Time -Travel on the Internet

Pranita Pingale[1], Prithvi Prathapan[2], Sahil Mulay[3], Siddhesh Patil[4], Dhananjay Sharma[5]

[1]Assistant Professor, Dept. of Computer Engineering, Pillai College of Engineering, Maharashtra, India.
[2-6]UG Student, Dept. of Computer Engineering, Pillai College of Engineering, Maharashtra, India.

Abstract – *The World Wide Web is a unique, widely used information tool that is used worldwide. Content is disappearing at an alarming rate, not only endangering our digital cultural memory but also organizational accountability. An effective technical solution to such a problem is the introduction of the Internet Archiving System aka, The "Digital Preservation Handbook". Web archiving is a process of gathering parts of the World Wide Web to ensure that information is archived by Digital Forensic archives, future researchers, historians, and the public. We focus on building a robust and software-based tool that stores and evaluates certain web clips / videos, using Web Crawling and data analysis to help individuals monitor and analyze online data. In Layman's terms, our software visits a set of URLs called seeds and identifies all hyperlinks to that seed, copying and saving information as its "crawls" seeds. This information is stored as an overview of the WARC file format so when it is played again with the WARC viewer, it appears as captured. With this project we strive to meet the primary goal of providing a consistent, easy-to-use, manageable and secure system that allows users with limited technical knowledge to easily download web content for archiving purposes. Software will be available free of charge for the benefit of the international community that maintains web history.*

***Key Words*: Web crawler, Information retrieval, WARC, , Web Scraping, Data gathering.**

1.INTRODUCTION

Web archive is a process of collecting data recorded on the World Wide Web, storing it, ensuring that the archives are archived, and making the collected data available for future research. The purpose of a web archive is to compile web content and store it for longer to deliver valuable data for generations. The volume of content from the digital channel, including the web, is growing exponentially, and the need to suspend digital media storage was needed. The space data consulting committee (CCSDS) has developed an open reference archive (OAIS) reference system, which is ISO 14,721 for long-term digital record keeping, and many web archiving programs are up to standard. However, the OAIS reference model does not provide a method of operation but imposes a requirement to ensure OAIS-compliance. There is a weakness in content integrity against unauthorized practices, and there is no way to guarantee long-term content integrity, although the OAIS reference model states long-term preservation of digital records. Also, the user is able to track, monitor and analyze data.

2. Literature Survey

A. An overview of Web Archiving: Jinfang Niu shared his views on some of the methods used, such as how the traditional concepts of archive management and theory can be applied to the organization and the definition of archived web resources. This approach is a review of the methods used in various universities, as well as international government libraries and archives, selecting, acquiring, defining and accessing web resources for their archives.

B. Accessing web archives at scale through a cloudbased interface: This paper introduces Archives Unleashed Cloud, a web-based visual interface and web archives on average. The current access paradigms, largely driven by the breadth and scope of web archives, often involve the use of a command line and a coding code.

C. Investigating websites and webpages: This chapter aims to show the searcher how websites work, how they can help with research and how to properly document website evidence. This chapter introduces the reader to the concepts of HTML, and its application in research. It also discusses what can be found on the website and how to view metadata in embedded images, texts, and videos.

D. The Wayback machine: Benefits of Web Storage: This paper focuses on Web archiving, a web archive campaign and shows how lost websites can be recovered using a back-up machine.

E. Design of an Enhanced Web Archiving System: The project proposes a BClinked archive program to maintain content integrity using blockchain technology and attempts to extend WARC file details.

2.1 Summary of Related Work

The summary of methods used in literature is given in Table 1.

Table 1 Summary of literature survey

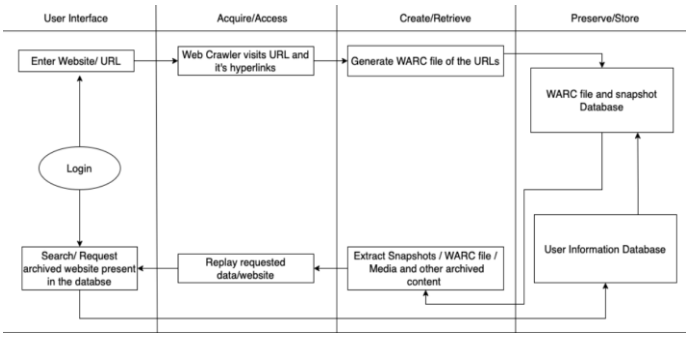
Literature	Advantages	Disadvantages
Jinfang Niu, 2012. [1]	This overview is based on a comprehensive review of literature that explains how web archiving is being done.	However, none of the literature directly addresses the knowledge and skills required by the professionals in the field who perform the daily routine of selecting, acquiring and cataloging web archives.
Nick Ruest , Samantha Fritz and Ryan Deschamps. January 2021 [2]	If the Archives Unleashed Cloud, unlocks the potential of web archives by developing and providing the tools, via a cloud service, for scholars with limited – but not none technical expertise to explore archived web content.	The problem encountered, like many digital humanities projects, is that of project sustainability
B T Sampath Kumar, January 2015. [4]	Due to the loss of web information due to several reasons like network failures, URL failure, etc. wayback machine can be used to recover some percentage of the missing web site.	N/A
ToddG. Shipley, Art Bowker,2017. (An Introduction to Solving Crimes in Cyberspace). [3]	With a little understanding the investigator can identify who owns the site, information about the contents, and potentially useful metadata from the source code as well as images and other items embedded on the site.	But scraping and using information of social media platforms or an individual's webpage without the owner's permission is not permissible by law.
Hwang, Hyun C., Jin G. Shon, and Ji S. Park 2020[5]	Blockchain technology is that connecting the previous node by using cryptograph y to secure all data in a blockchain, and it is the decentralizat ion system which can be shared via peer-topeer network.	However, in the BCLinked web archiving system because all blockchain node data should be rebuilt even if only one single content, it takes longer time.

3. Proposed Work

We focus on building a robust and software-based tool that stores and evaluates snippets / videos of specific websites, using Web Crawling and data analysis to help individuals monitor and analyze online data. In addition, a web archive system with authentic content authentication can provide a reliable web archive. We suggest a new blockchain-based web archive to solve this challenge, and records web archive and content archive activity in blockchain databases.

3.1Proposed System Architecture

The system architecture is given in Figure 1. Each block is described in this Section.



Content crawling: This block covers all crawling activities on the web. Crawler is a computer program that automatically searches documents on the web. Searches are scheduled for repeated actions to make browsing default. In our case, the searcher picks up a target URL and begins to identify the URL and all the hyperlinks embedded in it. While you visit the links, the searcher finds summaries of each page visited, and generates a WARC file for pages. All this information is stored in a web archive, and is later pushed to the blockchain domain.

Domain information storing: The first level of the Blockchain database is the Domain blockchain. This block enables domain information storage in the domain blockchain. Whenever a target URL is posted, a new node or block of Domain blockchain is generated with the same name as the URL domain. If the target domain already exists in the domain blockchain, incoming content is stored in the existing block. If the target domain is not available, a new block is created and added during the domain information storage. The blockchain blocks / node in this category contains only the domain name and general URL information. Nodes are linked to key content stored in Web Content Nodes.

Content information storing: The second level of Blockchain database is the Web Content blockchain. Building a new domain block in the previous phase, also leads to the creation of a web-based content block or node that stores domain-related content. The need for this block arises as all content in the WARC file can be too large to place in a domain blockchain due to block size limitations. Therefore only the amount that can identify content in a web content block is added to the domain block. Snapshots and WARC files found in the first section are stored in content blocks on the web.

Requirement Analysis

The implementation detail is given in this section.

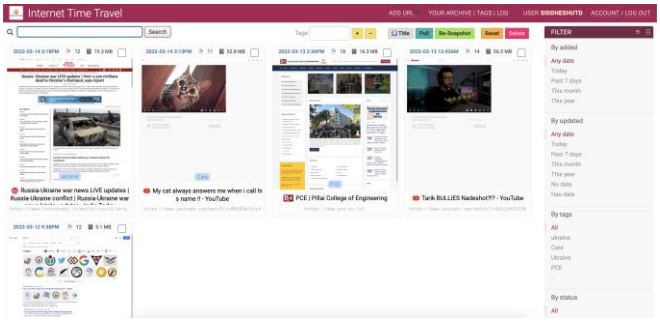
3.1 Software

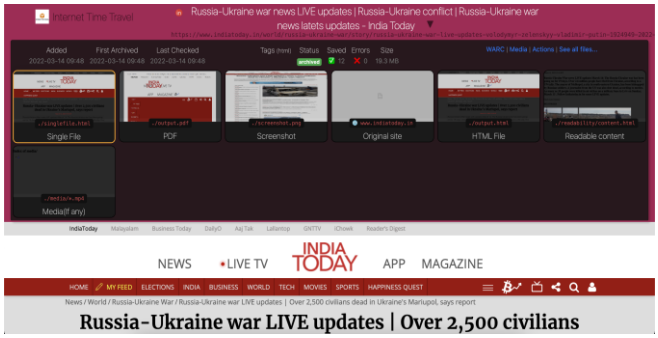
Operating System	Windows 10
Programming Language	Python

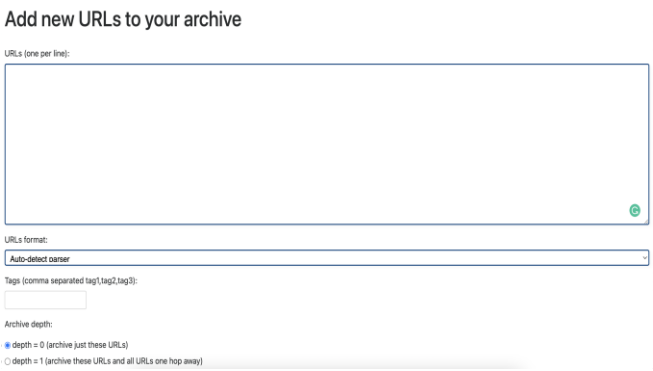
3.2 Hardware

Processor	3.5 GHz Intel
HDD	1 TB
RAM	8 GB

Implementation:







Acknowledge

It is our privilege to express our sincerest regards to our supervisor Professor Pranita Pingale for the valuable inputs,

able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this work. We deeply express our sincere thanks to our Head of the Department Dr. Sharvari Govilkar and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to present this work.

Conclusion

In this paper, we have conveyed the benefits of webscraping and analyzing the context of a single research center of interest, small firms innovating in the raw materials industry. We have provided a six-step approach to facilitate duplication of our processes through website analysis and social science communities. We hope this work provides a useful approach to the wider community of social science who is interested but who is not yet free to use online data. Although the Wayback Machine is an important resource for scholars in all fields, we believe that social scientists in particular have not yet been able to tap the full potential of online data resources. Going forward, there are significant opportunities to incorporate additional analytical tools and statistical models while simultaneously evaluating, refining, and reporting on the appropriateness of measures and results. However, one should keep in mind that both automatic and manual effort is required to create high quality information that can provide these benefits and overcome the limitations of archived information.

References

1. Niu, Jinfang, "An Overview of Web Archiving" (2012). School of Information Faculty Publications. 308.

2. Ruest, Nick & Fritz, Samantha & Deschamps, Ryan & Lin, Jimmy & Milligan, Ian. (2021). From archive to analysis: accessing web archives at scale through a cloud-based interface. International Journal of Digital Humanities. 10.1007/s42803-020-00029-6.

3. Yiu-ming Todd G. Shipley, Art Bowker, Chapter 13 - Investigating Websites and Webpages, Editor(s): Todd G. Shipley, Art Bowker, Investigating Internet Crimes, Syngress, 2014.

4. Sampath Kumar, B T. (2015). Wayback machine: A Boon for Preserving Web Sources.

5. Hwang, H.C.; Shon, J.G.; Park, J.S. Design of an Enhanced Web Archiving System for Preserving Content Integrity with Blockchain. Electronics 2020, 9, 1255. https://doi.org/10.3390/electronics9081255