Richtlinie zum „Software-Sprint“

Archivindex-Builder

Schlussbericht

Zuwendungsempfänger:

Travis Robert Brown

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS23S32 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Was war Deine Motivation? Welches Problem wolltest Du mit Deinem Projekt lösen? Wie war die geplante Vorgehensweise zur Problemlösung (auch Angabe der wichtigsten Meilensteine)?

Gemäß dem Vorschlag ermöglicht das Projekt Benutzer*innen, einen durchsuchbaren Index iterativ aus öffentlichen Webarchiven zu entwickeln. Die Benutzer*in stellt der Anwendung eine Reihe von URLs bereit, von denen die archivierten Inhalte heruntergeladen werden und dann verwandte URLs für eine zusätzliche Indizierung empfohlen werden. Die Inhalte können über eine Websuchschnittstelle durchsucht werden.

Diese Funktionalität ist wertvoll, da Webarchive wie die Wayback Machine des Internet Archive normalerweise nicht von herkömmlichen Suchmaschinen indiziert werden. Die Wayback Machine umfasst historische Schnappschüsse von über 835 Milliarden Webseiten. Diese Daten können eine wertvolle Ressource für Journalisten und andere Forscher*innen sein, aber selbst bei gezielten Untersuchungen kann es schwierig sein, sich darin zurechtzufinden. Der Versuch, mit diesen Daten in größerem Maßstab zu arbeiten, erfordert in der Regel einen erheblichen manuellen Aufwand.

In diesem Projekt haben wir ein System für den Zugriff auf Webarchivindizes, das Herunterladen von Schnappschüssen, die Indizierung dieser Schnappschüsse für die Volltextsuche und die Präsentation dieser indizierten Daten für den Benutzer*in über eine vertraute Websuchoberfläche entwickelt. Diese Schnittstelle kann auf vertrauenswürdige Benutzer*innen beschränkt werden, die über einen OAuth-Identitätsanbieter authentifiziert wurden (derzeit werden Google und GitHub unterstützt). Instanzen der Suchanwendung können über eine Reihe von Befehlszeilentools erstellt und verwaltet werden.

Beitrag des Projektes zu den Zielen der Förderinitative „Software-Sprint“

Welche Bezüge gibt es zu den Themenfeldern „Civic Tech“ und „Data Literacy“ des Software Sprints oder zu weiteren gesellschaftlich relevanten Zielen bzw. Lösungsansätzen?

Das Internet wird von verschiedensten Akteuren mit zum Teil böswilliger Absicht genutzt, um unter anderem durch Desinformation politischen Einfluss oder finanzielle Mittel zu erlangen. Beweise für ihre Handlungen werden oft verschleiert, sei es absichtlich oder durch den Lauf der Zeit und den Verfall von Links. Archive des offenen Webs spielen eine bedeutende Rolle bei der Verfolgung böswilliger Akteure, doch die Navigation darin erfordert häufig technisches Know-how, Fachwissen, sorgfältige Handarbeit und Glück.

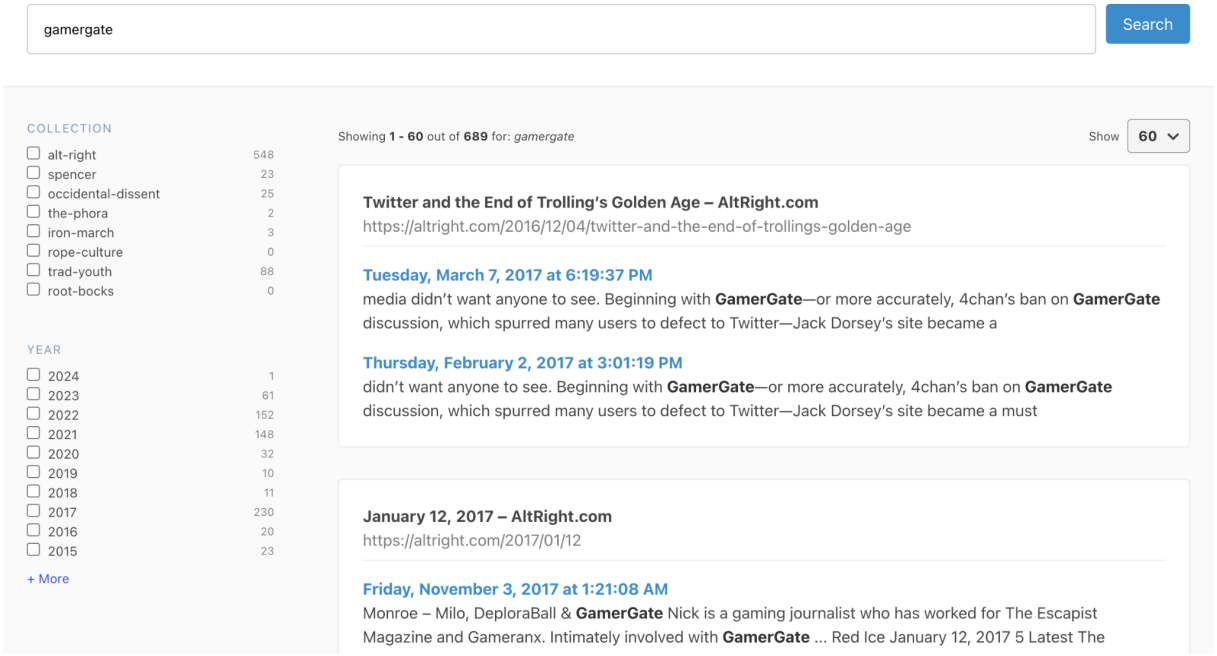
Ein konkretes Beispiel hierfür ist George Santos, ein ehemaliger Kongressabgeordneter der Vereinigten Staaten, über den erst nach seiner Wahl im Jahr 2022 viel Berichterstattung über angeblichen Betrug und Täuschung erfolgte. Rückblickend ist es zwar einfach, in Webarchiven Spuren seiner Vergangenheit zu finden, jedoch wurden die meisten dieser Untersuchungen erst nach Santos’ Amtsantritt durchgeführt.

Unsere entwickelte Anwendung zur Erstellung durchsuchbarer Indizes von Webarchivinhalten soll solche Recherchen für Journalist*innen und andere Forscher*innen praktikabler gestalten.

Ausführliche Darstellung der Ergebnisse

Welche konkreten Ergebnisse hast Du erzielt? Konnten alle Meilensteine erreicht werden? Welche zusätzlichen Erkenntnisse hast Du aus der Projektarbeit gewonnen, auch im Hinblick auf die Begleitung durch die Open Knowledge Foundation?

Das wichtigste Artefakt, das im Rahmen des Projekts entstanden ist, ist eine Web-Suchoberfläche. Diese ermöglicht es Benutzer*innen, den vollständigen Text archivierter Webseiten in einer intuitiven Benutzeroberfläche zu durchsuchen, die für jeden, der herkömmliche Web-Suchmaschinen verwendet hat, einfach zu bedienen ist (siehe Abbildung unten).



Über diese Benutzeroberfläche können Benutzer*innen ihre Suche entweder nach dem Jahr der Archivierung oder nach einer „Sammlung“ filtern. Letztere bezeichnet eine Reihe von Seiten, die ein gemeinsames URL-Präfix haben (normalerweise eine Website oder einen Abschnitt einer Website, z. B. eine einzelne Substack-Publikation oder Tweets, die von einem bestimmten Twitter-Konto gepostet wurden).

Die Benutzeroberfläche zeigt eine konfigurierbare Menge an Kontext für jedes Ergebnis an, und die Links in der Ergebnisliste führen den Benutzer*innen direkt zum archivierten Seiten-Snapshot (in der Regel gehostet von der Wayback Machine).

Die Darstellung der Ergebnisse unterscheidet sich von herkömmlichen Suchmaschinen dadurch, dass möglicherweise mehrere Ergebnisse für eine einzelne Seite angezeigt werden, wenn die Seite mehrmals archiviert wurde, möglicherweise über viele Jahre oder sogar Jahrzehnte hinweg. Einige Seiten enthalten möglicherweise Hunderte oder Tausende einzelner Schnappschüsse. Daher versucht das System, die relevantesten zu identifizieren und darzustellen. Benutzer*innen können jederzeit nach Datum filtern, um einen Drilldown zu einem bestimmten Zeitraum durchzuführen.

Die Schnittstelle basiert auf einer Open-Source-React-Bibliothek, die vom Elastic-Projekt entwickelt wurde, und nutzt einen Dienst, der von Tantivy unterstützt wird, einer Rust-Bibliothek, die es ermöglicht, einen Standardsatz an Suchoperationen bereitzustellen (z. B. Suche nach genauen Phrasen und boolesche Operatoren).

Der größte Teil der Entwicklungsbemühungen des Projekts konzentrierte sich jedoch nicht auf diese Schnittstelle, sondern auf die Infrastruktur, die für die zugrunde liegende Datenplattform erforderlich ist. Zunächst wurde die `/cdx/search`-API der Wayback Machine verwendet, um unseren Index zu erstellen, jedoch sind wir dann auf die neuere Timemap CDX-API umgestiegen, die einige Vorteile bietet, aber leider weniger gut dokumentiert ist und ein alternatives Paging-Schema verwendet.

Bei großen Websites kann der Zugriff auf den Index der verfügbaren Snapshots Stunden dauern. Daher ist es von Bedeutung, diese Arbeit nach Möglichkeit nicht mehrfach durchführen zu müssen.

Das Projekt stellt daher Werkzeuge zur Verfügung, um einen lokalen CDX-Indexdatenspeicher mit Apache Parquet als Speicherformat aufzubauen. Dies ermöglicht es anderen Tools, mit dem CDX-Index zu arbeiten, ohne erneut auf die Dienste des Internetarchivs zugreifen zu müssen, oder dass Administratoren mit einer Anwendung wie DuckDB interaktiv durch den Store navigieren können. Die Verwendung der Parquet-Bibliotheken, die vom Open-Source-Ökosystem Rust bereitgestellt werden, empfanden wir als etwas umständlich. Daher entwickelten wir zu Beginn des Projekts eine Reihe von Tools zur Parquet-Codegenerierung, um den Prozess zu vereinfachen. Weitere Einzelheiten finden Sie unter den untenstehenden Links.

Der Zugriff auf Wayback Machine-Snapshots für Sammlungen kann sogar noch länger dauern (Tage oder Wochen bei großen Websites), daher liegt der Schwerpunkt erneut darauf, die Notwendigkeit wiederholter Arbeiten zu reduzieren. Diese Herausforderung wird durch die Tatsache verschärft, dass das Internetarchiv diese Anfragen oft langsam bearbeitet oder stundenlang überhaupt nicht verfügbar ist. Wir haben daher mit der Spiegelung von Schnappschüssen experimentiert, sodass Benutzer*innen nicht direkt auf das Internetarchiv zugreifen müssen, um ihre Ergebnisse im Kontext zu sehen.

Während der Entwicklung haben wir mit Benutzer*innen zusammengearbeitet, deren Projekte eine nicht standardmäßige Suchindexfunktionalität erforderten, und das aktuelle System bietet einige benutzerdefinierte Indizierungen zur Unterstützung dieser Anwendungsfälle. Ein Beispiel ist die Indizierung von Bild-URLs. Herkömmliche Suchmaschinen indizieren im Allgemeinen nur Text, der auf der gerenderten Seite sichtbar ist. In manchen Kontexten kann es jedoch sinnvoll sein, nach nicht sichtbaren Inhalten zu suchen. Beispielsweise möchte ein Forscher*in, der eine Instanz zur Indizierung einer Sammlung extremistischer Foren eingerichtet hat, möglicherweise auf mehreren Websites nach Kommentaren einer bestimmten Person suchen, von denen einige möglicherweise die Gravatar-Plattform für Benutzeravatare verwenden. Die Möglichkeit, nach Gravatar-Bild-URLs zu suchen, macht dies viel praktischer als die einfache Suche nach Benutzernamen, die oft mehrdeutig sind. Wir bieten eine ähnliche benutzerdefinierte Indizierung für Seiten, die Tweets enthalten, und das System ist so konzipiert, dass es bei Bedarf in Zukunft praktisch ist, zusätzliche benutzerdefinierte Indizierungen hinzuzufügen.

Das System ist als Sammlung halbunabhängiger Komponenten implementiert, die es ermöglichen, dass einzelne Funktionalitäten potenziell in anderen Kontexten als Bibliotheken verwendet werden können. Im Allgemeinen erreicht das umgesetzte Projekt die im Vorschlag dargelegten hochrangigen Meilensteine, obwohl einigen Bereichen, basierend auf dem, was wir im Verlauf des Projekts gelernt haben, mehr Aufmerksamkeit geschenkt wurde, anderen weniger (Einzelheiten finden Sie im übernächsten Abschnitt), und Einige ungeplante Funktionen wurden implementiert.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Welcher Nutzen ergibt sich für die Zielgruppe aus den Ergebnissen Deines Projekts? Welche weiter-gehenden Effekte ergeben sich aus der Open-Source-Stellung der Ergebnisse? Gibt es Ideen für die Weiterentwicklung Deiner Lösung und Pläne zu deren Umsetzung?

Hat die Arbeit in dem Projekt Dich in Deiner persönlichen, fachlichen Weiterentwicklung unterstützt?

Einige Komponenten des Systems wurden bereits in journalistischen Ermittlungen eingesetzt, und wir bereiten eine Instanz für die Nutzung durch dieselbe Gruppe vertrauenswürdiger Aktivist*innen und

Forscher*innen vor, die derzeit Zugriff auf die Dienste haben, die in unserem Hassreden-Tracker-Projekt (unterstützt vom Prototype Fund in 2022) entwickelt wurden.

Das Projekt ermöglichte es mir, Fachwissen in einigen wichtigen Technologien zu entwickeln, die für mich neu waren, und hat bereits zu mehreren neuen Arbeitsbeziehungen mit Forscher*innen und Journalist*innen geführt.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Gab es Arbeiten bzw. Lösungsansätze, die nicht weiter verfolgt wurden? Was waren die Hinter-gründe, und wie bist Du alternativ vorgegangen?

Der Vorschlag beinhaltete Unterstützung für den Zugriff auf und die Indizierung von Archive Today, einem weiteren weit verbreiteten Webarchiv, das nicht mit dem Internet Archive verbunden ist. Wir haben mit einer solchen Unterstützung experimentiert, aber Einschränkungen des automatisierten Zugriffs machten es schwierig, und die aktuelle Version der Software unterstützt Archive Today nicht, obwohl wir hoffen, diese Integration in der zukünftigen Entwicklung abzuschließen.

Der Vorschlag besagt auch, dass Möglichkeiten zur Verwaltung von Instanzen „über eine Befehlszeilenschnittstelle und möglicherweise eine Webschnittstelle“ bereitgestellt werden sollen. Während wir mit der Arbeit an der administrativen Webschnittstelle begonnen haben, ist nur die Befehlszeilenschnittstelle fertig.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Wo können sich Interessenten detailliert über Deine Projektergebnisse informieren (z.B. Webseite, GitHub, Veröffentlichungen)?

- Haupt-Code-Repository: https://github.com/travisbrown/archivindex-builder
- Werkzeuge für die Codegenerierung und andere Parquet-Hilfsmittel: https://github.com/travisbrown/parquetry
- Werkzeuge zum Flusen und Erstellen von JSON-Schemas: https://github.com/travisbrown/json-schema-tools

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Gab es im Projektverlauf Ereignisse, die eine Anpassung der Planung erforderlich machten – z.B. Mehr- oder Minderaufwand bei der Bearbeitung von Teilaufgaben?

Die Arbeiten wurden im Allgemeinen planmäßig abgeschlossen, jedoch wurde mehr Zeit als ursprünglich geplant für Infrastrukturarbeiten verwendet, während weniger Zeit in die Entwicklung der Schnittstellen floss.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Gab es Entwicklungen anderer Personen oder Institutionen, die Einfluss auf Deine Arbeiten und die Zielsetzung hatten? Wenn ja, worin bestand dieser und wie bist Du damit umgegangen?

Wir haben die Entwicklungen in der Webrecorder-Community aufmerksam verfolgt, die sich auf Archivierungstools und -standards konzentriert. Diese stellen unserer Meinung nach eine sinnvolle Ergänzung zu unserem Fokus auf die Indizierung dar. Für unsere ersten Anwendungsfälle, vor allem in der Forschung und im Journalismus, ist die Möglichkeit, große Sammlungen von Wayback Machine-Schnappschüssen zu indizieren, von entscheidender Bedeutung und in der Regel ausreichend. Allerdings werden Websites und Plattformen wie X (ehemals Twitter) zunehmend zu Hindernissen für die öffentliche Archivierung durch das Internetarchiv. Daher wird es immer wichtiger, die Indizierung verteilter und privater Archive zu unterstützen, wie sie beispielsweise mit dem Browsertrix Crawler von Webrecorder erstellt werden können. Das System bietet derzeit eine minimale Interoperabilität über den WACZ-Import und -Export, ist jedoch darauf ausgerichtet, eine umfassendere Unterstützung in der Zukunft praktisch umzusetzen.