The Internet Archive

Derek D’Souza, A01266791

Computing, British Columbia Institute of Technology

COMP 7071: Database Applications Development and Optimization

Tejinder Randhawa

April 16th, 2025

Table of Contents

Derek D’Souza – A01266791

Overview

This document comes in two parts. The first is an essay about what the Internet Archive is, how it works, its importance in preserving digital history, the website’s legal position and battles, and recent security breaches, written from a third-person point of view. The second part documents the development of my Internet Archive Replica project, written from a first-person point of view. Including what I planned and considered in scope for the project, how that scope changed, and how I rationalized my design decisions.

Part I – What is the Internet Archive?

Introduction

The Internet Archive is a San Francisco-based non-profit digital library of webpages and files that is free to the public. It was founded by Brewster Kahle in 1996 on a mission to provide “universal access to all knowledge.” Users can upload files to the Archive’s digital collection, but most of its data is automatically collected through web crawlers for site archival. The Internet Archive also owns the Wayback Machine, which lets users view snapshots of websites from specific dates, and the Open Library, which lets one user borrow one copy of an e-book at a time (The Internet Archive, n.d.).

The Internet Archive stores its digital library amongst 4 data centers, all in California, between 745 nodes and 28,000 spinning disks. Its primary data center is in a former Christian Science Church in San Francisco. The Internet Archive’s custom-designed PetaBox system stores a grand total of 212 Petabytes, 57 for the Wayback Machine alone, 42 for books/music/video collections, accounting for a total of 99 in unique data, the other 113 keeping backup data (The Internet Archive, n.d.). A significant portion of the archived material is obsolete and unavailable through mainstream outlets such as digital or physical stores. Such as old software/video

games, analog digitization (such as from VHS tapes or vinyl records), TV airings, radio shows, and concert recordings. The Internet Archive publicly disclosed an annual budget of $37 million, from web crawling services, partnerships, donations, and grants (Wikipedia, 2020).

The Internet Archive is important because it combats digital censorship and preserves digital information and history that would otherwise be ephemeral. Aiding researchers, journalists, and activists in accessing information necessary for scientific progress and the documentation of current events and crises. Inevitably, the project has faced and lost several legal disputes over the years, as well as recently encountered cyberattacks. In Brewster Kahle’s own words, “It’s a research library. It’s there to record and make available an accurate version of the past. Otherwise, we’ll end up with a George Orwell world where the past can be manipulated and erased.” (Blistein, 2025)

DMCA Exemption

In October 2003, the Internet Archive received a Digital Millennium Copyright Act (DMCA) exemption to help archive vintage software. Due to the inevitable degradation of magnetic media, such as floppy disks, within 10 to 30 years (The Internet Archive, 2003). This status is renewed according to a rulemaking proceeding held by the Copyright Office to:

"Determine whether there are particular classes of works as to which users are, or are likely to be, adversely affected in their ability to make non-infringing uses due to the prohibition on circumvention of access controls."

This means asking the Internet Archive if digital rights management (DRM) is unfairly preventing legal use of creative works. The four classes of exempt software described by the official Statement of the Librarian of Congress Relating to Section 1201 Rulemaking:

1. *Compilations consisting of lists of Internet locations blocked by commercially marketed filtering software applications that are intended to prevent access to domains, websites or portions of websites, but not including lists of Internet locations blocked by software applications that operate exclusively to protect against damage to a computer or computer network or lists of Internet locations blocked by software applications that operate exclusively to prevent receipt of email.*

In short, exempt websites that are commonly blocked by internet filters (such as parental control software) but not by antivirus software.

2. *Computer programs protected by dongles that prevent access due to malfunction or damage and which are obsolete.*
3. *Computer programs and video games distributed in formats that have become obsolete and which require the original media or hardware as a condition of access*
4. *Literary works distributed in ebook format when all existing ebook editions of the work (including digital text editions made available by authorized entities) contain access controls that prevent the enabling of the ebook's read-aloud function and that prevent the enabling of screen readers to render the text into a specialized format.*
5. (U.S. Copyright Office, 2003)

When software is no longer supported by its copyright holder, it becomes abandonware, whether the publisher is out of business or simply no longer distributed. Abandonware could be widely used, but it is under threat of being lost media if not preserved. Websites like the Internet Archive and *My Abandonware* help preserve abandonware for anyone who wants to access it. By digitally preserving software from obsolete hardware and formats, those hardware and formats can be retired.

Hachette v. Internet Archive

In March 2020, the Internet Archive launched the National Emergency Library (NEL) program as an offshoot of their Open Library project to help students, professors, researchers, and readers access digital copies of books they need. However, unlike the Open Library, the NEL lifted the Open Library's one-user-one copy book borrowing rule. Due to backlash from book authors accusing the Internet Archive of perpetrating piracy with the NEL, the Internet Archive reinstated that rule within just 2 months of the program’s launch. However, publishing houses such as Hachette, HarperCollins, Penguin Random House, and Wiley filed a lawsuit. 3 years later, the Internet Archive lost the case; district court judge John G. Koeltl ruled that the Internet Archive created “derivative works” not within fair use/transformative (Nast, 2024). The financial settlement was not publicly disclosed.

Music Labels v. Internet Archive

In August 2023, Music labels, including Universal Music Group (UMG) Records and Sony Music Entertainment, filed a lawsuit against the Internet Archive. Targeting their Great 78 Project, their effort to digitize and preserve 78 rpm records, the once-dominant format of physical audio in shellac discs, between the 1890s and vinyl’s takeover in the 1940s and 1950s. Also, preserving music from long-defunct labels at risk of becoming lost media. Contracted audio preservationists, led by George Blood, also a defendant in the lawsuit, digitized 400,000 records from 2017 up to August 2023. To summarize the court case documented by Rolling Stone, the defendants accused the Great 78 Project of perpetrating “wholesale theft of generations of music”, rejecting the project’s intent to perverse and research 78 rpm records, and arguing that it “undermines the value” of the recordings by cutting off royalty and revenue streams to legitimate copyright holders. Blood defended the Great 78 Project’s claim by arguing that 95% of the material is not available anywhere else and is lost to time. The plaintiff music labels are

seeking $150,000 in damages per music record of 4,142 uploads, totaling a potential loss of $621 million as of September 2024, (Blistein, 2025) which is about 16 years' worth of the Internet Archive's annual budget of $37 million from 2019, a potentially devastating blow to the archive and digital history. This court case is still ongoing.

A Series of Security Breaches

In October 2024, the Internet Archive suffered two episodes of cyberattacks. The first was on the 9th, a data breach by an anonymous threat actor, and a distributed denial of service (DDoS) attack by the BlackMeta hacktivist group, shutting down the Wayback Machine. The data breach leaked the website's user authentication SQL database, 6.4 GB in size, with approximately 31 million user records, containing email addresses, usernames, password change timestamps, and bcrypt-hashed passwords, among other data (Abrams, 2024). A JavaScript alert was injected into the website by the hacker, announcing to users:

Have you ever felt like the Internet Archive runs on sticks and is constantly on the verge of suffering a catastrophic security breach? It just happened. See 31 million of you on HIBP! HIBP stands for *Have I Been Pwned*, a popular data breach notification service.

Brewster Kahle came forward with a statement on X the same night:

What we know: DDOS attack–fended off for now; defacement of our website via JS library; breach of usernames/email/salted-encrypted passwords. What we've done: Disabled the JS library, scrubbing systems, upgrading security. Will share more as we know it.

And the next morning:

Sorry, but DDOS folks are back and knocked http://archive.org and http://openlibrary.org offline. @internetarchive is being cautious and prioritizing keeping data safe at the expense of service availability. Will share more as we know it (Kahle, 2024).

BlackMeta took to X on October 9th and 13th to vocalize why they targeted the Internet Archive. In summary, they claimed that the website had ties to the government of the United States of America, Israel’s closest ally, and therefore is a pro-Israel organization. The group also accused the website of perpetuating piracy, citing lawsuits by publishers and music labels. (SN_BLACKMETA, 2024). However, the Internet Archive is a non-profit and not affiliated with the US government or any pro-Israel/Zionist organization. As pointed out by community notes, giving context to readers in the thread. Who also condemned the attack as counterintuitive to pro-Palestinian activism, as the Internet Archive is used to distribute documentation of Israel's occupation and genocide against Palestinians. The Internet Archive would suffer another security breach on October 20th. A group of anonymous hackers exploited unrotated ZenDesk APIs to breach its support platform. Exposing personal information of users who submitted support tickets. None of the attacks were financially motivated (Ahmed, 2024).

Conclusion

The Internet Archive could be at risk of shutting down – and with it, up to a hundred petabytes of Internet history lost. Battling both legitimate and malicious parties, from copyright holders to hackers. The Internet Archive is serving the public good by providing *universal access to all knowledge.* Copyright holders’ protection of their intellectual property and profits clashes with media accessibility and preservation, perpetuating an inequality in access to media among the general public. However, the Internet Archive is

still accountable to scrutiny about what and how media is distributed, and how they secure their systems.

Part II – The Internet Archive Replica

Introduction

This project’s purpose is to replicate the Internet Archive in a local environment. The front end would provide a visually reminiscent, though simplified, user interface and experience to the Internet Archive. User interface elements will be borrowed through Inspect Element if they can. The functionality will be showcased through create, read, update, and delete (CRUD) operations with locally stored files and metadata stored in a relational database.

Scope

I expect this project to handle up to 8 GB worth of files locally stored. Cloud storage would add unnecessary monetary, architectural, and performance costs. Such as paying for storage space itself and/or the API to upload and download files from the cloud, I had originally planned to use Minio for S3 object storage. With only a few gigabytes, unlike the many petabytes of the Internet Archive, there won’t be as much to narrow down for. I had originally also planned to replicate the Internet Archive’s Wayback Machine. It became clear early on that ambition was out of scope. For archiving and presenting snapshots of webpages, I would have to learn web crawling and at least one API for it. It was best to focus on just the Internet Archive. I also overestimated how scalable the project would be. However, due to time constraints, I had to focus on functionality in a local environment.

Derek D’Souza – A01266791

Choosing My Tech Stack

I couldn’t find what the Internet Archive’s official technological stack is for either its backend or frontend, the website is closed source. But being a website nearly 30 years old, they'd likely still use mostly traditional HTML/CSS and JavaScript for the front end. And plausibly raw SQL over object-relational mapping (ORM) tools. The Internet Archive Developer Portal provides developers with APIs, primarily written and implemented in Python, which could be a significant part of their backend. The Open Library is open source, and its code is publicly available on GitHub, which gave me more hints for what tech stack to use. It is primarily written in Python, HTML/CSS, and JavaScript. I prioritized choosing a stack based on what I’m already familiar with and what could scale well. I chose ReactJS, Python, and PostgreSQL for my frontend, backend, and database, respectively. However, I chose to use only vanilla CSS instead of a framework, it would help me replicate the website's older aesthetic design. I’m using the React framework as a safety net to keep my user interface and user experience (UI/UX) design from becoming too complex to manage. I had originally set up my project to run in a Docker container with the front end, back end, database, plus a Minio and Elastic Search clients for storage and a search engine, respectively. However, as I began to work on backend interactions, two things became obvious to me. One, Minio would be unnecessary for storing just a few gigabytes of data. And two, a user wouldn’t need the advanced metadata search options afforded by Elastic Search to narrow down an overall simple metadata schema.

Frontend

The dilemma I face in designing the front end for this replica project is between authenticity and simplicity. It would be more effective to capture the feeling of browsing an old website, such as the Internet Archive, by presenting a UI/UX that accurately reflects 2000s web design. However, the Internet Archive itself has updated its UI/UX over the years. As I use Inspect Element on the

website itself to help me design the replica's UI, due to the sheer quantity of HTML and CSS for just one webpage, it would be a project in and of itself to recreate it precisely. I implemented a grid view of search results like the website but without image previews. The Internet Archive itself offers savvy users an Advanced Search option to narrow down the content they're looking for. But my project's library won't be big enough to necessitate that as a feature.

Backend

I'm using a Python Flask server to expose RESTful APIs to handle CRUD operations for file archives in my application, specific endpoints such as "upload", "download", "search", "item", "edit", and "delete". The Flask server interacts with a PostgreSQL database for storing file metadata. My backend also manages the uploaded files in an "uploads" folder in my computer's Documents directory. Since I'm sticking to local storage, this will simplify file management and ensure tight coupling between metadata and content. Reducing the complexity and overhead of external storage services. However, for a scalable application, using an S3 object storage solution would be more appropriate.

Database

I only have one table in my database: "files," which stores metadata for files in my project. Columns: id, filename, file path, title, description, subjects, creator, upload date, collection, language, and file size. Each row is small and simple, especially compared to the extensive metadata schema used by the Internet Archive, as detailed in their developer portal. Metadata is itself semi-structured and can be represented by fixed/structured (for more quantitative data) or evolving/unstructured schemas (for more qualitative data) (AWS, n.d.). Because my schema is relatively simple, my choice of database technology wasn't a critical decision, so I chose the

user-friendly PostgreSQL. The more horizontally scalable NoSQL would be more appropriate for handling large-scale, distributed metadata storage. PostgreSQL offers vertical scalability, making it suitable for maximizing performance on a single server (Chandu, 2024). Since my metadata is structured and I don’t expect to deal with unstructured or rapidly evolving schemas—where NoSQL’s schema-less design excels—PostgreSQL is a more appropriate choice for my project.

Conclusion

I generally met my scope for this project. A user can upload a single file (or a zipped archive of multiple) up to 2 GB, search files by title, creator, keywords in description or subject, and collection (e.g., community texts or software), download them and edit their metadata on the website. I was able to design a landing and upload pages resembling their original Internet Archive counterparts. What I think I could’ve done differently is design my database before the backend. There were a few oversights I ended up making that I had to fix later, such as forgetting a file size column and mismatches between collection names. Had I planned better, those issues would’ve been prevented. If I intended to create a production build of the project, I would’ve added user authentication to tie uploads to users. What I achieved is to create an “Internet Archive Playground”, letting a user play around with file and metadata CRUD, more than a straight replica.

Works Cited

Abrams, L. (2024, October 9). *Internet archive hacked, data breach impacts 31 million users*. BleepingComputer. https://www.bleepingcomputer.com/news/security/internet-archive-hacked-data-breach-impacts-31-million-users/

Ahmed, D. (2024, October 21). *Internet archive (Archive.org) hacked for second time in a month*. Hackread - Latest Cybersecurity, Hacking News, Tech, AI & Crypto. https://hackread.com/internet-archive-archive-org-hacked-for-second-time/

Blistein, J. (2025, January 15). *Inside the $621 million legal battle for the 'Soul of the internet'*. Rolling Stone. https://www.rollingstone.com/music/music-features/internet-archive-major-label-music-lawsuit-1235105273/

Chandu, P. [Premchandu]. (2024, July 17). *Postgres horizontal scalability?* Medium. https://medium.com/@premchandu.in/postgres-horizontal-scalability-4e125c73aa2f

Internet archive gets DMCA exemption to help archive vintage software. (2003, October). Internet Archive: Digital Library of Free & Borrowable Texts, Movies, Music & Wayback Machine. https://archive.org/about/dmca.php

Internet Archive Projects. (n.d.). Internet Archive: Digital Library of Free & Borrowable Texts, Movies, Music & Wayback Machine. https://archive.org/projects/

Internet archive: About IA. (n.d.). Internet Archive: Digital Library of Free & Borrowable Books, Movies, Music & Wayback Machine. https://archive.org/about/

Internet archive: Operations. (2020, May). Wikipedia, the free encyclopedia. Retrieved April 16, 2025, from https://en.wikipedia.org/wiki/Internet_Archive#Operations

Khale, B. [@brewster_kahle]. (n.d.).

X. https://x.com/brewster_kahle/status/1844183111514603812

Nast, C. (2024, September 4). *The internet archive loses its appeal of a major copyright case*.

WIRED. https://www.wired.com/story/internet-archive-loses-hachette-books-case-appeal/

Petabox. (n.d.). Internet Archive: Digital Library of Free & Borrowable Texts, Movies, Music &

Wayback Machine. https://archive.org/web/petabox

Structured data vs unstructured data - Difference between collectible data - AWS. (n.d.).

Amazon Web Services, Inc. https://aws.amazon.com/compare/the-difference-between-structured-data-and-unstructured-data/

U.S. Copyright Office (www.copyright.gov). (2003, October). *Statement of the librarian of Congress relating to section 1201 rulemaking*. U.S. Copyright Office | U.S. Copyright

Office. https://www.copyright.gov/1201/docs/librarian_statement_01.html

[@Sn_darkmeta]. (2024, October 13).

X. https://x.com/Sn_darkmeta/status/1845518663975158043