Abb.

Webseitenarchivierung im Test

von Michael Cöln, Johannes Ehrengruber, Andreas Jüngling, Michael Korn, Jens Löffler, Gregor Patt, Dietmar Pertz, Tobias Schröter, Johannes Thomé

Einführung

Die Bundeskonferenz der Kommunalarchive beim deutschen Städtetag hat bereits 2002 festgestellt, dass sich das historische Erbe der Städte, Gemeinden und Landkreise nicht nur in Bauwerken, kulturellen, politischen und wirtschaftlichen Traditionen, sondern vor allem auch in der archivalischen Überlieferung und hier u. a. in „Unterlagen aus digitalen Systemen" konkretisiert.[1] Daraus folgt, dass auch Webseiten, denen eine besondere Bedeutung für das soziale und kulturelle Leben, die Wirtschaft sowie die Identität einer Kommune zukommt, archiviert werden müssen. Webseiten sind hoch aggregierte Quellen, die helfen,

- eine Basisüberlieferung zu der über die bzw. von der Webseite vertretenen Institution – sei es ein Sportverein, eine Firma oder eine Verwaltung – zu bilden;
- angesichts der stetig sinkenden Bedeutung von Printmedien und Printerzeugnissen auf unverwechselbare Art und Weise zu zeigen, wie sich eine Institution in der Öffentlichkeit präsentiert hat. Webseiten sind seit Ende der 1990er-Jahre in vielerlei Hinsicht die virtuellen „Schaufenster", Litfaßsäulen und Eingangstüren von Institutionen aller Art.

Diesbezüglich gilt es zunächst aber festzuhalten, was eine Webseite überhaupt ist. Laut dem einschlägigen Wikipedia-Artikel definiert man eine „Website" – mit anderen Worten die Webpräsenz, den Webauftritt, das Webangebot oder den Netzauftritt – als „die unter einer individuellen Webadresse erreichbare Präsenz eines Anbieters von Telemedien im weltweiten Netz (World Wide Web). Sie ist mit Webtechniken, beispielsweise HTML, erstellt und kann mit einem Nutzeragenten, beispielsweise einem Browser, wiedergegeben werden. Zur Webpräsenz gehören Webseiten und optional vorhandene herunterladbare Schriftstücke."[2] Abzugrenzen sind Webseiten somit von sozialen Medien und sozialen Netzwerken.

Auch wenn bei der Archivierung von Webseiten also ‚nur' von Produkten des sogenannten ‚Web 1.0' die Rede ist, erschließt sich recht rasch, dass die Archivierung eines solch komplexen Konstruktes nach dem heutigen Stand der elektronischen Langzeitarchivierung erhebliche Probleme bereitet. Die meisten gängigen Verfahren setzen technisch auf die Hilfe eines sogenannten „Web-Crawlers". Hierbei handelt es sich um eine Software, die das World Wide Web durchsucht und Webseiten analysiert. Ausgehend von einem oder mehreren Startpunkten arbeitet der Crawler eine Webseite nach vorgegebenen Regeln anhand der vorhandenen Links Unterseite für Unterseite ab. Bei Crawlern, die für die Web-Archivierung entwickelt wurden und eingesetzt werden, wird idealerweise von jedem erfassten Webseitenelement eine Momentaufnahme gespeichert.

Die Speicherung erfolgt häufig – aber nicht immer – im WARC-Format. WARC (WebARChive) ist ein Container-Format, das speziell für die Web-Archivierung entwickelt wurde.[3] WARC-Dateien können über spezielle Software in einem Browser ausgelesen werden und zeigen eine Webseite wieder (weitgehend) so an, wie sie zur Zeit der Archivierung im Internet zu finden war. Inzwischen auch als ISO-Standard zertifiziert (ISO 28500:2017), gilt WARC als etablierter Standard. Dennoch schließen Konstanze Weimer

1 Empfehlung der Bundeskonferenz der Kommunalarchive beim Deutschen Städtetag, Positionspapier: Das Kommunalarchiv, S. 1. Online unter: https://www.bundeskonferenz-kommunalarchive.de/empfehlungen/P_das_Kommunalarchiv_BV.pdf [Stand: 14.06.2022, gilt ebenfalls für alle nachfolgenden Hinweise auf Internetseiten].

2 https://de.wikipedia.org/wiki/Website.

3 Eine kurze Beschreibung des WARC-Formats findet sich auf der Webseite des nestor-Netzwerks. Konstanze Weimer / Astrid Schoger, Das Dateiformat WARC für die Webarchivierung, 2021: https://files.dnb.de/nestor/kurzartikel/thema_15-WARC.pdf.

und Astrid Schoger den entsprechenden NESTOR-Kurzartikel mit der Feststellung:

„Fragen der Langzeitarchivierung [bleiben] offen, da das Containerformat unterschiedlichste von Obsoleszenz bedrohte Dateiformate enthalten kann. Neben der Migration einzelner im WARC-Container enthaltener Dateiformate wird [deshalb] das Verfahren der Emulation früherer Browsertypen als Langzeitarchivierungsstrategie erprobt."[4]

Technisch gesehen ist die Archivierung von Webseiten daher zurzeit also in der Regel ‚nur' die „Speicherung einer möglichst originalgetreuen Kopie einer Webseite bzw. der sie konstituierenden Webseiten mit dem Ziel ihrer prinzipiell zeitlich unbefristeten Archivierung und Zugänglichkeit für eine ebenso möglichst originalgetreue Wiedergabe"[5]. Bei der Webseiten-‚Archivierung' geht es also häufig eigentlich darum, etwas zunächst so lange zu sichern, bis eine Möglichkeit zur elektronischen Langzeitarchivierung im eigentlichen Sinne gefunden ist.

Weder dies noch die Tatsache, dass sich bislang insbesondere landes- und bundesweit aktive große Bibliotheken auf dem Gebiet der Webseitenarchivierung engagieren, darf darüber hinwegtäuschen, dass die Verantwortung für die Sicherung originär kommunaler Seiten nur und ausschließlich bei den kommunalen Archiven liegen kann. Wenn die Generaldirektorin der Deutschen Nationalbibliothek (DNB) 2015 forderte, die Archivierung von Webseiten als Daueraufgabe aller damit beauftragten Einrichtungen zu definieren,[6] so sind hiermit zwangsläufig (auch) Kommunalarchive gemeint.

Es spricht für sich, dass es Ende November 2021 allein über 17 Millionen .de-Domains gab. Die Zahl archivwürdiger, nichtstaatlicher Webseiten und die damit zusammenhängende Datenmenge ist schlechterdings zu groß, als dass man die Sicherung derselben allein dem Engagement staatlicher Bibliotheken und privater Initiativen überlassen könnte. Nicht umsonst wird der in § 2 des Gesetzes über die Deutsche Nationalbibliothek (DNBG) definierte gesetzliche Auftrag der DNB, alle in Deutschland veröffentlichten „Medienwerke […] im Original zu sammeln"[7], einige Paragraphen später im gleichen Gesetz wieder eingeschränkt, um im Zweifel eine Überforderung der DNB abwenden zu können.[8]

Folglich bedarf es auch keiner näheren Begründung, warum sich unser regionaler Arbeitskreis zur digitalen Langzeitarchivierung der Frage stellte, wie es mit möglichst geringem Ressourceneinsatz gelingen kann, die für die historische Überlieferung einer Kommune maßgeblichen Webseiten zu bewahren.[9] Dies sollte geschehen, indem einige gängige Hilfsmittel und Tools einem Nutzungstest unter möglichst realitätsnahen Bedingungen im Arbeitsalltag rheinischer Kommunalarchivarinnen und -archivare unterzogen wurden. Neben den Testkandidaten sind weitere kostenfreie wie kostenpflichtige Tools erhältlich (z. B. das Offline Web Archive der Firma oia, die auch für die DNB arbeitet).

Bereits bei der Auswahl der für die Tests herangezogenen Seiten zeigte sich, wie wichtig es ist, die Aufgabe der Webseitenarchivierung schon heute und ungeachtet der noch ungelösten technischen Probleme anzugehen. Von den drei ursprünglich für den Vergleich ausgewählten Webseiten – der Seiten des Flugplatzes Hangelar (www.edkb.de, www.flugplatz-hangelar.de), eines Pizza-Lieferdienstes (www.speed-pizza.com) sowie der Stadtverwaltung Sankt Augustin (www.sankt-augustin.de) – durchlief eine während der Tests eine grundlegende Überarbeitung. Während die einen Tools anhand der relativ statischen, alten Seite der Stadtverwaltung Sankt Augustin getestet wurden, arbeiteten andere Tools plötzlich mit der sehr modernen, dynamischen Folgeversion. Die über Jahre maßgebliche Seite der Stadtverwaltung war plötzlich aus dem Netz verschwunden. Für die Tests bedeutete dies, dass die ursprünglich ins Auge gefasste Begrenzung auf festgelegte Seiten zumindest zum Teil aufgegeben werden musste. Die Mitglieder des Arbeitskreises ergänzten daher die Arbeit mit den oben genannten Seiten um ‚Experimente' mit selbst ausgewählten Webseiten.

Um die Tests im Rahmen der alltäglichen Arbeit bewältigen zu können, testete jedes Mitglied des Arbeitskreises jeweils nur ein Tool.

Ausgehend von einer allgemeinen Einführung zu demselben – im Folgenden als Basisinformationen bezeichnet – folgt eine kurze Einschätzung zu (Schwierigkeiten bei) Installation und Nutzerfreundlichkeit sowie eine Darstellung und Einordnung der Testergebnisse. Handlungsleitend waren bei der Testauswertung eine möglichst offene Einschätzung zu Vor- und Nachteilen sowie die Frage, für welche Art von Webseiten bzw. für welche Einsatzbereiche ein Tool geeignet sein könnte. Wenn dabei zum Teil unter erheblichem Zeitdruck gearbeitet und entschieden werden musste, so wurde dies aus methodischen Gründen bewusst in Kauf genommen. Die Gruppe ist von der Prämisse ausge-

4 Weimer/Schoger, Das Dateiformat WARC (wie Anm. 3), S. 3.
5 Reinhard Altenhöner / Achim Oßwald, Im Fokus: Webarchivierung in Bibliotheken, in: Zeitschrift für Bibliothekswesen und Bibliographie 62/3–4 (2015), S. 139–143, hier S. 139.
6 Elisabeth Niggemann, Im weiten endlosen Meer des World Wide Web: Vom Sammelauftrag der Gedächtnisorganisationen, in: Zeitschrift für Bibliothekswesen und Bibliographie 62/3–4 (2015), S. 153–159, hier S. 58.
7 Gesetz über die Deutsche Nationalbibliothek (DNBG), § 2.
8 DNBG, § 20: „Zur geordneten Durchführung der Pflichtablieferung und um einen nicht vertretbaren Aufwand der Bibliothek sowie um Unbilligkeiten zu vermeiden, wird das für Kultur und Medien zuständige Mitglied der Bundesregierung ermächtigt, durch Rechtsverordnung zu regeln: 1. die Einschränkung der Ablieferungs- oder der Sammelpflicht für bestimmte Gattungen von Medienwerken, wenn für deren Sammlung, Inventarisierung, Erschließung, Sicherung und Nutzbarmachung kein öffentliches Interesse besteht […]".
9 Der Arbeitskreis gründete sich im Januar 2020 auf Initiative einiger Archivarinnen und Archivare des Rhein-Sieg-Kreises und versteht sich als Austauschforum in praktischen Fragen der digitalen Langzeitarchivierung. Derzeit gehören dem Arbeitskreis Michael Cöln (Stadtarchiv Hürth), Johannes Ehrengruber (Stadtarchiv Troisdorf), Andreas Jüngling (Stadtarchiv Meckenheim), Michael Korn (Stadtarchiv Sankt Augustin), Jens Löffler (Stadtarchiv Bornheim), Dietmar Pertz (Stadtarchiv Rheinbach) sowie Tobias Schröter (Interkommunales Archiv Lohmar) an. Der Arbeitskreis wird vom LVR-Archivberatungs- und Fortbildungszentrum in Person von Gregor Patt unterstützt.

gangen, dass Webseitenarchivierung in kommunalen Archiven nur dann als regelmäßige Aufgabe in den Kanon der Fachaufgaben aufgenommen werden kann und wird, wenn sich die entsprechenden Tools intuitiv bedienen lassen und Einarbeitungszeiten auf ein Minimum beschränkt werden können. Anders als in großen (National-)Bibliotheken wird Webseitenarchivierung in kleinen und mittleren Kommunalarchiven immer eine Aufgabe sein, für die nur ein sehr begrenztes Zeit- und (gegebenenfalls) Kostenbudget zur Verfügung steht.

Heritrix

(Johannes Ehrengruber, Stadtarchiv Troisdorf)

Basisinformationen

Die Webarchivierungssoftware Heritrix,[10] die bei einer bedeutenden Anzahl von Archiven, Bibliotheken und Webarchivierungsprojekten im Einsatz ist, ist ein klassischer Web-Crawler, der 2004 vom Internet Archive entwickelt wurde und seitdem als kostenlose Open-Source-Software unter Anwendung der sogenannten *Apache-Lizenz* zur Verfügung steht. Heritrix sucht auf Grundlage einer Start-URL die jeweilige Webseite ab und verfolgt alle Links auf den gefundenen Seiten und Unterseiten. Die Software basiert dabei auf der Programmiersprache Java und speichert die erfassten Seiteninhalte sowie einige Meta-Daten (Abfragezeit, http-Header) im WARC-Format. Enthalten ist der integrierte, lokale Webserver Jetty, über den die *Jobs*[11] (Archivierungsvorgänge) geplant und ausgeführt werden. Die Software bietet also eine webbasierte Benutzeroberfläche, die mit einem Webbrowser zugänglich ist, um die Crawls zu steuern und zu überwachen. Ferner besitzt Heritrix durch verschiedene ersetzbare Module eine hohe Erweiterbarkeit und crawlt in einem adaptiven Tempo, sodass die normalen Webseiten-Aktivitäten nicht gestört werden. Analog zu Suchmaschinen respektiert die Software in der Standardkonfiguration die Datei *robots.txt* sowie sog. *NoFollow Tags* in hohem Maße. Es erfasst somit entsprechend charakterisierte Bereiche und Inhalte einer Webseite nicht und lässt deren Inhalte unberücksichtigt.

Installation und Nutzerfreundlichkeit

Heritrix ist eine Linux-Software, d. h., eine Installation auf einem Windows-System wird offiziell nicht unterstützt. Trotzdem ist eine Installation auf Windows möglich. Die Installation erfordert dabei allerdings ein gewisses technisches Know-how. Es gibt keine benutzerfreundliche und selbsterklärende Installationsoberfläche, daher werden u. a. Kenntnisse im Umgang mit Kommandozeilentools (z. B. Windows Eingabeaufforderung) benötigt. Gerade für IT-Neulinge kann dies zur Herausforderung werden. Positiv hervorzuheben ist jedoch, dass es ein sehr informatives englisches Heritrix-Wiki[12] gibt, das umfangreiche Hilfestellung anbietet. Ferner ist eine detaillierte, deutschsprachige Abhandlung von Marc Malwitz (LWL.IT Service Abteilung, Münster) zur Installation und Anwendung von Heritrix unter Windows 10 vorhanden.[13]

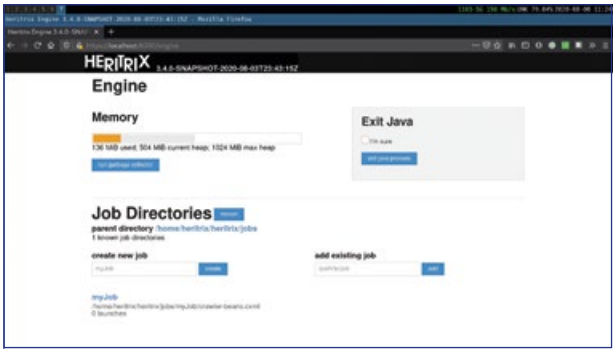

Abb. 1: Administrationsoberfläche von Heritrix. Quelle: Wikipedia (Tomskyhaha, veröffentlicht unter Lizenz CC BY-SA 4.0, https://commons.wikimedia.org/wiki/File:Heritrix_3.4.0_Web_UI.png)

Testergebnisse

Die Praxistests haben gezeigt, dass Heritrix vor allem gut bei der Erfassung und Speicherung von Seiten mit relativ großem statischem Content funktioniert. Probleme treten bei dynamischeren Seiten (Animationen, Mouse-Over, dynamische Menüstrukturen etc.) und auch bei nutzerkommentierten Seiten auf. Auch werden des Öfteren Stylesheets (eine Art Formatvorlage für visuelle Informationen) nicht richtig erfasst, Schriften und Iconsätze (Web-Fonts, Dateityp ttf oder woff) werden durch Standardschriften und Symbole oder Icons durch Kästchen ersetzt. Bilder, pdf-Dateien, Hintergrundgrafiken, eingebettete Video-Streams, Datenbankinhalte etc. fehlen manchmal vollständig.

Vorteile:

- Kostenlose Open-Source-Software
- Speicherung im WARC-Format

Nachteile:

- Wenig benutzerfreundlich, sehr aufwendige Installation und Inbetriebnahme
- Dynamische Seiten werden oft fehlerhaft dargestellt
- Einige Seitenelemente werden bei der Speicherung nicht berücksichtigt

Web Curator Tool

(Johannes Thomé, LVR-Archivberatungs- und Fortbildungszentrum)

Basisinformationen

Das Web Curator Tool (WCT)[14] ist eine Open-Source-Software zur automatischen Erstellung von Abbildern ausgewählter Webseiten. Das Programm wurde von den Nationalbibliotheken Neuseelands und der Niederlande entwickelt. Zur Gewährleistung der vollständigen Funktionalität sollte es in einer Serverumgebung eingerichtet und betrieben werden. Das WCT ist für Linuxsysteme optimiert

10 https://github.com/internetarchive/heritrix3/wiki.
11 Für jede zu sichernde Webseite muss der Benutzer einen sog. Job anlegen.
12 Vgl. https://github.com/internetarchive/heritrix3/wiki.
13 Die Anleitung ist bisher nicht veröffentlicht, kann aber über das DiPS. kommunal-Support-Team des LWL angefordert werden.
14 https://webcuratortool.org/.

und auch gut dokumentiert, wobei nach Aussage der Entwickler auch ein Betrieb unter Windows möglich ist. Es nutzt die Technologie von Heritrix und legt die gecrawlten Webseiten als WARC-Dateien ab.

Installation und Nutzerfreundlichkeit

Auch wenn die Entwicklerinnen und Entwickler für sich in Anspruch nehmen, dass die Benutzung des Programms keine umfassende IT-Expertise erfordert, so ist die erstmalige Installation doch für „normale" Anwenderinnen und Anwender wenig intuitiv. Es steht allerdings zu Demonstrationszwecken eine vorkonfigurierte Installation in einer virtuellen Umgebung zur Verfügung, die ohne größere Umstände verwendet werden kann und für die vorliegende Ausarbeitung getestet wurde. Das umfangreiche Handbuch ist ebenso wie das Programm in englischer Sprache geschrieben. Mit dem WCT können in einem mehrstufigen Verfahren automatische Prozesse anlegt werden. Diese dienen dazu, Webseiten abzurufen und zu speichern. Die Abfragen können dabei präzise konfiguriert werden, sodass entweder alle oder nur ausgewählte Unterseiten abgerufen werden. Für jede Zielseite lassen sich somit automatische Zeitpläne erstellen, nach denen die Seite abgerufen und gespeichert wird.

Die Vielzahl der möglichen Einstellungen erschwert die Bedienung durch unerfahrene Nutzerinnen und Nutzer erheblich. Mit Hilfe des Handbuchs ist es jedoch innerhalb eines angemessenen Zeitraums möglich, die wesentlichen Funktionen des Programms zu erfassen und zu verwenden. Einstiegshürden ergeben sich auch dadurch, dass das WCT mit nicht unmittelbar ersichtlichen Abfragen im Hintergrund prüft, ob eine Eingabe gültig ist. So ist es etwa nur möglich, einen automatischen Harvestingprozess anzustoßen, wenn zuvor eine entsprechende Genehmigung durch den Webseitenbetreiber im Programm hinterlegt wurde. Nach Verständnis der grundlegenden Abläufe im und der Arbeitsweisen mit dem Programm ist die Verwendung jedoch ohne große Probleme möglich.

Testergebnisse

Zu Testzwecken wurden automatische Prozesse für die Webseiten des Flugplatzes Hangelar (https://edkb.de/), des Pizza-Lieferdienstes Speed Pizza (https://www.speed-pizza.com/) sowie der Vereinigten Adelsarchive im Rheinland (https://adelsarchive-rheinland.de/home.html) angelegt und gestartet. Abgeschlossene Harvestingprozesse können in der Nutzeroberfläche überprüft und als Webseite angesehen, mit vorherigen Ergebnissen verglichen, genehmigt und schließlich exportiert werden. Darüber hinaus werden umfangreiche Log-Dateien für jeden Harvestingprozess angelegt, die für diesen Test jedoch nicht durchgesehen wurden.

Bei der Ansicht der gespeicherten Version der Webseite der Vereinigten Adelsarchive kam es zu Problemen mit dem Cookiebanner: Während dieses auf der Originalseite nach erstmaligem Auswählen einer Option nicht mehr angezeigt wird, erscheint es bei der gespeicherten Version beim Aufruf jeder Unterseite erneut. Davon abgesehen ist die gespeicherte Seite voll funktionsfähig. Auch auf der Seite direkt hinterlegte pdf-Dokumente (hier: Publikationen in der Reihe „Rheinische Adelsgeschichte digital") wurden im Testdurchlauf automatisch gespeichert, ohne dass diesbezügliche gesonderte Einstellungen vorgenommen wurden. Es ist jedoch möglich, den Download von solchen Dokumenten durch eine entsprechende Konfiguration zu verhindern.

Auch die Webseite von Speed Pizza wurde mitsamt aller Unterseiten gespeichert und ist auf den ersten Blick nicht von der Originalseite zu unterscheiden. Dynamische Elemente, Hintergrundbilder und farbliche Hervorhebungen werden korrekt wiedergegeben. Zu Problemen kommt es jedoch in Fällen, in denen sich beim Klicken auf Buttons ein Pop-up-Fenster öffnet – in der vom WCT gespeicherten Version werden diese Fenster aus ungeklärten Gründen nicht geöffnet. Ob sich dies durch eine Anpassung der Einstellungen korrigieren lässt, ließ sich im Zuge der Tests nicht herausfinden.

Grundsätzlich gilt für alle drei überprüften Ergebnisse, dass das WCT gute Ergebnisse liefert, die die Webseiten originalgetreu wiedergeben. Lediglich dann, wenn auf den Seiten Elemente von anderen Webseiten eingebunden sind, die nicht auf dem Server des Seitenbetreibers hinterlegt sind, werden diese (verständlicherweise) nicht gespeichert und können auch nicht wiedergegeben werden. In diesen Fällen erscheinen Fehlermeldungen und Links führen ins Leere – so etwa auf der Seite des Flugplatzes Hangelar. Davon abgesehen wurden aber auch hier alle Unterseiten und hinterlegte Dokumente zuverlässig gespeichert.

Das Problem der wiederholt auftauchenden Cookiebanner bestand nur auf der Seite der Vereinigten Adelsarchive im Rheinland – ob die Ursachen in der auf dieser Seite verwendeten Technik oder an unpassenden Einstellungen im WCT lagen, ließ sich anhand der Testseiten nicht überprüfen, da nur die Seite der Adelsarchive überhaupt ein solches Banner anzeigte. Pop-up-Fenster, die nicht hätten funktionieren können, gab es nur auf der Seite des Pizzalieferdienstes.

Nach der Qualitätsprüfung können die Ergebnisse des Harvestings aus dem WCT exportiert und auf einen anderen Server eingespeist werden. Diese Funktion ließ sich in der Demonstrationsumgebung allerdings nicht testen; auf welche Weise und mit wie viel Aufwand sich die gespeicherten Webseiten aus dem Programm heraus in ein digitales Langzeitarchiv überführen lassen, kann daher nicht beurteilt werden. Die Verwendung durch verschiedene Nationalbibliotheken lässt jedoch vermuten, dass das Exportieren der Daten ähnlich reibungslos verläuft wie das Speichern von Kopien der Seiten.

Vorteile:

- Kostenloses Tool, basierend auf Heritrix
- Viele Einstellungs- und Automatisierungsmöglichkeiten

- Umfangreiche Dokumentation und Anleitung
- Langzeitstabile Speicherung durch Export der Daten im WARC-Format in ein digitales Langzeitarchiv
- Gute Wiedergabequalität

Nachteile:

- Höherer Einrichtungs- und ggf. Einarbeitungsaufwand
- Einige Pop-up-Fenster werden nicht zuverlässig dargestellt

MirrorWeb

(Andreas Jüngling, Stadtarchiv Meckenheim)

Basisinformationen

Die Webarchivierungsanwendung MirrorWeb ist eine browserbasierte Softwarelösung der britischen Firma MirrorWeb aus Manchester, deren Produkt in Deutschland von der Firma Walter Nagel GmbH vertrieben wird.[15] Der technische Support erfolgt durch die britischen Entwickler. Die Sprache der Anwendung ist folglich Englisch. Die Anwendung basiert – wie auch das WCT – auf Heritrix. Die Archivierung selbst erfolgt durch die britische Firma, mit der vertraglich die Häufigkeit (temporäre Rhythmen) und Umfänge sowie die jeweils anfallenden Kosten geregelt werden müssen. Derzeit ist von Kosten von rund 500 bis 1.000 Euro pro Speicherereignis auszugehen, abhängig von der Größe der Seiten bzw. Datenmenge. Subdomains müssen separat aufgelistet, vorab auf Archivwürdigkeit geprüft und gegebenenfalls als Teil des Speicherungsumfangs und -rhythmus vereinbart werden. Laut Angaben der Firma Walter Nagel kann die Vertragsgestaltung flexibel erfolgen, sodass z. B. Extraspeicherungen bei besonderen Ereignissen vorgesehen werden können. Dies hat laut Auskunft von Walter Nagel allerdings Auswirkungen auf die Preisgestaltung.

Installation und Nutzerfreundlichkeit

Aufwände für eine Installation entfallen bei der von einem Dienstleister zur Verfügung gestellten kostenpflichtigen Lösung. Die Anwendung selbst präsentiert sich auf dem ersten Blick als komfortabel bedienbar. Die drei Funktionsfelder „Dashboard", „Web Archiving" und „Social Archiving" strukturieren die Gesamtübersicht sowie die Verwendungsbereiche von MirrorWeb. Unter „Dashboard" wird eine Übersicht aller gespeicherten Webseiten nach URL, Speicherzeitpunkt und Zugangslink zum gespeicherten Inhalt in Tabellenform sowie zusätzlich die Gesamtdatenmenge angeboten. „Web Archiving" ist untergliedert in „Web Archives" und „Archive Export". Bei der Planung von Archivierungsvorgängen ist allerdings zu beachten, dass die Zeitangaben je Speicherungsereignis nach Greenwich-Zeit erfolgen, d. h., es existiert eine Zeitverschiebung um eine Stunde.

MirrorWeb speichert die Webseiten im WARC-Format ab. Zur Identifikation und Authentifikation der WARC-Dateien werden beim Crawling nicht veränderbare Hash-Werte berechnet und gespeichert. Die Dateien werden mehrfach redundant in einer Cloud (aber auf deutschen Servern) vorgehalten. Damit soll auch sichergestellt werden, dass die WARC-Dateien im Falle von künftigen Änderungen des Standards in andere Formate migriert werden können. Es besteht keine Möglichkeit, die WARC-Dateien zu löschen oder zu bearbeiten. Grundsätzlich kann vertraglich die Übergabe der WARC-Dateien an das Archiv geregelt werden. Da die Nutzung der Dateien über MirrorWeb an die firmeneigene Browser-Plattform gebunden ist, empfiehlt sich die Datenhaltung beim Anbieter. Spätestens mit Vertragsende sollen die WARC-Dateien laut MirrorWeb bzw. der Firma Walter Nagel ohne zusätzliche Kosten dem Archiv zur Verfügung gestellt werden. Zur weiteren Nutzung der WARC-Dateien muss dann allerdings ein passendes Tool zum Lesen bzw. Wiedergeben von WARC-Dateien im Archiv vorhanden sein.[16]

Testergebnisse

Im Benutzerprofil werden unter dem Punkt „Web Archives" sämtliche Crawls nach dem Namen bzw. der URL gesammelt, die sich in einer nach der Crawl-ID (Hash-Wert) gegliederten Darstellung anzeigen lassen. Hier können auch weitere Metadaten eingesehen werden, wie z. B. Anfang, Ende und Dauer eines Speichervorgangs, eine graphische Analyse nach dem Mime-Type sowie die Log-Dateien mit den Metadaten des Crawlings als log-, txt- und gz-Dateien nebst einer umfangreichen Seitenanalyse. „Archive Export" bietet ferner eine tabellarische Aufstellung der aus den gespeicherten Webseiten generierten Exporte in Form von pdf- oder PNG-Dateien, die sich herunterladen lassen. Ein integrierter pdf-Reader erlaubt das Lesen, Drucken und Speichern der Datei. Webseiten werden entsprechend als einzelne pdf-Seiten in einer Datei angeboten oder als png-Bild der gesamten Seite. Das pdf-Format erweist sich als barrierefrei navigier- und nutzbar und ist für Sehbehinderte mit einem einschlägigen Leseprogramm geeignet.

Grundsätzlich verspricht MirrorWeb, alle Seiten und Unterseiten je Crawl inklusive integrierter pdf-Dateien zu speichern. Solche pdf-Dateien lassen sich öffnen, lesen und drucken. Ausgenommen sind externe Links sowie eingebettete Formate wie Videos von externen Plattformen (z. B. Youtube) und Programmen. Verknüpfungen zwischen den Unterseiten werden grundsätzlich funktionsfähig mit übernommen. Im Test ergaben auch dynamische Seiten mit automatischen Bildläufen keine Wiedergabe- oder Seitenarchitekturprobleme. Einzige Ausnahme waren Google-Karten, die nicht übernommen werden. Statische Seiten konnten vollkommen unproblematisch und fehlerfrei gespeichert werden.

Trotzdem zeigen sich in der praktischen Anwendung zum Teil gravierende Probleme. Zum einen ließen sich die gespeicherten Seiten auch nach wiederholten Versuchen nicht über das sogenannte „Replay" im sich gesondert öff-

15 https://www.walternagel.de/webarchivierung.

16 Z. B. der WARC-Player des Internet Archive: https://archive.org/details/WARCPlayer.

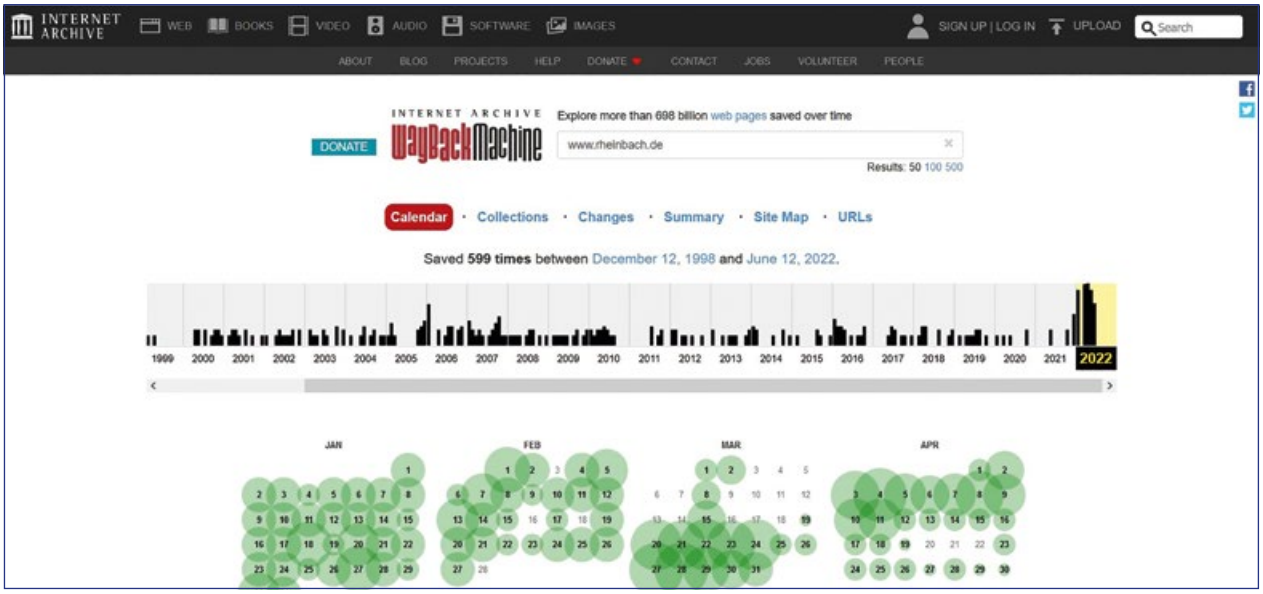

Abb. 2: Übersicht über die vorhandenen Versionen der Webseite der Stadt Rheinbach seit 1998

nenden Ansichtsfenster anzeigen. Die einzige Lösung des Problems war in solchen Fällen die Ab- und Neuanmeldung in der Anwendung. Die vorhandene Funktion „Disable Enhanced Replay" im Sichtfenster generiert hingegen bestenfalls eine Sitemap der jeweiligen Seite. Die Möglichkeit, innerhalb der bzw. zwischen den verschiedenen Zeitstufen der Speicherung zu navigieren, gelang in keinem Fall. Auch die „Zurück"-Funktion im Browser über den Cache konnte vorherige Seitenaufrufe nicht ermöglichen.

Die Qualität der Wiedergabe variierte von Aufruf zu Aufruf. Die Spannweite reichte von einem komplett problemfreien Navigieren auf allen Seiten bzw. Unterseiten bis zum teilweisen oder völligen Bruch der Seitenarchitektur. Möglicherweise wurden in solchen Fällen die CSS-Dateien (Layoutinformationen) nicht oder nur unvollständig aufgerufen. Auch der Wechsel auf andere Sprachseiten innerhalb der Testspeicherungen führte zum gleichen Problem. Gleichfalls unpraktisch ist die dysfunktionale Suchfunktion der gespeicherten Seiten. Lediglich integrierte Suchmaschinen wie Google funktionierten weiterhin. Für die Nutzung komplexer und informationsreicher Webseiten kann dies für den heutigen Anwender nur als unpraktisch bewertet werden. Weitere in der Speicherseite eingebettete Anwendungen wie die Druckfunktion sind nicht bedienbar – hier bietet sich aber der Ausweg über die Generierung von pdf- oder png-Dateien an. Weiterhin wurden im Kontext der mutmaßlichen CSS-Probleme Schriftgrößen und -formen nicht originalgetreu wiedergegeben.

Vorteile:

- Keine Installation, da basierend auf Webbrowser
- Bequemer Arbeitsablauf, da die Speicherungen durch den Dienstleister bzw. Programmanbieter erfolgen
- Kein eigener Speicherplatzbedarf
- Gute Wiedergabequalität bei einfachen, statischen Seiten
- Speicherung im WARC-Format

Nachteile:

- Abhängig von der Speichermenge regelmäßige (hohe) Kosten
- Je komplexer und dynamischer die Webseite, desto geringer die Qualität und Wiedergabeverlässlichkeit
- Keine individuellen Einstellungen möglich

Internet Archive: Wayback Machine

(Dietmar Pertz, Stadtarchiv Rheinbach)

Basisinformationen

Die Wayback Machine ist ein Onlinedienst zur Langzeitarchivierung von Webseiten. Er ist Teil des Internet Archive, einer US-amerikanischen Non-Profit-Organisation. Das Internet Archive wurde im Mai 1996 gestartet und hat seit 2007 den offiziellen Status einer Bibliothek. Über die Webseite http://web.archive.org/ lassen sich unterschiedliche Versionen der vom Internet Archive auf Servern in Kalifornien und Kanada archivierten Webseiten abrufen.

Installation und Nutzerfreundlichkeit

Eine Installation des Tools ist weder möglich noch vorgesehen. Nach einer Registrierung/Anmeldung als „Member" auf der Seite des Internet Archive kann man unter https://archive.org/web/ unter dem Punkt „Save Page now" zu archivierende Seiten manuell auswählen, sichern lassen und zusätzlich in seinem Benutzerprofil in einem Ordner „My web archives" ablegen. Letzteres erhöht die Übersichtlichkeit bei der Verwaltung heruntergeladener Seiten.

Bei der Speicherung kann man angeben, ob dazugehörige Unterseiten ebenfalls gespeichert werden sollen. Die gespeicherten Seiten sind nur über einen Link abrufbar. Eine herunterladbare WARC-Datei kann nicht erstellt werden. Somit ist eine Speicherung der WARC-Dateien auf dem eigenen Rechner/Server oder gar die Langzeitarchivierung in einem eigenen digitalen Magazin nicht möglich. Nur die Metadaten zu den Webseiten, die durch die Wayback Ma-

chine gespeichert wurden, können als JSON- oder txt-Datei heruntergeladen werden.

Testergebnisse

Getestet wurde anhand der Seiten des Pizzadienstes Speed Pizza, des Flughafens Hangelar sowie der Stadt Rheinbach (www.rheinbach.de).

Die Seite der Stadtverwaltung Rheinbach wurde am 12. Dezember 1998 erstmals gespeichert. Seitdem hat das Internet Archive die Startseite nahezu sechshundertmal (zumeist automatisch) gesichert. Die Unterseiten wurden aber nicht immer mitgespeichert, sodass, wenn man die entsprechenden Links anklickt, u. U. ältere oder jüngere Versionen der Unterseite dargestellt werden. Das jeweilige Datum des Snapshots wird aber in der Kopfzeile übersichtlich wiedergegeben. Manche Unterseiten wurden in den zurückliegenden 24 Jahren überhaupt nicht gespeichert.

Bei der Speicherung der einfach aufgebauten Seite des Pizzadienstes Speed Pizza gab es keine Probleme. Die Sicherung der aktuellen Webseite der Stadt Rheinbach funktionierte ebenfalls ohne Schwierigkeiten. Die Sicherung der Webseite der Stadt Rheinbach mit allen Unterseiten dauerte aber sehr lange. Die Abbildungsqualität ist gut. Gleiches gilt für die Speicherung der Seite des Flugplatzes Hangelar.

Vorteile:

- Kostenlos
- Einfache Handhabung
- Keine Installation notwendig
- Gute Ergebnisse

Nachteile:

- Keine eigene, lokale Speicherung in einem langzeitarchivfähigen Format möglich
- Datenhoheit des Archivs über die Crawls nicht gegeben

Conifer

(Michael Cöln, Stadtarchiv Hürth)

Basisinformationen

Conifer[17] ist das Ergebnis eines mehrjährigen Forschungs- und Entwicklungsprojekts zur Schaffung eines Webarchivierungsdienstes, der von 2015 bis 2020 unter dem Namen *Webrecorder.io* bekannt und beim Internetprojekt *Rhizome* gehostet wurde. Die in dieser Zeit entstandenen Open-Source-Komponenten bilden nun die Grundlage von Conifer, das weiterhin von Rhizome verwaltet, entwickelt und gehostet wird. Rhizome ist eine gemeinnützige Organisation, die 1996 auf Initiative des Künstlers Mark Tribe gegründet wurde und sich in erster Linie der Präsentation und Bewahrung digitaler Kunst und Kultur im Internet verschrieben hat. Dazu zählt unter anderem auch die Archivierung komplexer Webseiten. Viele Programme und Angebote im Bereich der Webseitenarchivierung funktionieren nach dem Prinzip des Harvestings, auch *crawlen* genannt. Das bedeutet, dass das Programm eine vorher eingegebene Domain ‚erntet', also je nach Einstellung automatisch Kopien von einzelnen Seiten und Unterseiten der Domain sowie ggf. implementierter Unterlagen erstellt. Im Gegensatz dazu ist Conifer eine benutzergesteuerte Plattform, die nur ausschließlich die Seiten archiviert, die die Anwenderinnen und Anwender aktiv ansteuern.

Installation und Nutzerfreundlichkeit

Um den Dienst nutzen zu können, ist eine kostenfreie Anmeldung auf der Seite von Conifer nötig. Da das Programm nicht auf einem Rechner installiert werden muss, fallen mögliche Kosten für Installation und Hosting weg, was positiv zu werten ist. Nach Freischaltung per Bestätigungsemail können die ersten Schritte mit Conifer unternommen werden.

Nach der Anmeldung befindet man sich in seiner persönlichen „Umgebung". Für jede zu archivierende Webseite muss hier eine „Collection" (Sammlung) angelegt werden. Um die Inhalte einer Seite zu archivieren, wird die URL im Schlitz „New capture" (Neue Erfassung) eingefügt. Dann wird die Sammlung ausgewählt, in der die Seiten abgelegt werden sollen. Schließlich muss noch der Browser, mit dem die zu archivierenden Seiten angesteuert werden, ausgewählt werden.

Testergebnisse

Im Zuge des Praxistests hat sich ergeben, dass mit dem Firefox-Browser alle drei Testseiten – Flugplatz Hangelar, Stadt Sankt Augustin, Pizza-Service Speed Pizza – problemlos archiviert werden konnten. Die hochdynamische Seite des Flugplatzes Hangelar konnte mit den Browsern Microsoft Edge und Google Chrome nicht fehlerfrei erfasst werden. Nachdem die Einstellungen vorgenommen wurden, wird mittels „Start Capture"-Button die „Session" (Sitzung) gestartet. Nun wird die Seite in einer Art Browser in der persönlichen Umgebung angezeigt. Conifer zeichnet die Seiten auf, die auf der Homepage vom Anwender angeklickt werden. Technisch erfasst Conifer die Interaktion zwischen dem Browser (Anfragen) und dem Server (Antworten), auf dem die Webseite gehostet wird. Das bedeutet, dass Conifer nur die Seiten und Unterseiten erfasst, die tatsächlich aktiv vom Nutzer angesteuert werden. Mittels des Buttons „Autopilot" ist es möglich, dass die aktuell angesteuerte Seite bzw. Unterseite von Conifer mittels automatischen Scrollens erfasst wird. Dies ist aber die einzige zeitliche Ersparnis, die geboten wird. Die Entwickler von Conifer scheinen sich des zeitlichen Aufwands bewusst zu sein, denn die Erfassung der Webseite kann jederzeit unterbrochen und zu einem späteren Zeitpunkt in einer neuen „Session", die der „Collection" zugeordnet wird, wiederaufgenommen werden.

Vor allem komplexe, dynamische Seiten wie die des Flughafens Hangelar konnten mit Hilfe von Conifer vollständig erfasst werden. Die Nutzung von Conifer ist sehr

17 https://conifer.rhizome.org/.

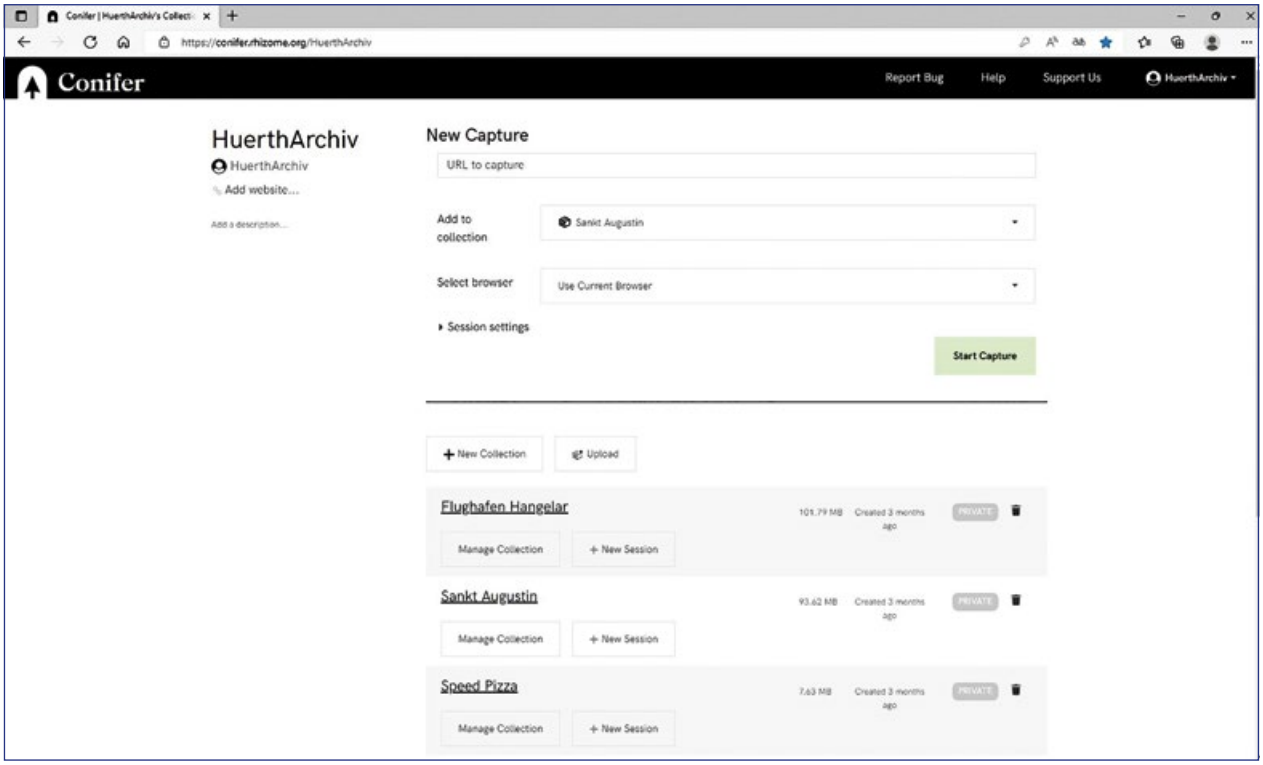

Abb. 3: Benutzeroberfläche von Conifer

intuitiv, sodass man keine Expertin oder Experte sein muss, um Erfolge zu erzielen.

Insgesamt stehen jedem Benutzerkonto 5 GB Speicherplatz zur Verfügung. Die moderne, dynamische Seite des Flughafens Hangelar hat einen Umfang von 101,79 MB, die vergleichsweise statische, einfach gehaltene Seite des Pizzadienstes hat mit 7,63 MB deutlich geringeren Speicherbedarf. Mit dem Speicherplatz von 5 GB lassen sich also problemlos eine ganze Reihe komplexer, hochdynamischer Seiten speichern. Unabhängig davon bietet Conifer auch die Möglichkeit, die einzelnen Collections als WARC-Datei herunterzuladen.

Vorteile:

- Kostenfreies Tool, das nicht installiert werden muss
- Einfacher Einstieg, kein großes Vorwissen nötig
- Erfasst hochdynamische und komplexe Webseiten
- Exportfunktion nach WARC

Nachteile:

- Sehr hoher zeitlicher Aufwand aufgrund der fehlenden automatischen Crawling-Funktion

PABLO

(Michael Korn, Stadtarchiv Sankt Augustin; Tobias Schröter, Interkommunales Archiv Lohmar)

Basisinformationen

Beim kostenpflichtigen Java-Tool *PABLO* der Firma startext GmbH[18] handelt es sich um eine Kombination aus Webcrawler und Archivierungswerkzeug. Das Programm crawlt die eingegebene URL und erstellt für jede Unterseite zwei Dateien in langzeitarchivfähigen Formaten: Zum einen fotografiert PABLO die Seitenoberfläche ab, womit das Aussehen der Seite bewahrt wird. Dieser Screenshot wird als Bilddatei in einem auswählbaren Format gespeichert. Zum anderen erzeugt PABLO eine XML-Datei, in der die als Text definierten Inhalte, die Metadaten, der Seitenaufbau sowie die Links zu anderen (Unter-)Seiten beschrieben werden. Dadurch sind die archivierten Seiten im Volltext durchsuchbar. Die ursprünglich vorhandenen Links werden über Pixelkoordinaten[19] auf den Screenshots positioniert. Die Strukturierung der gesamten Webseite wird ebenfalls in XML wiedergegeben.

Das Programm erstellt optional neben dieser Archivierungsfassung eine Präsentationsfassung, mit der die archivierte Seite über einen Browser abgerufen werden kann.

Durch die Kombination der Inhalte der XML-Datei mit den Screenshots ergibt sich die Möglichkeit, das „Look and Feel" der archivierten Seite in einer Weise zu bewahren, die es späteren Nutzerinnen und Nutzern erlaubt, sich durch die Seite durchzuklicken.

Die Firma startext bietet zu PABLO inzwischen zwei verschiedene Geschäftsmodelle an. Einerseits ist es möglich, die einmalige Archivierung einer Webseite komplett als Dienstleistung einzukaufen, d. h., startext übernimmt die vollständige Konfiguration sowie den Archivierungs-

18 https://www.startext.de/produkte/pablo.

19 Jedes digitale Bild basiert auf einer Menge an Bildpunkten (Pixel), die jeweils auf einer senkrechten und einer waagerechten Achse exakt lokalisiert werden können. Die üblicherweise als Bildauflösung verstandenen Werte (z. B. 1920x1080) geben dabei die Anzahl der Bildpunkte auf der waagerechten (1920 Pixel) und senkrechten (1080 Pixel) Achse des Bildes wieder. Anhand der Kombination eines Punktes auf der waagerechten und eines Punktes auf der senkrechten Achse ergibt sich wie in einem Koordinatensystem ein exakt bestimmbarer Pixel im digitalen Bild.

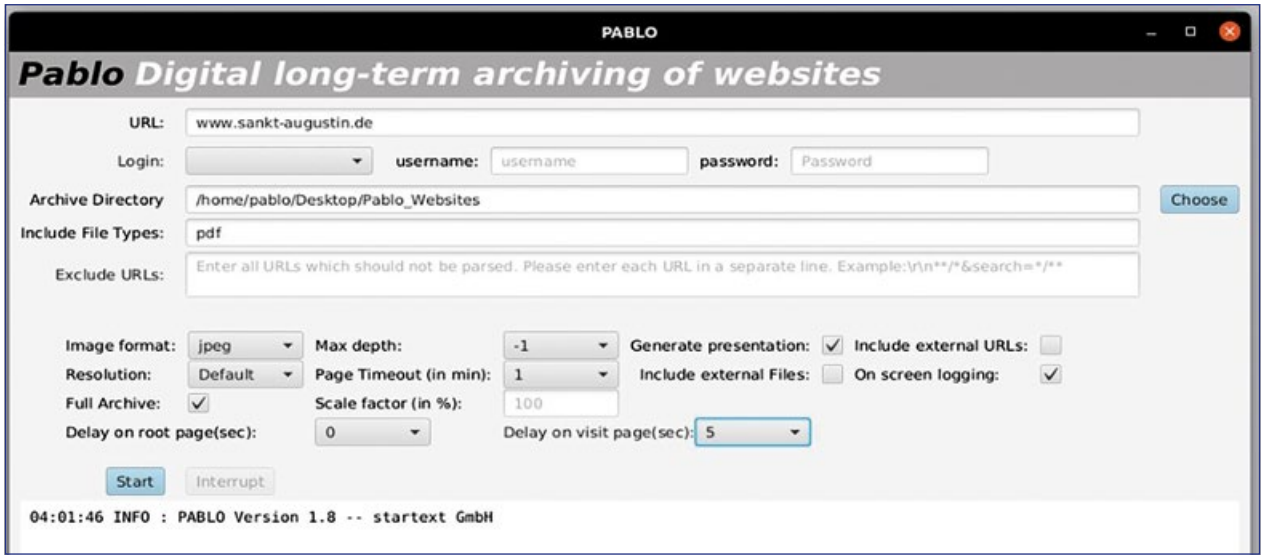

Abb. 4: Benutzeroberfläche von PABLO

vorgang mit PABLO und liefert das Ergebnis an die Kunden aus. In diesem Fall wird jeder Archivierungsvorgang als Dienstleistung bezahlt.

Alternativ ist es nach wie vor möglich, eine PABLO-Lizenz zu erwerben. In diesem Fall wird PABLO als Anwendung innerhalb einer virtuellen Maschine[20] für die Archivierung einer bestimmten Webseite vorkonfiguriert und ausgeliefert. Dabei wird die Lizenz sowie für jede zu archivierende Webseite der einmalige Konfigurationsvorgang bezahlt. Hinzu kommen die üblichen jährlichen Wartungs- und Pflegekosten. Dafür ist es jedoch möglich, jede vorkonfigurierte Seite so oft wie gewünscht zu archivieren, ohne dass jeder Archivierungsvorgang bezahlt werden müsste.

Installation und Nutzerfreundlichkeit

Da das Programm als Software entwickelt wurde, die durch Archive eigenständig eingesetzt werden kann, ist das Programm sehr einfach in der Handhabung und läuft auf fast allen Rechnern. Eine Installation ist nicht notwendig. Für die Nutzung stehen wenige, aber ausreichende Einstellungsparameter über eine grafische Benutzeroberfläche zur Verfügung. Neben der zu archivierenden Internetadresse/URL können z. B. der Speicherort für das Ergebnis, die Speichertiefe (d. h., wie viele Ebenen unterhalb der Hauptseite archiviert werden sollen), das Speicherformat, die Auflösung der archivierten Bilddateien, ein etwaiger Einschluss der jeweils ersten externen Seite oder der Ausschluss bestimmter Unterseiten (z. B. bei eingebauten Kalendern relevant) ausgewählt werden. Zudem besteht die Möglichkeit, fast beliebig viele in die Webseite eingebundene Dateien im jeweiligen Ausgangsformat (z. B. pdf-Dateien, Fotos, Filme) zusätzlich herunterzuladen.

Testergebnisse

Nach Eingabe der zu archivierenden Internetadresse/URL und Einstellung der oben angesprochenen Parameter wird ein Job (Archivierungsvorgang) gestartet. Je nach Umfang und Komplexität der Webseite sowie des zur Verfügung

```
    </FContent>
  </file>
</fileGrp>
<fileGrp ID='FILEGRP_2' USE='page_snapshot'>
  <file ID='FID_00000003' MIMETYPE='image/jpeg' USE='page_snapshot'>
    <FLocat LOCTYPE='URL' xlink:href='index.html.jpeg' />
    <FContent>
      <xmlData>
        <wa:metadata xmlns:wa='http://www.startext.de/WebArchiver'>
          <wa:link id='LINK_0000000000'>
            <wa:href>https://www.speed-pizza.com/</wa:href>
            <wa:text></wa:text>
            <wa:width unit='pixel'>450</wa:width>
            <wa:height unit='pixel'>92</wa:height>
            <wa:leftOffset unit='pixel'>281</wa:leftOffset>
            <wa:topOffset unit='pixel'>4</wa:topOffset>
          </wa:link>
          <wa:link id='LINK_0000000001'>
            <wa:href>http://www.pizza-onlineshop.com/</wa:href>
            <wa:text>Pizza Shopsystem</wa:text>
            <wa:width unit='pixel'>90</wa:width>
            <wa:height unit='pixel'>15</wa:height>
            <wa:leftOffset unit='pixel'>1207</wa:leftOffset>
            <wa:topOffset unit='pixel'>1</wa:topOffset>
          </wa:link>
          <wa:link id='LINK_0000000002'>
            <wa:href>https://www.speed-pizza.com/</wa:href>
            <wa:text>HOME</wa:text>
            <wa:width unit='pixel'>56</wa:width>
            <wa:height unit='pixel'>21</wa:height>
            <wa:leftOffset unit='pixel'>281</wa:leftOffset>
            <wa:topOffset unit='pixel'>120</wa:topOffset>
          </wa:link>
          <wa:link id='LINK_0000000003'>
            <wa:href>https://www.speed-pizza.com/285/Pizza+bestellen/</wa:href>
            <wa:text>PIZZA</wa:text>
            <wa:width unit='pixel'>53</wa:width>
            <wa:height unit='pixel'>21</wa:height>
            <wa:leftOffset unit='pixel'>341</wa:leftOffset>
            <wa:topOffset unit='pixel'>120</wa:topOffset>
```

Abb. 5: Abbildung der Seitenstruktur mit Pixelkoordinaten in XML

stehenden Arbeitsspeichers kann ein Job zwischen einigen Sekunden und mehreren Tagen dauern. PABLO sollte nach Möglichkeit außerhalb von behördlichen Netzwerken eingesetzt werden, um zu vermeiden, dass (Unter-)Seiten durch strenge Firewalleinstellungen geblockt werden. Bisher ist zudem keine Stapelverarbeitung mehrerer Internetadressen/URLs möglich, d. h., dass für jede zu archivierende URL ein eigener Archivierungsvorgang manuell angestoßen werden muss.

PABLO ist v. a. für die Sicherung des statischen Webs 1.0 geeignet. Interaktive Elemente – v. a. Pop-ups usw. – kön-

20 Eine virtuelle Maschine emuliert eine ganz konkrete Hardware- und Softwareumgebung (in diesem Fall einen Linux-Rechner). Auf diese Weise kann PABLO in einer ‚kontrollierten Umgebung' betrieben werden, die Anpassung des Programms an unzählige mögliche Kombinationen unterschiedlicher Betriebssysteme und Browserversionen entfällt dadurch.

nen hingegen aufgrund des Screenshot-Prinzips kaum gesichert werden. Manche Webseiten werden gar nicht oder nur zu geringen Teilen gesichert; bei hoher Komplexität einer Seite kommt es nicht selten vor, dass ein Job wegen produzierter Dubletten manuell abgebrochen werden muss.

Auch wenn PABLO eigentlich recht einfach zu bedienen ist, konnte die Nutzung bisher je nach Webseite sehr zeitintensiv und mühselig sein, da es bei länger laufenden Jobs ratsam ist, hin und wieder das Zwischenergebnis zu prüfen, um im ungünstigen Fall den Job manuell abbrechen zu können. Aufgrund des neuen Vertriebsmodells ist es wahrscheinlich, dass der notwendige Zeitaufwand hierfür deutlich sinkt.

Insgesamt kann man zu PABLO ein positives Urteil abgeben, wenn man die Einschränkungen aufgrund der Funktionsweise des Tools beachtet und keine Wunderwaffe zur Webarchivierung erwartet. Trotz der Einschränkungen in Hinblick auf die dynamischen Inhalte modernerer Webseiten konnten bei umfangreichen Tests an mehreren hundert Seiten[21] in den letzten Jahren 50 bis 70 Prozent der Webseiten in sehr guter bis passabler Qualität gesichert werden. Aus archivischer Sicht sind vor allem die sehr gute Langzeitarchivierbarkeit und Nutzbarkeit der archivierten Ergebnisse große Vorteile. Im Gegensatz z. B. zur Archivierung mithilfe von WARC-Dateien ist jederzeit ersichtlich, auf welchen Datenformaten die archivierte Webseite beruht, sodass selbst eine Migrationsstrategie für die mit PABLO archivierten Webseiten realistisch erscheint. Zudem kann die Nutzungsversion einfach mit Hilfe eines üblichen Browsers aufgerufen werden und erfordert keine besonderen Vorkenntnisse oder Tools.

Vorteile:

- Speicherung der Inhalte in nur zwei von vorneherein langzeitarchivfähigen Datenformaten (XML und ein wählbares Bildformat)
- Inhalte ohne spezielle Software les- und durchsuchbar
- Einfaches, übersichtliches Tool
- Optional zusätzlicher Download der eingebundenen Ausgangsformate wie Bilder oder Filme möglich

Nachteile:

- Kostenpflichtig
- Je interaktiver und komplexer die Seite, desto fehleranfälliger und lückenhafter wird das Ergebnis
- Zeitlicher Aufwand je Job bei umfangreicheren Webseiten im Vorfeld (bisher) kaum kalkulierbar
- Ein kleinerer Teil an Webseiten wird gar nicht bearbeitet
- Keine Möglichkeit zur Stapelverarbeitung mehrerer Webseiten

HTTrack

(Jens Löffler, Stadtarchiv Bornheim)

Basisinformationen

Der Webcrawler HTTrack[22] bietet die Möglichkeit, lokale Kopien von Webseiten zu erstellen. Layout und Navigation der Seiten bleiben dabei erhalten. Das Programm existiert bereits seit 1998. Die aktuelle Version 3.49–2 wurde im Mai 2017 veröffentlicht. Es handelt sich um Freeware unter allgemeiner Veröffentlichungsgenehmigung (*GNU General Public License*). Somit ist das Programm kostenlos nutzbar. Es gibt eine Windows- und eine Linux-kompatible Programmversion sowie eine Version, die unter Android läuft.

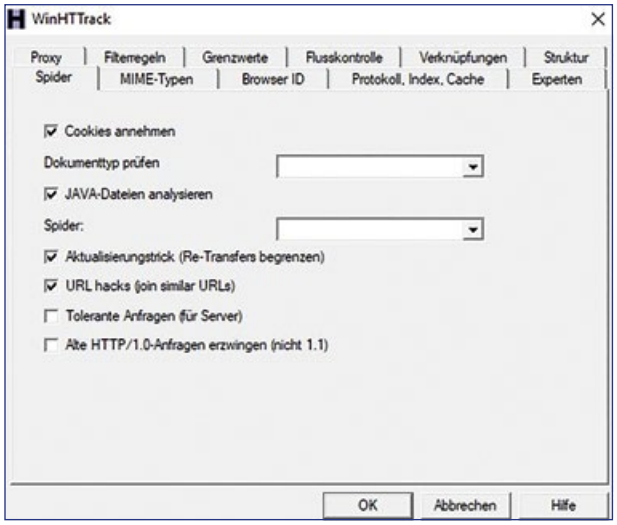

Abb. 6: Benutzeroberfläche von HTTrack

Installation und Nutzerfreundlichkeit

Die Installation kann unter Windows über einen Installer erfolgen. Das Programm kann aber auch ohne Installer über das Entpacken einer ZIP-Datei installiert werden. Nach Eingabe der zu spiegelnden URL lassen sich die Einstellungen für den Crawl über eine Benutzeroberfläche in mehreren Registerkarten festlegen. Einzelne Dateitypen können gezielt aus- oder eingeschlossen werden. Die Linktiefe des Crawls kann ebenso angepasst werden wie die Möglichkeit, externe Links zu verfolgen. Zusätzlich gibt es eine Reihe weiterer Einstellungen wie z. B. zur Anpassung der Browser-ID, zur Speicherstruktur der Kopie, zu Cookies, JAVA und vielem mehr. Im Gegensatz zum Heritrix-Crawler besteht bei HTTrack auch die Möglichkeit, in den Einstellungen festzulegen, ob robots.txt-Regeln ignoriert oder beachtet werden sollen. Die Vielzahl der möglichen Einstellungen kann mitunter verwirren, in der Nutzercommunity lassen sich dabei nicht immer klare Aussagen über die Auswirkungen der getroffenen Einstellungen auf den Crawl finden. Das macht eine Fehlersuche in den Einstellungen, insbesondere für ungeübte Nutzer, mitunter sehr mühsam.

21 Das Stadtarchiv Sankt Augustin setzt PABLO bereits seit 2016 selbst ein.
22 https://www.httrack.com/.

Technisch gesehen bildet HTTrack die Webseiten als lokale Kopie in Struktur und Dateiformat nach. Im Ergebnis bekommt man also eine über html-Dateien verknüpfte Dateisammlung mit einer entsprechenden Vielfalt an Dateiformaten. Eine direkte Ausgabe in einem etablierten Archivierungsformat wie WARC ist bei HTTrack nicht möglich, soll aber mit einem Zusatztool (HTTrack2warc)[23] erreicht werden können. Die Nutzung dieses Tools erfordert allerdings noch etwas detaillierteres technisches Wissen.

Betrachtet werden können die Crawls mit jedem gängigen Browser.

Testergebnisse

Im Praxistest lieferte HTTrack durchaus respektable Ergebnisse. Die Internetseite der Stadt Sankt Augustin wurde problemlos gespiegelt. Bilder wurden korrekt angezeigt und auch dynamische Elemente wie Slider und farbliche Hinterlegungen der angewählten Menüpunkte korrekt abgebildet. Entsprechend mussten im „Look and Feel" der gespiegelten Seite keine Abstriche gemacht werden. Auch die auf der Seite bereitgestellten pdf-Dateien konnten wieder ganz normal geöffnet werden.

Bei der Spiegelung der Internetseite des Flughafens Hangelar hatte HTTrack lediglich Probleme mit einem großformatigen, responsiven Slider auf der Startseite. Möglicherweise hängt dies mit dem Alter der aktuellen Programmversion (2017) zusammen, die die neueren Entwicklungen der Web-Technologie seit 2017 u. U. (noch) nicht abbilden kann. Alle anderen Unterseiten wurden jedoch korrekt dargestellt. Weitere dynamische Teile der Seite, wie z. B. Ziehharmonika-Elemente oder dynamisch durchlaufender Text, stellten kein Problem dar.

Nur die Seite des Lieferdienstes Speed Pizza konnte in Ihrer Gesamtheit nicht erfolgreich gecrawlt werden. Hier scheint es beim Crawl zu einem Umleitungsfehler gekommen zu sein. Aus ungeklärten Gründen ergänzte HTTrack beim Aufruf von Links zu Unterseiten stets „/index.html" in den Adressen der Unterseiten, weshalb nur die Startseite erfolgreich gespiegelt werden konnte. In der Vielzahl der möglichen Einstellungsmöglichkeiten ließ sich – trotz mehrerer Versuche – keine Lösungsmöglichkeit für dieses Problem finden.

Vorteile:

- Gute Ergebnisse
- Einfache Installation
- Kostenlos

Nachteile:

- Keine Speicherung in einem langzeitarchivfähigen Format
- Vielzahl der Einstellungsmöglichkeiten schränkt Nutzerfreundlichkeit ein und kann zu zeitaufwändiger (und u. U. ergebnisloser) Fehlersuche führen
- Letzte Programmversion aus 2017

Zusammenfassung

Von den sieben getesteten Tools sind fünf kostenfrei. Zwei Tools verursachen Lizenz- bzw. Servicekosten. Abgesehen davon, dass für die Speicherung der Daten auch im Falle der Freeware-Programme Kosten anfallen werden, ist insgesamt festzustellen, dass die Anwendung der kostenfreien Lösungen aufgrund ihrer geringeren Nutzerfreundlichkeit in der Regel komplizierter ist und damit auch zeitaufwändiger ausfällt. Die kostenpflichtigen Lösungen PABLO und MirrorWeb bieten „Webarchivierung as a Service" an, was wiederum die personellen Ressourcen schont. Auch bei diesen Lösungen muss allerdings ausreichend Zeit für die Qualitätssicherung eingeplant werden.

In Hinblick auf die Langzeitarchivierbarkeit ist zwischen drei technischen Verfahren zu unterscheiden. PABLO beruht auf der Kombination von Screenshots und einer in XML gespeicherten Seitenstruktur, während die meisten anderen Tools die Crawls in WARC-Dateien ablegen – mit den in der Einführung erwähnten diesbezüglichen Einschränkungen. Keine integrierte Speicherungsmöglichkeit in einem langzeitarchivfähigen Format bietet hingegen HTTrack.

Eine Sonderstellung nimmt die Wayback-Machine ein. Diese liefert zwar gute Ergebnisse und beruht ebenfalls auf WARC. Sie kommt für eine OAIS-konforme Webarchivierung jedoch nicht in Frage, da die Datenhoheit des Archivs über die Crawls nicht gegeben ist und das Archiv keine Möglichkeit hat, die Archivierungsergebnisse technisch zu validieren, zu schützen und im Zweifelsfall anderweitig zu sichern.

Letztlich bleiben vor allem zwei Dinge festzuhalten: Es gibt bereits jetzt eine Vielzahl an funktionalen Tools für die Sicherung und Archivierung von Webseiten. Die „eierlegende Wollmilchsau", die alles kann, dabei möglichst einfach zu bedienen und (quasi) kostenlos ist, existiert allerdings nicht (und wird mutmaßlich auch nie existieren).

Jedes Archiv muss also anhand der eigenen personellen Kapazitäten, seiner technischen Fähigkeiten sowie anhand des eigenen Budgets entscheiden, welche Lösungen zum Einsatz kommen können. Möglicherweise wird sich dabei herauskristallisieren, dass die Kombination mehrerer Tools zum Erfolg führt. Ein Mangel an Möglichkeiten ist jedenfalls kein Argument mehr gegen einen schnellen Einstieg in die Webarchivierung. ■

Arbeitskreis digitale Langzeitarchivierung (dLZA) Rhein-Sieg

Michael Cöln
Stadtarchiv Hürth
mcoeln@huerth.de

Johannes Ehrengruber
Stadtarchiv Troisdorf
ehrengruberj@troisdorf.de

Andreas Jüngling
Stadtarchiv Meckenheim
andreas.juengling@meckenheim.de

23 https://github.com/nla/httrack2warc.

Michael Korn
Stadtarchiv Sankt Augustin
michael.korn@sankt-augustin.de

Jens Löffler
Stadtarchiv Bornheim
jens.loeffler@stadt-bornheim.de

Dr. Gregor Patt
LVR-Archivberatungs- und Fortbildungszentrum, Pulheim
gregor.patt@lvr.de

Dietmar Pertz M. A.
Stadtarchiv Rheinbach
archiv@stadt-rheinbach.de

Tobias Schröter
Interkommunales Archiv Lohmar
tobias.schroeter@lohmar.de

Johannes Thomé B. A.
LVR-Archivberatungs- und Fortbildungszentrum, Pulheim
johannes.thome@lvr.de