Webarchivierung im Auftrag - Erfahrungen des Bibliotheksservice-Zentrums Baden-Württemberg

Webarchivierung bis 2017 – dezentrales Harvesting

Das Bibliotheksservice-Zentrum Baden-Württemberg (BSZ) bearbeitet das Thema Webarchivierung schon seit Beginn der 2000er-Jahre aktiv. 2004 bot das BSZ erstmalig für Bibliotheken und Archive einen separaten Dienst zur Archivierung von Webauftritten an. Dabei handelte es sich um eine eigenentwickelte Webanwendung, die eine maskenunterstützte Erfassung hierarchischer Metadatenstrukturen, Crawlingparameter und Jobsteuerungselemente erlaubte. Jobs konnten gestartet, überwacht und beendet und Logs teilweise eingesehen werden. Zunächst wurden alle Spiegelungen auf der Basis von HTTrack angeboten, ab 2012 parallel auch mit Heritrix. Die Präsentation erfolgte innerhalb der Applikation selbst (HTTrack) sowie über die Anbindung einer Open Wayback Machine (WARC).

Technischer Betrieb, Weiterentwicklung der Anwendung sowie technischer und bibliothekarischer Support lagen beim BSZ, die Kundeneinrichtungen parametrisierten die Jobs selbst und kontrollierten ihre Ergebnisse autonom. Durchgeführt wurde – und wird auch heute – ein selektives Harvesting ausschließlich institutioneller Websites und Events. Dabei liegt der Fokus nicht auf einer möglichst großen Menge gespiegelter Websites, sondern vielmehr auf einer größtmöglichen Vollständigkeit und Tiefe, sowie einem hohen Authentizitätsgrad der Archivkopien ausgewählter Webauftritte. Dies betrifft das Look-and-feel und die Funktionalität der gespiegelten Site, im selben Maße aber auch die Verfügbarkeit der Inhalte. Die Archivkopien waren - und sind bis heute - in der Regel öffentlich zugänglich.

2016 entschloss sich das BSZ zur Ablösung seiner nur noch unter größerem Ressourceneinsatz entwicklungsfähigen Software, angestrebt wurde die Nutzung eines Fremdsystems. In einer Evaluationsphase wurden verschiedene Angebote untersucht, unter anderem das Web Curator Tool (WCT, damals noch mit Heritrix Versionen 1.x), die NetarchiveSuite (NAS) und Edoweb. Ziel war eine Anwendung mit hohem Nutzungskomfort und einer realistischen technischen Zukunftsperspektive, die ein Crawling unter der bisher schon im BSZ eingesetzten Heritrix-Generation 3.x erlaubte. Weiterhin war eine Lösung für eine integrierte Präsentation oder eine Migration der vorhandenen HTTrack-Crawls nach WARC erforderlich. Es zeigte sich, dass zum damaligen Zeitpunkt allein Archive-It, ein kommerzieller Webarchivierungsdienst des Internet Archivs in San Francisco, diese Kriterien vollständig erfüllte. Das BSZ entschied sich daher, ab 2017 diese Dienstleistung zu nutzen.

Es bot sich an, im selben Zuge auch das Servicemodell anzupassen - vom dezentralen Spiegeln durch Kundeneinrichtungen hin zu einem zentralen Crawlingservice im BSZ.

Webarchivierung seit 2017 – zentrales Harvesting im BSZ

Seit 2017 bietet das BSZ im Rahmen seiner Dienstleistung SWBregio unter Nutzung des Angebots von Archive-It einen zentralen Service für die Webarchivierung an, welcher kommunalen, regionalen und Kreisarchiven offensteht. Derzeit nutzen rd. 20 Archive aus dem gesamten Bundesgebiet dieses Angebot. Für die Saarländische Landes- und Universitätsbibliothek Saarbrücken (SULB) führt das BSZ darüber hinaus die Archivierung der im Rahmen ihres Pflichtauftrags gesammelten saarländischen Websites durch.

Insgesamt werden aktuell rd. 680 Domains laufend durch das BSZ gespiegelt, in der Regel ein- oder zweimal jährlich. Die Ressourcen sind zumeist öffentlich zugänglich, in wenigen Fällen aber auch eingeschränkt (Modell Lesesaal-Nutzung) oder ausschließlich intern.

Die Aufgaben der Kundeneinrichtungen sind bei diesem Modell im Vergleich zu früher sehr stark geschrumpft. Die Archive treffen heute einmalig allgemeine Festlegungen zur Darstellung und Strukturierung ihrer Sammlungen bei Archive-It. Sie nehmen die inhaltliche Auswahl der zu spiegelnden Webauftritte vor sowie, soweit erforderlich, die Einholung der Spiegelungserlaubnisse beim Anbieter. Die Beauftragung des BSZ zur Durchführung konkreter Crawls erfolgt standardisiert mit Hilfe eines Formulars, welches unter anderem die zu spiegelnde URL, Metadaten zur Website, Archivierungsintervall und den späteren Zugriff auf die Archivkopie festhält.

Das BSZ übernimmt die Pflege der Sammlungen bei Archive-It sowie die Erfassung und Pflege der Webauftritte und ihrer Metadaten. Erstmalig zu spiegelnde Websites werden umfassend analysiert und daraufhin die erforderlichen Spiegelungsparameter festgelegt. Es erfolgen – zum Teil iterativ – Testcrawls, Sichtprüfungen und inhaltliche Abnahme anhand der vorliegenden Reports und Logs. Falls sich schwierige Fälle nicht im BSZ klären lassen, wird das Helpdesk von Archive-It einbezogen. Die Abwicklung und Überwachung aller regelmäßig laufenden Crawls einschließlich einer durchgehenden Qualitätskontrolle werden vom BSZ durchgeführt.

Erfahrungen und Bewertung

Die Einführung der zentralen Archivierung brachte eine deutliche Aufgabenverschiebung von den Archiven hin zum BSZ mit sich. Auf dem Hintergrund der bisherigen Erfahrungen mit dem autonomen Crawling durch die Kundeneinrichtungen zeigen sich jedoch deutliche Vorteile dieses Servicemodells.

Das frühere, dezentrale Spiegeln band sowohl bei den Archiven als auch beim BSZ sehr viele personelle Kapazitäten. Insbesondere die Aufwände für Kommunikation und Support waren beträchtlich. Es waren immer wieder individuelle Schulungen durchzuführen und Anwenderdokumentationen zu erstellen - nicht zuletzt, weil die zur Archivierung eines hoch dynamischen Contents erforderlichen Kenntnisse bei zum Teil nur sporadischer Anwendung schwer gefestigt werden konnten. Personelle Fluktuation in den Archiven führte zu Wissensabfluss.

Das BSZ erhielt aus den Kundeneinrichtungen Anfragen zu nicht zufriedenstellenden, unvollständigen oder misslungenen Spieglungen, die im Nachhinein zu analysieren sich sehr aufwendig gestaltete. Oft lagen diese Probleme an einer fehlerhaften oder nicht vollständigen Job-Parametrisierung. Das BSZ versuchte zum Teil, die Ergebnisse nachträglich nachzubessern und zu vervollständigen. Die Qualitätskontrolle wurde durch die eigene Software nicht tiefgehend unterstützt und bereitete daher einigen Kundeneinrichtungen sichtlich Schwierigkeiten.

Trotz aller Bemühungen waren die Spiegelungen qualitativ sehr inhomogen. Dies zeigte sich später auch im Zuge der durch Archive-It durchgeführten Migration der HTTrack-Ressourcen nach WARC, bei welcher unvollständige Ergebnisse auffielen und für die weitere Verarbeitung nachträglich aufbereitet werden mussten.

Die Erfahrungen mit der zentralen Archivierung im BSZ sind dem gegenüber sehr erfreulich. Es entstehen spürbar weniger Aufwände für Support und Beratung, so dass personelle Ressourcen zielgerichteter und effizienter eingesetzt werden können. Die Durchführung von Schulungen oder die Erstellung von Anwenderdokumentationen sind nicht mehr erforderlich.

Der Service aus einer Hand bietet darüber hinaus weitere Vorteile. Alle Aufgaben können innerhalb eines einheitlichen Workflows sehr effizient bearbeitet werden. Im BSZ wurden über die Jahre umfassende Erfahrungen mit Websites, ihren Strukturen und zu erwartenden Problemen beim

Crawling gesammelt, so dass die Parametrisierung heute sehr routiniert erfolgen kann. Schwierigkeiten können bereits bei der Sichtung eines Webauftritts erkannt und von vorne herein umgangen werden. Die Parametrisierung der Crawler (Heritrix oder Brozzler) erfolgt nach einheitlichen Mustern und Methoden, wodurch eine hohe Homogenität der Ergebnisse erzielt wird. Da die Spiegelungen durchgehend Qualitätskontrollen unterzogen werden, besitzen sie eine hohe Qualität und Vollständigkeit. Die teilnehmenden Einrichtungen werden ihrerseits spürbar von Aufwänden entlastet, da sie die Durchführung der Webarchivierung vollständig ans BSZ abgeben.

Nach nunmehr 6 Jahren zentraler Webarchivierung ziehen das BSZ als Dienstleister ebenso wie die teilnehmenden Einrichtungen ein durchweg positives Fazit.