Russia and pausing active engagement with current Russian customers. We are ceasing the promotion of Russian universities or Russia as a study destination.

Figure 16. Journalists use web archives to view deleted content. InsideHigherEd showed deleted content from Quacquarelli Symonds about Russian university rankings exclusions between March 7 and 24, 2022.

However, linking to one memento alone is not sufficient. In order to show how content has changed on a page over time, there needs to be a link to the previous version of the page, the current version of the page on the live web, and a snapshot of the current page. It is best practice to link to the live web version of the page, so that the original address is not obscured. Linking to the live web version of the page also makes sense in this case because the articles show that journalists have a strong disposition for referencing the original resource. By linking to the live web version, the journalists wanted to link to the version of the page at that moment in time. But because web pages change over time, some of the live web versions have content that does not match what is referred to in the articles. Content drift is the major reason why each journalist should have created a snapshot of the current page version at the time of their article. They could have used services such as the Internet Archive's Save Page Now tool or Archive.today.

Robust Links [84], [135] provide a way to link all of these versions of the page in a clear way. Robust Links contain a link to the original resource, a link to a snapshot of the current version, and the datetime of that snapshot. By including the datetime, alternate comparable snapshots can be located if necessary. Using Robust Links helps to ingrain a process for preserving functional hyperlinks: creating a snapshot of the live version of the page, as well as linking to resources that can be used to recover the information if the link becomes broken.

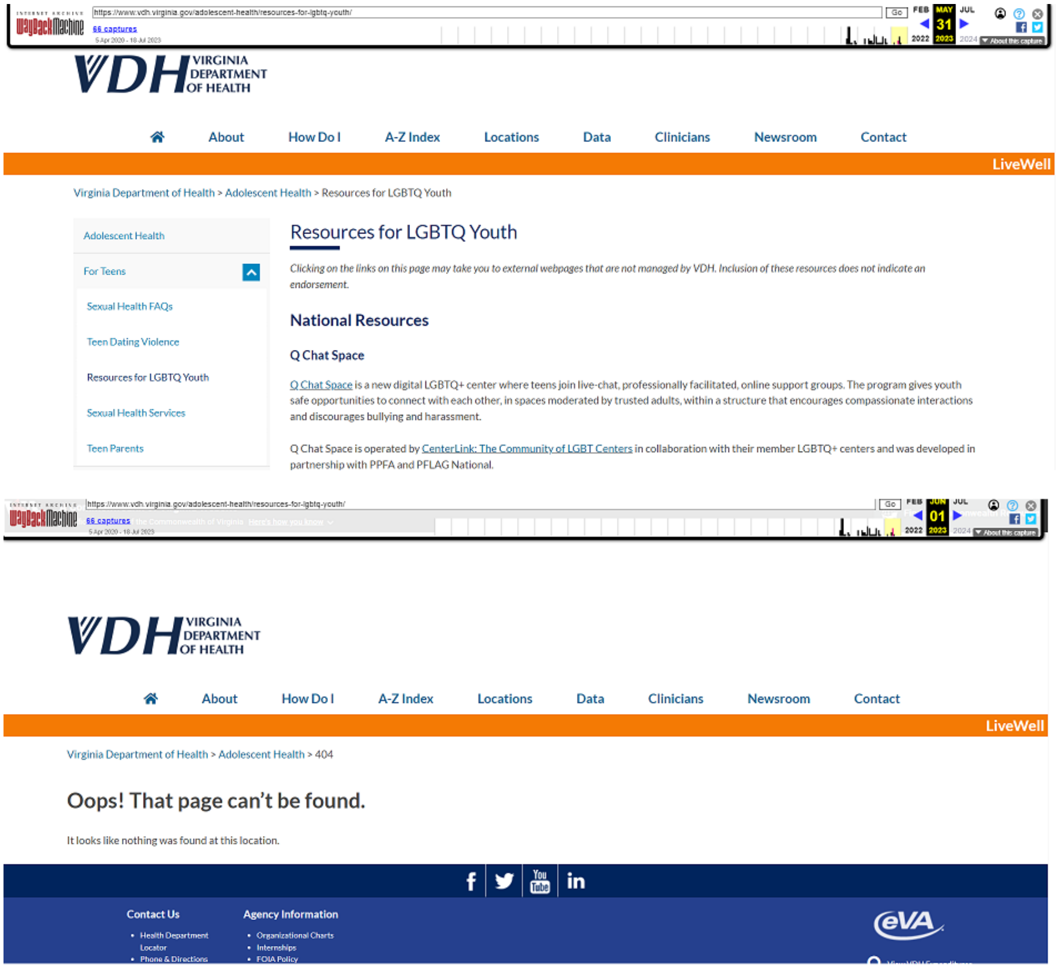
NYPost.com used web archives to compare the current version of the US-Taiwan fact sheet with the previous version from August 2018 in their article China blasts US over wording change on State Department's Taiwan website.[3] The journalist linked to both the live version of the page as well as the memento with the previous version. However, the live version of the page referenced in the article refers to the fact sheet updated on May 5, while the fact sheet has been updated again as of May 28, so the live version of the page no longer contains the content referenced in the article. The May 5 version of the page has been archived, so it is possible to view the page as it was at the time of the article being written. However, this version is not linked in the article, so viewing the current live version of the page would be confusing to readers. This example shows how content drift negatively affects linking to live resources, and why there should be links to both the live page as well as a snapshot of that version.

4.2 SLIDE USER NAVIGATION TASKS

There are other users besides journalists who use web archives to view change over time.

[3] nypost.com/2022/05/10/china-rips-us-over-wording-change-on-state-dept-taiwan-website

AlNoamany et al. [13] coined these users Slide Users. Slide users utilize web archives to view one page (URI-R) over time (mementos at multiple datetimes). Figure 17 shows a page that has been deleted, which may be of interest to a slide user.



Virginia Department of Health > Adolescent Health > Resources for LGBTQ Youth

Resources for LGBTQ Youth

Clicking on the links on this page may take you to external webpages that are not managed by VDH. Inclusion of these resources does not indicate an endorsement.

National Resources

Q Chat Space

Q Chat Space is a new digital LGBTQ+ center where teens join live-chat, professionally facilitated, online support groups. The program gives youth safe opportunities to connect with each other, in spaces moderated by trusted adults, within a structure that encourages compassionate interactions and discourages bullying and harassment.

Q Chat Space is operated by CenterLink: The Community of LGBT Centers in collaboration with their member LGBTQ+ centers and was developed in partnership with PPFA and PFLAG National.

Virginia Department of Health > Adolescent Health > 404

Oops! That page can't be found.

It looks like nothing was found at this location.

Figure 17. Slide users view multiple versions of a page over time. A web archive user viewing the LGBTQ youth resources page on the Virginia Department of Health website over time could see the page was deleted between May 31 and June 1, 2023.

Figure 18 shows how a user might navigate between page versions using the Internet Archive Wayback Machine’s banner navigation buttons.

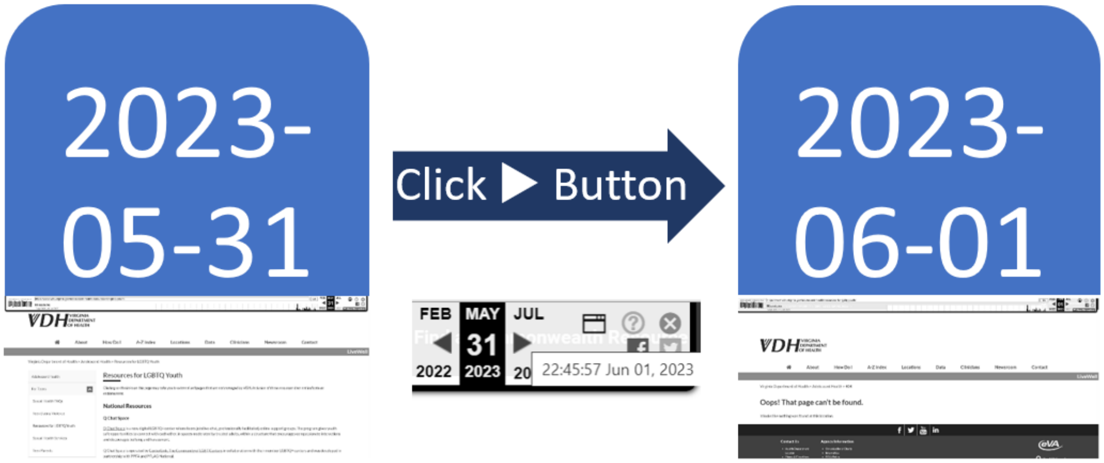
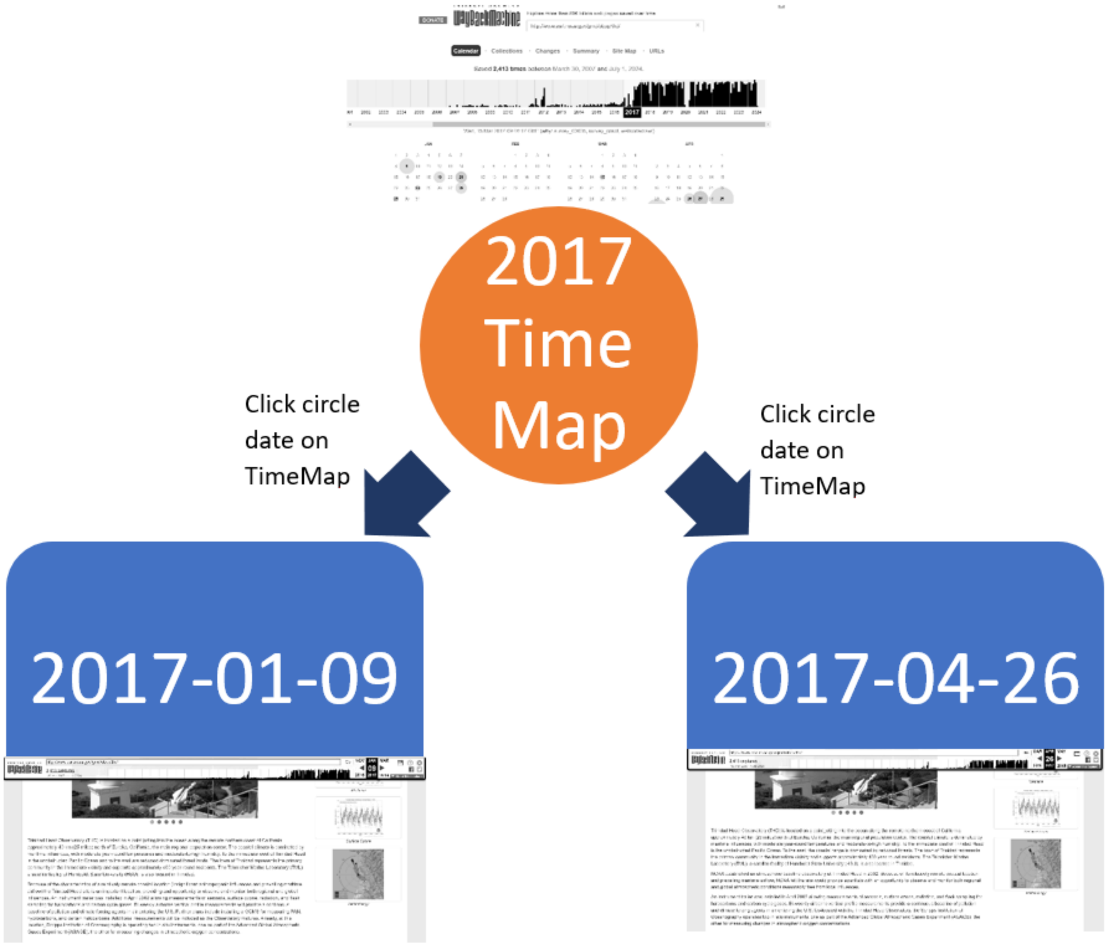


Figure 18. Slide users navigate between page versions via banner buttons. A web archive user viewing the LGBTQ youth resources page on the Virginia Department of Health website over time can use the banner buttons to view when the page was deleted.

Figure 19 shows how a user might navigate between page versions using the TimeMap, specifically the Internet Archive Wayback Machine’s calendar interface.

In this section, we analyze the prevalence of different navigation patterns of slide users and identify slide user tasks.



2017 Time Map

Click circle date on TimeMap

Click circle date on TimeMap

2017-01-09

2017-04-26

Figure 19. Slide users navigate between page versions via the TimeMap calendar page. A web archive user viewing the Trinidad Observatory page on the NOAA website over time can use the TimeMap to view when the term anthropogenic was deleted.

4.2.1 Extracting Information About How Slide Users Navigate in Web Archives

The Internet Archive provided a full day server log of Wayback Machine accesses from 2019 [74]. This log was pre-processed by the Internet Archive for IP address anonymization to protect user privacy. We aimed to only analyze the properties of the pages that would not violate a user's right to open inquiry without scrutinization, similar to how books are examined for damage that would necessitate repair or replacement when they are returned to a library, but the content itself is not scrutinized.

Using the slide user IP address list from Jayanetti et al. [74], we could extract all lines matching each IP address for further analysis. The first part of the analysis was to determine the URI of each slide page accessed by each user. Some users accessed multiple pages in a slide pattern. Then we were able to filter for users who accessed government websites (.gov) in a slide pattern. Anonymizing IP addresses protects user privacy, while only analyzing government websites respects website content creators' data rights, since US government websites are in the public domain. We extracted the timestamps of each page accessed in a slide pattern to identify if the user was browsing forwards in time, backwards in time, or mixed. We also computed the Levenshtein distance [95] of each pair of pages to have a quantitative measure of the amount of change over time between the versions.

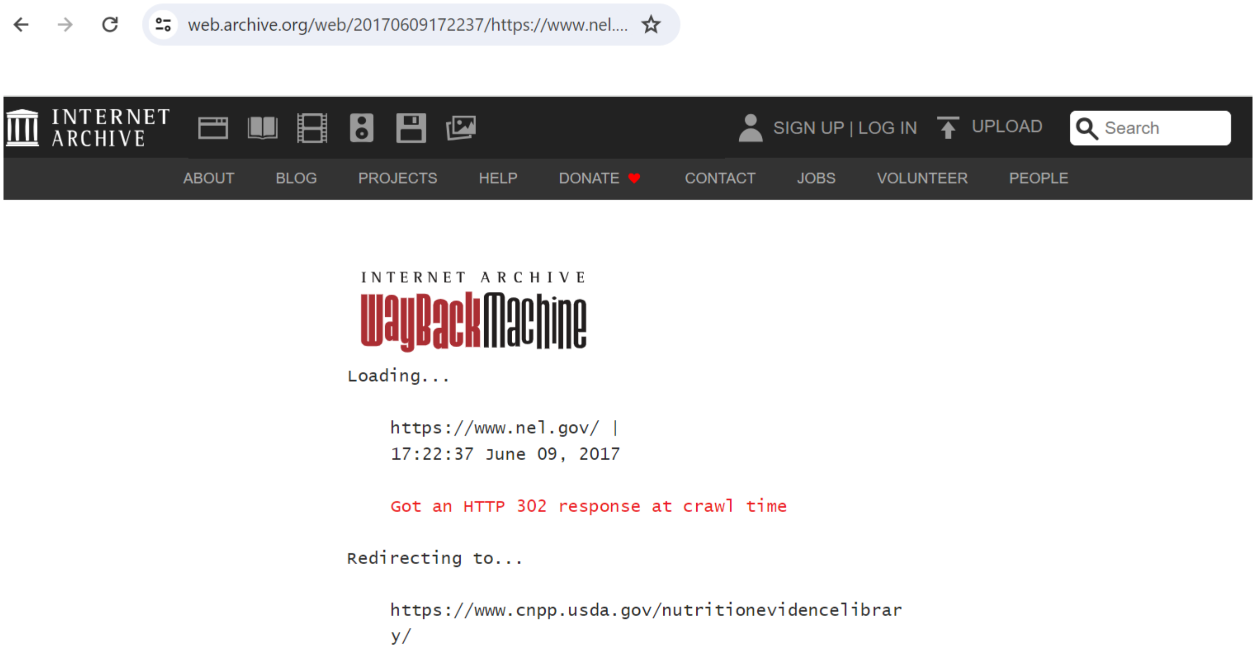
Some of the additional properties of the pages that we were able to extract from the log include the domain of the HTTP referrer and the HTTP status codes of the pages. The log itself did not contain useful status codes, because it contained the server response code rather than the archived status code. For example, when a page redirects in the Wayback Machine, a 200 found page is served with the message "HTTP 3xx at crawl time," so the server log would record a 200 status code while the archived status code is 3xx. Listing 4.1 shows a top level domain (which is broad enough to not require further anonymization) from the log listed with a 200 status code, but Figure 20 clearly shows this page is actually a redirect. This is because the redirect page shown to the user uses the meta and http-equiv HTML elements to force a redirect on refresh.

Listing 4.1. Log Snippet, Highlighting Shows (A) Redirect Listed with HTTP 200 Status Code and (B) TimeMap Referrer Indicated by Asterisk

```
0.126.141.113 web.archive.org - [07/Feb/2019:14:04:39 +0000] "GET
/web/20170609172237/https://www.nel.gov/ HTTP/2.0" 200 8171
"https://web.archive.org/web/20170215000000*/www.nel.gov"
```

We needed to query the individual archived pages for their status codes rather than using the status codes in the log. There were many pages with non-successful status codes being viewed by slide users, which is important information for determining these users' tasks and needs.

Finally, we were able to extract information from the log to determine each user's navigational style. Some slide users went back and forth between the TimeMap and different versions of pages, which is a "hub and spoke" navigation style, and shown in Figure 19. Other users used the arrow buttons in the replay banner to navigate in a "linear" style, as shown in Figure 18. We noticed that these users had trouble with this navigation interface because they had to hit the next button many times in some cases for the next version of the page with changes to be replayed. Users had additional difficulties using the arrow buttons when one of the page versions had a 3xx HTTP status code, because the default behavior in the Wayback Machine is to give a technical message ("HTTP 3xx at crawl time" as shown in Figure 20) and redirect automatically, rather than to give a message in plain language and let them choose if they want to follow the redirect or not. We also



```
Loading...

    https://www.nel.gov/ |
    17:22:37 June 09, 2017

    Got an HTTP 302 response at crawl time

Redirecting to...

    https://www.cnpp.usda.gov/nutritionevidencelibrar
    y/
```

Figure 20. The nel.gov memento from 2017-06-09 is a redirect, but is listed with a 200 status code in the 2019 Internet Archive access log. This is because the page uses the meta and http-equiv HTML elements to force a redirect on refresh.

observed a few users using “birds eye” navigation via the timeline in the replay banner, but this style of navigation was much less frequent than the other two types.

We could identify these three navigation types based on the HTTP referrer as well as the original request. For the referrer, navigation buttons have the referrer as the same memento with URL canonicalization, while TimeMap navigation referrers include an asterisk. Listing 4.1 shows a TimeMap navigation pattern, as indicated by the asterisk in the referrer. Anonymized Listing 4.2 shows a button navigation pattern, as shown with the referrer being another visited version of the requested page. Navigation arrow buttons have 14-digit exact datetimes in the URL while timeline navigation requests have a URL with a datetime ending in six zeroes representing non-exact hours, minutes, and seconds. Anonymized Listing 4.3 shows a timeline navigation pattern, as shown with the requested page having six zeros with a 3xx status code and the referrer being another version of the requested page. In addition to observing hub and spoke navigation between TimeMaps and mementos, we also observed users engaging in hub and spoke navigation on a top level archived webpage to find subpages that had, for example, moved locations.

Listing 4.2. Log Snippet - Page Anonymized, Highlighting shows Banner Button Navigation via same page referrer

```
0.104.42.170 web.archive.org - [07/Feb/2019:18:20:39 +0000] "GET
/web/20111208183700/http://www.census.gov/popest/states/
NST-ann-est.html HTTP/2.0" 200 15858 "https://en.wikipedia.org/"
0.104.42.170 web.archive.org - [07/Feb/2019:18:21:06 +0000] "GET
/web/20120107121625/http://www.census.gov:80/popest/states/
NST-ann-est.html HTTP/2.0" 404 13581
"https://web.archive.org/web/20111208183700/
http://www.census.gov/popest/states/NST-ann-est.html"
```

Listing 4.3. Log Snippet - Page Anonymized, Highlighting shows Banner Timeline Navigation indicated by six zeroes

```
0.116.87.58 web.archive.org - [07/Feb/2019:17:35:56 +0000] "GET
/web/20070901000000/http://www.epa.gov/acidrain/ HTTP/2.0" 302 0
"https://web.archive.org/web/20090305011035/
http://www.epa.gov/acidrain/"
```

Based on these page properties, we were able to identify overall goals and more specific tasks.

The goals and tasks are similar to some of the tasks identified earlier for journalists, but there are additional tasks for these specific users. The two major goals we identified are viewing unavailable content and viewing change over time, which are both goals that were identified for the journalists. Some of the additional tasks we identified beyond those that were common for the journalists were accessing historical data and viewing pages where at least one memento has considerable damage [26]. In all, we analyzed the navigational patterns of 50 random slide users of government webpages in the log.

4.2.2 Tasks of Slide Users

Below, we present examples of some of the tasks slide users engage in. In addition to anonymizing the IP addresses of users to protect their privacy, it is also necessary and customary to anonymize the pages they were viewing to protect their right to open inquiry without scrutinization. For example, Reyes Ayala [122] also followed this anonymization technique. We have replaced the pages that users viewed with comparable pages from the Environmental Data Governance Initiative's 2016-2020 federal websites changes dataset or other federal webpages. Comparable pages have similar properties: the same HTTP status codes, a similar Levenshtein distance, and so on, but the actual content of the pages has no effect on whether or not a set of pages is comparable.

Viewing changes on deep links

We found that 24% of the 50 slide users that we analyzed accessed two or more pages on the same site in a slide pattern. Further analysis showed hub and spoke navigation involving the TimeMap as well as a "Dive" pattern [13]. A Dive pattern is when a user views a memento for a new page by clicking on a link on the current memento. The new memento should have a similar datetime as the current memento. The purpose of this navigation pattern was to locate the previous version of a page with an archived error code. Figure 21 shows an example of this navigation pattern. The EPA Current Legislation page had an archived HTTP 404 status in 2011. In order to find the previous address for the page, the user can go to the main EPA.gov page and use the navigation links on the page in a Dive pattern to find the prior page address.

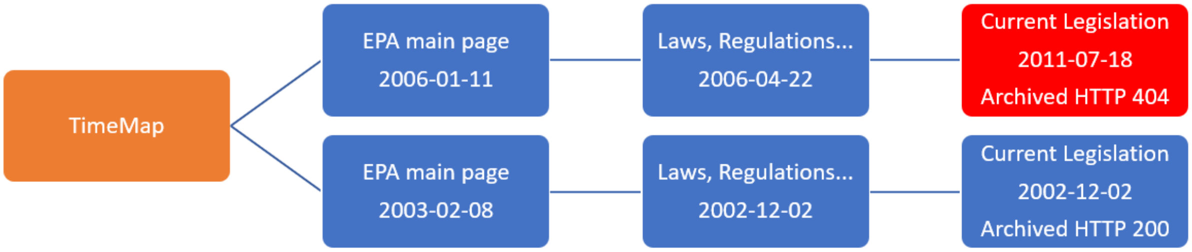
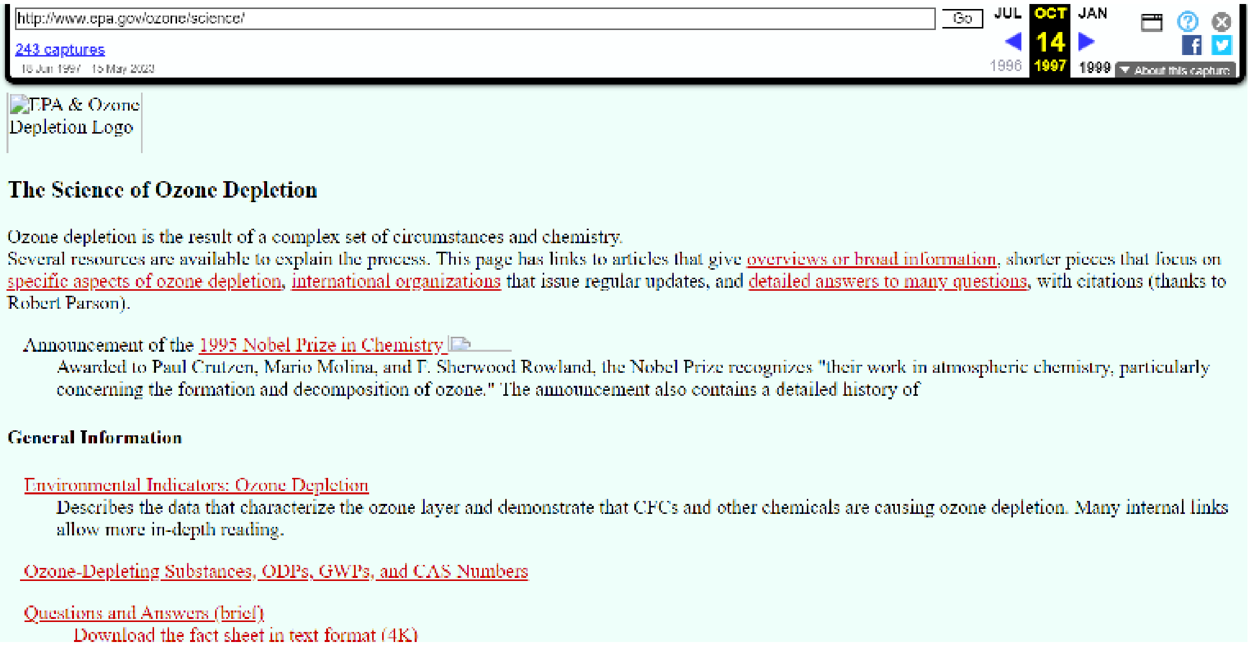

TimeMap
EPA main page 2006-01-11
Laws, Regulations... 2006-04-22
Current Legislation 2011-07-18 Archived HTTP 404
EPA main page 2003-02-08
Laws, Regulations... 2002-12-02
Current Legislation 2002-12-02 Archived HTTP 200

Figure 21. Slide users find deep links that have moved. A web archive user viewing the EPA Current Legislation page over time can use the navigation links on the main EPA page over time to find the previous address of the page.

Viewing less damaged mementos

We found that 10% of slide users analyzed viewed multiple versions of the same page because the first version viewed had embedded images or videos missing, also known as memento damage. Figure 22 shows an example of two versions of a page, one with damage (left) and one without (right).



EPA & Ozone Depletion Logo

The Science of Ozone Depletion

Ozone depletion is the result of a complex set of circumstances and chemistry.
Several resources are available to explain the process. This page has links to articles that give overviews or broad information, shorter pieces that focus on specific aspects of ozone depletion, international organizations that issue regular updates, and detailed answers to many questions, with citations (thanks to Robert Parson).

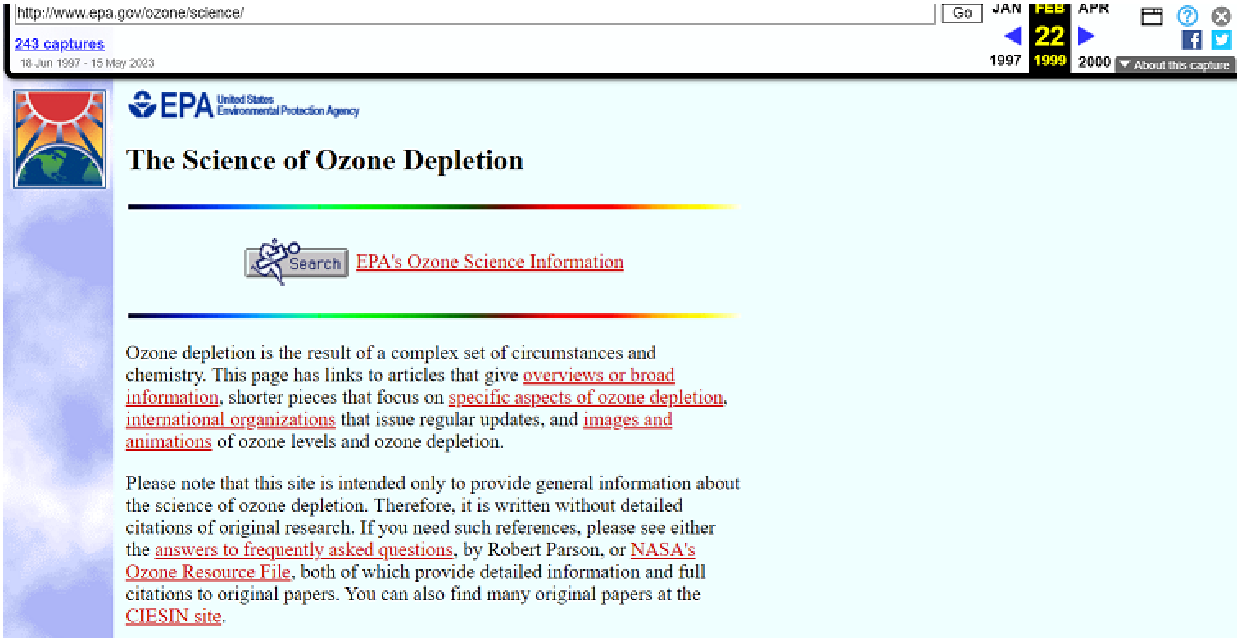
Announcement of the 1995 Nobel Prize in Chemistry
Awarded to Paul Crutzen, Mario Molina, and F. Sherwood Rowland, the Nobel Prize recognizes "their work in atmospheric chemistry, particularly concerning the formation and decomposition of ozone." The announcement also contains a detailed history of

General Information

Environmental Indicators: Ozone Depletion
Describes the data that characterize the ozone layer and demonstrate that CFCs and other chemicals are causing ozone depletion. Many internal links allow more in-depth reading.

Ozone-Depleting Substances, ODPs, GWPs, and CAS Numbers

Questions and Answers (brief)
Download the fact sheet in text format (4K)


http://www.epa.gov/ozone/science/
Go
243 captures
18 Jun 1997 - 15 May 2023
JAN
FEB
22
APR
1997
1999
2000
About this capture
EPA United States Environmental Protection Agency
Search

The Science of Ozone Depletion

EPA's Ozone Science Information

Ozone depletion is the result of a complex set of circumstances and chemistry. This page has links to articles that give overviews or broad information, shorter pieces that focus on specific aspects of ozone depletion, international organizations that issue regular updates, and images and animations of ozone levels and ozone depletion.

Please note that this site is intended only to provide general information about the science of ozone depletion. Therefore, it is written without detailed citations of original research. If you need such references, please see either the answers to frequently asked questions, by Robert Parson, or NASA's Ozone Resource File, both of which provide detailed information and full citations to original papers. You can also find many original papers at the CIESIN site.

Figure 22. Slide users want undamaged mementos. A web archive user viewing the EPA Science of Ozone Depletion page from 1997 will notice that it is damaged because its embedded images are not archived, while the version from 1999 is undamaged.

Viewing consecutive mementos to view change over time

We found that 50% of the analyzed slide users viewed at least three versions of the same page over time. Many of these users were navigating with the banner buttons. They continued viewing page versions without changes until finally a version with changes appeared, or they gave up. The banner would function better for these users if the arrows went to the next changed version of the page instead of the next capture. Figure 23 shows the 2021 Changes calendar for the EPA Radiation Advisory Committee webpage. A slide user would need to press the next button four times in order to see a version with changes. While the Changes calendar gives the user an indication about which mementos should be examined, the navigation button gives no such indication.

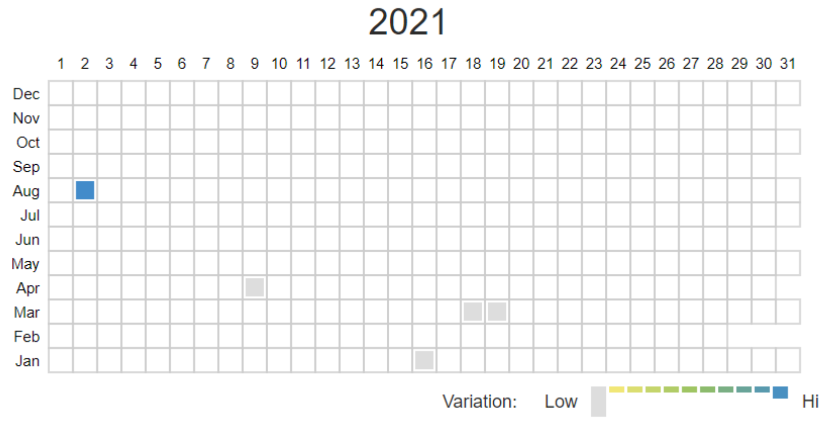
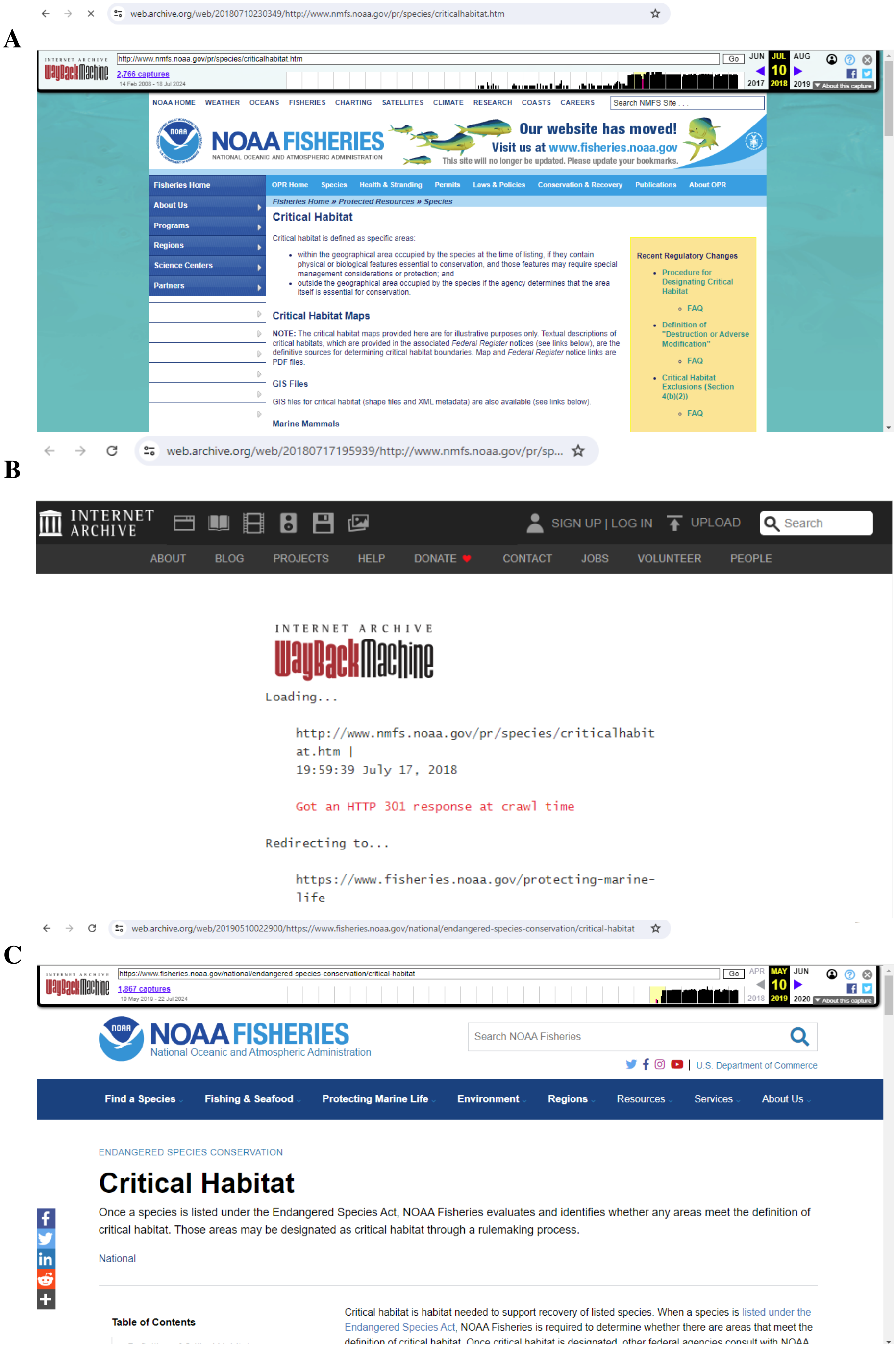


Figure 23. Wayback Machine Changes Calendar for a page with a low change rate. A slide user would have to press the next button four times to see a changed version of the EPA Radiation Advisory Committee webpage in 2021.

Unsuccessful archived HTTP status codes

We found that 34% of slide users analyzed viewed multiple versions of the same page where at least one version did not have an HTTP 200 status code. As previously shown in Figure 17, sometimes the non-successful archived HTTP status code is 404 not found. Other times, as previously shown in Figure 20, sometimes the non-successful archived HTTP status code is 3xx redirect. With a 404 status code, the banner navigation arrows still function so the user can move backwards in time to find versions of the page before it was deleted. However, with a 3xx status code, the banner navigation buttons no longer function since the page address changes. This makes it hard for the user to analyze change over time. Figure 24 shows how a user might try to navigate the NOAA Fisheries website in 2019. The last memento with an HTTP 200 status code at the subdomain nmfs.noaa.gov was captured in July, 2018. Pressing the next navigation button results in a redirect. No navigation arrows are present on the redirect page. The new location for the page at the fisheries.noaa.gov subdomain was first archived in May, 2019. The navigation button to go backwards is greyed out, even though previous captures of this page exist at the previous address.



web.archive.org/web/20180710230349/http://www.nmfs.noaa.gov/pr/species/criticalhabitat.htm

http://www.nmfs.noaa.gov/pr/species/criticalhabitat.htm

2,766 captures

14 Feb 2008 - 18 Jul 2024

Our website has moved!

Visit us at www.fisheries.noaa.gov

This site will no longer be updated. Please update your bookmarks.

Critical Habitat

Critical habitat is defined as specific areas:

- within the geographical area occupied by the species at the time of listing, if they contain physical or biological features essential to conservation, and those features may require special management considerations or protection; and
- outside the geographical area occupied by the species if the agency determines that the area itself is essential for conservation.

Critical Habitat Maps

NOTE: The critical habitat maps provided here are for illustrative purposes only. Textual descriptions of critical habitats, which are provided in the associated *Federal Register* notices (see links below), are the definitive sources for determining critical habitat boundaries. Map and *Federal Register* notice links are PDF files.

GIS Files

GIS files for critical habitat (shape files and XML metadata) are also available (see links below).

Marine Mammals

Recent Regulatory Changes

- Procedure for Designating Critical Habitat
  - FAQ
- Definition of "Destruction or Adverse Modification"
  - FAQ
- Critical Habitat Exclusions (Section 4(b)(2))
  - FAQ

web.archive.org/web/20180717195939/http://www.nmfs.noaa.gov/pr/sp...

```
Loading...

    http://www.nmfs.noaa.gov/pr/species/criticalhabit
    at.htm |
    19:59:39 July 17, 2018

    Got an HTTP 301 response at crawl time

Redirecting to...

    https://www.fisheries.noaa.gov/protecting-marine-
    life
```

web.archive.org/web/20190510022900/https://www.fisheries.noaa.gov/national/endangered-species-conservation/critical-habitat

https://www.fisheries.noaa.gov/national/endangered-species-conservation/critical-habitat

1,867 captures

10 May 2019 - 22 Jul 2024

Critical Habitat

Once a species is listed under the Endangered Species Act, NOAA Fisheries evaluates and identifies whether any areas meet the definition of critical habitat. Those areas may be designated as critical habitat through a rulemaking process.

Critical habitat is habitat needed to support recovery of listed species. When a species is listed under the Endangered Species Act, NOAA Fisheries is required to determine whether there are areas that meet the definition of critical habitat. Once critical habitat is designated, other federal agencies consult with NOAA

Figure 24. It is not possible to navigate across URL changes for the NOAA Critical Habitat page. *(A)* Last HTTP 200 page capture from July, 2018, at nmfs.noaa.gov. *(B)* Pressing the next button results in a redirect. No navigation buttons are present on the redirect. *(C)* The first capture of this page on fisheries.noaa.gov from May 2019 has a greyed out previous button, hindering navigation.

4.3 SUMMARY

In this chapter, we established the unmet needs of web archive users when searching and navigating. First, we provide evidence that users such as journalists want to view changes in webpages over time. Next, we provide evidence that users who view more than one version of a webpage need navigation buttons that link between changes, rather than between set temporal intervals. We use these formative investigations to guide our interface design changes in the next chapter.

CHAPTER 5

A CHANGE-TEXT SEARCH INTERFACE FOR WEB ARCHIVES

In this chapter, we first address Research Question 1, How can we make changes in webpages discoverable and understandable? The formative investigation from the previous chapter about journalists' tasks inspired the design of a change text search engine for web archives. We explain the constructs necessary for the backend of such a system, and then describe the implementation of the frontend which also includes an animated changes replay tool. Portions of this chapter are based on our work published at ACM/IEEE JCDL 2023 [52].

We also address Research Question 2, How can we increase efficiency in web archive user navigation for viewing change over time? The formative investigation from the user log analysis in the previous chapter inspired a redesign of the Internet Archive Wayback Machine's replay navigation banner. We describe the added functionality of navigation buttons that skip identical versions, as well as a new landing page for archived redirects.

5.1 CHANGE-TEXT SEARCH ARCHITECTURE

The architecture for the change text search engine consists of three levels, as shown in Figure 25. First, webpages with multiple versions must be acquired. Next, the webpages are indexed for both text content and replay. The changes in the text content are calculated after the initial indexing. Finally, the user interface for a search engine provides the user with a way to discover the changes in the webpages, and replay those changes in context.

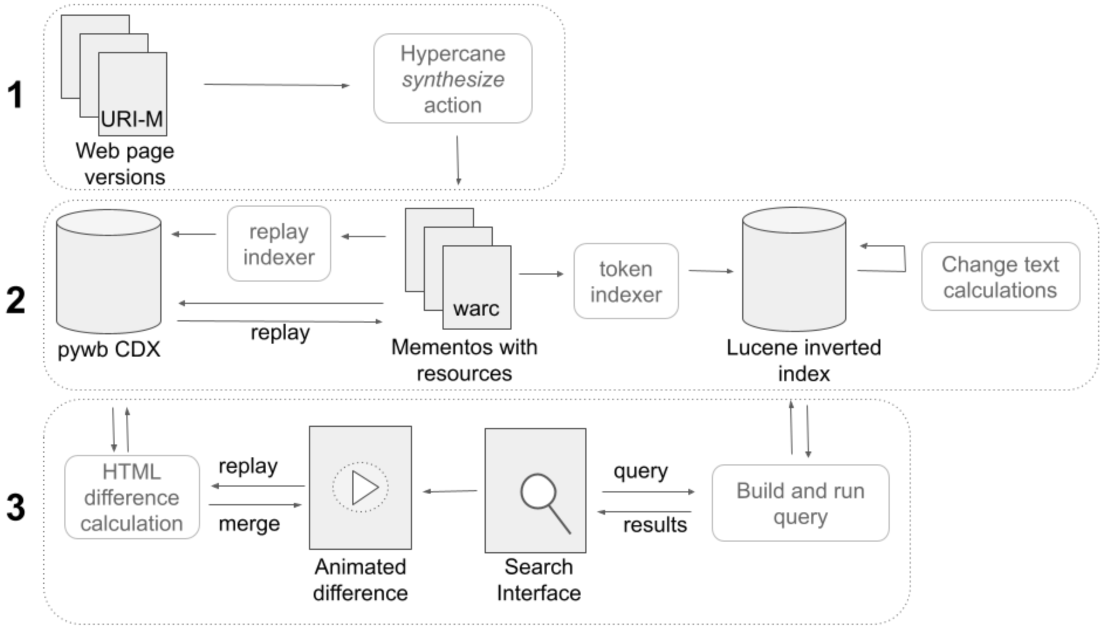


Figure 25. The architecture of the change text search engine. Level 1 consists of document acquisition. Level 2 consists of the documents and indices. Level 3 consists of the user interface. ©2023 IEEE [52]

5.1.1 Document Acquision

The documents for the change text search engine need to be in WARC format. While users can replay public web archives' holdings, they cannot access the original WARC files. Indexing these public holdings into the change text search engine is therefore not possible without a tool that can turn a URI-M into a WARC file. Hypercane [83] is a tool for interacting with web archive collections. One feature of the Hypercane tool is to *synthesize* a WARC from a given URI-M. The WARC file output by Hypercane includes the original HTML of the web page in raw form [78], [79], along with all available embedded resources. The HTML is necessary for creating an inverted index of the changes in page content, while the embedded resources are necessary for merged replay showing the differences between page versions in context. Users and organizations with existing WARC collections would not need to perform any additional document acquisition.

5.1.2 Solr for Versioned Document Collections

A collection of any type of documents, including WARCs, is only useful if it can be both accessed and searched. WARCs can be searched if they are indexed. We indexed the WARCs into Lucene. Solr, a platform built on top of Lucene, is a good basis for a WARC search platform. However, Lucene has not been used in conjunction with the web archive validity range concept before, so we made changes to the Solr XML configuration to allow for the Lucene index to properly hold the change text calculations.

We used a Solr date range field to support the temporal validity ranges for each document version. The validity range field and its type were added to the Solr XML schema file, as shown in Listing 5.1.

Listing 5.1. Solr XML schema field types

```
<field name="validity_range" type="dateRange" />

<fieldType name="dateRange" class="solr.DateRangeField" />
```

To compute change text across versions, we created three fields (*additions, deletions,* and *semi-deletions*). Each of these fields represents a set of terms, so phrase queries on those fields are nonsensical. Note that even though direct phrase queries over these term fields are disabled, it is possible to construct a Lucene query that supports phrase searches over term changes by using a combination of the term field and the content field, since the content field does support phrase queries. A deleted term and an added term is defined as all instances of the term being present/absent on the next version, while a semi-deletion is defined as a differing, positive term count between consecutive versions. Semi-additions would follow easily as an additional field that could be added in future implementations. Table 2 shows examples of the four types of change text.

Table 2. Examples of the four types of change text for the term "forget" using text from Jane Austen's *Persuasion*. Additions start with a zero count of the term and end with a positive count. The other three start with a positive count of the term, and end with zero, more, or fewer of the term.

Version n	Version $n+1$	Change type
We certainly do not **forget** you, so soon as you **forget** us.	We certainly do not **dismiss** you.	deletion
We certainly do not **forget** you, so soon as you **forget** us.	We certainly do not **forget** you, so soon as you **dismiss** us.	semi-deletion
We certainly do not **dismiss** you, so soon as you **dismiss** us.	We certainly do not **forget** you, so soon as you **forget** us.	addition
We certainly do not **forget** you, so soon as you **forget** us. It is, perhaps, our fate rather than our merit.	We certainly do not **forget** you, so soon as you **forget** us. That we do not **forget** is our fate rather than our merit.	semi-addition

We added these three fields to the Solr XML schema file as shown in Listing 5.2.

Listing 5.2. Solr XML schema fields

```
<field name="deleted_term" type="text_general" indexed="true"
termOffsets="false" stored="false" termPositions="false"
termVectors="false" multiValued="false"/>

<field name="added_term" type="text_general" indexed="true"
termOffsets="false" stored="false" termPositions="false"
termVectors="false" multiValued="false"/>

<field name="semi_del_term" type="text_general" indexed="true"
termOffsets="false" stored="false" termPositions="false"
termVectors="false" multiValued="false"/>
```

5.1.3 Computing Temporal Validity Ranges Using Lucene

We implemented the calculation of temporal validity ranges in Java. After indexing the entire collection, it is possible to compute temporal validity ranges documents with non-empty titles and non-empty content. Listing A.6 (Appendix A) shows how the versions are ordered.

Iteration over the entire Lucene index is necessary to identify the valid documents and to order their versions. A hashmap holds each document and its versions. We used the normalized URL of each document as its hashmap key. We used insertion into a tree set to temporally order the versions for each document, as shown in Listing A.1 (Appendix A).

Once we have ordered all of the versions, we make a pass over the hashmap to link each version to its successor. While a tree set is efficient for ordering the versions, we needed a linked list to compute the validity ranges. We then make a second pass over the hashmap to compute the validity ranges using Java's Calendar from its utilities library (shown in Listing A.2, Appendix A) from the timestamps as well as the document version ordinal.

The output from this process is the validity ranges in JSON. Users of our change text search system can insert this JSON into the index using the Solr dashboard. We also enabled a flag that automatically splits the JSON output into segments of one thousand postings to better interface with the Solr dashboard posting capabilities.

For a test data set that included 100,000 files in the WARCs indexed, about 10,000 were identified as valid documents and the algorithm to compute the validity ranges took 1.98 seconds on a Windows machine. For a second test data set that included 200,000 files in the WARCs indexed, about 20,000 were identified as valid documents and the algorithm to compute the validity ranges took 3.95 seconds.

After running the initial indexing from the UKWA WARC indexer on a Windows machine, garbled unicode characters appeared. HTML entities were also present. We converted characters that were supposed to be punctuation to their ASCII equivalents prior to computing validity ranges. We fixed the unicode characters using Java,[1] and we replaced the HTML entities using built in PHP HTML functions. Then, we reposted the text to the Lucene index in the Solr dashboard as JSON. In addition, the UKWA WARC indexer does not remove standard ports when normalizing URLs. We removed standard ports using the built in Java java.net.URI library and then re-posted before computing validity ranges.

[1] https://gist.github.com/xijo/d4bad3953f7b9979dd91

5.1.4 Computing Change Text Using Lucene

Concurrently with computing the temporal versions, the change text of each document is computed with respect to its next version. First, we populated each document with its set of terms and term counts by tokenizing its content field with Lucene's Standard Analyzer. Next, we used the set difference operation to compute the additions and deletions as shown in Listing A.3 (Appendix A). In theory, deletions are defined by two versions: the version before and the version after the term disappears. In practice, the deleted term should only be assigned to one of these versions, and in this implementation it was assigned to the version before the deletion of the term, as shown in Figure 26.

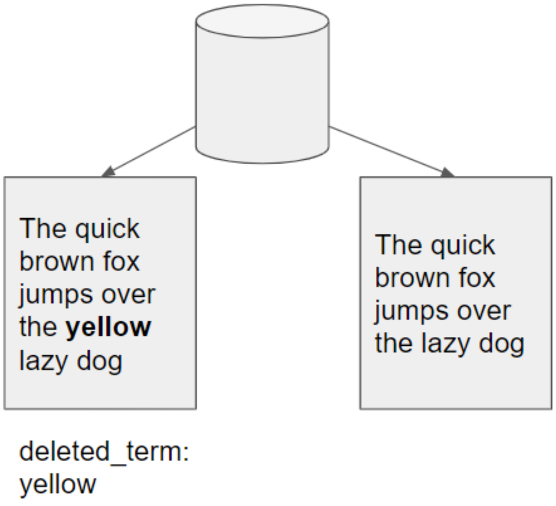

The quick brown fox jumps over the **yellow** lazy dog
The quick brown fox jumps over the lazy dog
deleted_term: yellow

Figure 26. Assignment to the deleted terms field. The term yellow has been deleted from this document. The term yellow is assigned to the deleted terms field for the version directly preceeding the deletion.

Finally, we use the term counts to compute the semi-deletions as shown in Listing A.4 (Appendix A). All terms are considered additions on the first version of a document.

In order to implement rankings where a page with more instances of a term deleted is ranked higher than a page with fewer instances of that term deleted, we needed to post the term counts to the index as well. We accomplished this by posting a stream of terms to the deleted_terms field with each term repeated the number of times it was deleted.

5.1.5 Querying Change Text Using Lucene

The backend now supports querying for a single term deletion by field. The Lucene query is shown in Table 3. The standard result of this query is one version of the page, namely the version directly prior to the term being deleted.

Table 3. The Lucene index can be queried by field for deleted and semi-deleted terms and phrases.

Target	Query
Deleted term	deleted_term:TERM
Deleted phrase	text:"PHRASE TERMS" deleted_term:PHRASE deleted_term:TERMS
Semi-deleted term	semi_del_term:TERM
Semi-deleted phrase	semi_del_term:(PHRASE OR TERMS) text:"PHRASE TERMS"

The backend also supports querying for a phrase deletion by field indirectly. The Lucene query is shown in in Table 3. This query searches for the text containing the phrase, as well as both terms deleted. The query contains an implicit "and" boolean operator for the deleted terms. This will match pages where both terms have been completely deleted. The deleted_term field is not directly searchable by phrase, because it is a random ordering of the set of all terms.

Terms and phrases can also be semi-deleted, which is when some but not all of the terms on the page are removed. The Lucene query for single term queries is shown in in Table 3. The Lucene query for phrase queries is also shown in in Table 3. This query will match pages where one term was completely deleted but the other was not, where the phrase was removed and both terms are still present but in a smaller quantity than before the phrase was deleted, as well as when the phrase is still present but in a smaller quantity than before.

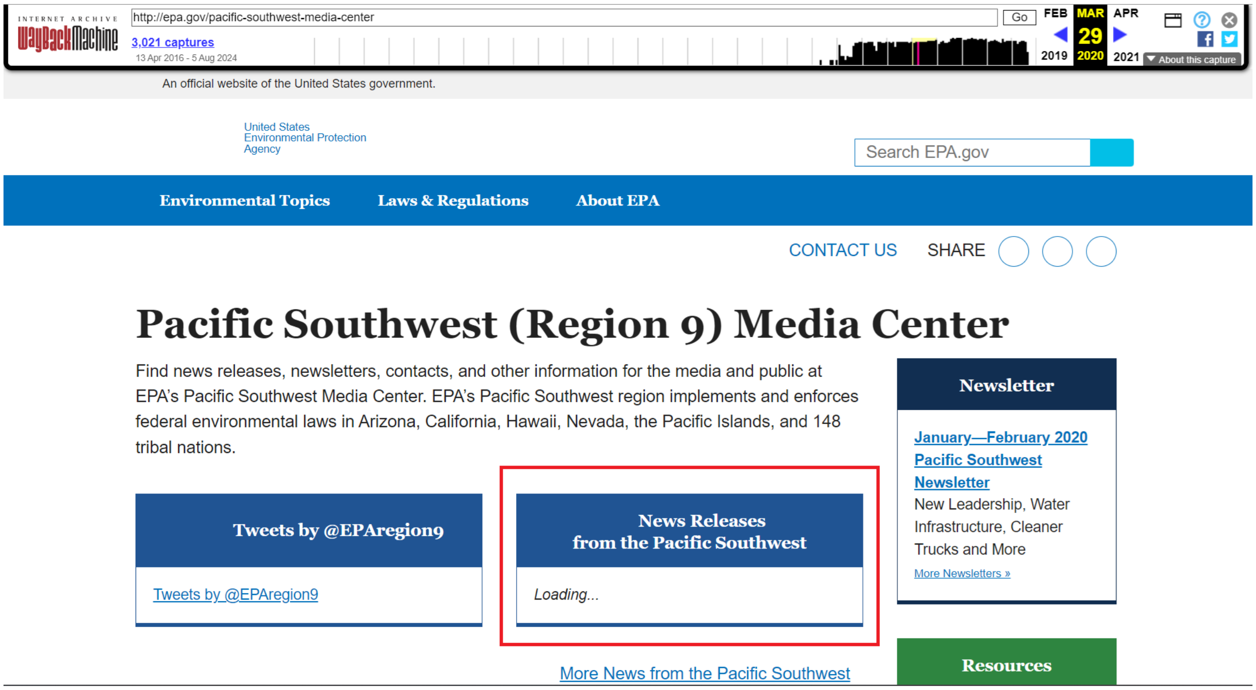
The semi-deleted phrase query does not result in a set of definite matches for a semi-deleted phrase, but rather results in a superset that contains all possible matches. To be a semi-deleted phrase, the phrase must show up $m > 1$ times in the previous version, and $0 < n < m$ times in the next version. The query guarantees that the phrase was in the previous version, and that at least one of the terms is calculated as a semi-deleted term, but does not guarantee that the phrase was kept at least one time on the next version. For a full deletion, the deletion of one term is sufficient to show that the phrase has been deleted. However, the semi-deletion of one term may not involve the phrase. It is possible that one term appears in another place on the page, and that is the term that was removed to result in the semi-deletion calculation being positive.

Thus, our system must evaluate each potential match from the Lucene query to ensure that it is an instance of a semi-deleted phrase. Counting occurrences of phrases with an arbitrary number of terms is not available directly through Lucene. However, it is possible to filter the query after execution one by one in the search engine results page.

5.1.6 Ranking Change Text Search Results

Currently we use the default Lucene score to rank documents with one exception. When manually examining results for deleted terms, it became clear that some pages were included in the results because all of the content on the page had been deleted. Non 200 HTTP status codes were excluded from our indexing process, so that means that these pages where all of the content was deleted were soft 404s or indexing errors of dynamically loaded content. Surprisingly, soft 404s occurred at a much lower frequency than indexing errors of dynamically loaded content. Figure 27 shows an example of a page with dynamically loaded content that was falsely categorized as a deletion by the change text algorithm. This page appears to have had over half of its content deleted, but the memento is damaged because the dynamically loaded content was not archived and will not replay.

Since soft 404s and dynamically loaded content errors do not represent true deletions in the way that the other results do, we chose to rank these items last. AlNoamany et al. [12] found that soft 404s can be automatically detected when the older version of the webpage is at least ten times larger than the newer version. We incorporated this calculation into the Java backend, as shown in Listing A.5 (Appendix A), for use with the PHP frontend.



An official website of the United States government.

Pacific Southwest (Region 9) Media Center

Find news releases, newsletters, contacts, and other information for the media and public at EPA's Pacific Southwest Media Center. EPA's Pacific Southwest region implements and enforces federal environmental laws in Arizona, California, Hawaii, Nevada, the Pacific Islands, and 148 tribal nations.

Tweets by @EPAregion9

Tweets by @EPAregion9

News Releases from the Pacific Southwest

Loading...

More News from the Pacific Southwest

Newsletter

January—February 2020 Pacific Southwest Newsletter

New Leadership, Water Infrastructure, Cleaner Trucks and More

More Newsletters »

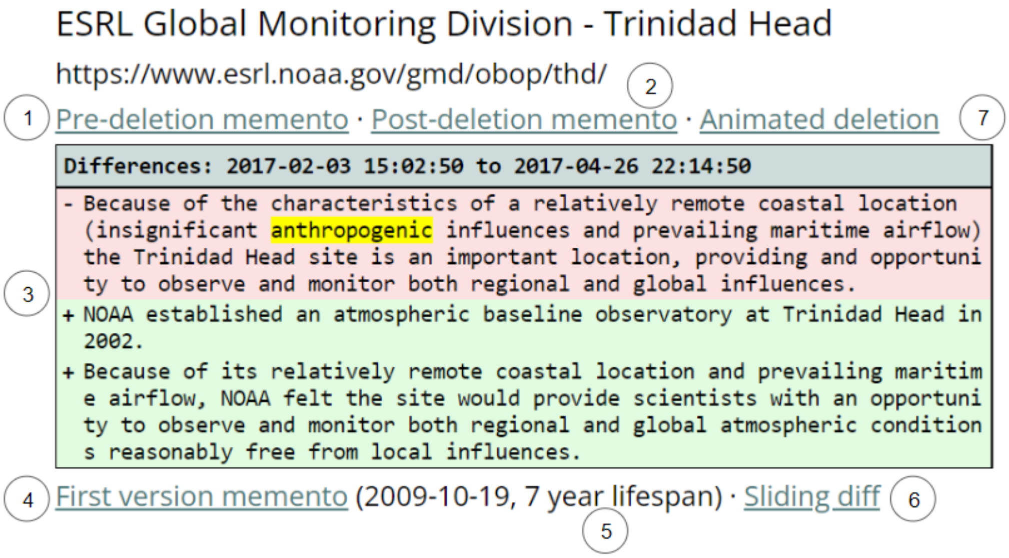
Resources

Figure 27. Dynamically loaded content is a false positive deletion in the change text index. The EPA Pacific Southwest Media Center page switched from statically loaded content in 2016 to dynamically loaded content by 2020. The size of the page appears to be reduced by over half even though the content was not deleted.

5.2 CHANGE-TEXT SEARCH INTERFACE DESIGN

A user can interact with the change text term index through a search interface, which we built using Solarium,[2] a PHP Solr interface. A user may query for a deleted term, a deleted phrase, an added term, or an added phrase. The query page is shown in Figure 30.

The search results are displayed in a search engine results page, shown in Figure 28. The interface includes links to the pre and post deletion mementos along with their datetimes, a sliding difference, an animated deletion, and a diff showing the changes. Figure 29 shows the mementos that match this SERP and the diff shown.



ESRL Global Monitoring Division - Trinidad Head

https://www.esrl.noaa.gov/gmd/obop/thd/

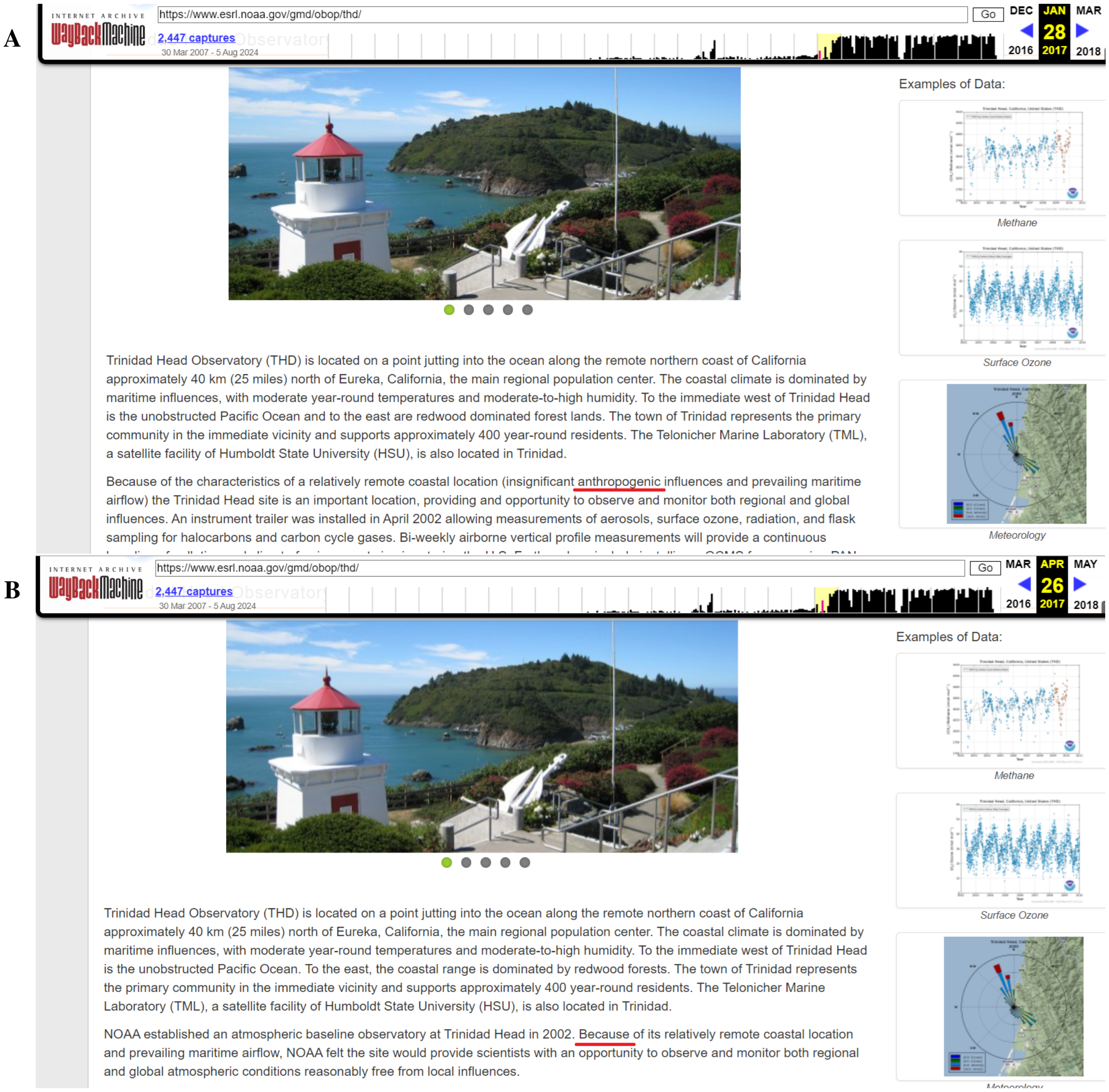
Pre-deletion memento · Post-deletion memento · Animated deletion

```
Differences: 2017-02-03 15:02:50 to 2017-04-26 22:14:50
- Because of the characteristics of a relatively remote coastal location
  (insignificant anthropogenic influences and prevailing maritime airflow)
  the Trinidad Head site is an important location, providing and opportuni
  ty to observe and monitor both regional and global influences.
+ NOAA established an atmospheric baseline observatory at Trinidad Head in
  2002.
+ Because of its relatively remote coastal location and prevailing maritim
  e airflow, NOAA felt the site would provide scientists with an opportuni
  ty to observe and monitor both regional and global atmospheric condition
  s reasonably free from local influences.
```

First version memento (2009-10-19, 7 year lifespan) · Sliding diff

Figure 28. Change text search interface. 1, 2, and 4: individual replay links to page mementos; 3: the diff between the pre and post deletion versions; 5: content lifespan calculation; 6: the link to the sliding diff viewer across all indexed versions of the page; 7: the link to the deletion animation ©2023 IEEE [52]

[2] https://github.com/solariumphp/solarium



Trinidad Head Observatory (THD) is located on a point jutting into the ocean along the remote northern coast of California approximately 40 km (25 miles) north of Eureka, California, the main regional population center. The coastal climate is dominated by maritime influences, with moderate year-round temperatures and moderate-to-high humidity. To the immediate west of Trinidad Head is the unobstructed Pacific Ocean and to the east are redwood dominated forest lands. The town of Trinidad represents the primary community in the immediate vicinity and supports approximately 400 year-round residents. The Telonicher Marine Laboratory (TML), a satellite facility of Humboldt State University (HSU), is also located in Trinidad.

Because of the characteristics of a relatively remote coastal location (insignificant anthropogenic influences and prevailing maritime airflow) the Trinidad Head site is an important location, providing and opportunity to observe and monitor both regional and global influences. An instrument trailer was installed in April 2002 allowing measurements of aerosols, surface ozone, radiation, and flask sampling for halocarbons and carbon cycle gases. Bi-weekly airborne vertical profile measurements will provide a continuous

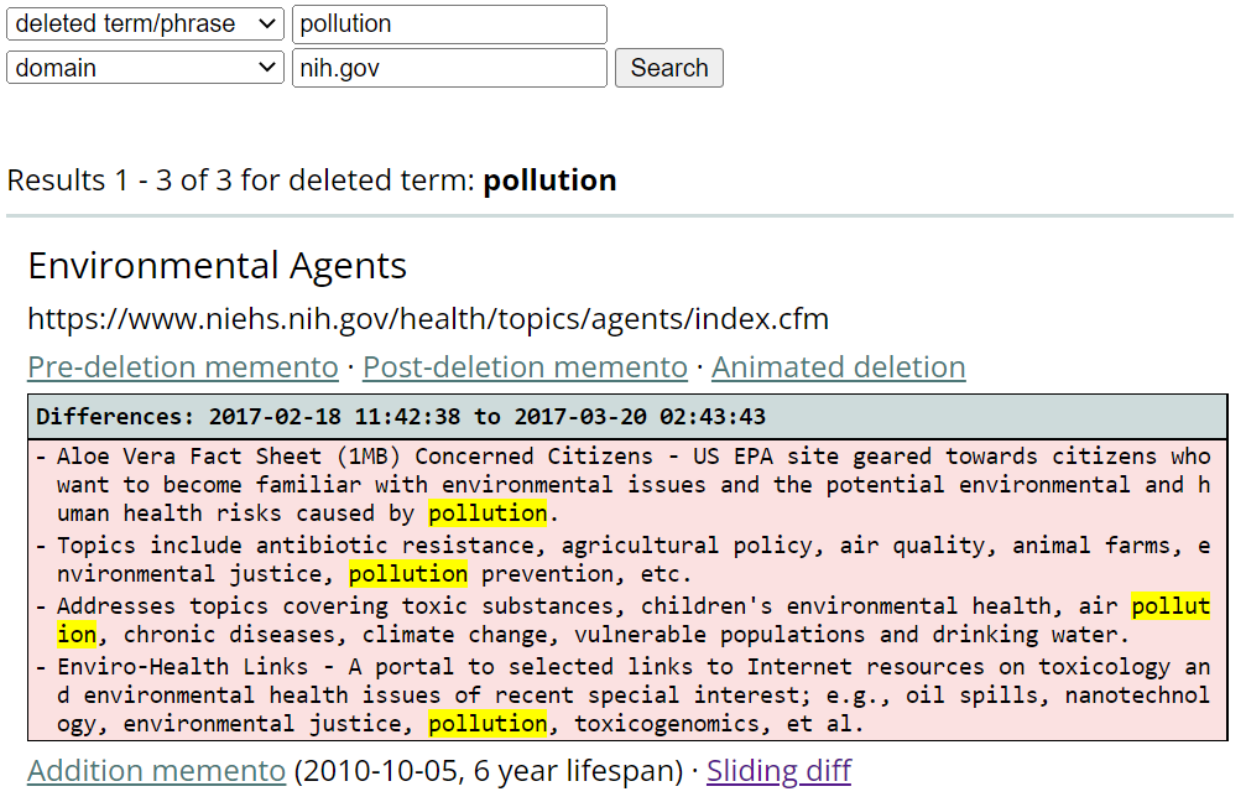
Trinidad Head Observatory (THD) is located on a point jutting into the ocean along the remote northern coast of California approximately 40 km (25 miles) north of Eureka, California, the main regional population center. The coastal climate is dominated by maritime influences, with moderate year-round temperatures and moderate-to-high humidity. To the immediate west of Trinidad Head is the unobstructed Pacific Ocean. To the east, the coastal range is dominated by redwood forests. The town of Trinidad represents the primary community in the immediate vicinity and supports approximately 400 year-round residents. The Telonicher Marine Laboratory (TML), a satellite facility of Humboldt State University (HSU), is also located in Trinidad.

NOAA established an atmospheric baseline observatory at Trinidad Head in 2002. Because of its relatively remote coastal location and prevailing maritime airflow, NOAA felt the site would provide scientists with an opportunity to observe and monitor both regional and global atmospheric conditions reasonably free from local influences.

Figure 29. The term anthropogenic was deleted from the NOAA Earth System Research Laboratory page. *(A)*The term is present in January, 2017. There is another capture of this page at the Library of Congress web archive, but this web archive is still unavailable as of June 2024. *(B)* The term is absent by April 26, 2017, which matches the SERP shown in Figure 28.

5.2.1 Change-Text Query Page

Users are able to query for a single term deletion by field. The PHP user interface implementing Solarium provides the user with a drop down menu, as shown in Figure 30 to indicate their intention to search the deletions field, alongside a text box to type the deleted term. This combination of the drop down menu and the user's term "TERM" is then translated into the Lucene query specified in in Table 3.



deleted term/phrase: pollution
domain: nih.gov

Results 1 - 3 of 3 for deleted term: **pollution**

Environmental Agents

https://www.niehs.nih.gov/health/topics/agents/index.cfm

Pre-deletion memento · Post-deletion memento · Animated deletion

```
Differences: 2017-02-18 11:42:38 to 2017-03-20 02:43:43
- Aloe Vera Fact Sheet (1MB) Concerned Citizens - US EPA site geared towards citizens who
  want to become familiar with environmental issues and the potential environmental and h
  uman health risks caused by pollution.
- Topics include antibiotic resistance, agricultural policy, air quality, animal farms, e
  nvironmental justice, pollution prevention, etc.
- Addresses topics covering toxic substances, children's environmental health, air pollut
  ion, chronic diseases, climate change, vulnerable populations and drinking water.
- Enviro-Health Links - A portal to selected links to Internet resources on toxicology an
  d environmental health issues of recent special interest; e.g., oil spills, nanotechnol
  ogy, environmental justice, pollution, toxicogenomics, et al.
```

Addition memento (2010-10-05, 6 year lifespan) · Sliding diff

Figure 30. Drop down menus in change text search engine results page. In the change text search engine results page, the versions of the page `https://www.niehs.nih.gov/health/topics/agents/index.cfm` that match the addition and deletion of the query term "pollution" are indicated clearly. ©2023 IEEE [52]

Users can query for a deleted phrase as well. The drop down menu indicating a deleted term/phrase search along with the user's phrase is translated into the Lucene query specified in in Table 3. The PHP code to query is shown in Listing B.1 (Appendix B).

5.2.2 Change-Text Search Engine Results Page

In the search results shown by the Solarium implementation, multiple versions of the same page are combined to create one meaningful search result. The version before and after the deletion, along with the version containing the term addition, are all shown as one result, along with links to replay these versions using PyWB. The system calculates the content lifespan, or the difference between the timestamps of the version with the term addition and the version after the term deletion, and shows it in the search results as well. The text snippet that is presented in the search results is a "diff" between the indexed text of the pre and post deletion versions of the page. The system filters the diff to only show lines that contain the search term, and highlights that term. The system also includes a link to show the full diff over time. In the search results, there is also a link to replay an animation of the deletion. Figure 28 shows the layout of these items on the search engine results page.

Since the result of a deleted phrase query is one version of the page, namely the version directly prior to the term being deleted, the next version with the deletion needs to be included in the search result as well. Listing B.2 (Appendix B) shows how the validity range is used to find the next version of the page. Because the validity ranges were implemented with inclusive ranges, there is an overlap of one second between consecutive page versions in the index.

Besides making a different Lucene query, we needed to write additional code to support showing the highlighted phrase. We calculated the diff using a PHP diff library [30]. Because the diff is thus not a Lucene field, Solarium is not capable of highlighting it. This was not problematic when the search query was a single term, but search phrases can be separated by any token in the resulting text. This is because the text being searched is tokenized, while the text being displayed is the original text. A common example of this is hyphenated phrases: a user may enter the phrase “year round” but it may appear in the document as “year-round.” Each search term is highlighted individually to sidestep this issue. Listing B.3 (Appendix B) shows how the diff is highlighted for the query terms.

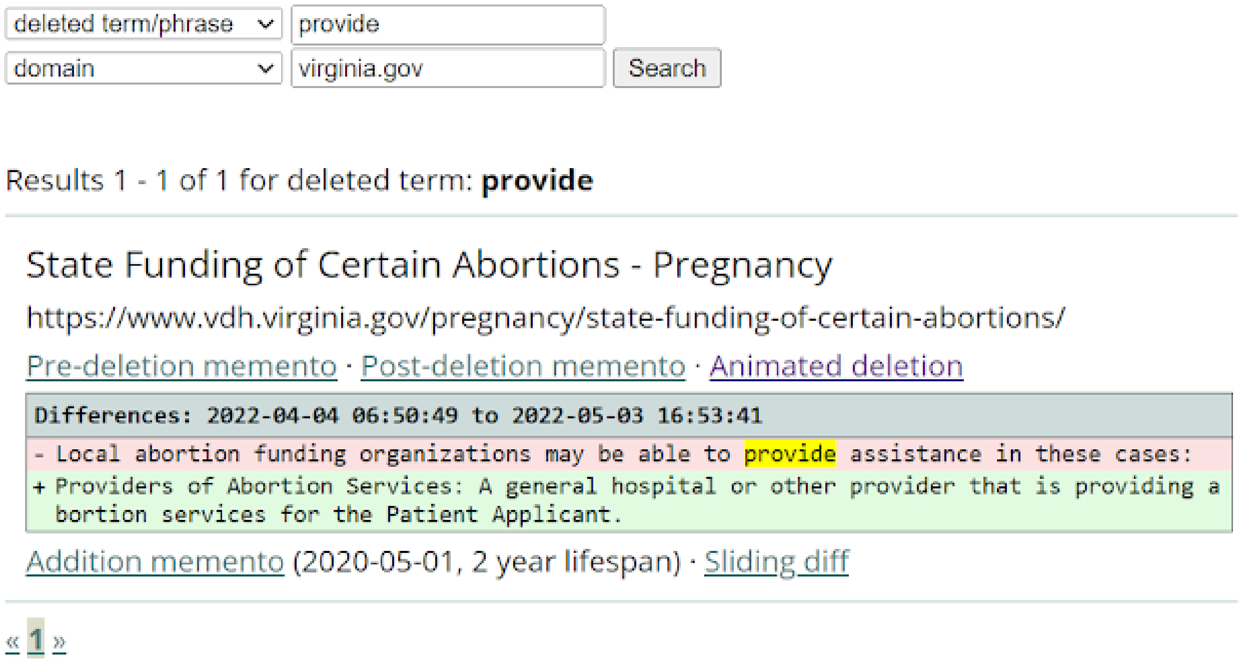
In order to calculate content lifespan for a deleted phrase, we first query for the latest page version with any of the terms labeled as an addition. The date of this page result is used as a lower bound to find the first occurrence of the phrase. The code is similar to Listing B.1 (Appendix B) except the query is for an added term instead of a deleted term.

Users can also query for semi-deleted terms and phrases. After querying for semi-deleted phrases, the system evaluates each potential match from the Lucene query specified in in Table 3 to ensure that it is an instance of a semi-deleted phrase. UAX #29 [38] has a PHP implementation,[3] which we used to tokenize the text field to enable counting occurrences of phrases with an arbitrary number of terms, which is not available directly through Lucene. The system shows each page that has at least one but less than the previous version’s count of that phrase as a result for the user’s query. The code to count phrases is shown in Listing B.7 (Appendix B) and the code to filter the results is shown in Listing B.6 (Appendix B).

Users can also filter the results by top level domain, as shown in Figure 31. Entering a top level domain appends an additional part to the query:

[3] `https://emptyheap2019.github.io/posts/parse-words-php/`

```
domain:TLD.COM
```



deleted term/phrase: provide

domain: virginia.gov

Results 1 - 1 of 1 for deleted term: **provide**

State Funding of Certain Abortions - Pregnancy

https://www.vdh.virginia.gov/pregnancy/state-funding-of-certain-abortions/

Pre-deletion memento · Post-deletion memento · Animated deletion

```
Differences: 2022-04-04 06:50:49 to 2022-05-03 16:53:41
- Local abortion funding organizations may be able to provide assistance in these cases:
+ Providers of Abortion Services: A general hospital or other provider that is providing a
  bortion services for the Patient Applicant.
```

Addition memento (2020-05-01, 2 year lifespan) · Sliding diff

Figure 31. Users can filter by top level domain. In this example, the user has filtered by the top level domain virginia.gov.

The domain field is pre-populated by Solrwayback and the UK WARC Indexer. A user can also enter a URL. We translates the URL normalization algorithm into PHP from the UK WARC Indexer.[4] The Java normalization algorithm starts with the standard canonicalization algorithm written and used by the Internet Archive in tools like Heritrix and Open Wayback. The scheme *https* is converted to *http* and the subdomain www is removed. The slash is removed from most pages and added to top level domain addresses. UTF-8 is also escaped. While the standard Internet

[4] https://github.com/ukwa/webarchive-discovery

Archive canonicalization algorithm has been ported to Python,[5] it was never implemented in PHP. In order to implement both the canonicalization and the normalization algorithms in PHP, we used the standard PHP parse_url function to break the original URL into its parts. This is similar to the URI class in Java. The scheme *https* is converted to *http*. We ported the regular expression to remove *www* subdomains directly. The trailing slash is added or removed appropriately. This code is shown in Listing B.8 (Appendix B).

To implement the ranking scheme that reranks soft 404s behind other results, we used the addSort Solarium function in conjunction with a Solr function query as shown in Listing B.4 (Appendix B).

PHP and Solarium have been configured to run on port 8888. Port 8888 is the default port suggested in the Solarium documentation.

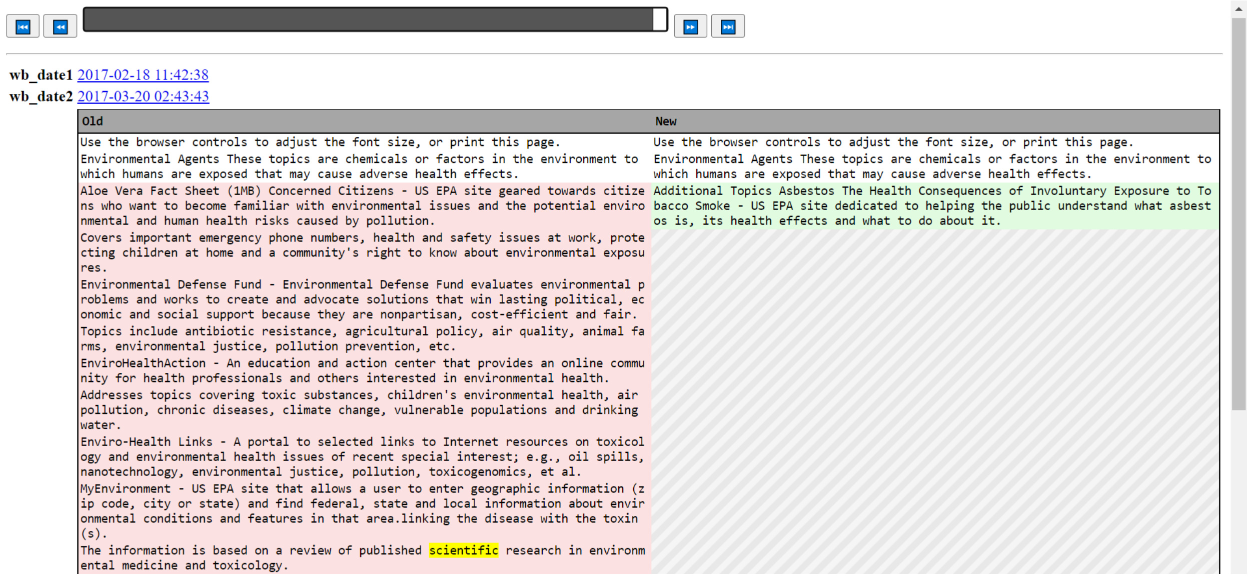
5.2.3 Viewing Sliding Differences

Each search result is an aggregation of multiple versions of the same page. Jackson et al. [70] noted that one of the disadvantages of grouping multiple versions of the same page is that it can hide the ways that pages change over time. Thus, there needs to be a way for the user to ungroup these versions. One method of presentation that would allow the user to view how each version has changed is to show each version's diff with the next version.

Each search engine result includes a link to compare all page versions between the addition and deletion with this sliding diff method as shown in Figure 32. Users can navigate between each set of differences using a slider. Henley [61] showed this technique to be effective for comparing source code differences over time. The diff is shown in the side-by-side format. The search term

[5] `https://github.com/internetarchive/surt`

is also highlighted in the diff.



wb_date1 2017-02-18 11:42:38
wb_date2 2017-03-20 02:43:43

Old	New
Use the browser controls to adjust the font size, or print this page.	Use the browser controls to adjust the font size, or print this page.
Environmental Agents These topics are chemicals or factors in the environment to which humans are exposed that may cause adverse health effects.	Environmental Agents These topics are chemicals or factors in the environment to which humans are exposed that may cause adverse health effects.
Aloe Vera Fact Sheet (1MB) Concerned Citizens - US EPA site geared towards citizens who want to become familiar with environmental issues and the potential environmental and human health risks caused by pollution.	Additional Topics Asbestos The Health Consequences of Involuntary Exposure to Tobacco Smoke - US EPA site dedicated to helping the public understand what asbestos is, its health effects and what to do about it.
Covers important emergency phone numbers, health and safety issues at work, protecting children at home and a community's right to know about environmental exposures.	
Environmental Defense Fund - Environmental Defense Fund evaluates environmental problems and works to create and advocate solutions that win lasting political, economic and social support because they are nonpartisan, cost-efficient and fair.	
Topics include antibiotic resistance, agricultural policy, air quality, animal farms, environmental justice, pollution prevention, etc.	
EnviroHealthAction - An education and action center that provides an online community for health professionals and others interested in environmental health.	
Addresses topics covering toxic substances, children's environmental health, air pollution, chronic diseases, climate change, vulnerable populations and drinking water.	
Enviro-Health Links - A portal to selected links to Internet resources on toxicology and environmental health issues of recent special interest; e.g., oil spills, nanotechnology, environmental justice, pollution, toxicogenomics, et al.	
MyEnvironment - US EPA site that allows a user to enter geographic information (zip code, city or state) and find federal, state and local information about environmental conditions and features in that area.linking the disease with the toxin (s).	
The information is based on a review of published scientific research in environmental medicine and toxicology.	

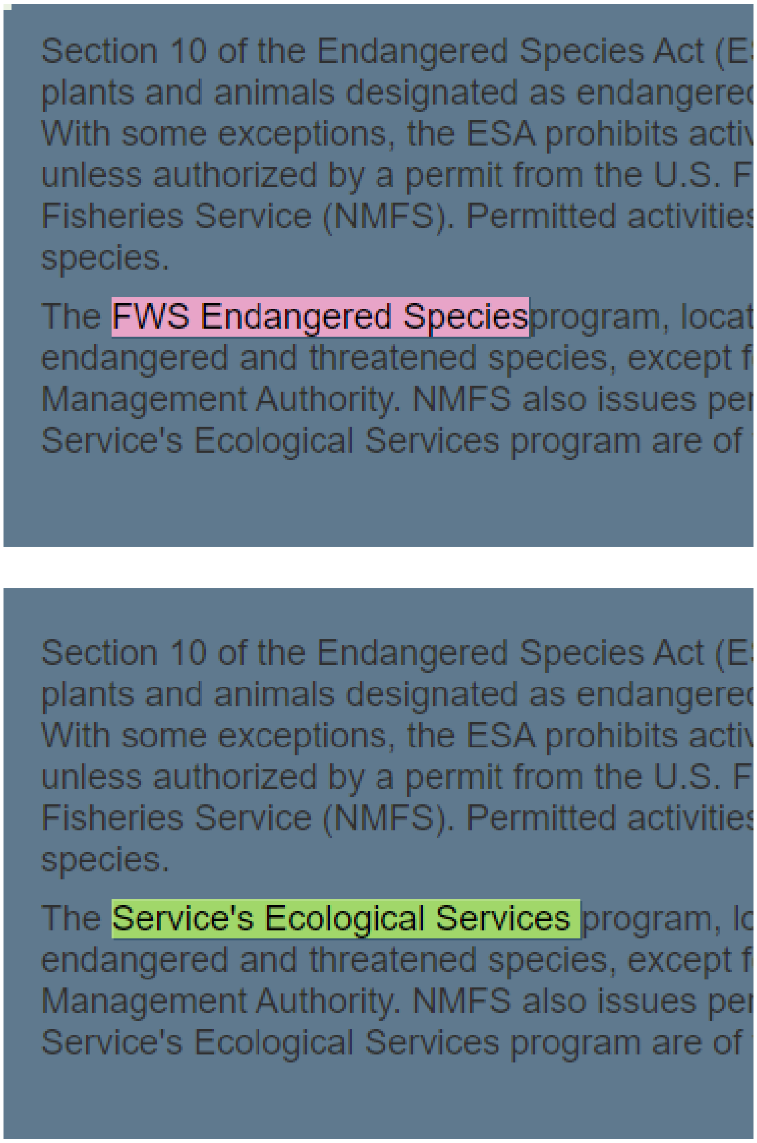
Figure 32. The sliding difference viewer shows the term ‘scientific’ was deleted from the page `https://www.niehs.nih.gov/health/topics/agents/index.cfm` in 2017, along with the context of the deletion. ©2023 IEEE [52]

With the sliding difference tool, the user can skip past identical mementos with the fast forward and rewind buttons. We created this sliding difference viewer by using the plaintext indexed content from Lucene. The viewer functions using a Solarium query, the PHP difference library [30], and some additional JavaScript. The date of addition and deletion is input to the sliding difference viewer via URL parameters, as shown in Listing B.5 (Appendix B). This can be used to query the index directly as shown in Listing C.1 (Appendix C). All page versions are highlighted with the system showed in Listing B.3 (Appendix B) and then provided as input to the JavaScript via hidden elements as shown in Listing C.2 (Appendix C). The JavaScript then accesses each element as shown in Listing C.3 (Appendix C).

Functionality to skip past identical versions is shown in Listing C.4 (Appendix C). Because the JavaScript stores the plaintext of each page, consecutive versions can be easily compared to determine if they are identical or not.

5.2.4 Viewing Animated Deletions

In addition to the sliding diff viewer, the search engine result also includes a link to replay the pre- and post-deletion versions simultaneously as shown in Figure 33. This dual replay tool shows an animation of the difference in context, and is shown in action at `https://youtu.be/qHSVvcubuYo`. The differences in the animation use hues. Both the animation itself along with the highlighted text colors lend themselves to pre-attentive processing of the changes. The animation jumps to each change in turn, in contrast to the static Changes Tool on the Wayback Machine. The animation is also meant to give the illusion of the changes happening in real time.



Section 10 of the Endangered Species Act (E
plants and animals designated as endangered
With some exceptions, the ESA prohibits activ
unless authorized by a permit from the U.S. F
Fisheries Service (NMFS). Permitted activities
species.

The FWS Endangered Speciesprogram, locat
endangered and threatened species, except f
Management Authority. NMFS also issues pe
Service's Ecological Services program are of

Section 10 of the Endangered Species Act (E
plants and animals designated as endangered
With some exceptions, the ESA prohibits activ
unless authorized by a permit from the U.S. F
Fisheries Service (NMFS). Permitted activities
species.

The Service's Ecological Services program, lo
endangered and threatened species, except f
Management Authority. NMFS also issues pe
Service's Ecological Services program are of

Figure 33. The animation shows the deletion of the phrase “endangered species" on the page

`http://www.fws.gov/ENDANGERED/permits/index.html`. ©2023 IEEE [52]

In order to create this animation, we used EDGI's Python HTML difference library.[6] Using this library, we calculate and combine the differences in a static context, and then extended the code to generate the HTML and JavaScript to animate the merged pages. We support successful difference calculation between page versions that originated from different web archives by using bannerless replay with PyWB. To draw the user's attention to each change, the page jumps to each change one-by-one, animates one change at a time, and pauses before jumping to the next change. Listing D.1 (Appendix D) shows the implementation of the pre-attentive processing algorithm. We chose to animate letter by letter for the first three words, and then word by word thereafter.

When using the EDGI HTML difference library to calculate the changes, it calculates all changes, not just the changes related to the query term. We removed deletions, for both cases of a query term and a query phrase, before performing the animation, as shown in Listing D.2 (Appendix D). Changes containing the query were then labeled for the animation.

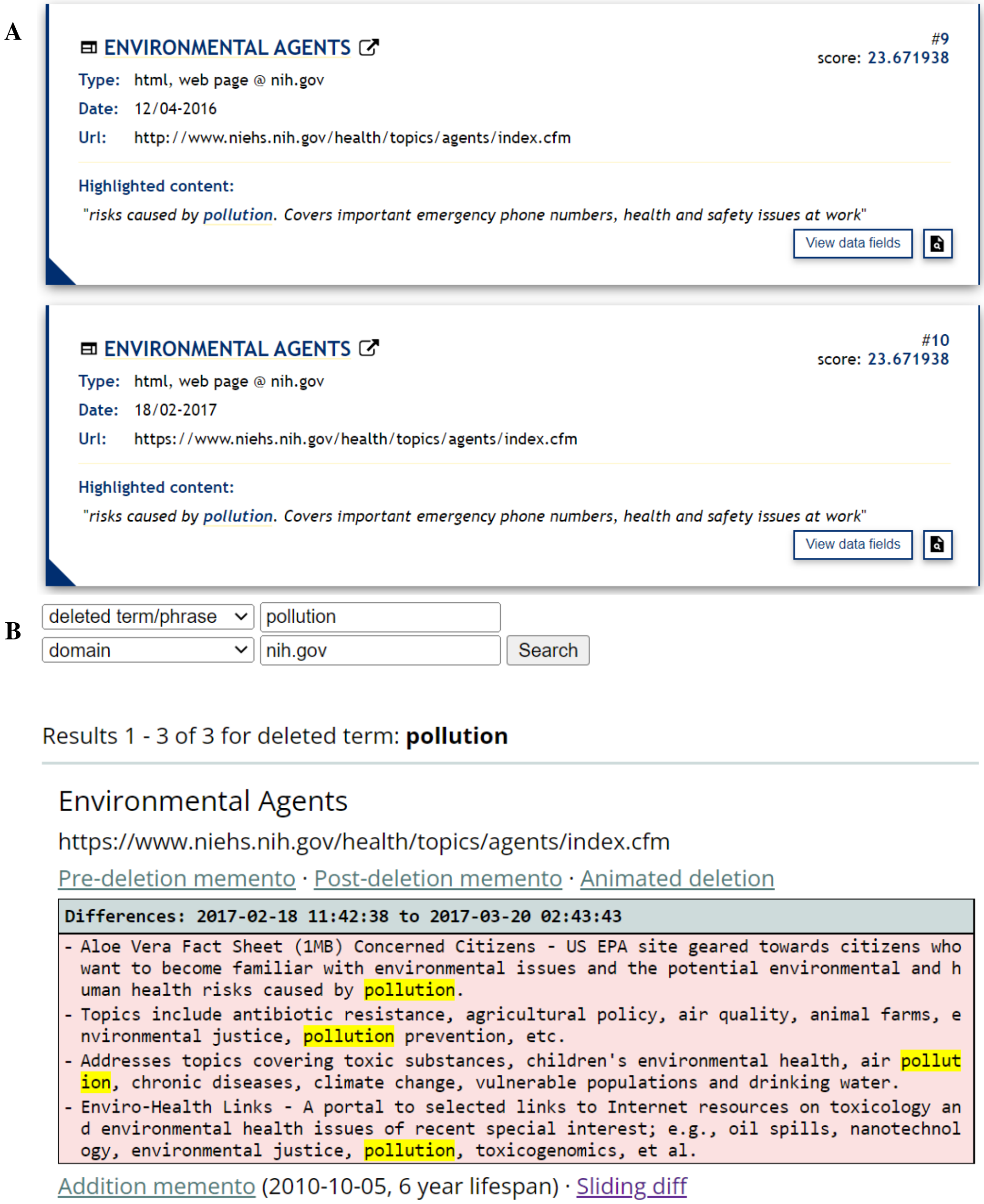
We also needed to remove additions not related to the query term. Listing D.3 (Appendix D) shows the implementation. Differences that remain but are not numbered are removed via regular expression match. Then the remaining differences are renumbered to ensure the animation algorithm functions properly.

Finally, the animation needed some additional HTML modifications to function properly. Because we added anchor links to control the animations, deletions with additional links needed to be modified so the anchor link would come first. Another change to the HTML that needed to be made was to add full addresses to relative links for images. This is a task that PyWB typically implements, but since the replay is outside of the PyWB environment, we implemented this ourselves.

[6] `https://github.com/edgi-govdata-archiving/web-monitoring-diff`

5.2.5 Discussion

The change text search engine results page has increased functionality compared to other temporal search engines for web archives, like SolrWayback, as well as compared to the web monitoring strategy used by EDGI. The change text search engine groups multiple versions of a page in a way that allows the user to make sense of the differences between them, as shown in Figure 34. In SolrWayback, multiple identical versions of the same page will be shown in the search results page, and the version right after the deletion will not be a part of the search results. In contrast, the change text search engine results page only shows versions with meaningful change of the query term. The search engine results page can also be used to examine changes in context because of the diff snippet, and further detail is available in the sliding diff and animation tools linked in each result. Examining changes in context is not possible with the web monitoring strategy.



ENVIRONMENTAL AGENTS

#9
score: 23.671938

Type: html, web page @ nih.gov
Date: 12/04-2016
Url: http://www.niehs.nih.gov/health/topics/agents/index.cfm

Highlighted content:
"risks caused by *pollution*. Covers important emergency phone numbers, health and safety issues at work"

ENVIRONMENTAL AGENTS

#10
score: 23.671938

Type: html, web page @ nih.gov
Date: 18/02-2017
Url: https://www.niehs.nih.gov/health/topics/agents/index.cfm

Highlighted content:
"risks caused by *pollution*. Covers important emergency phone numbers, health and safety issues at work"

deleted term/phrase: pollution
domain: nih.gov

Results 1 - 3 of 3 for deleted term: **pollution**

Environmental Agents
https://www.niehs.nih.gov/health/topics/agents/index.cfm
Pre-deletion memento · Post-deletion memento · Animated deletion

```
Differences: 2017-02-18 11:42:38 to 2017-03-20 02:43:43
- Aloe Vera Fact Sheet (1MB) Concerned Citizens - US EPA site geared towards citizens who
  want to become familiar with environmental issues and the potential environmental and h
  uman health risks caused by pollution.
- Topics include antibiotic resistance, agricultural policy, air quality, animal farms, e
  nvironmental justice, pollution prevention, etc.
- Addresses topics covering toxic substances, children's environmental health, air pollut
  ion, chronic diseases, climate change, vulnerable populations and drinking water.
- Enviro-Health Links - A portal to selected links to Internet resources on toxicology an
  d environmental health issues of recent special interest; e.g., oil spills, nanotechnol
  ogy, environmental justice, pollution, toxicogenomics, et al.
```

Addition memento (2010-10-05, 6 year lifespan) · Sliding diff

Figure 34. In SolrWayback, multiple versions of the same page are shown in the search results without any indication of how they are different. In the change text search engine results page, the versions of the page https://www.niehs.nih.gov/health/topics/agents/index.cfm that match the addition and deletion of the query term “pollution" are indicated clearly. *(A)* SolrWayback SERP *(B)* Change text SERP ©2023 IEEE [52]

5.3 NAVIGATIONAL REPLAY BANNER FOR CHANGES

We have shown that the navigation banner in the Internet Archive Wayback Machine's replay system is used, but is also not functioning at maximum utility for users. We propose functionality edits to the navigation banner based on the identified user needs from the log analysis. We suggest changes to how the arrow icons link to other mementos, incorporate an icon for the Changes Tool, prevent errors by warning users about redirects, and propose a new redirect landing page with more user autonomy.

5.3.1 Task: View Change Over Time

The user who is trying to view change over time likely has high familiarity with web archives. They use the TimeMap, timeline, or navigation buttons to manually identify changes on webpages. The timeline and navigation buttons are in the banner on the replay system. The current navigation banner used by the Internet Archive Wayback Machine is shown in Figure 35. The left red box is the timeline. The right red box shows a forward navigation button, but this box actually contains three links. The top link "MAY" contains the link to a memento approximately one month from the current memento, the middle button contains a link to the next consecutive capture of the page, and the bottom link "2011" contains a link to a memento approximately one year from the current memento.

The main problem with the navigation buttons on the banner are that they link to other versions of the page that are not meaningful for users trying to view consecutive changes. Based on the analysis of the log, users want to view the next version of the webpage with changes. Currently, the buttons may link to an identical version, or they may skip versions with changes. The buttons

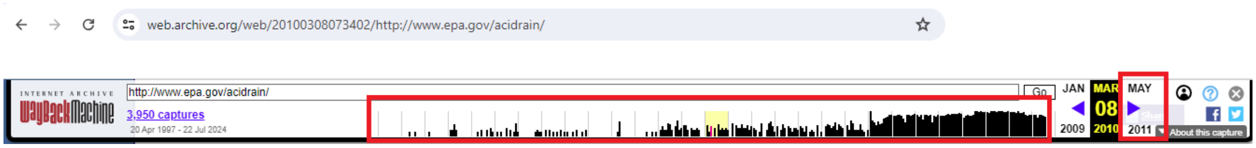

web.archive.org/web/20100308073402/http://www.epa.gov/acidrain/
INTERNET ARCHIVE
WaybackMachine
http://www.epa.gov/acidrain/
Go
3,950 captures
20 Apr 1997 - 22 Jul 2024
JAN
MAR
MAY
2009
08
2010
2011
About this capture

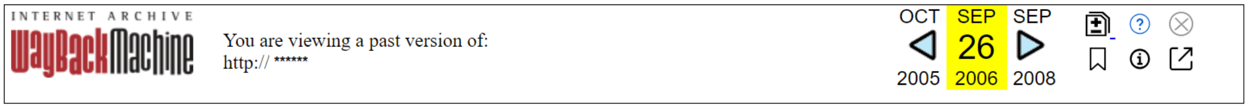
Figure 35. Current Internet Archive Wayback Machine banner navigation buttons. The timeline is the left red box, and the navigation buttons are in the right red box. The right box actually contains three links to nearest captures by day, month, and year.

should link to the previous and next versions of the page, not the previous and next captures, and should not skip any versions either. By changing where these buttons link, Slide Users can more easily identify changes. In addition, the Changes Tool is only linked from the TimeMap, not from the replay system. There should be a link in the replay banner to the Changes Tool so that users can see the changes highlighted in a diff-style, rather than having to manually identify the changes by examining both mementos.

In Figure 36, these functionality aspects have been incorporated into a new banner prototype. The buttons now link to the previous and next versions of the page with changes. There is also an icon, top row left, that allows the user to access the Changes tool quickly for difference between the current and previous versions of the page. The previous version is the last version in the previous validity range, and the next version is the first version in the next validity range. In the next version of the prototype, we would add a drop down below the memento's date containing a list of all of the mementos in the current validity range.

5.3.2 Task: View Unavailable Page

When a user tries to view a page with a redirect, they currently have a hard time going to a



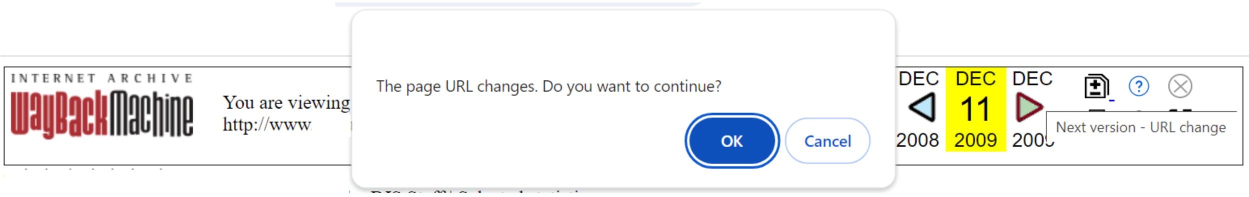
You are viewing a past version of:
http:// ******

Figure 36. Proposed banner navigation buttons. The buttons link to the previous and next versions of the page with changes. There is also an icon, top row left, to compare the current and previous versions with the Changes tool. The help and close icons, also on the top row, function the same as in the current interface. The bottom row replaces current "save to my web archive" window shaped button with a bookmark button, more consistent with current iconography on social media sites. The current "about this capture" label is replaced with an info button. Individual social media buttons are replaced with an aggregated share icon.

valid version of the page because the navigation bar disappears and they are forcefully redirected. The redirects are confusing to the user. Sometimes, the user is linked to a redirect page for their first memento, and they have no way of navigating back in time to a valid version because the navigation bar has disappeared. Another problem is that the redirect should go to the new location of the page, even if the machine readable redirect goes somewhere else, such as in the case where it goes to the main page of the site at the new address rather than the correct subpage. For example, the NOAA Fisheries website moved from nmfs.noaa.gov to fisheries.noaa.gov before 2020, and most of the pages redirect to a top level page instead of the sub page at the new location.

In the new proposed interface, the user retains the navigation bar, and they are given 3 choices: to return to the last working (HTTP 200 status code) version of the page, to follow the redirect (with information about text similarity), or to go to a suggested alternate that may better represent the subpage (with information about text similarity).

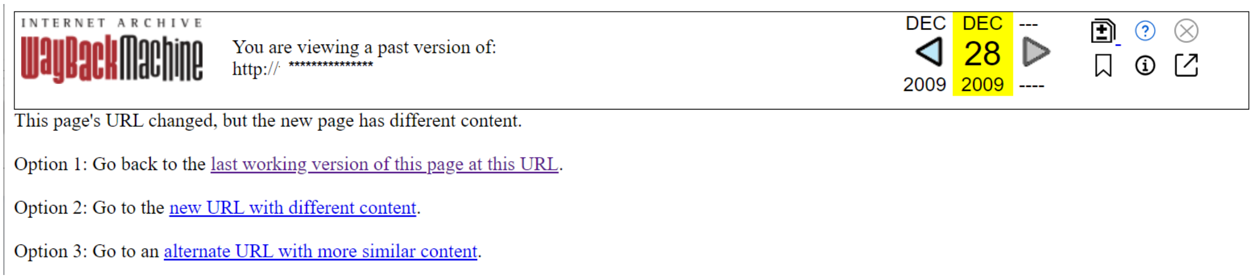
Figure 37 shows the proposed changes to the banner for when the page's URL changes, and has a redirect 3xx HTTP status code for the next capture. These design changes directly follow Nielsen's Usability Heuristics [109]. First, the arrow icon is the same color (green) as is used for redirects on the TimeMap. Note that the arrow icon for HTTP 200 mementos is blue, which is also the same color as is used on the TimeMap. Not shown, 404 status codes would be the same light orange color as on the TimeMap as well. This follows the principle of consistency. Next, there is a warning when hovering over the arrow icon that explains why it is a different color. In addition, when the arrow icon is clicked, the user is presented with a pop-up confirming the choice. These design choices follow the heuristic of error prevention.


The page URL changes. Do you want to continue?

Next version - URL change

Figure 37. Proposed banner redirect warnings. The arrow icon is the same color as on the TimeMap for redirects (green), there is a warning on hover for the arrow icon, and there is a pop-up confirmation box after clicking the arrow icon.

Figure 38 shows the proposed redirect landing page. The current redirect landing page is shown in Figure 20. One thing to note is that the current landing page actually has a 200 OK HTTP status itself. The proposed redirect page would have a 300 Multiple Choices HTTP status code. The language is user-friendly instead of jargon heavy, which follows the heuristic of match between system and real world. This redirect landing page also gives the user autonomy. While the previous landing page automatically redirected, this landing page gives the user at least two choices. In the more straightforward case of a correct machine readable redirect, the user would be able to choose between going back to the last version of the page before the redirect or following the redirect. In the less straightforward and fairly prevalent case where the machine readable redirect is incorrect, the user could be presented with three options: previous version, follow the redirect, or go to a page different from the machine readable redirect that has a higher text similarity match.



You are viewing a past version of:
http:// ***************

This page's URL changed, but the new page has different content.

Option 1: Go back to the last working version of this page at this URL.

Option 2: Go to the new URL with different content.

Option 3: Go to an alternate URL with more similar content.

Figure 38. Proposed redirect landing page. The user has autonomy to follow the redirect or go to a different page if the machine readable redirect is incorrect.

Implementing this new redirect page would require additional infrastructure. Identifying whether or not the redirect has similar text to the current page can be implemented using existing text similarity algorithms. However, if the redirect is below the chosen similarity threshold, finding a better match via full text is not a task that web archives currently support. There are two main possibilities: the machine readable redirect is incorrect, but the page exists at a different URL. This situation is common when websites are reorganized. One approach to this problem by Zhu et al.

[142] is to use a combination of live web search along with any successful redirects in web archives for pages with similar addresses to learn the new URL patterns. Klein et al. [89] investigated how well text-based lexical signatures, titles, link neighborhood lexical signatures, and social media generated tags perform for finding relocated webpages. They found that titles perform well, especially given their ease to compute and index compared to the other choices. The Internet Archive Wayback Machine has indexed metadata including titles for all of its holdings, so this solution could be viable for a large web archive.

Another possibility is that the page is not available on the live web or in the archive. In this case, the task becomes recommending a page with similar content rather than finding the new location. Prior work on this topic has focused on scalable solutions that only analyze metadata such as titles and URIs. Alkwai et al. [11] developed a method to recommend these "lost" pages that was successful when the URI had sufficient descriptors included. Additional work would be necessary to identify recommendations for pages with nondescriptive URIs.

5.4 SUMMARY

In this chapter, we addressed Research Question 1, How can we make changes in webpages discoverable and understandable? We created a change text search backend and frontend so that users can search for changes in webpages. We also created two tools to visualize these changes, a changes animation and a sliding differences tool. We also addressed Research Question 2, How can we increase efficiency in web archive user navigation for viewing change over time? We proposed a prototype for a navigation banner that considers changes in webpages instead of only linking in set temporal intervals. We also link the existing Wayback Machine Changes Tool directly in the banner for quicker access for the user to view the specific changes on the page.

CHAPTER 6

EVALUATION AND DISCUSSION: CHANGES ON US FEDERAL SITES

In this chapter, we address Research Question 3, How can aggregated webpage changes of a corpus be used computationally to provide compelling evidence for edit intentions? We evaluate the change text search index with the EDGI [110] federal environmental websites dataset, which covers one presidential term from 2016 to 2020. We first analyze term changes during this presidential term. We validate previously found deleted terms and contribute new frequently deleted corpus terms. Next, we align the EDGI dataset with the ORCAs dataset [19]. We find that environmental query terms were deleted from these webpages during this 2016-2020 timespan. Portions of this chapter are based on our work published at ACM/IEEE JCDL 2023 [52] and our work published at ACM CIKM 2024 [53].

6.1 TERM CHANGES DURING ONE PRESIDENTIAL TERM

Federal webpages with changes between 2016 and 2020, as calculated by EDGI, form an excellent data set for evaluation of a change text search engine. The EDGI data set consists of about 40,000 web page addresses. Approximately 10,000 of these web pages have versions in 2016 and 2020 at the Internet Archive. First, we expanded the data set to include more of the original pages by considering additional web archives. Memgator [9] queries about 20 web archives. Next, we filtered the results so that only pages with successful HTTP status codes (HTTP 200 OK) are kept for indexing. Finally, we prioritized pages with known term changes, and we identified additional versions of these pages inside of the four-year window with salient changes for indexing.

EDGI examined each monitored page to determine if a capture from both the first half of 2016 and the first half of 2020 existed, which they defined as a paired-page sample. Since only about 10,000 pages of the 40,000 monitored pages had paired mementos at the Internet Archive, there are about 30,000 pages to examine for paired mementos using other web archives. We generated a TimeMap for each of these 30,000 web page addresses using MemGator, a memento aggregation service. Examining the TimeMaps, about 8,500 of the 30,000 pages do have newly found paired mementos. Interestingly, about half of these new pairs exist at the Internet Archive. It is possible that these mementos were added after some kind of delay or embargo period. The other pairs directly rely on web archives besides the Internet Archive for at least one of the mementos in the 2016/2020 pair.

In the original set of approximately 10,000 paired mementos, about 75% of the pairs have successful HTTP status codes for both 2016 and 2020, as shown in the top of Figure 39. In the additional set of approximately 8,500 paired mementos, about 38% of the pairs have successful HTTP status codes for both 2016 and 2020, as shown in the bottom of Figure 39. In all, there are about 11,000 pairs of 2016/2020 mementos that are candidates for indexing, as shown in figure 40. EDGI collected their data in a way that minimized non-successful HTTP status codes by using the Internet Archive's CDX API.[1] For the new pair set, other web archives do not have public CDX APIs and TimeMaps do not have any status code information included, which is what led to more pages in the new pair set having non-successful HTTP status codes.

[1] `https://github.com/internetarchive/wayback/tree/master/wayback-cdx-server`

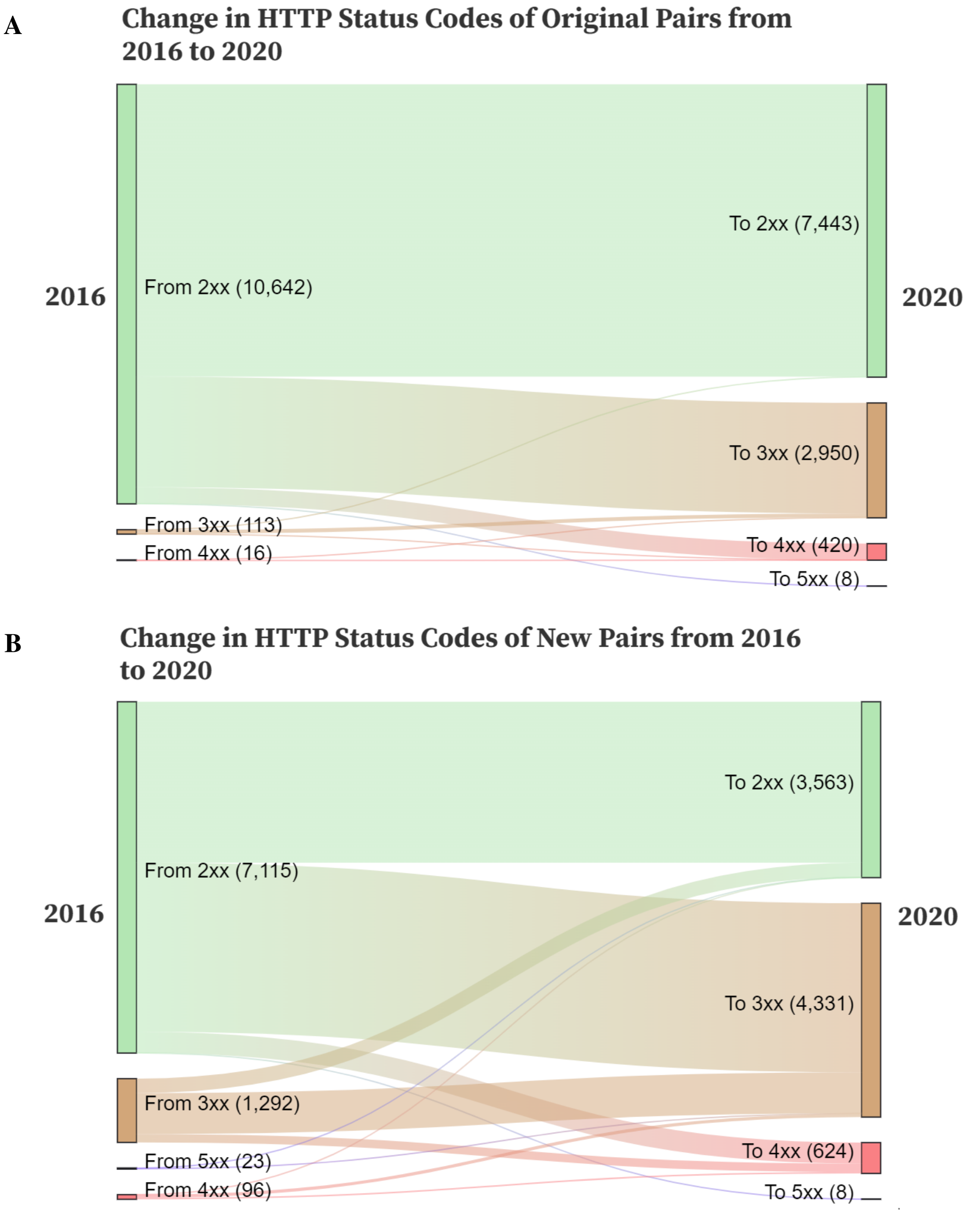


Figure 39. Out of about 40,000 seed URI-Rs in the EDGI data set, approximately 11,000 have successful (200) HTTP status codes in both 2016 and 2020. We found an additional 3,500 mementos from multiple web archives with successful status codes. *(A)* About 75% of the paired mementos identified by EDGI have successful (200) HTTP status codes in both 2016 and 2020. *(B)* In the new pairing set, an additional 3,500 paired mementos have successful (200) HTTP status codes in both 2016 and 2020. ©2023 IEEE [52]

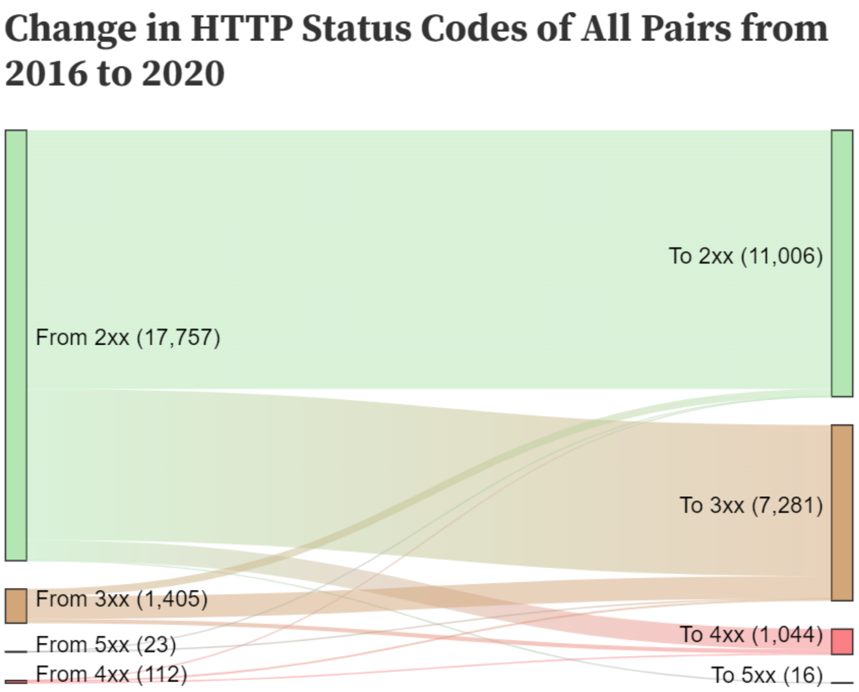


Figure 40. Combined chart showing HTTP status codes for all pairs.

The Wayback Machine holds both paired mementos in the original EDGI set. In contrast, the new pair set has mementos located at multiple web archives. In this new set, there are 3,563 mementos with successful status codes in both 2016 and 2020. Table 4 shows the web archives that hold these mementos, according to their TimeMaps. Some URI-Rs have multiple valid mementos in the time range of interest, which is the first half of 2016 and 2020. In the table, these URI-Rs are counted once for each web archive. The percentages represent the number of URI-Rs with at least one memento found in the TimeMap in the specified time range, divided by 3,563. The percentages allow for comparison of holdings between archives and across years. Since each URI-R may have a memento at multiple archives, the sums of the columns are non-meaningful.

Indexing the paired 2016/2020 mementos is a starting point for determining additional versions

Table 4. Percentage of the URI-Rs found at each archive for mementos in the first half of 2016 and 2020 for the 3,563 pages in the new pairing set with 200-to-200 status codes, as shown in the bottom of Figure 39. ©2023 IEEE [52]

Web Archive	% (2016)	% (2020)
webarchive.loc.gov	59.81	60.93
web.archive.org	34.41	71.15
wayback.archive-it.org	12.41	10.36
arquivo.pt	5.95	3.59
perma.cc	0.39	0.39
web.archive.org.au	0.11	0.08
swap.stanford.edu	0.11	0.06
wayback.vefsafn.is	0.11	0.20
waext.banq.qc.ca	0.08	0.08
www.webarchive.org.uk	0.03	0.08
archive.md	0.03	0.06

of each page to index. The change text calculation script will determine all terms that have been added and removed between the two versions for each page indexed. Then, a binary search over the other versions of the page can be used to increase the temporal granularity of when a term was changed. Pages that were already identified as containing a term or phrase deletion by EDGI are prime candidates for early indexing. Since EDGI only tracked about 50 terms and phrases, additional meaningful terms can be identified from the change text calculations.

The initial indexing consisted of a small set of 100 pairs of mementos. EDGI calculated that these pairs had complete or partial deletions of the terms *sustainability, pollution, anthropogenic,* or the phrase “endangered species.” The next indexing set was larger, consisting of 1,000 pairs, or 10% of the original EDGI matched pairs. These pairs had complete or partial deletions of the terms and phrases: “toxic", “clean energy", “climate change", and “global warming". Many of the trends that emerged from the initial small sample persisted into the larger sample, so these trends are likely to persist throughout the entire data set.

Downloading the 1,000 pairs using Hypercane, part of Level 1 on Figure 25 (Chapter 5), took 41 hours. Initial Lucene indexing of the WARCs using the UKWA WARC indexer took 2.5 hours, and PyWB indexing took 40 minutes. The change text calculation script took no more than 10 seconds to generate the JSON updates to post to the Lucene index, and posting the data took no more than 5 seconds. These steps correspond to Level 2 on Figure 25.

While the original EDGI study only tracked 56 terms and phrases with environmental motivations, additional common deleted terms are now discoverable, as shown in Table 5. The original term list included terms like “safety", “transparency", “regulation", and “jobs", but “public", “access", “action", “development" and “science" are terms with similar meaning that were also commonly deleted. Many of the top deleted terms were stop words or temporal terms. Two of the original EDGI terms, “change" and “state", are technically stop words, but that have domain-related meaning within a federal environmental dataset.[2] The list of newly found terms, in order by most deletions, is shown in Table 6.

[2] `https://countwordsfree.com/stopwords`

Table 5. Categorization of top 100 deleted terms in the 1,000 paired memento index sample. ©2023 IEEE [52]

Term Type	Examples	Count
Stop words	about, more, which, state	49
Temporal	2015, 2, one, year	13
EDGI	climate, clean, impacts, water	11
Newly found	public, development, access, science	29

6.1.1 Discussion

The effectiveness of the change text search index leads to implications for researchers examining change in a corpus, such as how the webpage data should be collected from web archives. The Library of Congress web archive holds mementos for many pages that are not available in the Wayback Machine in the first half of 2016. Future researchers examining U.S. federal websites should utilize the Library of Congress web archive in addition to the Wayback Machine. Secondly, it appears that Arquivo.pt conducted a crawl of U.S. federal websites in May 2016, and mementos with similar datetimes are not available in the Wayback Machine or at the Library of Congress. Using MemGator to query for aggregated listings of mementos across all web archives not only increases the amount of historical evidence available to analyze, but also brings light to entire collections and the possible motivations behind the creation of those collections.

Another implication of the change text index for researchers involves how text is extracted from the pages. One of the major differences between indexing text content with Lucene and the

Table 6. Count of top 100 deleted terms in the 1,000 paired memento index sample.

Terms	Count Range
national, support	at least 100
public	90 - 99
program, resources	80 - 89
process, data, including, u.s., development, united, learn, department, action, access, work, impact, tools	70 - 79
areas, search, laboratory, technology, efforts, include, natural, science, planning, address, open	60 - 69

deleted terms calculation completed by EDGI using Python is that the methods have completely different boilerplate removal techniques. Generally, the Lucene boilerplate removal strips more from the page than the EDGI technique. This meant that some terms were identified as deleted according to EDGI, but not according to the Lucene index. Other pages in Lucene indexed poorly due to the boilerplate removal, which is another difference in the deleted terms calculation. Some of the terms identified as deleted by EDGI were in various navigation page sections that were not stripped during boilerplate removal, so whether or not these should be named as deletions depends on the individual researcher's boilerplate removal preferences. Additionally, indexing all of the content with Lucene provides access to all of the deleted terms, while the EDGI web monitoring technique can only track pre-defined terms. The ability to find deleted terms without pre-defining them is powerful, which may outweigh poor indexing for some researchers.

A third implication of the change text index for researchers is use of the index to discover changed terms corpus wide. Additional discovery of deleted phrases is possible by examining the search results for the new deleted terms. For example, “public" is a new deleted term comparable to the original term “transparency." Manual inspection of the search results for “public" showed that the term often comes up in the context of a deleted phrase. “Public comment" was a common context for the deleted term “public," and this usage is similar to the original term “transparency." Another more frequent context was the phrase “public health," which is actually more similar to the original term “safety" than “transparency."

There are also implications from the change text search interface, and how changes should be shown to users. The two additional tools we have provided for examining changes in detail, the sliding diff and the animation as discussed in Chapter 5, each have advantages for different contexts. The sliding diff tool can show more than two versions of a page, while the animation can only show two versions of a page at a time. The sliding diff also shows the changes in a persistent manner, while the deleted text in the animation view disappears by nature. Another benefit of the sliding diff tool is that it loads extremely quickly, because only content is compared. The animation must load both mementos through replay, compute the HTML difference, and then load and show the animation, so it takes much longer. The sliding diff shows the blocks of text with changes over time well, but diffs hide content that has not changed. If the user wants to view the entire page in context, the animation shows the entire page, including text and images that have not changed. The animation also makes the differences pop out more strongly than the sliding diff tool. Both tools suffer when there are too many changes on the page, because there is too much highlighted and the animation is too slow.

Finally, the EDGI federal webpages dataset was both valuable and appropriate for evaluating this change text search engine. Documents created by the government are not entitled to privacy; in fact, these documents should be archived and made discoverable for the people. However, researchers must take great care when indexing web archive collections. Heterogeneous web archive collections may contain material that should be subject to stronger privacy protections, such as websites made by minors. Lin et al. [96] speculated that people are unaware their websites are archived because they never gave explicit consent, and other people rely on a lack of a web archive search function to provide them with privacy. Mackinnon [100] confirmed both of these hypotheses in a user study. Both Lin et al. and Mackinnon suggest anonymization as a tool for researchers to study overall web archive collections without compromising the privacy of the content creators.

6.2 CHANGES IN RETRIEVABILITY DURING ONE PRESIDENTIAL TERM

In the previous section, we showed the utility of the changes search engine to discover deleted terms in the environmental dataset. In this section, we consider the impact of these deletions on retrievability.

We use the change-text search engine to determine which query terms were deleted on the pages between 2016 and 2020. We align the EDGI dataset [110] with ORCAS, a real-world user query click dataset [19] created during the same time period. We found a pattern of document manipulation that removed query terms, resulting in lower retrievability. Below, we detail how we aligned the EDGI and ORCAS datasets, computed page lifespans and deleted terms, and finally determined the prevalence of queries with deleted terms.

6.2.1 Datasets

Nost et al. [110], on behalf of the EDGI, monitored changes on 30 US federal environmental agency websites between 2016 and 2020. They compared the change in 56 pre-chosen environmental terms and phrases on 40,000 webpages using the web archive holdings at the Internet Archive. The Open Resource for Click Analysis in Search (ORCAS) [35] is a dataset with Microsoft Bing queries and clicks from 2017 to 2020. The dataset employs k-anonymity, so each query in the dataset was made by a large number of different users. This protects user privacy, but also ensures that the queries in the dataset are popular. Since the datasets cover the same time period, it makes sense to compute their intersection to analyze how the webpage changes an the user queries are related.

6.2.2 Intersection Between ORCAS and EDGI

A simple lexicographical intersection query is not enough to determine which US federal environmental agency webpages are present in ORCAS, because multiple similar URLs can resolve to the same webpage. SURT resolves this problem. After transforming all URLs in ORCAS and EDGI, we computed the intersection. There are 573 webpages in the intersection of both datasets, 71 of which have queries for deleted terms tracked by EDGI. The intersection represents a real world dataset of the pages that were important enough to track for changes in real-time but also had enough clicks to uphold users' privacy rights.

In all, the median number of queries per page in the intersection is 12 queries, and the mid-spread of the queries ranges from 4 to 34 queries. The distribution skews right: there are 62 pages with over 80 queries. Figure 41 shows this skewed distribution with a logarithmic scale. Because of the way ORCAS was constructed to maximize user privacy, each of these queries has a large minimum number of associated actual users and corresponding clicks.

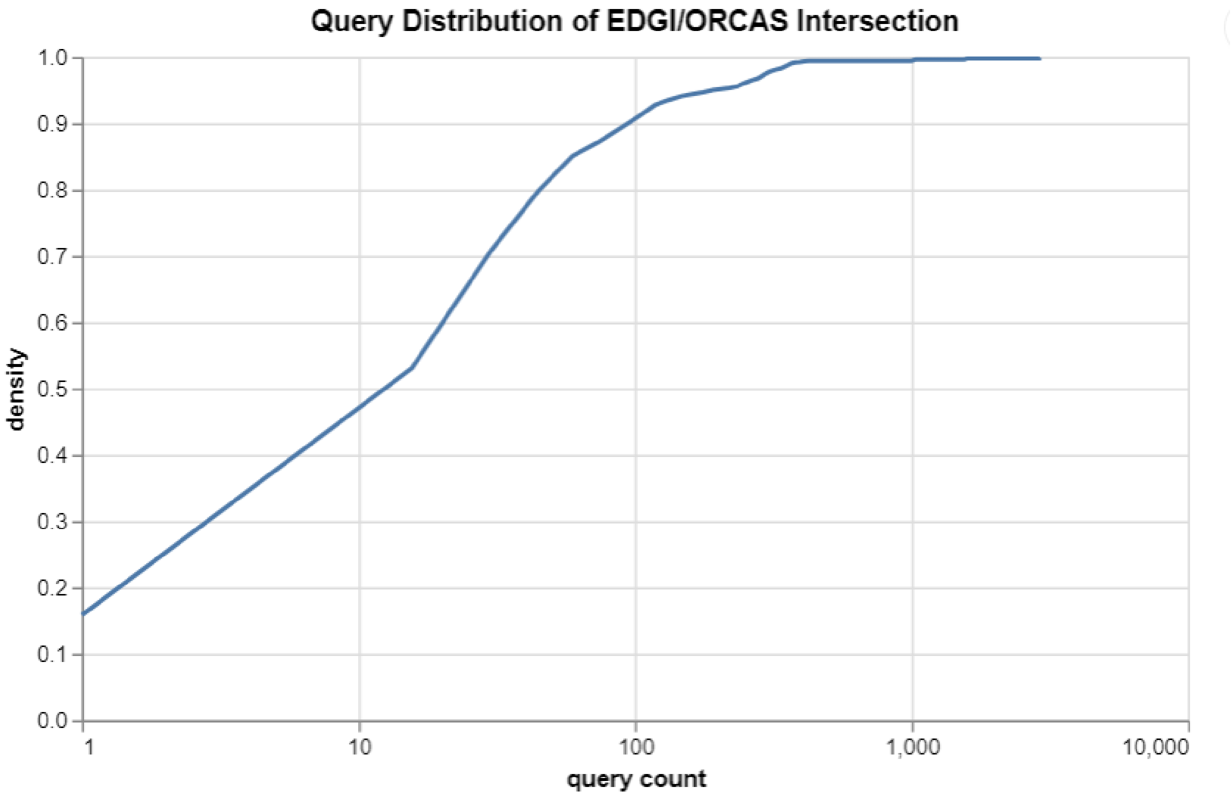


Figure 41. The Query Distribution of the EDGI/ORCAS Intersection Skews Right

6.2.3 Examining Lifespans of Pages

The ORCAS dataset was created from clicks between November 2017 and January 2020. However, not all webpages were functioning for that entire period of time. In order to calculate the lifespan of a webpage, the Internet Archive provides an API for its crawl indices (CDX).The CDX API takes a URL as input, and outputs a listing that includes all captures, their timestamps, and the status code of the page at the time of capture.

Of the 573 pages in the intersection, only 483 had a valid memento between January and June in both 2016 and 2020. These are called paired mementos [110]. By 2020, 13 pages were not found, 89 pages were moved and found, and 381 pages were never moved.

6.2.4 Comparing Known Term Changes to Queries

Nost et al. on behalf of EDGI [110] contributed a dataset with two matrices, one for 2016 and one for 2020, with the 40,000 rows representing the webpages and the 56 columns representing the term counts. To determine which pages have changes per term, the two matrices must be put together and categorized by the change in term count. A page that never contained the term is excluded. EDGI used 999 as an error code in the term count, so those pages must also be excluded. A *full deletion* or *full addition* is a page that had a 0 count of the term to end or start, respectively. A *semi-deletion* or *semi-addition* is a page with a change in term count without being zeroed out. A *static* page contains the term with no count change.

For each of the 56 terms, the EDGI dataset can be queried to determine if the term count changed between 2016 and 2020, and ORCAS queries for those pages and terms can be looked up directly.

6.2.5 Indexing More Term Changes

We used the change text search index discusssed in Chapter 5 to analyze the corpus created by the intersection. The ORCAS dataset is queried for each page in the EDGI dataset, then each term of each query is searched in the change-text index.

6.2.6 Computing Changes on Redirected Pages

Analyzing the CDX files, some pages did not retain a successful (200) HTTP status code through 2020. For each page with a redirect (3xx HTTP status code), a script followed the machine readable redirect and recorded the resulting URL. This set of paired URLs could then be queried against the list of all pages in the Lucene index to determine their IDs. In the change text calculation script, documents with the same canonicalized URLs are automatically linked. Using these ID pairs as additional input, the redirected paired pages were also linked for change text calculations.

6.2.7 Finding Queries with Deleted Terms

For each of the 483 pages with paired mementos, the change terms from the change-text index were compared with that page's queries from ORCAS. When processing the queries for deleted terms analysis, stopwords were removed.Both dictionary and non-dictionary terms were considered, because many of the queries included agency acronyms like USGS (United States Geologic Survey) and place names (such as names and abbreviations of states). For example, the most common query term was "OSHA." There were typos included in the queries, for example, (mis)spellings of "climate change" appear in the queries for climate.nasa.gov, such as "cimate" and "chage." Because we considered non-dictionary terms in our analysis, these typos were included; however, none of these typos appear as common deleted query terms corpus-wide.

6.2.8 Identifying More Environmental Queries

EDGI only tracked 56 environmental terms and phrases, but there are more than 56 terms and phrases that represent environmental concepts. In order to identify which additional deleted query terms were environmental, we used the Environmental Ontology [27]. We exported the terms from WikiData using a SPARQL query[3] for all terms with an Environmental Ontology ID (P3859). The query is shown in Listing 6.1.

[3] https://query.wikidata.org/

Listing 6.1. SPARQL query for environmental terms

```
SELECT
  ?item ?itemLabel
  ?value ?valueLabel
WHERE
{
  ?item wdt:P3859 ?value
  SERVICE wikibase:label {
  bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

6.2.9 Searching for Deleted Terms

In 81 of the 483 paired EDGI pages found in the ORCAS query dataset, all of the queries that returned the page contained terms that were partially or completely removed from the page. One example is the EPA's Air Pollution Challenges webpage, `epa.gov/clean-air-act-overview/air-pollution-current-and-future-challenges`. EDGI monitored the term "pollution", which was *semi-deleted* on this page and was present in 75% of the queries for this page in ORCAS. Example queries include "fire pollution" and "pollution levels". The remaining 25% of queries all contain the term "air". Example queries include "clean air act" and "air quality by state." This term "air" was not monitored by EDGI, but can be queried using the change-text index: it is also a semi-deleted term for this page. Another example is the Global Change Research Program's page on US Impacts, `globalchange.gov/browse/reports/global-climate-change-impacts-united-states`.

EDGI monitored the term “climate” that was present in 100% of ORCAS queries for this page and was *semi-deleted* from this page. Since standard ranking algorithms typically take into account term counts, it is reasonable to assume that removing occurrences of all of the actual query terms used to access these pages would lower their retrievability. Figure 42 shows the prevalence of deleted queries by page for the 483 pages. More than two-fifths of the pages had deleted terms in at least half of their queries.

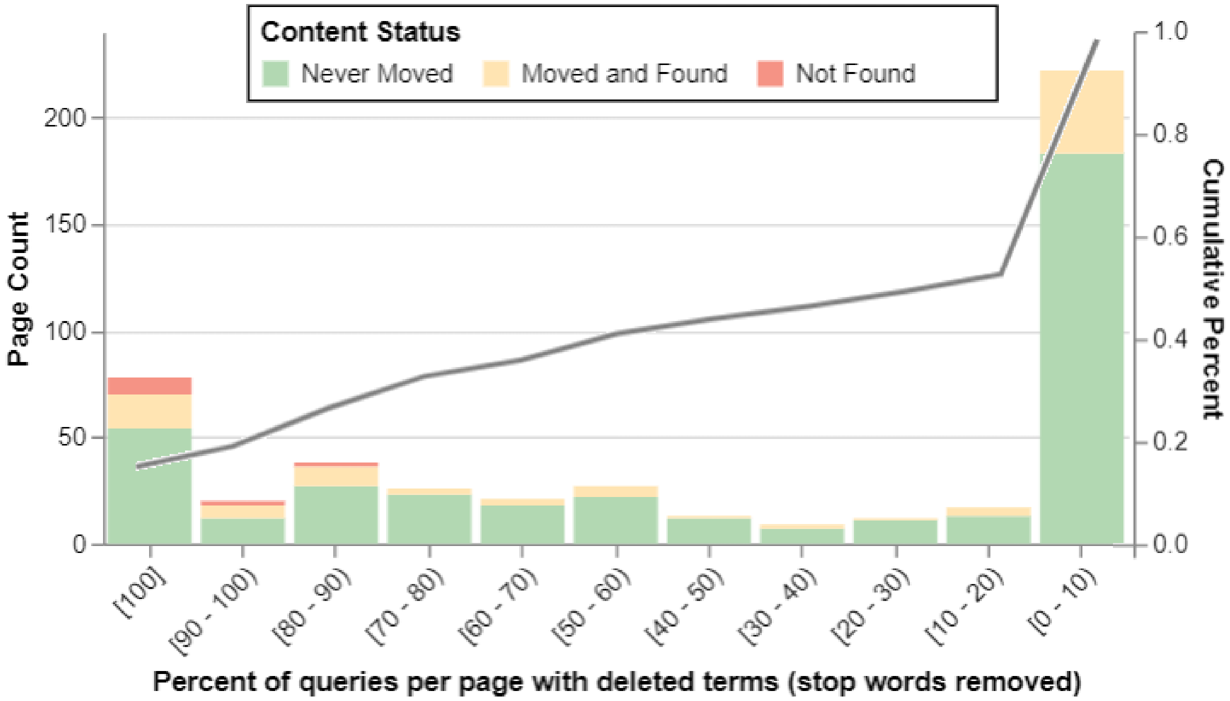


Figure 42. Over Two-Fifths of Pages had Deleted Terms in at Least Half of Their Queries.

Deleted and semi-deleted query terms common across the 483 pages include “OSHA” (30 pages), “energy” (23), “data” (16), “safety” (14), “health” (13), “air” (13), “gas” (12), “office” (12), “water” (11), and “EPA” (11). Of these terms, only “safety” is an EDGI tracked term; “energy” (clean energy), “air” (air quality), “gas” (unconventional gas) and “water” (water quality) are EDGI tracked phrases.

Beyond EDGI’s tracked terms, 22% of pages contained a query for an exact match to an Environmental Ontology term or phrase. 34% of pages contained a query for at least one environmental term, and 59% of pages with at least one deleted query term had a deleted environmental query term. 21% of deleted queries contain an environmental term. For example, the commonly deleted term “data” is not an environmental term, but “solar” is. Solar was queried for but deleted on 3 pages and is not a term or part of a phrase tracked by EDGI, but can be now identified and analyzed by using the Environmental Ontology.

6.2.10 Changes in the Corpus’ Query Terms Ratio

Vasilisky et al. [138] took the count of the query terms in a document divided by the term count in the document, and compared these ratios on different versions of the documents. They defined this comparison difference as the QueryTermsRatio measure. The average relative QueryTermsRatio [138] from 2016 to 2020 was 0.02. The positive ratio indicates that the 2016 versions have more query terms present than the 2020 versions. This is the opposite result expected for a typical corpus with pages undergoing search engine optimization, where pages tend to increase search terms over time [138]. A positive average relative query terms ratio also indicates lower retrievability of the pages in the corpus.

The results from the histogram in Figure 42 show that about half of pages did not have deleted query terms. The corpus can be divided into two distinct parts: pages that underwent traditional search engine optimization, and pages whose query terms were deleted. Figure 43 shows the percent of queries with deleted terms versus the average QueryTermsRatio by agency, for agencies

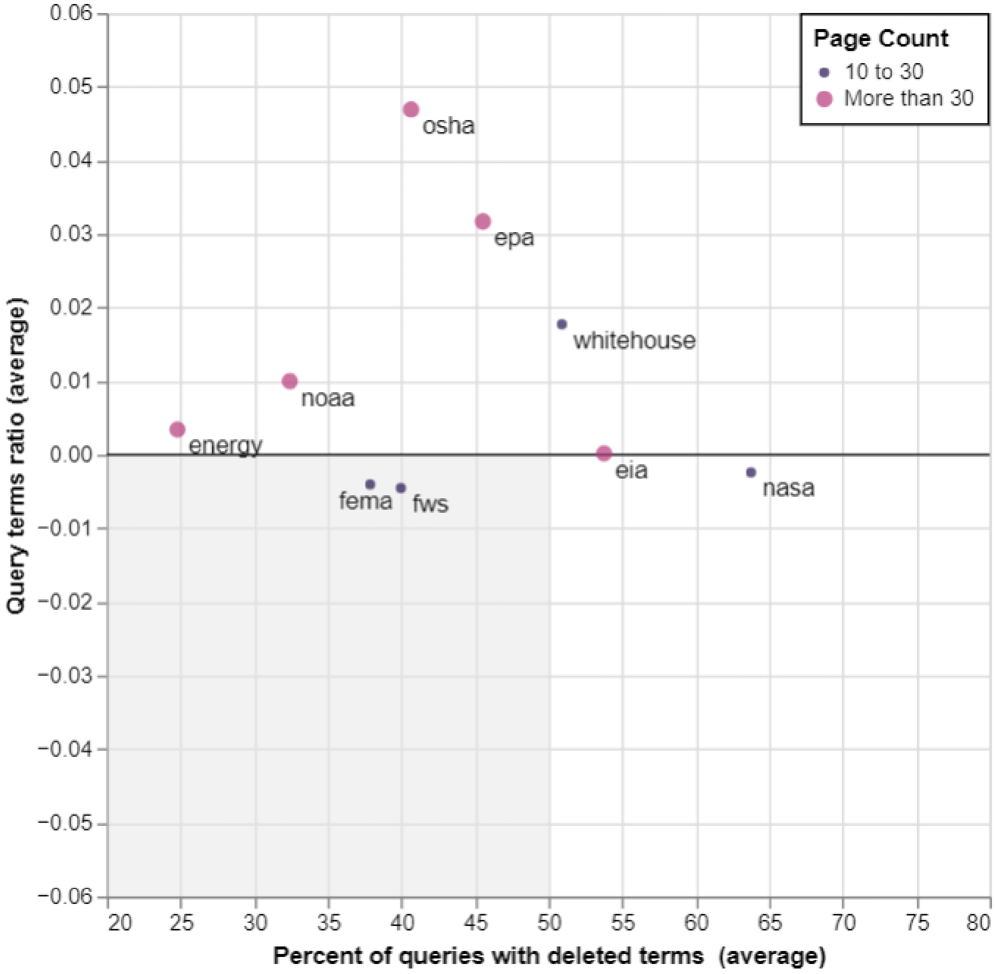


Figure 43. Multiple agencies' query terms were removed over time (positive values indicate term removal).

with at least 10 pages. This figure shows that some agency websites in the shaded region, like fema.gov, underwent typical search engine optimization and had a marginally negative average relative QueryTermsRatio, while other agency websites outside of the shaded region, like epa.gov, had query terms deleted, as shown by their strong positive relative average QueryTermsRatio. Vasilisky et al. [138] found that the average QueryTermsRatio of their corpus ranged between -0.08 and -0.03 during a 12 month period, so QueryTermsRatios in that same magnitude but positive are significant.

6.2.11 Hypothesized Editing Motivations

Search engines use more complicated algorithms than simple term matching when serving and ranking results. In addition to checking the page for matching query terms, a search engine might check for synonyms, rank pages with recent edits higher, or prioritize pages with more incoming links [24]. Some editing motivations, such as copyediting [140], are not rank incentivized. Below we consider some additional motivations for the editing patterns discovered in this data set to hypothesize the intent behind the edits.

It is possible that the tendentious editing on US federal sites was meant to raise the rankings of the pages for certain queries while lowering them for others. For example, EDGI found evidence that instances of “climate change” were replaced with “extreme weather events.” There is exactly one query in ORCAS for “extreme weather events” that resulted in clicks to a .gov page, which was for the EPA’s Climate Change Indicators page. It is just one of 47 queries in ORCAS for this page, 25 of which include “climate change”. In fact, the query is actually “extreme weather events climate change”. So, there is no evidence that editing the pages for synonyms reduced the clicks for the old phrase or increased the clicks for the new phrase in this instance. Edits of this type also do not improve clarity.

Another possibility is that pages were edited to give them a more recent edit time. However, it is not necessary to remove environmental query terms to accomplish this purpose. Vasilisky et al. [138] showed that query terms are typically added to a page, not removed, when editing a page with a ranking incentive.

Finally, pages might have been edited in an attempt to exploit the relationship between the link graph and the rankings. However, Nost et al. [110] found that large swaths of federal environmental websites were deleted entirely, which lowered the number of inbound links to the remaining pages due to the highly interconnected nature of the federal website link graph. Since none of these possibilities have proven probable, a more likely motivation for editing was to reduce public access to environmental information.

6.3 SUMMARY

In this chapter, we addressed Research Question 3, How can aggregated webpage changes of a corpus be used computationally to provide compelling evidence for edit intentions? We examined the claim made by EDGI that environmental terms were deleted from US federal websites between 2016 and 2020, and used the change text search index to validate the terms from Nost et al.'s study [110] as well as surface additional environmental terms deleted during the time period. We also examined EDGI's claim that access to environmental pages was lowered between 2016 and 2020. We aligned the EDGI dataset with the ORCAS click-query dataset and used the change text search index to determine the relationship between the deleted terms and the query terms. We found that users were searching for deleted terms on these websites, which lowers the retrievability of the websites. This finding strengthens EDGI's claim that access to these websites was lowered.

CHAPTER 7

CONCLUSIONS

In this final chapter, we lay the groundwork for future work, and wrap up the thesis with final thoughts.

7.1 FUTURE WORK

There are a variety of interesting directions for future work. Much work has been done on presenting deleted terms and phrases effectively, but presenting added terms and phrases has not been as thoroughly pursued. Additional work with the change text search interface to support finding of added terms and phrases will require additional backend and frontend work. On the backend, new algorithms will be needed to detecting added phrases when both terms are already present on the page separately, and identify partially added terms. On the frontend, brief analysis of current data suggests users looking for added terms only look for terms that have not yet been deleted. Querying for only pages where the most recent version still has the term and presenting these results may require a slightly different approach than is used with the deletions.

Future work with evaluations will involve a few parts. First, only one presidential term has been considered. Creating and analyzing a dataset that covers multiple presidential terms will yield new insights into both added and deleted terms over time. As the prior US president was in office beginning in 2008, many strategies will be needed to assemble a dataset of tripled mementos that span from 2008 through 2020. Next, the evaluation presented in this thesis covers the effectiveness of the backend, so a user study to evaluate the effectiveness of the frontend interface would further solidify the contributions presented. A user study would evaluate whether users such as journalists can more successfully complete their information seeking tasks. Another aspect of future work will include not simply measuring, but semantically categorizing the amount of change to a webpage. This work would make it possible to distinguish between different types of changes, such as an entire page rewrite aligned with an organization's new goals versus removals of blocks of content that remove public access to vital information. We could also expand discovery beyond keywords to phrases or entire page semantic meanings. We also need to improve the system so that false positives, such as pages with memento damage from dynamically loaded content, are not presented to the user as bona fide deletions.

Finally, we showed that non-200 HTTP status codes, such as 3xx redirects, cause information seeking problems for users who are trying to view change over time using web archives. Future work will investigate alternatives to using a URI as a key, so that when a page's address changes, lookup is not negatively affected.

7.2 CONTRIBUTIONS

Existing temporal search engines for web archives do not allow for users to query for change over time. In order to address Research Question 1, "How can we make changes in webpages discoverable and understandable?", this thesis presents a change text search engine, which allows users to find and view the changes in webpages. The search engine results page groups multiple versions of a page together without hiding the changes between these versions. In fact, the changes between the versions are the core reason why the grouping can occur, since the grouping represents the lifespan of a term on the page. A deletion animation shows changes in context, and a sliding difference viewer enables quick examination of the differences between many versions a page.

We also created a prototype for a new navigation banner to address Research Question 2, “How can we increase efficiency in web archive user navigation for viewing change over time?” We showed that the current banner in use by the Internet Archive Wayback Machine is inefficient for viewing changes, and proposed new functionality in the navigation buttons to jump between changed versions of a webpage instead of jumping based on set temporal units.

In order to address Research Question 3, “How can aggregated webpage changes of a corpus be used computationally to provide compelling evidence for edit intentions?”, we examined the change text index for the EDGI dataset. The inverted index contains valuable information about the most frequently deleted terms in the corpus. We also investigated EDGI’s claim that changes on US federal environmental sites resulted in less public access to information, and provide new evidence for the claim based on real user query data. We combined the EDGI 2016-2020 US federal environmental sites dataset and the ORCAS 2017-2020 dataset, then analyzed the queries for pages in both datasets. By indexing the change text, we found that over two-fifths of pages had deleted terms in at least half of their queries. We also found that the average relative query terms ratio from 2016 to 2020 was 0.02, indicating that fewer instances of query terms appeared in 2020 than in 2016. The positive relative terms query ratio shows retrogressive document manipulation of user queries, and further supports EDGI’s claim of tendentious editing of government websites between 2016 and 2020.

REFERENCES

[1] S. Abrams, A. Antracoli, R. Appel, C. Caust-Ellenbogen, S. Denison, S. Duncan, and S. Ramsay, “Sowing the seeds for more usable web archives: a usability study of Archive-It,” *The American Archivist*, vol. 82, no. 2, pp. 440–469, 2019. DOI: `10.17723/aarc-82-02-19`.

[2] S. Abrams, Z. Collier, E. Colón-Marrero, keondra bills freemyn, N. Krabbenhoeft, M. E. Wertheimer, and A. Wickner, *2022 web archiving survey report*, https://www.diglib.org/results-of-the-2022-ndsa-web-archiving-survey-report-now-available/, 2022.

[3] E. Adar, “User 4xxxxx9: Anonymizing query logs,” in *Proceedings of the Query Log Analysis Workshop, International Conference on World Wide Web*, 2007.

[4] E. Adar, “Temporal-Informatics of the WWW,” Ph.D. dissertation, University of Washington, 2009.

[5] E. Adar, M. Dontcheva, J. Fogarty, and D. S. Weld, “Zoetrope: Interacting with the Ephemeral Web,” in *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology*, 2008, pp. 239–248. DOI: `10.1145/1449715.1449756`.

[6] E. Adar, J. Teevan, S. T. Dumais, and J. L. Elsas, “The web changes everything: Understanding the dynamics of web content,” in *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, 2009, pp. 282–291. DOI: `10.1145/1498759.1498837`.

[7] S. Ainsworth, *Web Archiving in Popular Media*, Sep. 2016. [Online]. Available: `https://ws-dl.blogspot.com/2016/09/web-archiving-in-popular-media.html`.

[8] S. Alam, M. Kelly, M. C. Weigle, and M. L. Nelson, *A survey of archival replay banners*, Presented at the ACM/IEEE JCDL 2018 Workshop on Web Archiving and Digital Libraries (WADL), 2018.

[9] S. Alam and M. L. Nelson, “MemGator - A Portable Concurrent Memento Aggregator: Cross-Platform CLI and Server Binaries in Go,” in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, 2016, pp. 243–244. DOI: 10.1145/2910896.2925452.

[10] L. Alkwai, M. L. Nelson, and M. C. Weigle, “Comparing the archival rate of arabic, english, danish, and korean language web pages,” *ACM Transactions on Information Systems (TOIS)*, vol. 36, no. 1, 1:1–1:34, Jul. 2017. DOI: 10.1145/3041656.

[11] L. Alkwai, M. L. Nelson, and M. C. Weigle, “Making recommendations from web archives for "lost" web pages,” in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Aug. 2020, pp. 87–96. DOI: 10.1145/3383583.3398533.

[12] Y. AlNoamany, M. C. Weigle, and M. L. Nelson, “Detecting off-topic pages in web archives,” in *Proceedings of the International Conference on Theory and Practice of Digital Libraries*, 2015, pp. 225–237. DOI: 10.1007/978-3-319-24592-8_17.

[13] Y. A. AlNoamany, M. C. Weigle, and M. L. Nelson, “Access patterns for robots and humans in web archives,” in *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2013, pp. 339–348. DOI: 10.1145/2467696.2467722.

[14] A. AlSum and M. L. Nelson, “Thumbnail summarization techniques for web archives,” in *Proceedings of the European Conference on Information Retrieval*, 2014, pp. 299–310. DOI: 10.1007/978-3-319-06028-6_25.

[15] M. Aturban, M. Klein, H. Van de Sompel, S. Alam, M. L. Nelson, and M. C. Weigle, "Hashes are not suitable to verify fixity of the public archived web," *PLOS ONE*, vol. 18, no. 6, pp. 1–49, Jun. 2023. DOI: 10.1371/journal.pone.0286879.

[16] L. Azzopardi, R. English, C. Wilkie, and D. Maxwell, "Page Retrievability Calculator," in *Proceedings of the 36th European Conference on IR Research on Advances in Information Retrieval - Volume 8416*, 2014, pp. 737–741. DOI: 10.1007/978-3-319-06028-6_85.

[17] R. Baeza-Yates, C. Hurtado, M. Mendoza, and G. Dupret, "Modeling user search behavior," in *Proceedings of the Third Latin American Web Congress (LA-WEB'2005)*, 2005. DOI: 10.1109/LAWEB.2005.23.

[18] J. Bailey, A. Grotke, K. Hanna, C. Hartman, E. McCain, C. Moffatt, and N. Taylor, *Web archiving in the United States: A 2013 survey*, https://blogs.loc.gov/thesignal/2014/10/results-from-the-2013-ndsa-u-s-web-archiving-survey/, 2013.

[19] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, *et al.*, *MS MARCO: A Human Generated MAchine Reading COmprehension Dataset*, 2018. arXiv: 1611.09268.

[20] Z. Bar-Yossef, A. Z. Broder, R. Kumar, and A. Tomkins, "Sic transit gloria telae: Towards an understanding of the web's decay," in *Proceedings of the 13th International Conference on World Wide Web (WWW)*, 2004, pp. 328–337. DOI: 10.1145/988672.988716.

[21] K. Berberich, S. Bedathur, T. Neumann, and G. Weikum, "A Time Machine for Text Search," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007, pp. 519–526. DOI: 10.1145/1277741.1277831.

[22] T. Berners-Lee, R. T. Fielding, and L. M. Masinter, *Uniform Resource Identifier (URI): Generic Syntax*, 2005. DOI: `10.17487/RFC3986`.

[23] T. Berners-Lee, L. M. Masinter, and M. P. McCahill, *Uniform Resource Locators (URL)*, 1994. DOI: `10.17487/RFC1738`.

[24] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[25] A. Broder, "A taxonomy of web search," vol. 36, no. 2, pp. 3–10, 2002. DOI: `10.1145/792550.792552`.

[26] J. F. Brunelle, M. Kelly, H. SalahEldeen, M. C. Weigle, and M. L. Nelson, "Not all mementos are created equal: Measuring the impact of missing resources," *International Journal on Digital Libraries*, vol. 16, pp. 283–301, 2015. DOI: `10.1007/s00799-015-0150-6`.

[27] P. L. Buttigieg, N. Morrison, B. Smith, C. J. Mungall, S. E. Lewis, and E. Consortium, "The environment ontology: Contextualising biological and biomedical entities," *Journal of Biomedical Semantics*, vol. 4, pp. 1–9, 2013. DOI: `10.1186/2041-1480-4-43`.

[28] N. Caplan-Bricker, " Preservation Acts: Toward an ethical archive of the web.," *Harper's magazine*, 2018. [Online]. Available: `https://harpers.org/archive/2018/12/preservation-acts-archiving-twitter-social-media-movements/`.

[29] E. Chapman, "Crowdsourced Archiving of the January 6th US Capitol Insurrection: An r/DataHoarders Case Study," Ph.D. dissertation, Marquette University, 2021.

[30] J. Cherng and C. Boulton, *Jfcherng/php-diff*, version v6.13.0, Jan. 2023. [Online]. Available: `https://github.com/jfcherng/php-diff`.

[31] F. Chevalier, P. Dragicevic, A. Bezerianos, and J.-D. Fekete, "Using Text Animated Transitions to Support Navigation in Document Histories," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010, pp. 683–692, ISBN: 9781605589299. DOI: 10.1145/1753326.1753427.

[32] J. Cho and H. Garcia-Molina, "Effective page refresh policies for web crawlers," *ACM Transactions on Database Systems*, vol. 28, no. 4, pp. 390–426, 2003, ISSN: 0362-5915. DOI: 10.1145/958942.958945.

[33] Congress, US, *Digital millennium copyright act*, 1998.

[34] M. Costa and M. J. Silva, "Characterizing Search Behavior in Web Archives.," in *Proceedings of the Temporal Web Analytics Workshop (TWAW)*, 2011, pp. 33–40.

[35] N. Craswell, D. Campos, B. Mitra, E. Yilmaz, and B. Billerbeck, "ORCAS: 18 million clicked query-document pairs for analyzing search," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 2983–2989. DOI: 10.1145/3340531.3412779.

[36] D. Cruz and D. Gomes, "Adapting search user interfaces to web archives," in *Proceedings of the 10th International Conference on Preservation of Digital Objects*, vol. 17, 2013.

[37] Z. Dalal, S. Dash, P. Dave, L. Francisco-Revilla, R. Furuta, U. Karadkar, and F. Shipman, "Managing distributed collections: Evaluating web page changes, movement, and replacement," in *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, Tuscon, AZ, USA, 2004, pp. 160–168, ISBN: 1-58113-832-6. DOI: 10.1145/996350.996387.

[38] M. Davis and L. Iancu, "Unicode text segmentation," Unicode, Tech. Rep. 29, 2018.

[39] M. Day, "Preserving the fabric of our lives: A survey of web preservation initiatives," in *Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)*, 2003, pp. 461–472. [Online]. Available: `https://doi.org/10.1007/b11967`.

[40] L. Denoue, S. Carter, and M. Cooper, "SlideDiff: Animating Textual and Media Changes in Slides," in *Proceedings of the ACM Symposium on Document Engineering*, 2018, pp. 1–4. DOI: `10.1145/3209280.3229107`.

[41] S. Dori-Hacohen, E. Yom-Tov, and J. Allan, "Navigating Controversy as a Complex Search Task," in *Proceedings of the Supporting Complex Search Task Workshop at European Conference on Information Retrieval*, 2015.

[42] F. Douglis, T. Ball, Y.-F. Chen, and E. Koutsofios, "The AT&T Internet Difference Engine: Tracking and viewing changes on the web," *World Wide Web*, vol. 1, no. 1, pp. 27–44, 1998. DOI: `10.1023/A:1019243126596`.

[43] J. E. G. Coffman, Z. Liu, and R. R. Weber, "Optimal robot scheduling for web search engines," *Journal of Scheduling*, vol. 1, no. 1, pp. 15–29, 1998.

[44] D. Edwards, "Lost in hyperspace: Cognitive mapping and navigation in a hypertext environment," *Hypertext: Theory into practice*, 1989.

[45] T. Egense, T. Eskildsen, and A. K. Myrvoll, *Solrwayback*, 2022. [Online]. Available: `https://github.com/netarchivesuite/solrwayback`.

[46] M. Farrell, E. McCain, M. Praetzellis, G. Thomas, and P. Walker, *Web archiving in the United States: A 2017 survey*, https://ndsa.org/2018/12/12/announcing-publication-of-ndsa-s-2017-web-archiving-survey-report.html, 2018.

[47] D. Fetterly, M. Manasse, M. Najork, and J. Wiener, “A large-scale study of the evolution of web pages,” in *Proceedings of the 12th international conference on World Wide Web*, Budapest, Hungary, 2003, pp. 669–678, ISBN: 1-58113-680-3. DOI: 10.1145/775152.775246.

[48] R. Fielding, M. Nottingham, and J. Reschke, *RFC 9110: HTTP semantics*, 2022. DOI: 10.17487/RFC9110.

[49] C. Fiesler, “Ethical considerations for research involving (speculative) public data,” in *Proceedings of the ACM International Conference on Supporting Group Work*, 2019, pp. 1–13. DOI: 10.1145/3370271.

[50] C. Fiesler and N. Proferes, ““Participant” Perceptions of Twitter Research Ethics,” *Social Media + Society*, vol. 4, no. 1, 2018. DOI: 10.1177/2056305118763366.

[51] L. Frew, *Web Archiving in Popular Media II: User Tasks of Journalists*, Aug. 2022. [Online]. Available: https://ws-dl.blogspot.com/2022/08/2022-08-04-web-archiving-in-popular.html.

[52] L. Frew, M. L. Nelson, and M. C. Weigle, “Making Changes in Webpages Discoverable: A Change-Text Search Interface for Web Archives,” in *Proceedings of the 23rd ACM/IEEE-CS Joint Conference on Digital Libraries*, ©2023 IEEE. Reprinted, with permission., 2023, pp. 71–81. DOI: 10.1109/JCDL57899.2023.00021.

[53] L. Frew, M. L. Nelson, and M. C. Weigle, “Retrogressive Document Manipulation of US Federal Environmental Websites,” in *Proceedings of the 33rd ACM International Conference on Information & Knowledge Management*, 2024.

[54] M. Fröbe, C. Akiki, M. Potthast, and M. Hagen, "Noise-Reduction for Automatically Transferred Relevance Judgments," in *Procedings of the 13th International Conference of the CLEF Association (CLEF 2022)*, 2022, pp. 48–61. DOI: 10.1007/978-3-031-13643-6_4.

[55] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts, "Visual comparison for information visualization," *Information Visualization*, vol. 10, no. 4, pp. 289–309, 2011. DOI: 10.1177/1473871611416549.

[56] D. Gomes, E. Demidova, J. Winters, and T. Risse, Eds., *The Past Web: Exploring Web Archives*. Springer, 2021. DOI: 10.1007/978-3-030-63291-5.

[57] D. Gomes, J. Miranda, and M. Costa, "A survey on web archiving initiatives," in *Proceedings of Theory and Practice of Digital Libraries (TPDL)*, 2011, pp. 408–420, ISBN: 978-3-642-24468-1.

[58] M. Graham, *More than 9 million broken links on Wikipedia are now rescued*, https://blog.archive.org/2018/10/01/more-than-9-million-broken-links-on-wikipedia-are-now-rescued/, 2018.

[59] R. C. Gur and E. R. Hilgard, "Visual imagery and the discrimination of differences between altered pictures simultaneously and successively presented," *British Journal of Psychology*, vol. 66, no. 3, pp. 341–345, 1975.

[60] C. G. Healey, K. S. Booth, and J. T. Enns, "High-Speed Visual Estimation Using Preattentive Processing," *ACM Transactions on Computer-Human Interaction*, vol. 3, no. 2, pp. 107–135, Jun. 1996. DOI: 10.1145/230562.230563.

[61] A. Z. Henley and S. D. Fleming, “Yestercode: Improving code-change support in visual dataflow programming environments,” in *Proceedings of the 2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, 2016, pp. 106–114. DOI: 10.1109/VLHCC.2016.7739672.

[62] D. S. Hirschberg, “Algorithms for the longest common subsequence problem,” *Journal of the ACM (JACM)*, vol. 24, no. 4, pp. 664–675, 1977. DOI: 10.1145/322033.322044.

[63] J. W. Hunt and M. D. MacIlroy, *An algorithm for differential file comparison*. Bell Laboratories Murray Hill, 1976.

[64] L.-D. Ibáñez and E. Simperl, “A Comparison of Dataset Search Behaviour of Internal versus Search Engine Referred Sessions,” in *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, 2022, pp. 158–168. DOI: 10.1145/3498366.3505821.

[65] International Internet Preservation Consortium Access Working Group, “Use Cases for Access to Internet Archives,” International Internet Preservation Consortium, Tech. Rep., 2006. [Online]. Available: https://digital.library.unt.edu/ark:/67531/metadc1457746/.

[66] International Organization for Standardization Technical interoperability committee, “28500:2009 Information and documentation-WARC file format,” International Organization for Standardization, Tech. Rep., 2009.

[67] P. Jaccard, “The distribution of the flora in the alpine zone,” *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[68] A. Jackson, *The provenance of web archives*, https://anjackson.net/2015/11/20/provenance-of-web-archives/, 2015.

[69] A. Jackson, *Web Archive Discovery - WARC Indexer*, 2022. [Online]. Available: https://github.com/ukwa/webarchive-discovery/tree/master/warc-indexer.

[70] A. Jackson, J. Lin, I. Milligan, and N. Ruest, "Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities," in *Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2016, pp. 103–106. DOI: 10.1145/2910896.2910912.

[71] B. J. Jansen, "Search log analysis: What it is, what's been done, how to do it," *Library & information science research*, vol. 28, no. 3, pp. 407–432, 2006. DOI: 10.1016/j.lisr.2006.06.005.

[72] S. Jaschik, *QS Ranks Russian Universities, Contrary to Original Plan*, https://www.insidehighered.com/admissions/article/2022/06/13/qs-ranks-russian-universities-despite-vow-not, 2022.

[73] A. Jatowt, Y. Kawai, S. Nakamura, Y. Kidawara, and K. Tanaka, "Journey to the Past: Proposal of a Framework for Past Web Browser," in *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia*, 2006, pp. 135–144. DOI: 10.1145/1149941.1149969.

[74] H. Jayanetti, K. Garg, S. Alam, M. L. Nelson, and M. C. Weigle, "Robots Still Outnumber Humans in Web Archives, But Less Than Before," in *Proceedings of the Theory and Practice of Digital Libraries Conference*, Sep. 2022, pp. 245–259. DOI: 10.1007/978-3-031-16802-4_19.

[75] H. R. Jayanetti, K. Garg, S. Alam, M. L. Nelson, and M. C. Weigle, “Robots still outnumber humans in web archives in 2019, but less than in 2015 and 2012,” *International Journal on Digital Libraries*, 2024. DOI: 10.1007/s00799-024-00397-2.

[76] J. Jiang, D. He, and J. Allan, “Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time,” in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014, pp. 607–616. DOI: 10.1145/2600428.2609633.

[77] E. S. Jo and T. Gebru, “Lessons from archives: Strategies for collecting sociocultural data in machine learning,” in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 306–316. DOI: 10.1145/3351095.3372829.

[78] S. Jones, *Mementos in the Raw*, Apr. 2016. [Online]. Available: https://ws-dl.blogspot.com/2016/04/2016-04-27-mementos-in-raw.html.

[79] S. Jones, *Mementos in the Raw, Take Two*, Aug. 2016. [Online]. Available: https://ws-dl.blogspot.com/2016/08/2016-08-15-mementos-in-raw-take-two.html.

[80] S. Jones, M. Klein, M. C. Weigle, and M. L. Nelson, *MementoEmbed and Raintale for web archive storytelling*, Presented at the ACM/IEEE JCDL 2020 Workshop on Web Archiving and Digital Libraries (WADL), Aug. 2020.

[81] S. M. Jones, H. Van de Sompel, H. Shankar, M. Klein, R. Tobin, and C. Grover, “Scholarly context adrift: Three out of four URI references lead to changed content,” *PLOS ONE*, vol. 11, no. 12, 2016. DOI: 10.1371/journal.pone.0167475.

[82] S. M. Jones, M. Klein, H. Van de Sompel, M. L. Nelson, and M. C. Weigle, “Interoperability for accessing versions of web resources with the Memento protocol,” in *The Past*

Web: Exploring Web Archives, Springer International Publishing, 2021, ISBN: 978-3-030-63290-8.

[83] S. M. Jones, M. C. Weigle, M. Klein, and M. L. Nelson, "Hypercane: Intelligent Sampling for Web Archive Collections," in *Proceedings of the 21st ACM/IEEE-CS Joint Conference on Digital Libraries*, 2021, pp. 316–317. DOI: 10.1109/JCDL52503.2021.00049.

[84] S. M. Jones, M. Klein, and H. Van de Sompel, "Robustifying links to combat reference rot," *Code4Lib Journal*, vol. 50, 2021.

[85] B. Jules, E. Summers, and V. Mitchell, "Documenting the now-white paper: Ethical considerations for archiving social media content generated by contemporary social movements: Challenges, opportunities, and recommendations," *Documenting the Now*, 2018. [Online]. Available: https://www.docnow.io/docs/docnow-whitepaper-2018.pdf.

[86] N. Kanhabua, P. Kemkes, W. Nejdl, T. N. Nguyen, F. Reis, and N. K. Tran, "How to search the internet archive without indexing it," in *Proceedings of the Theory and Practice of Digital Libraries Conference*, 2016, pp. 147–160. DOI: 10.1007/978-3-319-43997-6_12.

[87] J. Kiesel, A. P. de Vries, M. Hagen, B. Stein, and M. Potthast, "WASP: web archiving and search personalized," in *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems*, 2018, pp. 16–21.

[88] M. Klein and M. L. Nelson, "Revisiting lexical signatures to (re-) discover web pages," in *Proceedings of the International Conference on Theory and Practice of Digital Libraries*, 2008, pp. 371–382. DOI: 10.1007/978-3-540-87599-4_38.

[89] M. Klein and M. L. Nelson, "Moved but not gone: An evaluation of real-time methods for discovering replacement web pages," *International Journal on Digital Libraries*, vol. 14, no. 1, pp. 17–38, 2014. DOI: 10.1007/s00799-014-0108-0.

[90] M. Klein and M. L. Nelson, "Investigating the Change of Web Pages' Titles Over Time," in *Proceedings of the First International Workshop on Innovation in Digital Preservation*, 2009.

[91] I. Kreymer, *Pywb - web archiving tools for all*, 2022. [Online]. Available: https://github.com/ikreymer/pywb.

[92] O. Kurland and M. Tennenholtz, "Competitive Search," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 2838–2849. DOI: 10.1145/3477495.3532771.

[93] Y.-L. Lai and K.-L. Hui, "Internet opt-in and opt-out: Investigating the roles of frames, defaults and privacy concerns," in *Proceedings of the 2006 ACM SIGMIS CPR conference on computer personnel research: Forty four years of computer personnel research: achievements, challenges & the future*, 2006, pp. 253–263. DOI: 10.1145/1125170.1125230.

[94] M. Lanna, "Spotting the Difference," Ph.D. dissertation, University of Ottawa, 2009.

[95] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics - Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[96] J. Lin, I. Milligan, D. W. Oard, N. Ruest, and K. Shilton, "We could, but should we? Ethical considerations for providing access to GeoCities and other historical digital collections," in *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, 2020, pp. 135–144. DOI: 10.1145/3343413.3377980.

[97] O. Lynskey, “Control over Personal Data in a Digital Age: Google Spain v AEPD and Mario Costeja Gonzalez,” *The modern law review*, vol. 78, no. 3, pp. 522–534, 2015.

[98] A. Mabe, D. Patel, M. Gunnam, S. Shankar, M. Kelly, S. Alam, M. L. Nelson, and M. C. Weigle, *TMVis: Visualizing webpage changes over time*, Presented at the ACM/IEEE JCDL 2020 Workshop on Web Archiving and Digital Libraries (WADL), Aug. 2020.

[99] S. MacAvaney, C. Macdonald, and I. Ounis, “Reproducing Personalised Session Search Over the AOL Query Log,” in *Proceedings of the 44th European Conference on IR Research (ECIR 2022), Part I*, 2022, pp. 627–640. DOI: `10.1007/978-3-030-99736-6_42`.

[100] K. Mackinnon, “Databound: Histories of Growing Up on the World Wide Web,” Ph.D. dissertation, University of Toronto (Canada), 2022.

[101] J. Maddock, K. Starbird, and R. M. Mason, “Using historical Twitter data for research: Ethical challenges of tweet deletions,” in *Proceedings of the Workshop on ethics for studying sociotechnical systems in a Big Data World at Conference on Computer-Supported Cooperative Work & Social Computing*, 2015.

[102] D. Major, *It Takes a Village to Raise an Archive*, Jan. 2023. [Online]. Available: `https://webarchivingrt.wordpress.com/2023/01/18/it-takes-a-village-to-raise-an-archive-how-the-use-of-web-sources-fosters-collaboration/`.

[103] G. Marchionini, *Information seeking in electronic environments*. Cambridge University Press, 1995, ISBN: 9780511626388.

[104] F. Melo, H. Viana, D. Gomes, and M. Costa, “Architecture of the Portuguese web archive search system version 2,” Arquivo.pt-The Portuguese Web Archive, Tech. Rep., 2016. [On-

line]. Available: https://sobre.arquivo.pt/wp-content/uploads/architecture-of-the-portuguese-web-archive-search-1.pdf.

[105] M. L. Nelson, *Memento-Datetime is not Last-Modified*, http://ws-dl.blogspot.com/2010/11/2010-11-05-memento-datetime-is-not-last.html, 2011.

[106] M. L. Nelson and H. Van de Sompel, "Adding the dimension of time to HTTP," in *SAGE Handbook of Web History*, SAGE Publishing, 2019, ISBN: 9781473980051.

[107] T. N. Nguyen, N. Kanhabua, W. Nejdl, and C. Niederée, "Mining Relevant Time for Query Subtopics in Web Archives," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 1357–1362. DOI: 10.1145/2740908.2741702.

[108] C. Niederer, H. Stitz, R. Hourieh, F. Grassinger, W. Aigner, and M. Streit, "TACO: visualizing changes in tables over time," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 677–686, 2017. DOI: 10.1109/TVCG.2017.2745298.

[109] J. Nielsen, *10 usability heuristics for user interface design*, 1994. [Online]. Available: https://www.nngroup.com/articles/ten-usability-heuristics/.

[110] E. Nost, G. Gehrke, G. Poudrier, A. Lemelin, M. Beck, S. Wylie, and on behalf of the Environmental Data and Governance Initiative, "Visualizing changes to US federal environmental agency websites, 2016–2020," *PLOS ONE*, vol. 16, no. 2, pp. 1–27, Feb. 2021. DOI: 10.1371/journal.pone.0246450.

[111] J. Ogden, E. Summers, and S. Walker, "Patterns of Use: Conceptualising the role of web archives in online discourse," in *4th RESAW (Research Infrastructure for the Study of Archived Web Materials) conference*, Jun. 2021.

[112] A. Overwijk, C. Xiong, and J. Callan, “ClueWeb22: 10 Billion Web Documents with Rich Information,” in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 3360–3362. DOI: 10.1145/3477495.3536321.

[113] E. Pariser, *The filter bubble: How the new personalized web is changing what we read and how we think.* Penguin, 2011.

[114] I. Perez-Messina, C. Gutierrez, and E. Graells-Garrido, “Organic Visualization of Document Evolution,” in *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, 2018, pp. 497–501. DOI: 10.1145/3172944.3173004.

[115] M. Phillips, D. Chudnov, and J. Jacobs, *Exploratory Analysis of the End of Term Web Archive: Comparing two collections.* Presented at the ACM/IEEE JCDL 2016 Workshop on Web Archiving and Digital Libraries (WADL), 2016.

[116] M. E. Phillips and K. K. Phillips, “End of Term 2016 Presidential Web Archive,” *Against the Grain*, vol. 29, no. 6, p. 10, 2017. DOI: 10.7771/2380-176X.7874.

[117] W. Phillips, “On the distinction between sensory storage and short-term visual memory,” *Perception & Psychophysics*, vol. 16, pp. 283–290, 1974.

[118] P. Pirolli and S. Card, “Information foraging,” *Psychological review*, vol. 106, no. 4, p. 643, 1999.

[119] M. Ras and S. van Bussel, “Web archiving user survey,” National Library of the Netherlands (Koninklijke Bibliotheek), Tech. Rep., 2007.

[120] J. Reimer, S. Schmidt, M. Fröbe, L. Gienapp, H. Scells, B. Stein, M. Hagen, and M. Potthast, "The Archive Query Log: Mining Millions of Search Result Pages of Hundreds of Search Engines from 25 Years of Web Archives," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)*, Jul. 2023, pp. 2848–2860. DOI: `10.1145/3539618.3591890`.

[121] R. A. Rensink, "Change detection," *Annual review of psychology*, vol. 53, no. 1, pp. 245–277, 2002. DOI: `10.1146/annurev.psych.53.100901.135125`.

[122] B. Reyes Ayala, "When expectations meet reality: Common misconceptions about web archives and challenges for scholars," *International Journal of Digital Humanities*, vol. 2, no. 1, pp. 89–106, 2021. DOI: `10.1007/s42803-021-00034-3`.

[123] T. Seneca, A. Grotke, C. N. Hartman, and K. Carpenter, "It takes a village to save the web: The End of Term Web Archive," *Documents to the People (DttP)*, vol. 40, p. 16, 2012.

[124] T. Sherratt and A. Jackson, *Glam-workbench/web-archives*, version v1.1.0, Apr. 2022. DOI: `10.5281/zenodo.6450762`.

[125] M. Shibata, *Uniseg: A python package to determine unicode text segmentations*, 2024. [Online]. Available: `https://pypi.org/project/uniseg/`.

[126] K. Sigurðsson, M. Stack, and I. Ranitovic, *Heritrix user manual: sort-friendly URI reordering transform*, 2006. [Online]. Available: `http://crawler.archive.org/articles/user_manual/glossary.html#surt`.

[127] D. Smiley, E. Pugh, K. Parisa, and M. Mitchell, *Apache Solr enterprise search server*. Packt Publishing Ltd, 2015.

[128] I. Soboroff, “Dynamic Test Collections: Measuring Search Effectiveness on the Live Web,” in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 276–283. DOI: 10.1145/1148170.1148220.

[129] K. Sparck Jones, “A statistical interpretation of term specificity and its application in retrieval,” *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972. DOI: 10.1108/eb026526.

[130] E. Summers, *Memento Bisect*, https://inkdroid.org/2023/09/14/memento-bisect/, 2023.

[131] J. Tam and S. Greenberg, “A framework for asynchronous change awareness in collaborative documents and workspaces,” *International Journal of Human-Computer Studies*, vol. 64, no. 7, pp. 583–598, 2006. DOI: 10.1007/978-3-540-30112-7_7.

[132] J. Teevan, ““Where’d it go?”: How people ask after lost Web information,” *Proceedings of the American Society for Information Science and Technology*, vol. 44, no. 1, pp. 1–19, 2007. DOI: 10.1002/meet.1450440269.

[133] J. Teevan, E. Adar, R. Jones, and M. A. Potts, “Information re-retrieval: Repeat queries in Yahoo’s logs,” in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 151–158. DOI: 10.1145/1277741.1277770.

[134] J. Teevan, S. T. Dumais, D. J. Liebling, and R. L. Hughes, “Changing how people view changes on the web,” in *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, 2009, pp. 237–246.

[135] H. Van de Sompel, M. Klein, and H. Shankar, “Towards robust hyperlinks for web-based scholarly communication,” in *Proceedings of the International Conference on Intelligent Computer Mathematics*, 2014, pp. 12–25. DOI: 10.1007/978-3-319-08434-3_2.

[136] H. Van de Sompel, M. Nelson, and R. Sanderson, *RFC 7089 - HTTP framework for time-based access to resource states–Memento*, 2013. [Online]. Available: https://tools.ietf.org/html/rfc7089.

[137] Z. Vasilisky, O. Kurland, M. Tennenholtz, and F. Raiber, “Content-Based Relevance Estimation in Retrieval Settings with Ranking-Incentivized Document Manipulations,” in *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, 2023, pp. 205–214. DOI: 10.1145/3578337.3605124.

[138] Z. Vasilisky, M. Tennenholtz, and O. Kurland, “Studying Ranking-Incentivized Web Dynamics,” in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 2093–2096. DOI: 10.1145/3397271.3401300.

[139] D. Wang, J. S. Olson, J. Zhang, T. Nguyen, and G. M. Olson, “DocuViz: Visualizing Collaborative Writing,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 1865–1874. DOI: 10.1145/2702123.2702517.

[140] D. Yang, A. Halfaker, R. Kraut, and E. Hovy, “Identifying semantic edit intentions from revisions in Wikipedia,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2000–2010. DOI: 10.18653/v1/D17-1213.

[141] Y. Zhang, W. Chen, D. Wang, and Q. Yang, “User-click modeling for understanding and predicting search-behavior,” in *Proceedings of the 17th ACM SIGKDD international con-*

ference on Knowledge discovery and data mining, 2011, pp. 1388–1396. DOI: 10.1145/2020408.2020613.

[142] J. Zhu, A. Nyayachavadi, J. Zhu, V. Ruamviboonsuk, and H. V. Madhyastha, “Reviving dead links on the web with fable,” in *Proceedings of the 2023 ACM on Internet Measurement Conference*, 2023, pp. 131–144. DOI: 10.1145/3618257.3624832.

APPENDIX A

CHANGE-TEXT INDEX CALCULATION IMPLEMENTATION

Developers can write Java code to directly interace with the Lucene index. In order to implement the change text calculations, we developed the TemporalDocument Java class that allows each document in the index with the same normalized URL to be sorted by crawl time, shown in Listing A.1.

Listing A.1. TemporalDocument.java: ordering documents by date

```
public class TemporalDocument implements
Comparable<TemporalDocument> {
    private long wayback_date;
    public int compareTo(TemporalDocument that) {
        if (this.wayback_date > that.wayback_date) return 1;
        else if (this.wayback_date < that.wayback_date) return -1;
        else return this.docnum - that.docnum;
    }
```

Once the documents are sorted, then validity ranges are calculated. This is shown in Listing A.2.

Listing A.2. TemporalDocument.java: computing validity ranges

```
public String getDateRange() {
    String nextDateStr;
    if (next == null) {
        Calendar c = Calendar.getInstance();

        c.setTime(this.getWaybackDate());

        c.add(Calendar.DATE, 30);

        nextDateStr = getSolrDateString( c.getTime() );

    }

    else {

        //inclusive

        nextDateStr = next.getWaybackDateStr();

    }
```

Each added and deleted term in the document is calculated after sorting the documents, as shown in Listing A.3.

Listing A.3. TemporalDocument.java: computing deleted and added terms

```
public String getDeletedTerms() {

    HashSet<String> value = next.terms;

    HashSet<String> valueP = this.terms;

    //deletions

    HashSet<String> deletedValues = new HashSet<String>(valueP);

    deletedValues.removeAll(value);

    //additions

    HashSet<String> atermset = new HashSet<String>(value);

    atermset.removeAll(valueP);
```

In addition to calculating full deletions, semi-deletions are also calculated using a different approach. This is shown in Listing A.4.

Listing A.4. TemporalDocument.java: computing semi-deleted terms

```
public String getSemiDeletedTerms() {
    HashMap<String, Integer> map = next.termCounts;
    HashMap<String, Integer> mapP = this.termCounts;

    HashSet<String> deletedValues = new HashSet<String>();
    for (Map.Entry<String, Integer> entry : mapP.entrySet()) {
        String key = entry.getKey();
        int value = entry.getValue();
        int valueN = (map.containsKey(key)) ? map.get(key) : 0;
        if (value > valueN && valueN > 0) {
            deletedValues.add(key);
        }
    }
```

We also implemented a calculation to identify soft 404s, shown in Listing A.5.

Listing A.5. TemporalDocument.java: Calculating soft-404s

```
s404log = Math.log10(this.content.length() * 1.0 /
  next.content.length());
```

The second part of the Java code we developed is shown in Listing A.6 and uses the TemporalDocument class when reading from the Lucene index. Using a TreeSet sorts the documents as they are read.

Listing A.6. ValidityRangeIndexWriter.java: Iteration over all documents in the index

```
DirectoryReader reader = DirectoryReader.open(FSDirectory.open
  (Paths.get("index")));
int nd = reader.maxDoc();
HashMap<String, TreeSet<TemporalDocument>> map = new
  HashMap<String, TreeSet<TemporalDocument>>();
for (int i = 0; i < nd; i++) {
```

APPENDIX B

CHANGE-TEXT SEARCH INTERFACE IMPLEMENTATION

We used PHP and Solarium to create our change-text search interface, including the query page and the search engine results page. Below, we show how we implemented the query page. The query page and the results page are both implemented in the file deletion_results.php as shown in Listing B.1.

Listing B.1. deletion_results.php: Querying for deleted terms

```
$deleted_term = $_GET[’dterm’];

$query_show_text = ’deleted_term:’.$deleted_term;

$client = new Solarium\Client($adapter, $eventDispatcher, $config);

$query = $client->createSelect();

$query->setQuery($query_show_text);
```

In Listing B.2, We show how we query with Solarium to find the pre-deletion and post-deletion versions of the page, so we can group them together in one search results.

Listing B.2. deletion_results.php: Finding the post-deletion version

```
$query2 = $client->createSelect();

$query2->setQuery(’url_norm:’.$document->url_norm);

$query2->createFilterQuery(’crawltime’)->setQuery(

  ’validity_range:[’.$document->get_next_wayback_date().’ TO ’

  .$document->get_next_wayback_date().’]’);

$query2->setDocumentClass(’TemporalDoc’);
```

```
$query2->addSort('id', $query::SORT_DESC);

$resultset2 = $client->select($query2);

$document2 = $resultset2->getIterator()->current();
```

Because the diff is not a Lucene field, we had to highlight the query terms in a different way from the built in Solarium highlighter. This is shown below in Listing B.3.

Listing B.3. deletion_results.php: Highlighting the search terms in the diff

```
$diff_out = $document->diff($document2, $deleted_term);

$diff_out = preg_replace('/\b('.$deleted_term.')\b/i',

'<span style="background-color: #FFFF00">$1</span>' , $diff_out);
```

Solarium allows for custom ranking, and we used the soft 404 field we calculated previously to rank soft 404s lower than other results, as shown in Listing B.4.

Listing B.4. deletion_results.php: Downranking soft 404s

```
$s404 = 'if(gt(field(text_log_d),1.0),-1,0)';

$query->addSort($s404, $query::SORT_DESC);

$query->addSort('score', $query::SORT_DESC);
```

We also added links in the search engine results to our custom sliding difference tool, as shown in Listing B.5.

Listing B.5. deletion_results.php: Link with input to diff slider

```
<a href="diff-slider.php?page='.urlencode($document->url_norm).

'&wbdate1='.$slid_diff_page->wayback_date.'&wbdate2='.

$document2->wayback_date.'&dterm='.$deleted_term.'">
```

```
Sliding diff</a>;
```

Another page, semi_del_results.php contains the algorithms used to query for semi-deleted terms. This code is shown in Listing B.6.

Listing B.6. semi_del_results.php: Filtering query results for phrases

```
if (!$is_deleted_phrase || ($is_deleted_phrase &&
$document->compare_phrase_freq($document2, $deleted_term) > 0)) {
```

We created a custom class, temporal_document.php, which allowed us to identify and highlight deleted phrases. This is shown in Listing B.7.

Listing B.7. temporal_document.php: Counting terms

```


public function content_joined_with_spaces() {
    return $this->text_joined_with_spaces($this->content);
}


//implementation of UAX29
public function text_joined_with_spaces($phrase) {
    //...
}


public function count_phrase_freq($phrase) {
    $text_joined_with_spaces = $this->content_joined_with_spaces();
    return substr_count($text_joined_with_spaces, $phrase);
}


public function compare_phrase_freq($doc2, $phrase) {
    return $this->count_phrase_freq($phrase) -
    $doc2->count_phrase_freq($phrase);
}
```

Finally, the url_norm.php script shown in Listing B.8 contains the URL canonicalization algorithm implementation in PHP.

Listing B.8. url_norm.php: URL canonicalization algorithm in PHP

```
<?php
function unparse_url($parsed_url) {
  //...
  $scheme = (strtolower($scheme) == ’https://’) ? ’http://’ : $scheme;
  //...
  $query = isset($parsed_url[’query’]) ?
     ’?’ . $parsed_url[’query’] : ’’;
  $url = strtolower("$scheme$user$pass$host$port$path$query$fragment");
  $url = preg_replace("~([a-z]+://)(?:www[0-9]*|ww2|ww)[.](.+)~",
     "$1$2", $url);
  $url = rtrim($url, ’/’);
  if (preg_match("~https?://[^/]+$~", $url)) {
    $url .= ’/’;
  }
  return $url;
}
function url_norm($url) {
  return unparse_url(parse_url($url));
}
?>
```

APPENDIX C

SLIDING DIFF TOOL IMPLEMENTATION

We implemented our sliding difference tool with a combination of PHP and JavaScript. We wanted to continue to use the PHP difference library, so we implemented the interface in PHP in a file called diff-slider.php. First, we wanted to collect the text content of all of the page versions between the addition of the query term and its eventual deletion. This is shown in Listing C.1.

Listing C.1. diff-slider.php: Query for all versions between addition and deletion

```
$query->setQuery(’url_norm:’.$url_norm);

$query->createFilterQuery(’delrange’)->setQuery(

’validity_range:[’.$wbdate1formatted.’ TO ’.$wbdate2formatted.’]’);

$query->addSort(’id’, $query::SORT_ASC);
```

We needed to use custom highlighting for the query term and make the output available to the JavaScript portion of our implementation, as shown in Listing C.2.

Listing C.2. diff-slider.php: Adding highlighted plaintext input for javascript

```
foreach($diff_out_arr as $key => $value){

    $slider_wb_js .= "diff_push(".json_encode($value).");\n";

}

$slider_wb_js .= ’updateValue();</script>’;
```

To create the sliding effect, we needed JavaScript to dynamically update the page contents. This JavaScript code is saved in a file slider.js. We collect the text of the document versions generated by the diff-slider.php script, shown in Listing C.3.

Listing C.3. slider.js: Accessing highlighted plaintext

```
function updateValue() {
  var rangeInput = document.getElementById("rangeInput").value;
  var wb_date1 = document.getElementById("wb_date1");
  wb_date1.innerHTML = wb_arr[rangeInput-1];
```

We implemented coalescing for identical versions in the JavaScript portion of the code as shown in Listing C.4.

Listing C.4. slider.js: Coalescing in Javascript

```
function navigateDiff(nav_type) {
    //...
    else if (nav_type == NAV_COAL_FORW) {
        var idx = rangeInputDom.value;
        idx = idx + 1;
        while(idx < wb_arr.length - 1 &&
        diff_arr[idx - 1].diff == ’Page versions are identical’) {
            idx = idx + 1;
        }
        rangeInputDom.value = idx;
    }
```

APPENDIX D

ANIMATED DELETION TOOL IMPLEMENTATION

The animated deletion tool uses Python and JavaScript, all contained in the web_diff.py script. The HTML difference library is written in Python, so our tool needed to be written in Python. We used JavaScript to create the animation effects. The animation effect JavaScript code is shown below in Listing D.1.

Listing D.1. web_diff.py: Animating changes one at a time

```
function printLetterByLetter(index, speed){
var anchor = ’<a class="wm-diff-anchor" id="wm-diff-del’ +
index +’"> </a>’;
window.location = window.location.origin + window.location.pathname +
window.location.search + ’#wm-diff-del’ + index
var interval = setInterval(function(){
    //...
    document.getElementById(destination).innerHTML = anchor +
    destText.substring(0, destText.length - j);
    //...
    sleepFor(400);
```

We iterated through all of the changes calculated by the HTML difference library and only kept the deletions that matched the query term. This is shown in Listing D.2.

Listing D.2. web_diff.py: Removing deletions that do not match the query term or phrase

```
if ’ ’ not in dterm:
    //...
else:
    #phrase search
    //UAX29 implementation
    if " " + dterm + " " not in text_joined_with_spaces:
        deletion.decompose()
    else:
        deletion[’id’] = ’wm-diff-del-wrapper’ + str(i)
        deletion.a[’id’] = ’wm-diff-del’ + str(i)
        deletion.a.string = ’ ’
        i = i + 1
```

We also needed to remove additions that did not match the query term, but this needed to be done in a different way than the deletions, as shown below in Listing D.3.

Listing D.3. web_diff.py: Removing additions that do not match the query term or phrase

```
comparison[’combined’] = re.sub(’<del class="wm-diff"
id="wm-diff-del-wrapper([0-9]+)(.+?)</del><ins(.+?)</ins>’,
’<del class="wm-diff" id="wm-diff-del-wrapper\g<1>\g<2></del>
<INS id="wm-diff-ins-wrapper\g<1>"\g<3></INS>’, comparison[’combined’])

comparison[’combined’] = comparison[’combined’].replace(
’<ins class="wm-diff">’, ’’)
```

```
comparison[’combined’] = comparison[’combined’].replace(’</ins>’, ’’)

diffids = re.findall(r’wm-diff[\-a-z]+[0-9]+’, comparison[’combined’])

lastid = 0

for i in range(len(diffids)):

    id = int(re.sub(’[^0-9]’, ’’, diffids[i]))

    notid = diffids[i].replace(str(id), ’’)

    if id < lastid:

        comparison[’combined’] = comparison[’combined’].replace(

        diffids[i], notid + str(lastid))

    elif id > lastid:

        lastid = id
```

APPENDIX E

CREATIVE COMMONS LICENSE

VITA

Lesley Frew

Department of Computer Science

Old Dominion University

Norfolk, VA 23529

EDUCATION

2021-2024 (Expected), Old Dominion University, M.S. Computer Science

2006-2011, University of Virginia, M.T. Elementary Education

2006-2011, University of Virginia, B.A. Mathematics

PROFESSIONAL EXPERIENCE

2022-2024, Northern Virginia Community College, Dual Enrollment Instructor

2013-2024, Fairfax County Public Schools, Teacher

2011-2013, New Kent County Public Schools, Teacher

PUBLICATIONS

1. Lesley Frew, Michael L. Nelson, and Michele C. Weigle, “Retrogressive Document Manipulation of US Federal Environmental Websites,” In Proceedings of ACM International Conference on Information and Knowledge Management (CIKM). October 2024.
2. Lesley Frew, Michael L. Nelson, and Michele C. Weigle, “Making Changes in Webpages Discoverable: A Change-Text Search Interface for Web Archives,” In Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL). June 2023.