Arbeitsbericht zur Archivierung von Netzressourcen im Staatsarchiv des Kantons Basel-Stadt

Kerstin Brunner, Olivier Debenath

Der vorliegende Bericht dokumentiert die Vorgehensweise zur und die Erkenntnisse aus der aktuellen Praxis der Archivierung von Netzressourcen im Staatsarchiv des Kantons Basel-Stadt. Der Text ist in zwei Teile gegliedert: Der erste Teil beschreibt die archivischen Überlegungen und Entscheidungen hinter dem gewählten Vorgehen. Der zweite Teil des Arbeitsberichts befasst sich mit der Unternehmensarchitektur der technischen Lösung zur Sicherung von Webseiten.

Archivische Aspekte zur Webarchivierung

Anfänge

Mit dem Thema der Archivierung von Netzressourcen beschäftigt sich das Staatsarchiv Basel-Stadt seit 2006. Erste Sicherungen fanden bereits Ende 2008 statt. Dies geschah im Hinblick auf eine bevorstehende Verwaltungsreorganisation, welche tiefgreifende Umstrukturierungen im Verwaltungsapparat mit sich brachte, was sich wiederum in der grundlegenden Überarbeitung oder gar Abschaltung diverser Webseiten niederschlug.

Aufgrund der geringen verfügbaren Ressourcen für die Archivierung von Webinhalten erfuhr das Thema erst ab Ende 2011 einen neuen Arbeitsschub. Über mehrere Monate hinweg wurden verschiedene Ansätze innerhalb der schweizerischen Bibliotheks- und Archivlandschaft studiert. Die Analyse von Intranet- und Internetseiten des Kantons führte zur Identifikation geschäftsrelevanter Inhalte. Eine hausinterne Sicherung ausgewählter Seiten erwies sich als notwendig, was eine Evaluation bestehender Arbeitsweisen und Konzepte zur Folge hatte.

Es wurde entschieden, zwischen 2014 und 2016 eine Pilotphase durchzuführen, anhand derer einerseits Erfahrungen für eine regelmässige, beständige und gut durchdachte Sicherungsstrategie für Netzressourcen gemacht werden sollten, andererseits Prozesse implementiert wurden. Kosten und Personalaufwand sollten bis zur Evaluation der Pilotphase möglichst gering bleiben.

Aufgrund der hohen Anzahl an verfügbaren Seiten und Inhalten wurde ein Bewertungsvorschlag für die Pilotphase in Arbeit genommen. Ausserdem wurden

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 1 — http://dx.doi.org/10.18755/iw.2018.12

verschiedene Open-Source-Tools für die Sicherung der Seiten getestet; der Entscheid fiel schliesslich auf das Web Curator Tool (WCT).[1]

Bewertung

Zum Kernbereich der Überlieferungsbildung gehören Netzpublikationen der kantonalen Verwaltung ohne den Einbezug weiterer ‚Anbieter' aus dem halbstaatlichen oder privaten Umfeld.[2] Eine Übernahme in Auswahl wird im Bereich der Schulen verfolgt (Vorgabe: alle drei Bezirke sollten vertreten sein sowie nach Möglichkeit laufende Bestrebungen zur Schulharmonisierungs-Aktion HARMOS mit abgebildet werden). Nach einer Prüfung der Intranet-Angebote wurde auf die Sicherung derselben verzichtet, da die Inhalte überwiegend organisatorischen und administrativen Zwecken dienen. In Bezug auf partnerschaftliche Organisationen der beiden Halbkantone Basel-Stadt und Basel-Landschaft wurde auf den Bewertungsentscheid bezüglich der Zuständigkeit für physische Unterlagen zurückgegriffen. Temporäre Internetangebote, Abschaltungen, Relaunches oder ausschliesslich im Internet vorhandene Angebote sichert das Staatsarchiv Basel-Stadt nach Möglichkeit, jedoch kann kein periodisches und umfassendes Monitoring solcher Seiten erfolgen.

Gesamthaft kamen so rund 150 Webseiten zusammen, welche periodisch einmal jährlich geharvestet werden. Im Zentrum steht primär die Sicherung von Inhalten; Fragen zu Urheber- und Persönlichkeitsrechten werden im Nachgang der Pilotphase diskutiert.

Verzeichnung

Unter der Abteilung ‚Sammlungen' haben wir einen Fonds ‚Webarchiv' angelegt. Unter diesem Zweig existiert pro Dienststelle ein Bestand, je nach URL werden Serien vergeben.

Abbildung 1: Verzeichnungsstruktur Webarchiv, Ausschnitt

1 Siehe http://webcurator.sourceforge.net/. (Sämtliche Weblinks wurden am 19.02.2018 zuletzt aufgerufen.)

2 Nicht berücksichtigt werden demnach öffentlich-rechtliche Körperschaften, welche nur bestimmte Aufgaben im Auftrag des Kantons erledigen; dasselbe gilt für Personen / Organisationen / Firmen etc., die den Kanton als solchen zwar ‚prägen', aber nicht der Verwaltung angehören oder dieser in irgendeiner Form angebunden sind.

Den Snapshot[3] verzeichnen wir in einem Dossier. Das zugehörige Formular wurde spezifisch für Netzressourcen generiert und enthält Metadaten-Felder für die WCT Target Instance ID, die Anzahl Dateien und das Dateivolumen.

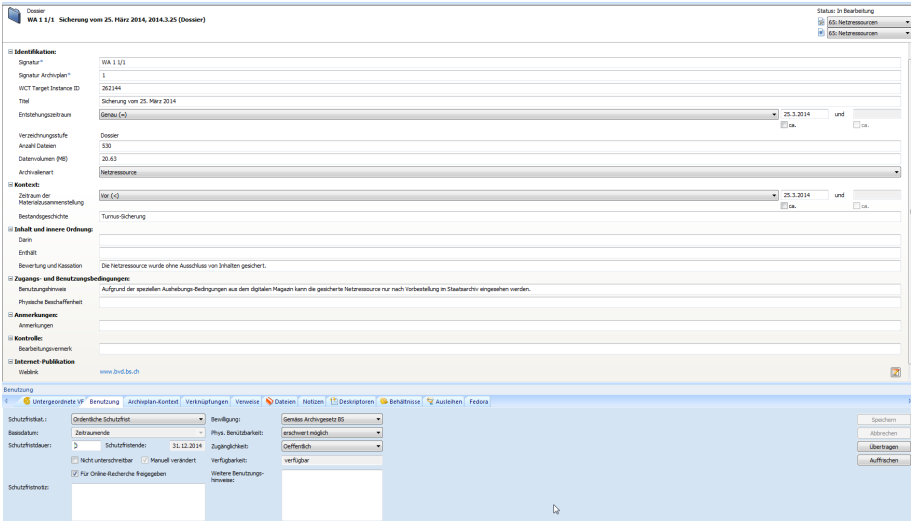

Abbildung 2: Web-Snapshot als Dossier verzeichnet

Die gesicherten Daten werden via Ingest in die Verzeichnungseinheit übernommen. Der Zugriffspfad auf die WARC-Daten in unserem Digitalen Magazin wird angelegt.

Benutzung

Die vom WCT zur Verfügung gestellte Oberfläche dient lediglich zur internen Qualitätssicherung. In Zukunft wird die Benutzung via den Digitalen Lesesaal stattfinden, der sich momentan noch im Aufbau befindet. Um die gesicherten Netzressourcen jedoch bereits jetzt nutzbar machen zu können, steht eine Übergangslösung zur Verfügung: In der Verzeichnungseinheit ist ein Weblink auf eine URL eingetragen, welche die Sichtung der Seite mittels Wayback-Server ermöglicht.

Aufwand in Zahlen

Den Harvest der 150 Seiten wickelt das System in einer Zeitspanne von ungefähr dreissig Stunden im Hintergrund ab. Zeitintensiv gestalten sich jedoch das Verzeichnen der gesicherten Webseiten und das Ermitteln der zugehörigen Metadaten. Hinzu kommen das Verlinken der WARC-Dateien, die daran anschliessende Durch-

3 Momentaufnahme einer Webseite, bzw. ein Screenshot einer kompletten Webseite.

führung des Ingests, das Setzen des Links auf den Wayback-Server sowie eine stichprobenmässige Qualitätskontrolle. Pro URL werden so grob geschätzt fünfzehn Minuten aufgewendet, für die gesamte jährliche Sicherung folglich rund vierzig bis fünfzig Stunden. Dazu kommen zusätzliche Aufwände für Problemfälle.

Aspekte zur Unternehmensarchitektur der Webarchivierung

Zusammenfassung

Die Archivierung von Webseiten im Staatsarchiv Basel-Stadt erstreckt sich über vier Bereiche. Zunächst geht es um das Auslesen der zu speichernden Webseite (Harvesting): Netzressourcen, die über eine URL erreichbar sind, werden in eine spezielle Archivdatei im Web Archive File Format WARC[4] geschrieben. Die gesicherten Inhalte liegen darin in serialisierter Form als XML-Struktur vor. Damit sie mit einem Browser betrachtet werden können, muss von der WARC-Datei ein Index erstellt werden (Indexing); dieser wird in eine CDX[5]-Datei geschrieben. Damit liegt die gesicherte Website vollständig vor. Jetzt erst wird sie archiviert. Dazu werden die WARC- und CDX-Dateien in ein OAIS-konformes AIP gepackt, verzeichnet und in ein Repository abgelegt (Verzeichnung/Ingest). Die gesicherte und archivierte Webseite kann jetzt benutzt werden (Benutzung).

Prozessdarstellung

Die im Staatsarchiv Basel-Stadt eingeführte und betriebene Archivierung von Webseiten lässt sich durch vier lose gekoppelte Prozesse darstellen. Die Bereiche Harvesting, Indexing, Ingest und Benutzung werden jeweils als separate Spur[6] abgebildet, innerhalb derer die zugehörigen Prozesse ablaufen. Diese Spuren sind als organisatorische Verantwortlichkeiten für die beinhalteten Prozesse zu verstehen.

4 Web ARChive file format, WARC: https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml.
5 CDX: https://archive.org/web/researcher/cdx_file_format.php.
6 In BPMN als swimlane dargestellt.

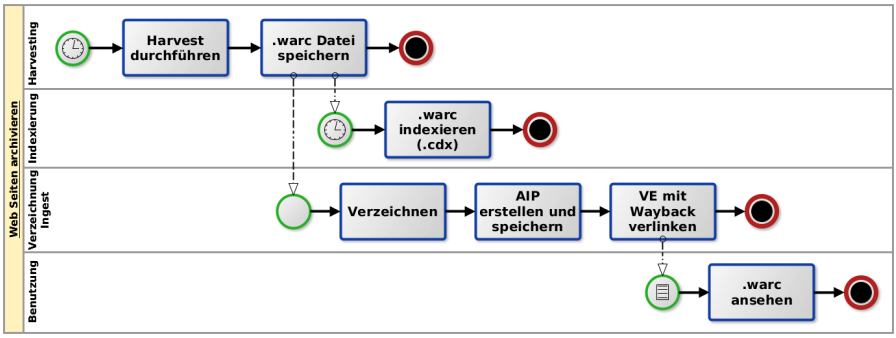

Abbildung 3: Prozessdarstellung

Die lose Koppelung zwischen den Prozessen legt zwar die Reihenfolge, nicht aber den exakten Zeitpunkt der Ausführungen fest. So wird beispielsweise der Ingest-/Verzeichnungsprozess je nach organisatorischer Ressourcenplanung irgendwann, auf jeden Fall aber immer erst nach dem zugehörigen Harvestprozess ausgeführt. Die Entkoppelung der Prozesse ermöglicht eine effiziente Organisation der verfügbaren Personalressourcen. Sie ist die Voraussetzung für die Skalierung hin zu einer breiten Serienproduktion.

Informationsarchitektur

Die beschriebenen Prozesse der Webarchivierung – dargestellt durch die abgerundeten Vierecke – stehen im Verhältnis zu verschiedenen Geschäftsobjekten (gelbe Vierecke) und Datenobjekten (blaue Vierecke). Während Geschäftsobjekte aus Sicht der Archivleitung eine strategische Bedeutung haben, sind die Datenobjekte für die technisch-operative Implementation wichtig.

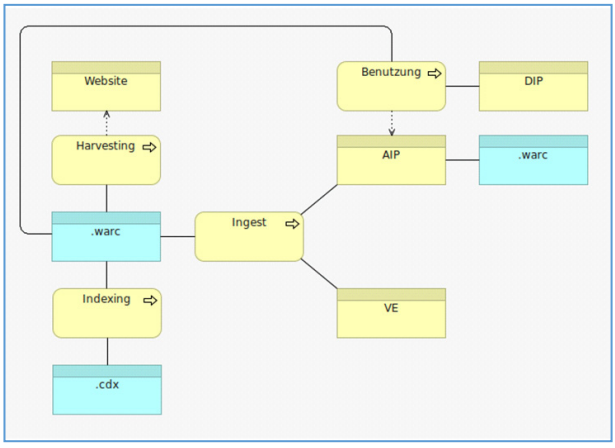

Abbildung 4: Informationsarchitektur

Die Objektdarstellung zeigt auf, dass eine als WARC-Datei gesicherte Netzressource zweimal vorgehalten wird, einmal durch den Benutzungsprozess und einmal durch den Ingestprozess. Diese Redundanz ist ein Kompromiss, der den Benutzungsansprüchen geschuldet ist: Die in einem AIP verpackte WARC-Datei muss im Falle einer Benutzung entpackt, indexiert und an einen geeigneten Ort kopiert werden. Diese Schritte können je nach Grösse und Komplexität der aufgerufenen Website sehr ressourcenintensiv sein und entsprechend lange dauern. Aus diesem Grund liegen bereits entpackte und indexierte Kopien der WARC- und CDX-Dateien für die Benutzung bereit. Die Verfügbarkeit dieser Arbeitskopien ist deutlich geringer gewichtet als jene der archivierten AIPs. Diese Benutzungskopien können jederzeit mit endlichem Aufwand aus den AIPs wieder hergestellt werden.

Anwendungsarchitektur

Um die charakterisierten Prozesse und Informationsobjekte zu implementieren, verwendet das Staatsarchiv Basel-Stadt folgende Anwendungskomponenten:

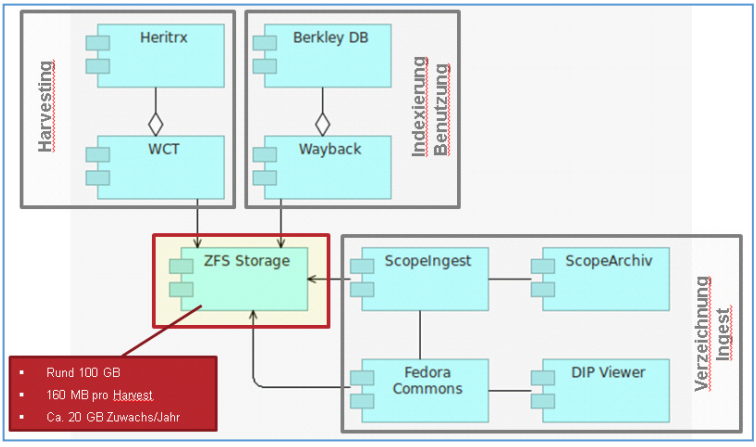

Abbildung 5: Anwendungsarchitektur

Der Harvesting-Prozess wird mit Heritrix[7] als Webcrawler und WCT[8] (Web Curator Tool) als Workflow Management Tool umgesetzt. Pro zu speichernde Webseiteninstanz wird ein Heritrix-Prozess ausgeführt. Die verschiedenen Heritrix-Prozesse werden mit Hilfe von WCT verwaltet und gesteuert. Für die Indexierung und Benutzung werden Berkley DB als Indexer und OpenWayback[9] als WARC-Viewer eingesetzt. Der WARC-Viewer ermöglicht es, eine gesicherte Webseite

7 Siehe: http://crawler.archive.org/index.html.
8 Siehe: http://webcurator.sourceforge.net/.
9 Siehe: https://github.com/iipc/openwayback/wiki.

unter Aufruf ihrer ursprünglichen URL in einem Webbrowser zu betrachten. Für die Verzeichnung und für den Ingest der zu archivierenden Webseite verwenden wir ScopeArchiv, ScopeIngest und Fedora Commons als Repository. Für die Speicherung der WARC- und CDX-Dateien sowie der AIPs wird eine redundante, ZFS[10]-basierte Speicherlösung verwendet.

Technologiearchitektur

Die Anwendungskomponenten werden vom Staatsarchiv Basel-Stadt in einer eigenen UNIX/Sparc-Umgebung betrieben. Gründe für die Wahl dieser Technologiearchitektur sind ihre grosse Ausfallsicherheit und ihre robuste Skalierbarkeit. Heritrix, WCT, OpenWayback, und Fedora Commons sind jeweils Java-Servlet[11]-Anwendungen. Sie werden in einem separaten Tomcat Servlet Container in einer separaten Solaris-Zone betrieben. Heritrix und WCT werden dabei aus Sicherheitsgründen in doppelter Ausführung vorgehalten: Einerseits für Webseiten ausserhalb der kantonalen Domäne – geschützt durch einen Forward Proxy – und andererseits für Webseiten innerhalb der bs.ch-Domäne. Die ZFS Storage Appliance ist ein dediziertes System. ScopeIngest muss innerhalb einer X86-Umgebung betrieben werden. Die Virtualisierung erfolgt dort über Oracle Virtual Box respektive Oracle VM (OVM).

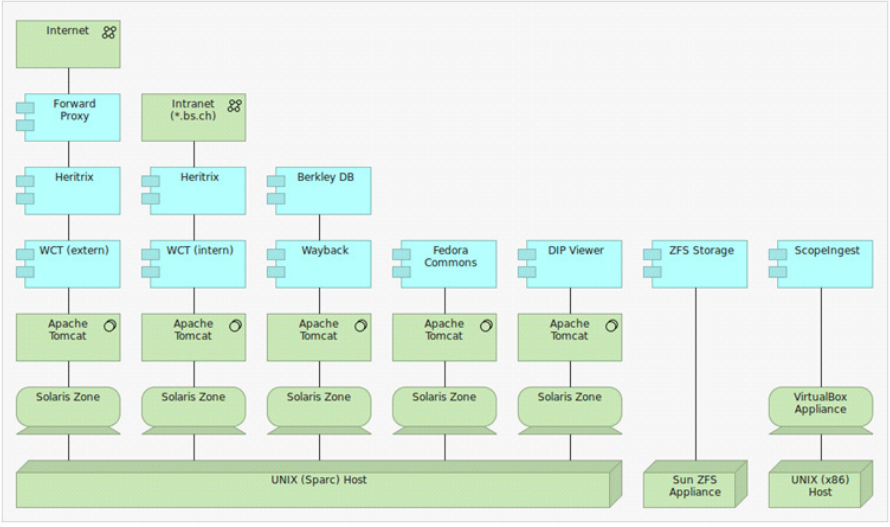

Abbildung 6: Technologiearchitektur

10 Siehe: https://de.wikipedia.org/wiki/ZFS_(Dateisystem).
11 Siehe Java Community Process JSR 315, JSR 340 resp. JSR 369.

Die abgebildete Technologiearchitektur wird tatsächlich parallel, in zwei unabhängigen und räumlich entfernten Standorten betrieben. Dadurch werden eine höhere Ausfallsicherheit[12] und die Trennung von Test und Produktion erreicht

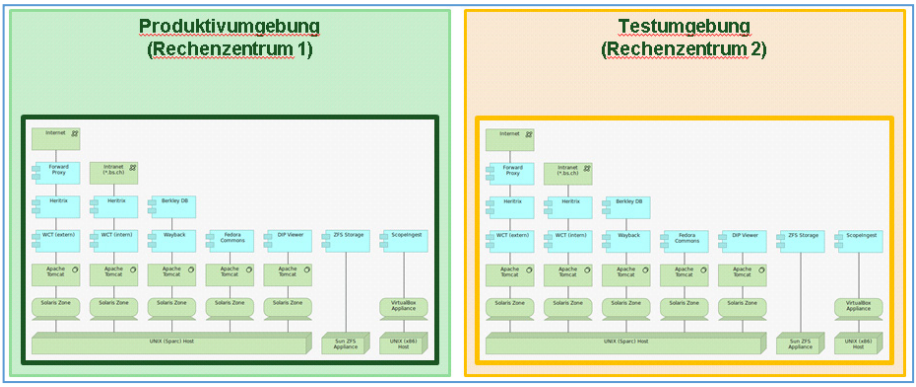

Abbildung 7: Produktiv- und Testumgebung

Fazit

Werden in einem Archiv bereits OAIS-konform digitale Inhalte archiviert, erfordert die Archivierung von Webseiten beziehungsweise Netzressourcen technologisch kein grosses zusätzliches Innovationspotential. Die bereits verwendeten Komponenten AIS und Ingestserver/Repository können gleichermassen eingesetzt werden. Lediglich das Harvesting, die Indexierung und die Benutzung erfordern zusätzliche Werkzeuge. Diese sind aber quelloffen verfügbar. Fachlich und organisatorisch erfordert die Archivierung von Netzressourcen dagegen erhebliche Planungs- und Koordinationsaufwände. Ihre Struktur, ihr Lebenszyklus und ihre Verfügbarkeit unterscheiden sich stark von dokumentartigem Archivgut.

12 Konkret: Reduktion der allfällig zu erwartenden Downtime und Data Loss Time.