Caroline Dupuis

Web-Archivierung an der Saarländischen Universitäts- und Landesbibliothek (SULB)

Web archiving at the Saarland University and State Library (Saarländische Universitäts- und Landesbibliothek, SULB)

DOI 10.1515/bd-2017-0055

Zusammenfassung: An der Saarländischen Universitäts- und Landesbibliothek (SULB) werden seit 2008 Webseiten archiviert. Zur Archivierung landeskundlicher elektronischer Dokumente steht das Repositorium SaarDok zur Verfügung. Eine gesetzliche Grundlage für die Ablieferung unkörperlicher Werke besteht seit Dezember 2015.

Schlüsselwörter: Langzeitarchivierung, Webseiten, SaarDok, Saarländisches Mediengesetz

Abstract: Since 2008, websites are being archived at the Saarland University and State Library (SULB). The repository SaarDok is available for archiving electronic documents concerning regional studies. A legal basis for depositing non-physical works exists since December 2015.

Keywords: long time archiving, websites, SaarDok, Saarland Media Law

1 Einleitung

Seit 2008 werden an der Saarländischen Universitäts- und Landesbibliothek Webseiten archiviert.[1]

1 http://www.sulb.uni-saarland.de/ [Zugriff: 01.03.2017].

Dr. Caroline Dupuis: c.dupuis@sulb.uni-saarland.de

Die SULB verfolgt einen selektiven Harvesting-Ansatz. Der Fokus liegt auf Webseiten ausgewählter Institutionen (z. B. Landesbehörden, Interessensverbänden, Kultureinrichtungen etc.) sowie auf Webseiten zu bestimmten Ereignissen (z. B. *Filmfestival Max-Ophüls-Preis*), und zu bestimmten Themen (z. B. *Erinnert Euch: saarländische Erinnerungsorte und Gedenkstätten über Widerstand und Verfolgung in der Nazi-Zeit*). Vereinzelt werden auch Webseiten privater Anbieter archiviert (z. B. *Saar-Nostalgie: Erinnerungen an frühere Zeiten im Saarland*).

Die Priorität der SULB liegt allerdings nicht auf der Archivierung von Webseiten, sondern vielmehr auf der Langzeitarchivierung von frei zugänglichen Online-Zeitschriften und Monographien (insgesamt ca. 1200 Titelaufnahmen, mit ca. 5600 Objekten untergeordneter Hierarchie).

2 Gesetzliche Grundlage

Mit der Novellierung des Saarländischen Mediengesetzes (SMG) im Dezember 2015 hat sich auch das Pflichtexemplarrecht wesentlich geändert.[2] Eine der wichtigsten Neuerungen ist die Abgabepflicht auch für digitalisierte (unkörperliche) Werke.

§ 14 Abs. 1: Von jedem Medienwerk, das im Saarland verlegt wird, ist unabhängig von der Art des Trägers und des Vervielfältigungsverfahrens von der Person, die das Medienwerk verlegt, unaufgefordert unmittelbar nach Beginn der Verbreitung unentgeltlich und auf eigene Kosten ein Stück (Pflichtexemplar) in marktüblicher Form an die Saarländische Universitäts- und Landesbibliothek abzuliefern. […]

§ 14 Abs. 3: Medienwerke in unkörperlicher Form müssen unter Einhaltung der von der Deutschen Nationalbibliothek für Pflichtexemplare festgelegten technischen Standards und Verfahren abgeliefert werden. [ ]

Am 24. November 2016 trat schließlich auch die Verordnung zur Durchführung des Saarländischen Mediengesetzes über die Ablieferung von Pflichtexemplaren (PflAV) in Kraft.[3] Darin wird auch das Ablieferungsverfahren für elektronische Dokumente näher geregelt.

2 Amtsblatt des Saarlandes Teil I vom 10. Dezember 2015, § 14.
3 Amtsblatt des Saarlandes Teil I vom 24. November 2016.

3 Rechtemanagement

Mit der Ablieferung eines Medienwerks in unkörperlicher Form erhält die SULB das Recht, das Werk in ihren Räumen zugänglich zu machen.

§ 14 Abs. 3: Mit der Ablieferung eines Medienwerkes auf einem elektronischen Datenträger oder eines Medienwerkes in unkörperlicher Form erhält die Bibliothek das Recht, das Werk zu speichern, zu vervielfältigen und zu verändern oder diese Handlungen in ihrem Auftrag vornehmen zu lassen, soweit dies notwendig ist, um das Medienwerk in die Sammlung aufnehmen, erschließen und für die Benutzung bereitstellen zu können sowie seine Erhaltung und Benutzbarkeit dauerhaft zu sichern. […]

Mit der Ablieferung eines Medienwerks in unkörperlicher Form erhält die Bibliothek das Recht, das Werk in ihren Räumen zugänglich zu machen.[4]

Da die SULB der breiten Öffentlichkeit die gespeicherten Webseiten zur Verfügung stellen möchte, wird die schriftliche Einverständniserklärung eingeholt, die Website über das Internet der Öffentlichkeit uneingeschränkt zugänglich machen zu dürfen. Die Rücklaufquote der unterschriebenen Formulare beträgt zur Zeit ca. 50%. Wird das Einverständnis verweigert oder der Anbieter der Website reagiert auf das Schreiben nicht, wird die Website dennoch archiviert und, wie es das Gesetz vorsieht, nur in den Räumen der Bibliothek (IP-basiert) zugänglich gemacht.

4 Technischer Hintergrund

Bereits 2003 startete an der Saarländischen Universitäts- und Landesbibliothek (SULB) ein Projekt zur Archivierung landeskundlicher elektronischer Dokumente. Die Software, die damals eingesetzt wurde, war allerdings nicht geeignet, Webseiten zu laden und zu archivieren. Ferner stellten sich frühzeitig Probleme beim Import der Daten in den Südwestdeutschen Bibliotheksverbund (SWB) ein. Aus diesen Gründen fand 2006 der Umstieg auf die Archivierungssoftware SWBcontent des Bibliotheksservice-Zentrum Baden-Württemberg (BSZ)[5] statt. Ein Vorteil der Software liegt im einfachen Import bibliographischer Daten aus dem SWB.

4 Amtsblatt des Saarlandes Teil I vom 10. Dezember 2015, § 14.

5 https://www.bsz-bw.de/mare/lza/index.html [Zugriff: 01.03.2017].

Eine Darstellung hierarchischer Strukturen ist ebenso möglich, wodurch Spiegelungen von Webseiten zusammenhängend dargestellt werden können. Entscheidend für den Umstieg war auch der einfache Nachweis aller Dokumente im SWB[6] und damit auch im Saarländischen Virtuellen Katalog (OPAC)[7] und KVK[8].

In den Anfängen der Archivierung erfolgte das Harvesting mit dem Crawler „HTTrack“, ab 2013 mit dem Web Crawler „heritrix“.

Als Archivformat wird das Webarchive File Format (WARC) genutzt, die Darstellung der WARC-Files erfolgt über eine Wayback Machine, eine eigene Applikation, die außerhalb des SWBcontent läuft.

Zur Archivierung der Sites steht seit Beginn des Harvestings das Repositorium SaarDok[9] zur Verfügung.

5 Workflow

Die Erschließung der Webseiten erfolgt noch vor der eigentlichen Archivierung in der Zeitschriftendatenbank (ZDB) gemäß RDA (bis 2015 RAK-WB/ZETA). Über den Datendienst ZDB-SWB sind die Webseiten so bereits nach wenigen Minuten als kostenfreie Online-Ressourcen im Katalog des Südwestdeutschen Bibliotheksverbundes (SWB) und mit dem gleichen Datensatz im Saarländischen Virtuellen Katalog, einer Teilsicht des SWB-Katalogs, der als WebOPAC der SULB dient, nachgewiesen.

Die MAB-Daten der zuvor katalogisierten Seite werden in SaarDok händisch geladen.

6 http://swb.bsz-bw.de/ [Zugriff: 01.03.2017].
7 http://swb2.bsz-bw.de/DB=2.340/ [Zugriff: 01.03.2017].
8 https://kvk.bibliothek.kit.edu/ [Zugriff: 01.03.2017].
9 http://saardok.sulb.uni-saarland.de/menu.do?start [Zugriff: 01.03.2017].

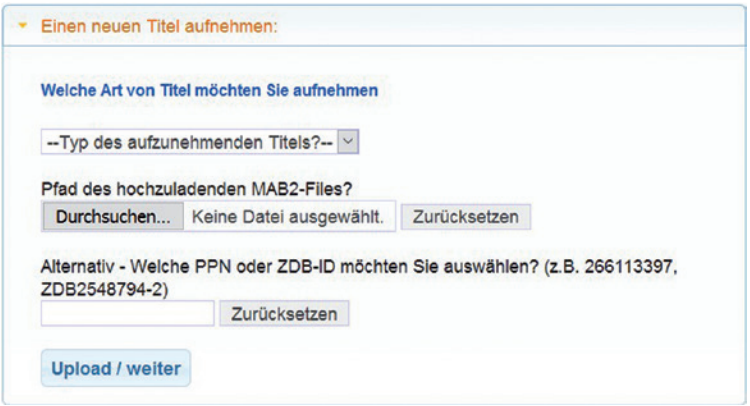

Abb. 1: Import der MAB-Daten.

Der Download der Webseite erfolgt in SaarDok über ein Fenster zur Konfiguration des Heritrix3 Crawls mittels User Interface, das konfiguriert werden kann.

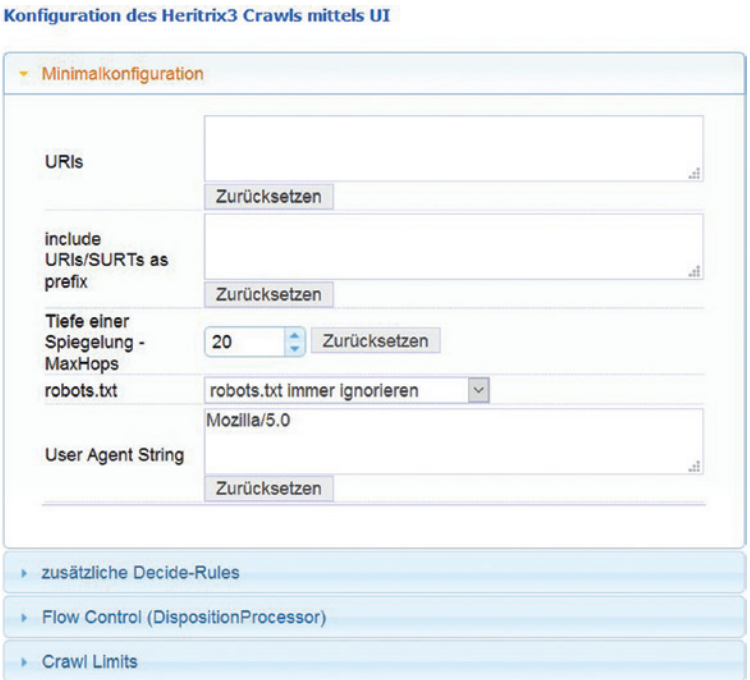

Abb. 2: Konfiguration des Heritrix-Crawls.

Es ist möglich eine oder mehrere URLs einzugeben. Ferner kann die Tiefe der Spiegelung bestimmt werden. Als Standard wird eine gegebenenfalls vorliegende robots.txt immer ignoriert. Über diese Datei versuchen Webmaster zu steuern welche Unterseiten ihrer Homepage nicht von Suchmaschinen indiziert werden sollen. Da die Betreiber der Seiten der SULB jedoch das Einverständnis zum Harvesting der gesamten Website vorher gaben, können diese Informationen bei der Archivierung übergangen werden.

Zusätzlich zu den eingegebenen URLs können Unterseiten eines Internetauftritts auch von der Archivierung ausgeschlossen werden („exclude URLs“).

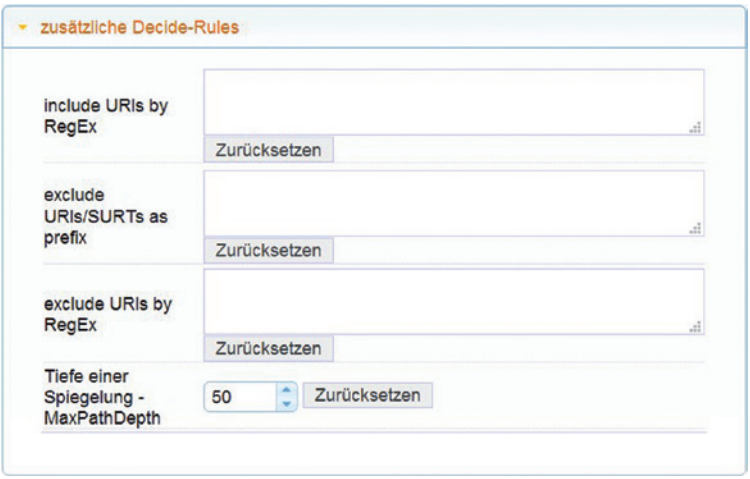

Abb. 3: Zusätzliche Decide-Rules.

Ferner kann auch eine Obergrenze und eine zeitliche Begrenzung für den Download eingegeben werden.

Abb. 4: Begrenzung des Downloads.

Da sich Webseiten immer wieder ändern, werden sie in regelmäßigen Abständen erneut eingesammelt. Die Sammelfrequenz kann in SaarDok für jede Webseite individuell eingestellt werden (*next_storage*), je nach Änderungshäufigkeit bzw. inhaltlicher Relevanz. Im Moment ist das Standardintervall jährlich. Ändert sich die Webseite im Laufe des Jahres nicht, wird das Intervall auf zwei Jahre hochgesetzt.

Soll eine Webseite nur in den Räumen der Bibliothek zugänglich sein, wird dies über das Feld *Object Access* gesteuert. Die Einschränkung erfolgt IP-basiert.

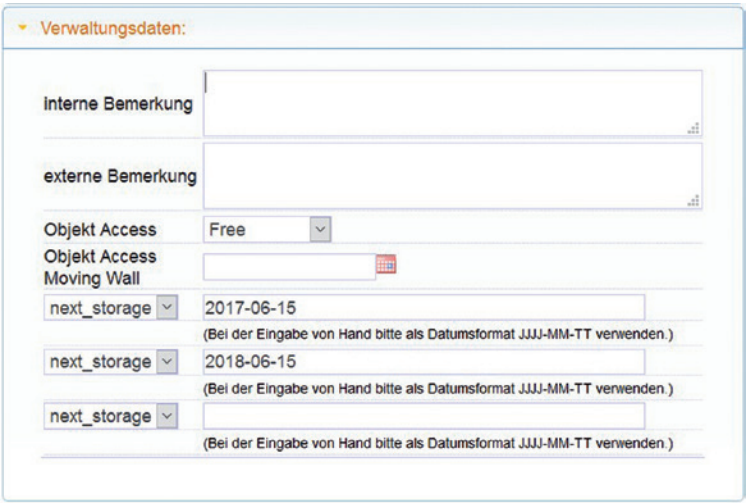

Abb. 5: Verwaltungsdaten.

Nach der Archivierung werden Webadressen, die Kennzeichnung der Langzeitarchivierung sowie der archivierte Bestand in der ZDB ergänzt. Die Adressenangaben erfolgen sowohl in persistenter Form über Uniform Resource Names (URN) als auch durch die Wiedergabe der proprietären URL-Adresse des Harvestingsystems. Titel, die über SaarDok weltweit angeboten werden, erhalten eine spezielle Kennzeichnung, so dass die Aufnahmen auch von anderen Bibliotheken (automatisch) nachgenutzt werden können.

Da ein regelmäßiger händischer Upload in SaarDok erfolgt, werden sowohl Titel, die sich bei Webseiten häufiger ändern, als auch URLs in der ZDB zeitnah aktualisiert.

Alle Webseiten werden im SWB für die saarländische Bibliographie[10] inhaltlich erschlossen.

Ist die Website nur in den Räumen der Bibliothek zugänglich, wird dies sowohl in der ZDB als auch in der Saarländischen Bibliographie vermerkt.

Probleme bereiten nach wie vor Inhalte, die in Datenbanken oder Content Management Systemen abgelegt sind. Diese werden durch den Harvester nicht erreicht. Ferner bereiten dynamische Elemente (Flash) erhebliche Probleme bei der Archivierung.

Die einzelnen Zeitschnitte der Seiten werden in Stichproben einer Qualitätskontrolle unterzogen. Sind die Seiten in großen Teilen oder gänzlich nicht durch den Harvester herunterzuladen, wird auf die Archivierung vorerst verzichtet. Es wird aber vermerkt, dass nach einem bis zwei Jahren erneut versucht wird die Seite zu archivieren.

Zur Zeit werden in SaarDok ca. 190 institutionelle, persönliche und thematische Webseiten archiviert.

SaarDok
Der Archivserver für das Saarland

Dr. Caroline Dupuis
Saarländische Universitäts- und Landesbibliothek
Postfach 15 11 41
66041 Saarbrücken
Deutschland
E-Mail: **c.dupuis@sulb.uni-saarland.de**

10 http://swb2.bsz-bw.de/DB=2.306/ [Zugriff: 01.03.2017].