Fachhochschule Graubünden
University of Applied Sciences

Churer Schriften
zur Informationswissenschaft

Herausgegeben von
Wolfgang Semar, Bernard Bekavac, Ivo Macek

Arbeitsbereich Master of Advanced Studies
in Information Science

Schrift 179

Webarchivierung im UZH Archiv

Erstellung einer Prozessbeschreibung und Erarbeitung von Empfehlungen für die Konzipierung eines Datenmodells sowie bezüglich der Wahl einer Preservation Planning-Strategie

Sandra Morach

Chur 2024

Churer Schriften zur Informationswissenschaft

Herausgegeben von Wolfgang Semar,
Bernard Bekavac, Ivo Macek

Schrift 179

Webarchivierung im UZH Archiv

Erstellung einer Prozessbeschreibung und Erarbeitung von Empfehlungen für die Konzipierung eines Datenmodells sowie bezüglich der Wahl einer Preservation Planning-Strategie

Sandra Morach

Diese Publikation entstand im Rahmen einer Thesis zum Master of Advanced Studies in Information Science.

Referent: Prof. Dr. Tobias Wildi

Korreferent: Prof. Dr. Ana Petrus

Verlag: Fachhochschule Graubünden

ISSN: 1660-945X

Ort, Datum: Chur, November 2024

Abstract

Die Arbeit befasst sich mit Fragestellungen rund um das Thema Webarchivierung. Im UZH Archiv (UAZ), dem Archiv der Universität Zürich, erfolgte eine Neuorganisation des Webarchivs. Das UAZ crawlt seit Januar 2024 den Webauftritt mit dem Webarchiv-Service Archive-It. Die Neuorganisation stellt das UAZ vor Herausforderungen und offene Fragen, die im Rahmen der Arbeit diskutiert werden. Im Zentrum der Arbeit steht einerseits der Gesamtprozess der Webarchivierung, wobei dem Thema Preservation Planning besondere Beachtung geschenkt wird. Andererseits geht es um die Erörterung der Frage, wie die auf Archive-It gesicherten Daten in das digitale Langzeitarchiv des UAZ überführt werden sollen.

Zur Bearbeitung der Thematik erfolgt im Rahmen der Arbeit eine eingehende Auseinandersetzung mit dem Prozess der Webarchivierung im UAZ, eine Beschäftigung mit dem Container-Format WARC, ein Abgleich der verschiedenen in den Prozess involvierten Komponenten – Archive-It, digitales Langzeitarchiv und Archivinformationssystem – und ihrer jeweiligen Elemente sowie eine Diskussion von möglichen Migrations- und Emulationsstrategien.

Aus der Arbeit resultieren eine Prozessbeschreibung sowie Empfehlungen in Zusammenhang mit der Konzipierung eines Datenmodells und bezüglich der Wahl einer Preservation Planning-Strategie.

Zum Abschluss der Arbeit erfolgt ein Wechsel vom konkreten Fall des UAZ auf eine allgemeinere Ebene. Erkenntnisse, Empfehlungen und zentrale Leitfragen sind für andere Gedächtnisinstitutionen, die sich ebenfalls mit der Webarchivierung befassen, zusammengetragen.

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

The World Wide Web at its best is a mechanism for people to share what they know, almost always for free, and to find one's community no matter where you are in the world [Brewster Kahle, director and founder of the Internet Archive].[1]

Das voranstehende Zitat von Brewster Kahle – dem Gründer des Internet Archive – veranschaulicht die grosse Bedeutung des World Wide Web. Das Web ermöglicht es, Wissen zu teilen und eine Verbindung zwischen Menschen zu schaffen, ganz unabhängig davon, wo sich diese auf der Welt befinden. Kahle verweist mit dieser Aussage auf zwei zentrale Komponenten, die das Web auszeichnen: das verfügbare Wissen sowie den verbindenden Charakter. Das Web ist aus der heutigen Zeit kaum mehr wegzudenken und gehört für viele Menschen zum Alltag. Es prägt die Gesellschaft, indem es die Art und Weise der Verbreitung und Beschaffung von Information sowie die Form der Kommunikation massgebend beeinflusst.

In Gedächtnisinstitutionen wie Archive und Bibliotheken besteht Einigkeit darüber, dass das Web ein wichtiges Zeitdokument darstellt, das es gilt für zukünftige Forschende, Historiker*innen und die interessierte Öffentlichkeit zu sichern. Die Realisierung dieser Aufgabe erfolgt mit der Webarchivierung, wessen sich das Internet Archive in San Francisco bereits seit 1996 annimmt[2]. Auch individuelle Institutionen (Archive und Bibliotheken) widmen sich dieser Aufgabe. Zu diesen Gedächtnisinstitutionen gehört auch das UZH Archiv (UAZ), das Archiv der Universität Zürich (UZH).

Das UAZ sichert seit 2012 Zeitschnitte von ausgewählten Websites der UZH. Die Seiten wurden bis anhin mit dem Tool *Offline Explorer* abgezogen und in der vorliegenden HTML-Struktur (im ZIP-Format) in das digitale Langzeitarchiv des UAZ überführt. Infolge einer Neuorganisation werden die archivwürdigen Seiten der UZH seit Januar 2024 neu mit dem Webarchiv-Service Archive-It gecrawlt. Die Neuorganisation stellt das UAZ vor Herausforderungen und offene Fragen.

Die vorliegende Arbeit widmet sich der Erarbeitung einer Prozessbeschreibung sowie der Auseinandersetzung mit relevanten Fragen in Zusammenhang mit der Erarbeitung eines Datenmodells sowie der Wahl einer Preservation Planning-Strategie. Fragestellungen in

[1] Kahle, Internet Archive Blogs 2021. Verfügbar unter: https://blog.archive.org/2021/07/21/reflections-as-the-internet-archive-turns-25/ [2.7.2024].

[2] vgl. Bragg/Hanna 2013: Seite 1.

Zusammenhang mit der Neuorganisation sollen geklärt – oder zumindest thematisiert und andiskutiert – werden.

- Die **Prozessbeschreibung** soll erstmals den gesamten Lebenszyklus sowie das konkrete Vorgehen hinsichtlich Webarchivierung im UAZ dokumentieren. Die Prozessbeschreibung soll sich an einem bestehenden Modell orientieren (Web Archiving Life Cycle Model vom Archive-It-Team), angepasst an die Gegebenheiten im UAZ.
- Nach einer eingehenden Diskussion von Fragen in Zusammenhang mit der Erarbeitung eines **Datenmodells** sowie bezüglich der Wahl einer möglichen **Preservation Planning-Strategie** sollen Empfehlungen für das weitere Vorgehen ausformuliert werden. Beim Datenmodell geht es dabei um die Frage, wie die archivierten Websites auf Archive-It in das digitale Langzeitarchiv des UAZ überführt werden sollen. Bezüglich Preservation Planning geht es darum, Migrations- und Emulationsstrategien zu eruieren.

Die Struktur der Arbeit gliedert sich wie folgt: Nach der vorliegenden Einleitung folgt der Abschnitt 2, der einige zentrale terminologische Bestimmungen festhält. In Abschnitt 3 folgt ein kurzer Abriss zur Geschichte und Bedeutung der Webarchivierung sowie den damit verbundenen Herausforderungen. Auf diese einführenden Abschnitte folgen die Abschnitte 4 bis 6, die den Hauptteil der Arbeit bilden. In Abschnitt 4 wird eine Prozessbeschreibung erarbeitet. Der Abschnitt 5 widmet sich der Diskussion von Fragen in Zusammenhang mit der Erstellung eines Datenmodells. Der Abschnitt 6 setzt sich mit dem Thema Preservation Planning auseinander. Schliesslich folgt in Abschnitt 7 ein Schlusswort sowie eine Zusammenstellung von institutionsunabhängigen wesentlichen Erkenntnissen, Empfehlungen und zentralen Leitfragen, die für andere Gedächtnisinstitutionen, die sich ebenfalls der Webarchivierung widmen, von Interesse sein könnten.

2 Terminologische Bestimmungen

Im Folgenden sollen einige der relevantesten Begrifflichkeiten in Zusammenhang mit der Webarchivierung erläutert werden. Vor allem die Unterscheidung zwischen Webpage und Website ist zentral. Diese beiden Begrifflichkeiten werden in der vorliegenden Arbeit fortwährend genannt. Auf die Erläuterung weiterer als der hier aufgeführten Begriffe wird – sofern sinnvoll – direkt an den entsprechenden Stellen in der Arbeit eingegangen.[3]

Crawler

Ein Crawler erkundet das Web und sammelt Daten über die vorgefundenen Inhalte. Der Crawler startet seinen Erfassungsprozess auf Basis einer Seed-Liste mit Einstiegs-URLs.[4]

Harvesting/Crawling

Beim Harvesting/Crawling werden Links verfolgt. Einzelne Dateien werden lokal gesichert und in eine Struktur überführt, welche den Gesamtkontext der Website wiedergibt.[5]

Webarchivierung

Die Webarchivierung ermöglicht das Sammeln, Archivieren und Bereitstellen von Websites. Bei diesen Websites handelt es sich um komplexe Angebote im Word Wide Web, die sich inhaltlich und technisch voneinander unterscheiden.[6]

Webpage (= Webseite)

Als eine Webpage wird eine einzelne Seite eines Internetanbieters im Web verstanden.[7] Das HTML-Dokument bildet den Rahmen einer Webpage. Es ermöglicht die Einbindung weiterer Webressourcen sowie eine Verlinkung. Die Verlinkung ermöglicht es, durch das Netz der Webpages zu navigieren.[8]

---

[3] Als ein sehr hilfreiches Instrument zum Nachschlagen von Begrifflichkeiten sei das «Glossary of Archive-It and Web Archiving Terms» genannt. Dieses Glossar ist im Archive-It Helpcenter integriert. Verfügbar unter: https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms [9.7.2024].

[4] vgl. Archive-It Help Center. Glossary. Stichwort: Crawler. Verfügbar unter: https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms [9.7.2024].

[5] vgl. Schoger/Beinert/Schmid/Donig/Eckl 2021: Folie 6.

[6] vgl. Schoger/Beinert/Schmid/Donig/Eckl 2021: Folie 6.

[7] vgl. Schweizerische Nationalbibliothek, Webarchiv Schweiz: Glossar. Version 1.7. 2024: Seite 6.

[8] vgl. KOST 2016: Seite 1, Fussnote 1.

Website (= Webauftritt)

Die Website besteht aus mehreren Webpages, die hierarchisch angegliedert sind. Die Hauptseite wird für gewöhnlich als Homepage bezeichnet.[9] Die Website umfasst die Menge aller Webpages. Diese bilden miteinander verlinkt thematisch eine Einheit. Die Webpages sind üblicherweise in einer bestimmten Domain zusammengefasst.[10]

[9] vgl. Schweizerische Nationalbibliothek, Webarchiv Schweiz: Glossar. Version 1.7. 2024: Seite 6.

[10] vgl. KOST 2016: Seite 1, Fussnote 1.

3 Kurze Übersicht zur Geschichte und Bedeutung der Webarchivierung sowie den damit verbundenen Herausforderungen

In einem kurzen Abriss soll an dieser Stelle auf die wichtigsten Meilensteine bezüglich der Webarchivierung sowie auf die damit verbundenen Herausforderungen eingegangen werden.

Weshalb ist Webarchivierung wichtig?

Die Website als historische Quelle gewinnt zunehmend an Bedeutung. Das Web stellt ein relevantes Kommunikations- und Publikationsmedium dar. Zudem wird im Web auch rechtsrelevante Information publiziert, die aus Beweisgründen für allfällige Verfahren zu sichern ist. Oft ist das Web die erste Anlaufstelle, über welche auf der Suche nach Information recherchiert wird. Teilweise werden Ressourcen im Web publiziert, die ausschliesslich in dieser Erscheinungsform verfügbar sind.

Für Julien Masanès findet die moderne Kultur im Web eine Ausdrucksform. Viele Aspekte, welche die Gesellschaft prägen, können im Web lokalisiert werden und spiegeln sich dort unter anderem in Publikationen, Debatten und sozialer Interaktion wider. Masanès betont, dass die Bewahrung des Webs eine kulturelle und historische Notwendigkeit darstellt.[11]

Wie bei allen anderen schriftlichen Erzeugnissen von öffentlich-rechtlichen Institutionen handelt es sich auch bei Websites schlicht um Unterlagen. Während früher Informationen in Akten ihren schriftlichen Niederschlag fanden, so fliessen diese gegenwärtig zunehmend in die Websites.[12] Die Website einer Behörde stellt einen integralen Bestandteil der Schriftgutproduktion dar. Die Darstellung im Web zeigt das Selbstverständnis von Behörden auf. Zudem dokumentiert das Web auch die Entwicklung der Informationsgesellschaft.[13] Die Website wird als Ergänzung zum Printmaterial mitgesichert. Auch können die archivierten Websites als Forschungsgrundlage dienen.[14]

Das Ziel der Webarchivierung besteht darin, dass die als archivwürdig bewerteten Websites für die zukünftige Nutzung durch Wissenschaft, Forschung und die breite Öffentlichkeit erhalten bleiben.[15]

---

[11] vgl. Masanès 2006: Seite 1.

[12] vgl. Geisler/Dannehl/Keitel/Wolf 2017: Seite 486.

[13] vgl. Keitel 2010: Seite 24.

[14] vgl. Mayr 2011: Folie 2.

[15] vgl. Schoger/Beinert/Schmid/Donig/Eckl 2021: Folie 6.

Kurze Übersicht zur Geschichte der Webarchivierung

Bestrebungen das Web zu archivieren gehen schon fast drei Jahrzehnte zurück. Im Jahr 1996 wurde das Internet Archive in San Francisco gegründet, welches Pionierarbeit in Sachen Webarchivierung leistete. Im Jahr 2002 veröffentlichte das Internet Archive den Crawler *Heritrix*, der auf Open Source basiert. Durch diese Software konnten Inhalte des Webs erfasst werden. Im Jahr 2006 führte das Internet Archive den Webarchiv-Service Archive-It ein. Abonniert eine Institution diesen Dienst, dann wird diese als Partnerorganisation von Archive-It bei der Erfassung, dem Aufbau und der Verwaltung von digitalen Sammlungen unterstützt. Das aus dem Crawlprozess mit *Heritrix* resultierende WARC-Format wurde im Jahr 2009 zum ISO-Standard für die Webarchivierung erklärt.[16] Auf das Dateiformat WARC wird eingehend in Abschnitt 5.1 eingegangen.

Bezogen auf den Standort Schweiz gehen die Bemühungen für die Archivierung des Webs ebenfalls bereits einige Zeit zurück. Im Jahr 2001 wurde durch die Schweizerische Nationalbibliothek (NB) das Projekt «e-Helvetica» lanciert. Das Ziel dieses Projektes bestand darin, eine Sammlung von Online-Helvetica anzulegen sowie ein digitales Langzeitarchiv aufzubauen, um den Erhalt dieser Sammlung sichern zu können.[17] Seit 2008 wird die Sammlung unter dem Namen «Webarchiv Schweiz» geführt und gepflegt. Das Webarchiv Schweiz basiert auf einem Zusammenarbeitsmodell zwischen der NB und den Kantonsbibliotheken. Die Kantonsbibliotheken sind zuständig für die Ermittlung, Anmeldung und Verzeichnung der Websites und die NB wiederum ist verantwortlich für die Einsammlung, Erschliessung, Archivierung sowie Bereitstellung der Websites. Das Webarchiv Schweiz umfasst eine Sammlung von landeskundlich relevanten Websites. Die Aufgabe, welche das Webarchiv Schweiz wahrnimmt, besteht darin, dass das geistige Online-Kulturgut der Kantone und der Schweiz archiviert und interessierten Institutionen und Personen zur Verfügung gestellt wird.[18]

Neben dem Internet Archive und der NB widmen sich auch zunehmend individuelle Gedächtnisinstitutionen wie Archive und Bibliotheken der Webarchivierung. Bibliotheken erfüllen damit ihren Sammelauftrag. Für Archive stellt die Archivierung des Webs eine wichtige Ergänzung der bisherigen Überlieferungspraxis dar. Im Web findet sich zentrale Information, die oft in konziser verdichteter Form vorliegt. Auch sind häufig Ressourcen online verfügbar, die ausschliesslich in dieser Form existieren und auf anderem Wege

[16] vgl. Bragg/Hanna 2013: Seite 1.

[17] vgl. Schweizerische Nationalbibliothek, Webarchiv Schweiz: Grundlagenpapier 2024: Seite 3.

[18] vgl. Schweizerische Nationalbibliothek, Webarchiv Schweiz: Grundlagenpapier 2024: Seite 4–5.

nicht in das Archiv gelangen, sodass die Archivierung des Webs einen erheblichen Mehrwert schafft.

Herausforderungen

Eine grosse Herausforderung besteht in der Volatilität des Webs. Diese stellt eine Gefährdung für das digitale kulturelle Gedächtnis dar. Die Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST) weist in ihrer Studie zur Webarchivierung darauf hin, dass sich der Inhalt auf einer Webpage sehr rasch verändern oder eine Seite ganz von der Bildfläche verschwinden kann. So kann es sein, dass ein späterer Zugriff auf die einmal gefundene Information innert kürzester Zeit verunmöglicht wird.[19] Die durchschnittliche Lebensdauer von Webpages ist kurz, gemäss Walter Nagel beträgt diese lediglich ca. 60–90 Tage.[20] Die Information auf dem Web muss zeitgerecht und mit einem vertretbaren Aufwand heruntergeholt und in einer langfristig erhaltbaren Form gesichert werden.[21] Eine weitere Herausforderung besteht darin, dass das Web einen immensen Umfang an Daten aufweist, der rapide anwächst und dadurch unübersichtlich ist.[22]

Brunner/Debenath weisen auf eine weitere Problematik hin. Sie machen darauf aufmerksam, dass sich die Struktur, der Lebenszyklus und die Verfügbarkeit von Netzressourcen sehr stark von dokumentartigem Archivgut unterscheiden. Idealerweise bereits in einem Archiv zur Verfügung stehende Komponenten wie ein Archivinformationssystem (AIS) sowie ein Ingestserver/Repository können zwar zur Webarchivierung genutzt werden, doch sind zusätzliche Tools für das Harvesting, die Indexierung und die Benutzung erforderlich.[23]

Herausfordernd für eine Gedächtnisinstitution kann es auch sein, dass es bezüglich der Webarchivierung nicht eine einzige Lösung gibt, um die Ressourcen auf dem Web abzuziehen und zu sichern. Es existieren verschiedene Methoden. Ein Archiv kann sich für eine einzelne Methode entscheiden (beispielsweise das Crawling) oder auch verschiedene Methoden miteinander kombinieren. Das Erzbischöfliche Archiv Freiburg verfolgt beispielsweise einen multimodalen Ansatz, bei welchem es je nach Inhalten eine

[19] vgl. KOST 2016: Seite 1.

[20] vgl. Walter Nagel GmbH & Co. KG. Webarchivierung. Verfügbar unter: https://www.walternagel.de/webarchivierung#:~:text=Die%20durchschnittliche%20%E2%80%9ELebensdauer%E2%80%9C%20von%20Webseiten,sichern%20damit%20alle%20relevanten%20Inhalte. [9.7.2024].

[21] vgl. Locher 2002/03: Seite 113.

[22] vgl. Locher 2002/03: Seite 112.

[23] vgl. Brunner/Debenath 2018: Seite 118.

Webpage crawlt, Datenbankschnitte übernimmt, Screenshots erstellt, eine Quellcodekopie anfertigt oder auch Sequenzen einer Seite abfilmt.[24]

Auch bezüglich Preservation Planning gibt es ungeklärte Fragen. Es ist zentral, sich früh Gedanken zum Preservation Planning zu machen und mögliche Lösungswege anzudenken. Auf diese Weise können wichtige Weichen gestellt werden, die langfristig Auswirkungen haben. Auf die Thematik Preservation Planning wird ausführlich in Abschnitt 6 eingegangen.

[24] vgl. Franzky 2024: Folie 4.

4 Prozessbeschreibung

Eine Prozessbeschreibung zur Webarchivierung im UAZ liegt bis anhin nicht verschriftlicht vor. Im Rahmen der Neuorganisation des Webarchivs soll die Gelegenheit genutzt werden, um den Prozess schriftlich festzuhalten. Die Prozessbeschreibung soll sämtliche Schritte von der Bewertung bis zum Preservation Planning berücksichtigen. Als Grundlage für den Prozess soll auch auf strategische Überlegungen eingegangen werden. Der Prozess soll ausreichend detailliert ausgeführt sein, damit er anhand der Prozessbeschreibung nachvollzogen werden kann. Ziel ist es, dass die Prozessbeschreibung als Leitlinie dienen kann, indem sie einen wertvollen Überblick über den Gesamtprozess schafft und beispielsweise auch für neue Mitarbeitende als Einführung in die Webarchivierung zur Verfügung steht. Die Dokumentation soll es zudem ermöglichen, potenzielle Schwachstellen ausmachen zu können, damit der Prozess bei Bedarf optimiert werden kann.

In Abschnitt 4.1 wird auf den Prozess bis Ende 2023 eingegangen. In Abschnitt 4.2 werden die Neuerungen nach der Neuorganisation ab 2024 vorgestellt. In Abschnitt 4.3 wird das Web Archiving Life Cycle Model (WALCM) vom Archive-It Team präsentiert, das als Grundlage für die Darstellung des Prozesses im UAZ dienen soll. Im darauffolgenden Abschnitt 4.4 wird die Prozessbeschreibung für das UAZ dargestellt. In Abschnitt 4.5 schliesslich wird auf weiterführende Gedanken und offene Fragen eingegangen.

4.1 Der Prozess bis Ende 2023

An dieser Stelle sollen zunächst ein paar allgemeine, einführende Details zur Website der UZH genannt werden. Die UZH verfügt seit 1993 über eine eigene Website.[25] Ab 2004 wurde die Website der Universität mit dem Content Management System (CMS) *Lenya* erstellt. Dieses wurde im Jahr 2015 durch das CMS *Magnolia* abgelöst.[26] Mitte April 2023 konnte das letzte Redesign des Webauftrittes abgeschlossen werden. Ziel war es, die Website nutzerfreundlicher zu machen und mit gut lesbarer Schrift zu gestalten,

[25] vgl. Messner, UZH Archiv Vitrine 2015. Verfügbar unter: https://www.archiv.uzh.ch/de/vitrine/aeltere_beitraege.html#Die_Dokumentation_universit%C3%A4rer_Publikationst%C3%A4tigkeit_im_Medienwandel [31.8.2024].

[26] vgl. (UAZ) PUB.010 «Webauftritt der Universität». Registerkarte Kontext. Verfügbar unter: https://mobile.cmistar.ch/webclients-r22/uzh/#/content/ec3f72cc92f94437bf2d28f488d43b50 [9.7.2024].

sodass die Website insbesondere auch auf mobilen Geräten optimal dargestellt werden kann.[27]

Das UAZ hat bereits im Jahr 2012 mit der Webarchivierung begonnen. Mitte 2012 wurden vom UAZ die ersten Websites der UZH gesichert. Unabhängig vom UAZ gab es aber auch schon vor 2012 Bestrebungen, um die Website der UZH zu sichern. Diese frühesten Archivierungsbemühungen erfolgten durch die Zentralbibliothek Zürich (ZB) im Rahmen von Webarchiv Schweiz. Wie in Abschnitt 3 erwähnt, arbeitet Webarchiv Schweiz eng mit den kantonalen Bibliotheken zusammen. Im Falle des Kantons Zürich schlägt die ZB der Nationalbibliothek vor, welche Websites als relevant zu deklarieren sind. 2010 erkundigte sich Silvia Bolliger, die damalige Leiterin des UAZ, bei der ZB ob Websites der UZH über das Webarchiv Schweiz bereits gesichert werden. Dies war zu diesem Zeitpunkt noch nicht der Fall. Die ZB hatte daraufhin die von Bolliger genannten zentralsten Domains der UZH zur Archivierung bei Webarchiv Schweiz angemeldet.[28]

Um der Relevanz des Webauftritts gerecht zu werden, entschied sich das UAZ späterhin dazu, die Websites zukünftig selbständig und unabhängig von Webarchiv Schweiz zu sichern. Eine selbständige Archivierung sollte es ermöglichen, dass die Auswahl der zu sichernden Websites unabhängig von der ZB und NB durch das UAZ getroffen werden konnte. Zudem schaffte dieses Vorgehen die Möglichkeit, dass das UAZ die vom Web abgezogenen Daten im eigenen Archiv sichern und darüber verfügen konnte. Damit eine solche selbständige Sicherung des Webauftritts realisiert werden konnte, hatte der Archivinformatiker Markus Kandlbinder verschiedene Tools geprüft und mit diversen Fachleuten Rücksprache gehalten, um die ideale Lösung für das UAZ zu finden. Basierend auf den Entscheid vom 5. November 2013 wurde die Website mit zwei verschiedenen Tools gesichert: mit dem Tool *Adobe Acrobat Pro* von der Firma Adobe Systems Incorporated und mit dem Tool *Offline Explorer* von der Firma MetaProducts. Diese beiden Tools wurden für die Webarchivierung miteinander kombiniert, da beide Vor- und Nachteile aufwiesen. *Adobe Acrobat Pro* lieferte das für das Archiv am besten geeignete PDF-Format. Ein Nachteil war es allerdings, dass nur kleinere Websites mit diesem Tool verarbeitet werden konnten. Da *Acrobat* das PDF vollständig im RAM erstellt, kam es bei umfangreicheren Seiten zu Memory-Problemen und damit zum Absturz des *Acrobat*. Aufgrund dieser Problematik wurden sämtliche zu archivierenden Webseiten gleichzeitig mit

[27] vgl. UZH News vom 23.3.2023 «Ein frischer Look für die UZH-Webseiten. Verfügbar unter: https://www.news.uzh.ch/de/articles/news/2023/neuer-webauftritt.html [9.7.2024].

[28] Internes Dokument (CMI G 2013-76). In der vorliegenden Arbeit wird laufend auf UAZ-interne Geschäfte im GEVER *CMI* verwiesen. Die Geschäftsnummer ist jeweils in Klammer am Ende der jeweiligen Fussnote eingetragen. Bei den internen Dokumenten handelt es sich um verschiedene Berichte, ein Handbuch sowie Korrespondenz.

dem *Offline Explorer* gesichert. Mit dem *Offline Explorer* konnten die Seiten im Original-HTML-Code gespeichert werden. Die Weblinks wurden in lokale Links übersetzt, sodass es möglich war, die gesicherten Websites offline aufzurufen. Durch die parallele Nutzung von *Acrobat Adobe Pro* und *Offline Explorer* konnte sichergestellt werden, dass die Website zusätzlich zum PDF-Format auch komplett als HTML mit sämtlichen Medieninhalten gesichert werden konnte.[29]

Um den manuellen Aufwand für das Bedienen der Programme und das Harvesting zu reduzieren, wurde im Jahr 2015 untersucht, ob und wie weit sich die beiden Programme *Adobe Acrobat Pro* und *Offline Explorer* automatisieren lassen.[30] Die Untersuchung ergab das folgende Resultat: Während sich der *Offline Explorer* weitgehend automatisieren liess und die Website in ihrer ursprünglichen Form zur Verfügung stellte, musste der *Adobe Acrobat* manuell bedient werden und wandelte die Website in eine PDF-Datei um. Das Handling von *Adobe Acrobat Pro* wurde als sehr aufwendig bewertet und es führte nicht zum erwünschten Ergebnis. Auch weitere Gründe sprachen dafür, zukünftig von einer Nutzung von *Adobe Acrobat Pro* abzusehen. Als problematisch beurteilt wurde die weiter oben im Abschnitt bereits erwähnte Memory-Problematik sowie die Unmöglichkeit des Herausfilterns von nicht gewünschten Bestandteilen der Website (mobile Seiten, englische Seiten bei Mehrsprachigkeit etc.). Weiterhin als Problem wurde thematisiert, dass multimediale Inhalte, wie Filme, Audiodateien oder auch Scripts und andere downloadbare Dateien (Programme, ZIP Files) nicht in das PDF integriert werden. Desweiteren wurde es als problematisch beurteilt, dass die in ein PDF umgewandelte Seite nicht dem Original der Website entspricht. Die signifikanten Eigenschaften, welche den Inhalt, das Layout, die Struktur sowie die Funktionalität einer Website umfassen, gehen durch eine Sicherung als PDF teilweise oder gänzlich verloren.[31] Aufgrund der Restriktionen, des Arbeitsaufwandes mit dem *Adobe Acrobat Pro* und den vorab erwähnten Problemen wurde in einem Entscheid vom 26. Juni 2015 die damals vorhandene Doppelspurigkeit beim Archivieren der Website aufgehoben. Ab diesem Zeitpunkt wurde auf ein Harvesting von Websites mit dem *Adobe Acrobat Pro* verzichtet und ganz auf den *Offline Explorer* gesetzt. Die bisher mit dem *Adobe Acrobat Pro* erstellten PDF-Dateien wurden ausnahmslos gelöscht.[32]

[29] Internes Dokument (CMI G 2012-240).

[30] Internes Dokument (CMI G 2012-240).

[31] Auf die signifikanten Eigenschaften wird in Abschnitt 6.1 etwas eingehender eingegangen.

[32] Internes Dokument (CMI G 2012-240).

Ab 2012 wurde die Hauptseite der UZH in einem halbjährlichen Rhythmus archiviert. Bei den Seiten von ausgewählten Abteilungen der Zentralen Dienste und den Fakultäten erfolgte eine Archivierung in einem Jahres- oder Zweijahresrhythmus. Ausserdem mitberücksichtigt wurden auch Websites von Kompetenzzentren und von den universitären und nationalen Forschungsschwerpunkten. Die Wahl des Crawl-Intervalls hing mit der begrenzten Kapazität zusammen. Zu dieser Zeit gehörte der Betrieb des Webarchivs in den Zuständigkeitsbereich einer einzelnen Person. Eine häufigere Sicherung der Websites war damals aufgrund des Aufwandes nicht möglich. Die Sicherung im Zweijahresrhythmus wurde aber als vertretbar beurteilt, da eine Abbildung der massgeblichen Veränderungen der Websites auf diese Weise trotzdem gewährleistet war. Am 6. September 2021 erfolgte eine Neubewertung bezüglich der Webarchivierung. Die Neubewertung beruhte darauf, dass neu auch die Seiten der an Fakultäten angegliederten Institute, Seminare und Kliniken bei der Archivierung mitberücksichtigt werden sollten, die bis dahin aus Kapazitätsgründen nicht abgezogen werden konnten. Die Erweiterung auf diese Websites wurde als sehr wertvoll eingestuft, da viele Informationen auf diesen Seiten zusammenkommen, die bis zu diesem Zeitpunkt nicht systematisch archiviert werden konnten. Nach der Neubewertung im September 2021 wurden die Subdomänen der UZH-Website mit wenigen Ausnahmen (Hauptdomäne sowie Studiengänge, Fakultäten, Seiten der Kommunikation, besondere Anlässe) standardmässig in einem Zweijahresrhythmus abgezogen. Die Hauptdomäne wurde auch nach der Neubewertung weiterhin zweimal jährlich gesichert, da auch Informationen zu ausserordentlichen Ereignissen wie beispielsweise Covid19 darin enthalten sein können. Für die Websites der Fakultäten wurde ab 2021 neu eine jährliche Archivierung gewählt, um die Studienprogramme und Studienordnungen zu dokumentieren. Für die Website der UZH Alumni und das längerfristige Projekt Stadtuniversität wurde neu ein Zweijahresrhythmus angesetzt. Websites von zeitlich begrenzten Anlässen (beispielsweise Jubiläen, Ausstellungen, etc.) sollten wie bis anhin einmalig abgezogen werden.[33]

Die abgezogenen Seiten wurden in der vorliegenden HTML-Struktur (im ZIP-Format) in das digitale Langzeitarchiv des UAZ überführt. In *CMI AIS* ist der archivierte Webauftritt der Abteilung «Publikationen mit offiziellem Charakter» angegliedert und unter dem Bestand «PUB.010 Webauftritt der Universität» erschlossen. Insgesamt liegen bis Ende 2023 Daten im Umfang von 1368.72 GB vor.[34]

[33] Internes Dokument (CMI G 2012-240).

[34] (UAZ) PUB.010 «Webauftritt der Universität». Verfügbar unter: https://mobile.cmistar.ch/webclients-r22/uzh/#/content/ec3f72cc92f94437bf2d28f488d43b50 [1.8.2024].

4.2 Neuerungen ab Januar 2024

2023 entschied sich das UAZ für eine Zusammenarbeit mit dem Internet Archive durch die Nutzung des Webarchiv-Service Archive-It. Die archivwürdigen Seiten der UZH werden nun seit Januar 2024 mit dem Webarchiv-Service Archive-It gecrawlt und im Dateiformat WARC gesichert.

Die Neuorganisation des Webarchivs wurde aus verschiedenen Gründen durchgeführt. Von grösseren Firmen in der Privatwirtschaft wird der *Offline Explorer* eingesetzt, doch in der Archivwelt stand das UAZ mit dieser Lösung alleine da.[35] Mit der Nutzung von Archive-It ist das UAZ nun eingebettet in die Welt der Archive, Bibliotheken, Museen und Universitäten. Beim Internet Archive handelt es sich um eine gemeinnützige Stiftung, die das grösste Webarchiv überhaupt betreut. Die Technologie basiert auf Open Source und die Community trägt zur Weiterentwicklung bei.[36] Ein weiterer Vorteil, der aus der Nutzung von Archive-It resultiert, ist, dass die Benutzerfreundlichkeit erhöht wird. Die über Archive-It gesicherten Zeitschnitte sind online abrufbar über die vom Internet Archive in Open Source zur Verfügung gestellte *Wayback Machine*. Die Lösung mit dem *Offline Explorer* war nicht benutzerfreundlich. Für eine Benutzung mussten die Mitarbeitenden des UAZ den entsprechenden Zeitschnitt aus dem digitalen Langzeitarchiv herausholen. Nach dem erfolgten Download des DIP wurde im Verzeichnis jeweils eine Datei mit dem Namen «Default.html» erzeugt. Die Website konnte durch einen Doppelklick auf diese Datei offline betrachtet werden.[37] Wenn Archivnutzende mehrere Zeitschnitte miteinander vergleichen wollten, dann mussten alle diese Zeitschnitte separat aus dem digitalen Langzeitarchiv heruntergeladen und für die Benutzung durch die Archivmitarbeitenden vorbereitet werden. Eine Ansicht über einen Zeitstrahl war nicht möglich.

Ein weiterer Grund, der für die Neuorganisation sprach, war die Unsicherheit bezüglich der Zukunft dieser bisherigen Speicherart. Als Standardformat für die Webarchivierung hat sich das Dateiformat WARC durchgesetzt. Viele Archive nutzen dieses Format. *Offline Explorer* mit dem HTML-Format entspricht nicht dem Standard.[38] Auch das Containerformat ZIP wurde als problematisch beurteilt. Das UAZ ingestierte die HTML-Daten in

[35] Eine Ausnahme bildet das Sächsische Staatsarchiv. Dieses Archiv arbeitet seit 10 Jahren mit dem Offline Explorer und wird sicher vorerst auch dabeibleiben. Vgl. Kortyla 2024: Folie 7 sowie Austausch mit Stephanie Kortyla (Sächsisches Staatsarchiv) vom 20. Juni 2024. In der vorliegenden Arbeit wird laufend auf Austausch/Korrespondenz mit verschiedenen Fachpersonen verwiesen. Im Anhang der Arbeit ist eine Tabelle mit einer Gesamtübersicht über die geführten Gespräche beigefügt.

[36] Internes Dokument (CMI G 2022-137).

[37] Internes Dokument (CMI G 2012-240).

[38] Austausch mit Markus Kandlbinder (Archivinformatiker UAZ) vom 13. Juni 2024.

komprimierter Form mit dem Containerformat ZIP in das digitale Langzeitarchiv. Das Containerformat ZIP wurde nicht als archivtaugliches Format bezeichnet und die Frage nach einer Lösung für die langfristige Erhaltung blieb dadurch offen. Durch die Nutzung von Archive-It kann nun mit dem Standardformat WARC gearbeitet werden.

Ein weiterer Vorteil, der aus der Nutzung von Archive-It resultiert, ist die Einsparung von Speicherplatz. Mit dem *Offline Explorer* wurden bis anhin immer Mastercrawls durchgeführt, also ein Gesamtabzug der Website. Archive-It ermöglicht es, dass nach einem Erstabzug (Gesamtabzug) anschliessend Deltacrawls durchgeführt werden können. Bei diesen Deltacrawls wird ausschliesslich abgezogen, was seit dem erfolgten Erstabzug neu hinzugekommen ist, respektive was sich in der Zwischenzeit verändert hat.[39] Die Möglichkeit mit solchen Deltacrawls zu arbeiten, kann auch das gewählte Crawl-Intervall beeinflussen. Die Deltacrawls schaffen die Möglichkeit, dass häufiger Crawls durchgeführt werden können, die sich automatisieren lassen, aber nicht mehr Speicherplatz benötigen im Vergleich zum bisherigen Vorgehen.

Geplant ist es, dass die auf Archive-It gesicherten WARC Files als Sicherheitskopien heruntergeladen und im digitalen Langzeitarchiv des UAZ gespeichert werden. Thematisiert wurde die Möglichkeit, die Altdaten, die sich (im ZIP-Format) im digitalen Langzeitarchiv des UAZ befinden, in WARC Files umzuwandeln und in den Archive-It Account des UAZ zu überführen, sodass den Archivnutzenden sämtliche Daten online über die *Wayback Machine* zur Verfügung stünden.[40] Davon wurde aber abgesehen. Das Internet Archive hat in der Vergangenheit unabhängig vom UAZ Abzüge von der UZH Website erstellt. Diese Crawls, die das Internet Archive bis 2023 durchgeführt hat, werden in den Archive-It Account des UAZ überführt (via Waybackfill Service). Denkbar ist es, dass allenfalls noch Lücken geschlossen werden können mit den Altdaten des UAZ. Dies in dem Falle, wenn zentrale Websites fehlen würden, von welchen also keine Abzüge des Internet Archive existieren.[41]

[39] Internes Dokument (CMI G 2022-137).

[40] Interne Dokumente (CMI G 2022-137).

[41] Austausch mit Inge Moser (stv. Leiterin UAZ) vom 5. Juli 2024.

4.3 Das Web Archiving Life Cycle Model (WALCM) vom Archive-It Team als Grundlage für die Darstellung des Prozesses im UZH Archiv

Im Beitrag «The Web Archiving Life Cycle Model» aus dem Jahr 2013 macht das Archive-It Team darauf aufmerksam, dass bis anhin ein gemeinsames Modell zur Webarchivierung fehlt. Das Modell von Archive-It schliesst diese Lücke und wirkt dem Mangel an Best Practices entgegen. Auch soll das Modell das Bewusstsein für die Relevanz der Webarchivierung erhöhen. Das Web Archiving Life Cycle Model (WALCM) hält gängige Arbeitsabläufe schriftlich fest und schafft ein Modell, auf das sich Institutionen beziehen können. Auf dieser Basis können Institutionen ihre Webarchivierungsvorhaben lancieren oder das bereits vorhandene Vorgehen optimieren. Archive-It ist der führende Web-Archivierungsdienst in der Community und hat das WALCM auf der Grundlage seiner Arbeit und aus der Erfahrung im Austausch mit verschiedenen Gedächtnisinstitutionen weltweit entwickelt.[42]

Die Prozessbeschreibung für das UAZ in Abschnitt 4.4 basiert auf dem WALCM. Dieses Modell wurde als Vorlage gewählt, da es sehr umfassend ist und alle wesentlichen Prozessschritte der gängigen Praxis ausführlich erläutert. Von Vorteil ist auch, dass es sich dabei um ein übergeordnetes Konzept handelt, welches nicht auf spezifische Software für die Ausführung in der Praxis verweist. Da die Publikation bereits etwas mehr als 10 Jahre zurückliegt (Publikationsjahr 2013) wurde abgeklärt, ob das Modell als Ganzes oder allenfalls Bestandteile davon veraltet sein könnten. Der Austausch mit diversen Fachleuten hat ergeben, dass das Modell weiterhin als aktuell beurteilt werden kann, da sich über die letzten 10 Jahre am Prozess nichts Grundlegendes geändert hat.[43] Auch eine Fachperson von Archive-It wurde dazu befragt. Gemäss Kody Willis (Ansprechpartner des UAZ bei Archive-It) entspricht das Modell nach wie vor dem neuesten Stand und es sind derzeit auch keine Revisionsarbeiten am Modell in Planung. Wirklich vergleichbare alternative Modelle sind nicht bekannt. Es gibt aber vereinzelt Gedächtnisinstitutionen, die ihre eigenen spezifischen Workflows online zur Verfügung stellen.[44] Die Bibliothèque nationale de France (BnF) hat beispielsweise ihren Workflow zur Einsicht für andere Gedächtnisinstitutionen online publiziert.[45] Der Beitrag der BnF ist sehr ausführlich und gehaltvoll. Da im Workflow der BnF aber auf spezifische Software und Tools

[42] vgl. Bragg/Hanna 2013: Seite 1–2.

[43] Unter anderem wurde die Aktualität des Modells thematisiert beim Austausch mit Angela Gastl (Webarchiv ETH) vom 25. März 2023 sowie beim Zwischengespräch zur Masterarbeit mit Tobias Wildi und Ana Petrus vom 10. Juni 2024.

[44] Korrespondenz mit Kody Willis (Archive-It) vom 4. Juni 2024.

[45] vgl. Le Follic/Stirling/Wendland 2012.

eingegangen wird, wurde das WALCM, welches sich neutraler und umfassender präsentiert, favorisiert, um als Vorlage für eine Prozessbeschreibung im UAZ zu dienen. Auch die Grafik, mit welcher das Archive-It Team beim WALCM arbeitet, erscheint übersichtlicher als diejenige Grafik der BnF.

Im WALCM wird auf die verschiedenen Schritte im Prozess eingegangen. Die einzelnen Schritte sind aber nicht gesondert, sondern in der Gesamtheit zu betrachten. Es kommt zu Überschneidungen. Die Gestalt des Modells ist kreisförmig, um den sich wiederholenden Charakter der Arbeitsschritte im Lebenszyklus anzudeuten.[46]

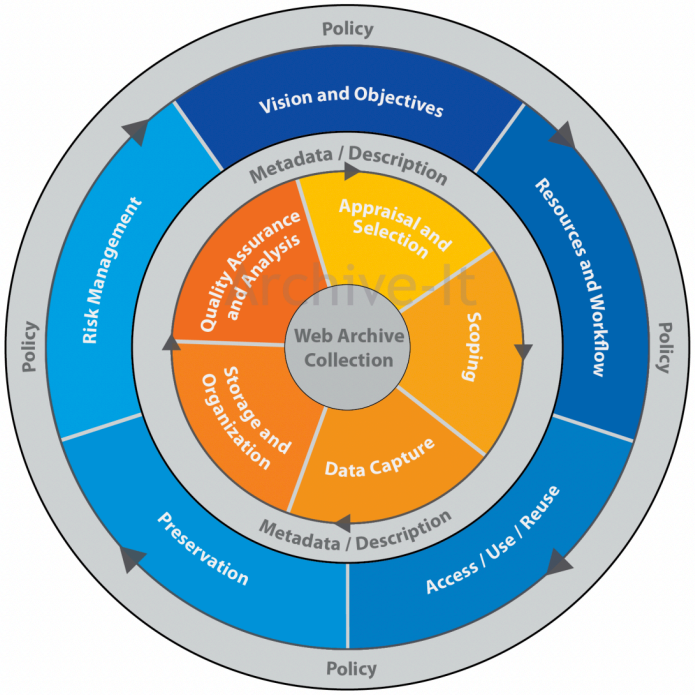

Abbildung 1: Web Archiving Life Cycle Model (Bragg/Hanna 2013: Seite 3)

In der voranstehenden grafischen Darstellung des WALCM sind zwei graue Umrahmungen zu sehen. Der graue Rahmen **Policy** bildet die äusserste Ebene des Lebenszyklus und umfasst diesen somit in seiner Gesamtheit. Diese Ebene beinhaltet grundsätzliche Strategieentscheidungen, welche sich auf den gesamten Prozess auswirken. In der Grafik ist dieser Umstand als durchgehendes graues Band entsprechend visualisiert. Der graue Rahmen **Metadata/Description** befindet sich in der Grafik näher im Kreisinnern. Auch hier wurde ein durchgehender Rahmen gewählt anstelle eines einzelnen Segmentes im Lebenszyklus. Diese Darstellung wurde so gewählt, da auch die Metadaten während des gesamten Prozesses relevant sind und nicht auf einen einzelnen bestimmten

[46] vgl. Bragg/Hanna 2013: Seite 2.

Prozessschritt reduziert werden können. Die Erstellung, der Import und der Export von Metadaten ist ein fortlaufender Prozess, der mit einer Reihe anderer Aktivitäten im Lebenszyklus einhergeht.[47] Der graue Kreis im Zentrum der Grafik stellt die **Web Archive Collection** dar. Dabei handelt es sich um den archivierten Webinhalt selbst, der das Resultat aller vorangehenden Prozessschritte und somit das Herzstück des Prozesses darstellt.[48]

Der äussere blaue Rahmen bestehend aus den Abschnitten **Vision and Objectives**, **Resources and Workflow**, **Access/Use/Reuse**, **Preservation** und **Risk Management** umfasst Entscheidungen auf höchster Ebene, die eine Institution bei der Einrichtung und Verwaltung ihres Webarchivs treffen muss. In den Prozessschritten des blauen Rahmens werden Ziele definiert, Ressourcen ausgemacht, Entscheidungen über den Zugang zu den Daten getroffen, Überlegungen zum Preservation Planning angestellt sowie das Risikomanagement thematisiert. Während der äussere blaue Rahmen die Entscheidungen auf höchster Ebene repräsentiert, umfasst der innere orange Rahmen mit den Abschnitten **Appraisal and Selection**, **Scoping**, **Data Capture**, **Storage and Organization** sowie **Quality Assurance and Analysis** die praktischeren Aufgaben. In den Prozessschritten des orangen Rahmens wird festgelegt, welche konkreten Websites gesammelt werden sollen. Der Umfang wird eingeschränkt und es wird definiert, wie häufig Crawls durchzuführen sind. Zudem wird entschieden auf welchem Speicher die Daten gesichert werden sollen und wie die Qualitätskontrolle zu erfolgen hat.[49]

Nachfolgend sind die jeweils zentralen Fragen zu den einzelnen Prozessschritten des WALCM zusammengetragen. Diese Fragen werden im nachfolgenden Abschnitt 4.4 spezifisch für die Verhältnisse im UAZ beantwortet.

Die beiden grauen Rahmen

- **Policy**[50]
    - Braucht es eine neue Policy spezifisch für die Webarchivierung?
    - Kann eine bereits vorhandene Policy mit den neuen Sammelaktivitäten ergänzt werden?

[47] vgl. Bragg/Hanna 2013: Seite 3.

[48] vgl. Bragg/Hanna 2013: Seite 5.

[49] vgl. Bragg/Hanna 2013: Seite 4–5.

[50] vgl. Bragg/Hanna 2013: Seite 3.

- **Metadata/Description**[51]
    - Wo überall (Archive-It, digitales Langzeitarchiv, AIS) und in welchem Detail sollen Metadaten erfasst werden?
    - Auf welcher Ebene sollen Metadaten in Archive-It erfasst werden (Collection, Seed, Document)?

Äusserer blauer Rahmen

- **Vision and Objectives** [52]
    - Was ist das Ziel der Webarchivierung?
    - Wie hängt die Webarchivierung mit dem übergeordneten Auftrag des Archivs zusammen?
    - Welche Fragen/Bedürfnisse sollen durch ausgewählte Collections beantwortet werden? Welche Themen sollen abgedeckt werden durch die Collections?
- **Resources and Workflow**[53]
    - Welche Ressourcen können für die Errichtung und den Betrieb eines Webarchivs genutzt werden (unter anderem: Finanzen, Personal)?
    - Welche (allenfalls bereits vorhandenen) Arbeitsabläufe können adaptiert werden und welche Arbeitsabläufe müssen neu definiert werden?
- **Access/Use/Reuse**[54]
    - Sollen die Inhalte des Webarchivs öffentlich zugänglich sein? Falls ja: Wie kann auf das Webarchiv zugegriffen werden?
    - Sollen die Collections beworben werden? Falls ja: Wie soll dies erfolgen?
- **Preservation**[55]
    - Welche Preservation Planning-Strategie soll verfolgt werden?
    - Ist es ausreichend, sich auf das Internet Archive für die Erhaltung der WARC Files und der dazugehörenden Metadaten zu verlassen oder sollen

[51] vgl. Bragg/Hanna 2013: Seite 20–21.

[52] vgl. Bragg/Hanna 2013: Seite 5–7.

[53] vgl. Bragg/Hanna 2013: Seite 8–12.

[54] vgl. Bragg/Hanna 2013: Seite 12–16.

[55] vgl. Bragg/Hanna 2013: Seite 17–18.

die Daten zusätzlich in ein eigenes digitales Langzeitarchiv überführt werden?

- **Risk Management**[56]
    - Soll Erlaubnis für die Archivierung bei den Betreibern der Website eingeholt werden? Falls ja: Inwiefern soll dies realisiert werden?
    - Ist das Einholen einer Erlaubnis überhaupt notwendig? Kann ein Archiv sich darauf berufen, dass es über das Recht und den Auftrag verfügt, öffentlich zugängliche Inhalte im Live-Web zu sichern?
    - Dürfen aufgrund des Archivauftrags robots.txt Files ignoriert werden, damit die archivierte Website vollständig gesichert und wiedergegeben werden kann?

Innerer oranger Rahmen

- **Appraisal and Selection**[57]
    - Welche spezifischen Websites sollen gesichert werden (Auswahl der URLs)?
    - Ist es sinnvoll, neben der übergreifenden Webpräsenz auch Sammlungen zu bestimmten Themen zu erstellen?
- **Scoping**[58]
    - Sollen ganze Websites archiviert werden oder nur bestimmte Bestandteile davon?
    - Zu welchem Zeitpunkt im Prozess soll der Umfang eingeschränkt werden (bevor die Website abgezogen wurde oder nach dem Harvest als Bestandteil der Qualitätsprüfung)?
    - Soll ein Tool verwendet werden, welches limitiert wie viel von einer Seite gecrawlt wird?
    - Soll die Dauer eines Crawls limitiert werden?

[56] vgl. Bragg/Hanna 2013: Seite 18–20.

[57] vgl. Bragg/Hanna 2013: Seite 22–23.

[58] vgl. Bragg/Hanna 2013: Seite 23–24.

- Soll der Crawl auf ein bestimmtes Format eingeschränkt werden (beispielsweise PDF)?

- **Data Capture**[59]
  - Wie häufig und zu welchem Zeitpunkt sollen Crawls erfolgen?
  - In welchen Fällen soll ein Crawl abgebrochen werden?
  - Sollen für unterschiedliche URLs unterschiedliche Crawl-Intervalle angesetzt werden?
  - Kann eine Zeitplanung aufgesetzt werden, wann und in welchem Intervall welche Websites gecrawlt werden sollen, sodass von einem Automatismus profitiert werden kann?
  - Soll anstelle eines produktiven Crawls zunächst ein Test Crawl erfolgen, damit sicherheitshalber überprüft werden kann, was alles von einer Website abgezogen wird, bevor die tatsächliche Sicherung erfolgt?

- **Storage and Organization**[60]
  - Wie sollen die archivierten Daten vorübergehend und langfristig aufbewahrt werden?

- **Quality Assurance and Analysis**[61]
  - Wie soll die Überprüfung bezüglich Qualität und Vollständigkeit der gesammelten Daten erfolgen?
  - Erfolgt die Qualitätskontrolle primär durch das Studium der Reports oder eher durch eine visuelle Überprüfung der gesicherten Website im Vergleich mit der Website auf dem Live-Web? Oder sollen beide Möglichkeiten miteinander kombiniert werden?
  - Soll vorwiegend bei der Erstellung von neuen Collections und beim Hinzufügen neuer Seeds in bestehende Collections eine eingehende Qualitätskontrolle erfolgen? Kann anschliessend der Aufwand etwas reduziert werden (beispielsweise auf Stichproben), sobald der Erfahrungsschatz angewachsen ist?

[59] vgl. Bragg/Hanna 2013: Seite 25–26.

[60] vgl. Bragg/Hanna 2013: Seite 4.

[61] vgl. Bragg/Hanna 2013: Seite 26–27.

4.4 Der Prozess der Webarchivierung im UZH Archiv

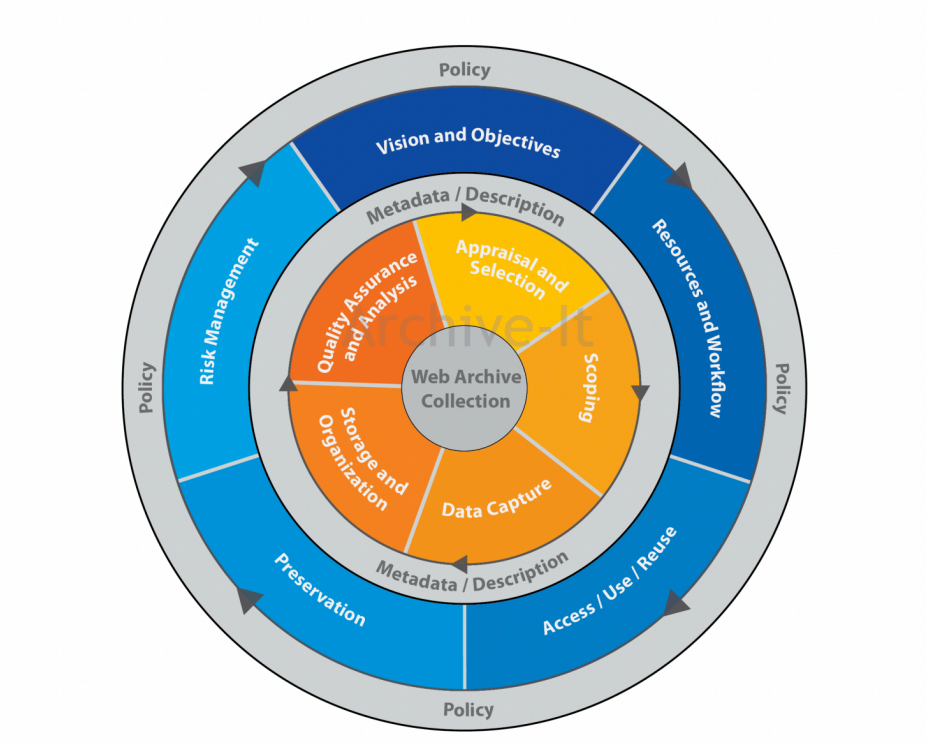

Abbildung 2: Web Archiving Life Cycle Model (Bragg/Hanna 2013: Seite 3)

Im vorliegenden Abschnitt wird auf der Basis des WALCM eine Prozessbeschreibung für das UAZ dargestellt. Zunächst erfolgt eine Erläuterung bezüglich den beiden grauen Rahmen **Policy** und **Metadata/Description**. Anschliessend wird auf die übergeordneten Schritte im äusseren blauen und zuletzt auf die praxisbezogenen Schritte im inneren orangen Rahmen eingegangen. Die im vorangehenden Abschnitt 4.3 aufgeführten Fragen werden für das spezifische Vorgehen im UAZ beantwortet.

Die beiden grauen Rahmen

Policy

Das UAZ verfügt über eine Policy Digitale Langzeitarchivierung. In der Policy ist festgehalten, dass das UAZ verantwortlich ist für die Archivierung von Unterlagen sämtlicher Organe der UZH. Dies umfasst Unterlagen der universitären Verwaltung sowie auch von Privatpersonen und Vereinen, die einen Bezug zur UZH aufweisen. Die archivwürdigen Unterlagen müssen zu rechtlichen, administrativen, kulturellen und wissenschaftlichen Zwecken aufbewahrt, erschlossen und vermittelt werden. Der gesetzliche Auftrag betrifft

sowohl analoge als auch digitale Unterlagen. Original digitale Unterlagen werden als solche übernommen, archiviert und zur Verfügung gestellt.[62]

In der Policy ist festgehalten, dass auch Inhalte aus dem Internet bei der Archivierung mitberücksichtigt werden müssen. Auch Hypertext wird als Unterlagenkategorie aufgeführt.[63] Websites gehören zu diesen Inhalten aus dem Internet, die archivwürdig sein können. Die Erstellung einer spezifischen Policy für die Webarchivierung ist folglich nicht notwendig, da Websites in der bestehenden Policy Digitale Langzeitarchivierung bereits mitberücksichtigt sind. Bezüglich der Webarchivierung könnten aber einige Anpassungen in der bestehenden Policy vorgenommen werden. So ist beispielsweise in der Policy an keiner Stelle explizit von «Webarchivierung» die Rede. Es wäre sicher sinnvoll, wenn der Ausdruck «Webarchivierung» explizit in der Policy vorkommt und nicht allgemein nur von «Internet» die Rede ist. Was allenfalls dereinst auch einer Anpassung bedarf ist die gewählte Preservation Planning-Strategie. Das UAZ setzt bis anhin ausschliesslich auf das Migrationsprinzip.[64] Falls dereinst in Zusammenhang mit der Webarchivierung für das Preservation Planning mit Browseremulation gearbeitet werden müsste, dann wäre dies entsprechend in der Policy zu ergänzen.

Das UAZ hat zudem eine Übersicht «Archivtaugliche Dateiformate» erarbeitet. Auch in diesem Dokument finden sich Websites erwähnt, dies in Zusammenhang mit dem Containerformat ZIP. ZIP wurde vom UAZ bis anhin als Containerformat für die Archivierung von Websites verwendet. Auch dieses Dokument wäre auf die neuen Verhältnisse anzupassen. Die Websites werden neu im Standardformat WARC gesichert und nicht mehr im ZIP-Format.[65]

Metadata/Description

Metadaten werden im UAZ nicht ausschliesslich im AIS, sondern auch auf der Plattform Archive-It sowie beim Ingest in das digitale Langzeitarchiv erfasst.

Auf das Vorgehen bezüglich der Erfassung von Metadaten für den Ingest in das digitale Langzeitarchiv wird im «Handbuch DLZA» des UAZ eingegangen. Die deskriptiven Metadaten werden in *docuteam packer* erfasst. Diese Daten werden via die Funktion

[62] vgl. UZH Archiv Policy Digitale Langzeitarchivierung Version 1.0 vom 1. Oktober 2022: Seite 4. Die Policy Digitale Langzeitarchivierung ist online auf der Webseite des UAZ verfügbar. Eine Übersicht zu sämtlichen verwendeten online verfügbaren Ressourcen des UAZ ist im Anhang der vorliegenden Arbeit aufgeführt.

[63] vgl. UZH Archiv Policy Digitale Langzeitarchivierung Version 1.0 vom 1. Oktober 2022: Seite 9.

[64] vgl. UZH Archiv Policy Digitale Langzeitarchivierung Version 1.0 vom 1. Oktober 2022: Seite 6–7.

[65] vgl. Archivtaugliche Dateiformate, aktualisiert am 18.10.2022.

«Import aus Ingest» in das *CMI AIS* überführt. Da mit dieser Importfunktion gearbeitet wird, werden die Metadaten bereits möglichst vollständig in *docuteam packer* erfasst. Da in *docuteam packer* nicht alle Pflichtfelder abgebildet werden, ist nach dem Import eine minimale nachträgliche Erschliessung im AIS notwendig.[66]

In der Tektonik in *CMI AIS* erscheinen die Metadaten zum archivierten Webauftritt in der Abteilung PUB «Publikationen mit offiziellem Charakter» und stellen damit einen Bestandteil der Amtsdruckschriftensammlung dar. Der Webauftritt ist im Bestand «PUB.010 Webauftritt der Universität» zusammengefasst. Die Archivnutzenden können online über den Archivkatalog die entsprechenden Metadaten einsehen. Die Grundlagen für die Erschliessung im AIS sind im Erschliessungshandbuch[67] festgehalten, so ist unter anderem darin definiert, welche Felder als Pflichtfelder jeweils auszufüllen sind. Interessant wäre es hier, zukünftig direkt über den Eintrag im Archivkatalog einen Link auf die entsprechende archivierte Website zu publizieren, wie es beispielsweise das Hochschularchiv der ETH tut. Das Hochschularchiv der ETH verlinkt die Einträge im Archivkatalog direkt auf die *Wayback* Calendar Page (Beispiel: Website vom Archiv für Zeitgeschichte 2022–2024)[68]: Beim verlinkten Eintrag auf den Archivkatalog ist unter «Link auf digitales Original» der Link auf die entsprechende Calendar Page nur einen Klick entfernt, was eine sehr benutzerfreundliche Lösung darstellt. Auf diese Weise können interessierte Personen direkt über den online publizierten Archivkatalog die archivierte Website auf der *Wayback* Calendar Page via dem in den Metadaten hinterlegten Link ansteuern.

Was die Erfassung von Metadaten auf der Plattform Archive-It angeht, ist das UAZ zurückhaltend. Es beschränkt sich auf die Angabe der wichtigsten Daten, da in *CMI AIS* die Metadaten zu den erschlossenen Websites ersichtlich sind respektive ersichtlich sein werden. Metadaten auf Archive-It werden auf der Ebene Collection und der Ebene Seed erfasst, wobei die Ebene Seed besonders zentral ist. Auf der Ebene der Collection werden Metadaten in den Feldern Creator, Description, Rights, Collector und Language erfasst. Diese Felder sind obligatorisch auszufüllen bei der Erfassung einer Collection. Zusätzlich werden Collection Topics ausgewählt. Auf der Ebene Seed werden die folgenden Metadatenfelder ausgefüllt: Title, Creator, Language und Collector. Auf der Ebene Document wird auf die Erfassung von Metadaten verzichtet.[69] Neben den deskriptiven Metadaten sind auch die technischen Metadaten zentral. Technische Metadaten (unter

[66] Internes Dokument (CMI G 2018-312).

[67] Erschliessungshandbuch UZH Archiv (UAZ), Version 1.1 vom 22. Februar 2021.

[68] vgl. Gastl 2022: Folie 11.

[69] Austausch mit Inge Moser (stv. Leiterin UAZ) vom 23. Mai 2024.

anderem: Datum/Uhrzeit des Crawls, verwendete Software etc.) werden beim Crawl direkt in das WARC File hineingeschrieben.[70] Archive-It stellt eine Liste mit Metadaten zur Verfügung, die heruntergeladen werden kann. Derzeit sind die Metadaten der ab 2024 mit Archive-It gecrawlten Websites in *CMI AIS* noch nicht erfasst, da noch keine digitalen Objekte in das digitale Langzeitarchiv übernommen wurden.

Äusserer blauer Rahmen

Vision and Objectives

Im Allgemeinen muss bezüglich dem Ziel der Webarchivierung das Bewusstsein vorhanden sein, dass eine vollständige Sammlung weder realistisch noch anstrebenswert ist. Zur Webarchivierung gehört ein bewusster Mut zur Lücke.[71]

Das UAZ hat gemäss Archivverordnung die Aufgabe, die archivwürdigen Unterlagen der UZH auszuwählen, zu bewerten und zu verzeichnen. Die Bestände müssen erhalten sowie die Benutzbarkeit gewährleistet werden.[72] Gemäss Archivverordnung § 10a besteht eine Ablieferungspflicht für Amtsdruckschriften.[73] Die Websites sind als eine solche Publikation öffentlichen Charakters zu verstehen.[74] Handelt es sich um eine Website, die als archivwürdig bewertet wurde, dann muss diese langfristig erhalten bleiben. Wie bereits unter dem weiter obenstehenden Abschnitt «Policy» erwähnt, sind archivwürdige Inhalte aus dem Internet gemäss der Policy Digitale Langzeitarchivierung zu sichern.[75] Das Ziel der Archivierungsbestrebungen besteht darin, die als archivwürdig bewerteten Bestandteile des Webauftritts der UZH zu erhalten. Da immer mehr Information ausschliesslich auf dem Web verfügbar ist, wird das Web historisch interessant. Gewisse Publikationen sind online vorhanden, die analog gar nicht mehr veröffentlicht werden. Für das UAZ ist insbesondere die Archivierung von Websites der Institute, Seminare und Kliniken wichtig. Dies da die Überlieferungsbildung auf Ebene der Institute anderweitig nicht flächendeckend realisiert werden kann. Diese Websites stellen eine zentrale Informationsquelle für die Verwaltung und Öffentlichkeit dar. Zudem enthalten sie Informationen zur Institutsgeschichte und -entwicklung. Auf den Websites sind ausserdem häufig Amtsdruckschriften

[70] vgl. Weimer/Schoger 2021: Seite 1.

[71] vgl. Beinert/Schrimpf/Wolf 2011a: Seite 1.

[72] vgl. Archivverordnung vom 9. Dezember 1998: §12, S. 3. Im Anhang der vorliegenden Arbeit ist eine Zusammenstellung der konsultierten relevanten gesetzlichen Grundlagen eingefügt.

[73] vgl. Archivverordnung vom 9. Dezember 1998: Seite 2–3.

[74] Internes Dokument (CMI G 2012-240).

[75] vgl. UZH Archiv Policy Digitale Langzeitarchivierung Version 1.0 vom 1. Oktober 2022: Seite 9.

wie kommentierte Vorlesungsverzeichnisse, Studienprogramme, Informationsbroschüren oder Hinweise auf Veranstaltungen ausserhalb des Vorlesungsverzeichnisses greifbar, die anderweitig selten den Weg ins Archiv finden.[76]

In Zusammenhang mit dem verfolgten Ziel der Webarchivierung sowie dem übergeordneten Auftrag des Archivs, ist es auch sinnvoll, sich mit dem Thema Redundanz auseinanderzusetzen. Es ist diesbezüglich wichtig, den Gesamtüberblick zu wahren, wenn es um die Überlieferungsbildung geht. Das Webarchiv soll nicht separat nur für sich allein betrachtet werden, sondern in Zusammenhang und in Abstimmung mit der Gesamtüberlieferung. Auf diese Weise können Dubletten vermieden werden, was in gewissen Bereichen sehr sinnvoll ist. Dies vor allem auch, wenn es um Dubletten geht, die viel Speicherplatz belegen würden, beispielsweise Videos. Das UAZ sichert bereits unabhängig vom Webarchiv Videoproduktionen der Multimedia und E-Learning Services (MELS). Diese Videos müssten folglich nicht noch zusätzlich von der Website (wo sie eingebunden sein können) abgezogen werden.[77] Ein gewisses Mass an Redundanz ist allerdings unvermeidlich, da nicht für jedes einzelne Video oder jede einzelne auf dem Web verfügbare Amtsdruckschrift überprüft werden kann, ob diese Ressource bereits anderweitig ins UAZ gelangt ist. In dieser Hinsicht sind sicherlich Doppelspurigkeiten vorhanden, die vom UAZ bewusst in Kauf genommen werden. Es gilt allerdings auch zu beachten, dass sich eine gewisse Redundanz teilweise durchaus lohnen kann. Beispielsweise können zusätzlich zu einer Statistik in Form einer Grafik auch die Rohdaten (beispielsweise aus einer Datenbank) übernommen werden, damit diese den Forschenden weiterhin zur Verfügung stehen. Dieselben Daten können einen völlig anderen Charakter aufweisen, je nachdem wie sie vorliegen. Mit der abweichenden Form geht auch einher, dass sich ein- und dieselbe Information auf unterschiedliche Art und Weise auswerten und nutzen lässt.[78]

An dieser Stelle soll nun noch auf die vom UAZ in Archive-It erfassten Collections eingegangen werden. Bis anhin wurden drei Collections erstellt. Die Collection «Websites der Universität Zürich 2024–» umfasst Crawls, welche das UAZ ab 2024 mit Archive-It durchgeführt hat. In den Collections «Websites der Universität Zürich 1997–2007» (betrifft

[76] Internes Dokument (CMI G 2012-240).

[77] In diesem Zusammenhang ist allerdings festzuhalten, dass die MELS-Videos nicht systematisch vom Crawling ausgeklammert werden können. Auch diese werden theoretisch mitgecrawlt. Es wurde aber festgestellt, dass *Heritrix* die Videos nicht vom Web abzieht, da sie auf der Plattform SWITCHcast MediaSpace (Powered by kaltura) eingebunden sind und deshalb gar nicht als Asset auf der Website hinterlegt wurden. Wenn nun also festgestellt wird, dass bei einem Crawl ein MELS-Video nicht abgezogen wurde, dann braucht sich das UAZ nicht nachträglich darum zu kümmern, dass dieses gesichert wird, da es anderweitig bereits ins UAZ gelangt. Die Vermeidung dieser Redundanz wirkt sich positiv auf die Speicherplatzbelegung aus.

[78] vgl. Loewenich 2024: Folie 5 (Script: Seite 7).

Main Host http://www.unizh.ch) und «Websites der Universität Zürich 2007–2023» (betrifft Main Host http://www.uzh.ch), welche den Zeitraum vor 2024 betreffen, werden via Waybackfill Service die vom Internet Archive erfassten Abzüge eingefügt.[79] Bei der Zusammenstellung der Collections auf Archive-It liegt der Fokus des UAZ auf dem Webauftritt der Gesamtuniversität, der Zentralen Dienste, der Fakultäten und der Institute, Seminare, Kliniken und Kompetenzzentren. Durch die Konsultation dieser Collections kann in Erfahrung gebracht werden, wie sich die Universität und ihre Fakultäten und Institute im Web präsentieren und auch wie sich die Darstellung mit der Zeit verändert hat. Den Kernbereich der Überlieferungsbildung in Bezug auf die Webarchivierung umfassen folglich die Netzpublikationen der Gesamtuniversität sowie ihrer Fakultäten und Institute. Thematische Spezialsammlungen wurden derweil noch keine angelegt. Die Collections würden aber die Möglichkeit bieten, thematische Schwerpunkte zu setzen. Mit diesen könnten dereinst unterschiedliche spezifische Forschungsfragen beantwortet werden. Das Hochschularchiv der ETH hat beispielweise einige thematische Schwerpunkte gesetzt und entsprechende Collections angelegt. Beispielsweise gibt es eine separate Collection zum Thema «Covid» (COVID-19 Coronavirus Collection). Allenfalls wären solche spezifischen Sammlungen auch für das UAZ für die Zukunft eine Option, als Ergänzung zur bisherigen Vorgehensweise. Es ist allerdings so, dass das UAZ bereits jetzt innerhalb der Collection der Websites der UZH auch Sites übernimmt, die themenspezifisch sind (beispielsweise von spezifischen Veranstaltungen wie die Master-Tage oder Jubiläen wie das 50 Jahre Jubiläum des Pädagogischen Instituts). Das UAZ arbeitet hier also auf der Ebene Seed themenspezifisch und nicht auf der Ebene Collection.

Resources and Workflow

Im UAZ waren bis anhin hauptsächlich zwei Personen in den Betrieb des Webarchivs involviert – der Archivinformatiker sowie die stellvertretende Leiterin des UAZ, die unter anderem im Bereich Digitale Langzeitarchivierung spezialisiert ist. Das UAZ sichert seit 2012 archivwürdige Websites. Es besteht also ein grosser Erfahrungsschatz mit dieser Thematik sowie auch personelle und finanzielle Ressourcen. Diese bereits vorhandenen Ressourcen können weiterhin genutzt werden. Eine Anpassung der Arbeitsabläufe war allerdings notwendig.

[79] Austausch mit Inge Moser (stv. Leiterin UAZ) vom 5. Juli 2024 sowie internes Dokument (CMI G 2022-137). Die Daten sind unterdessen bereits mit dem Archive-It-Account des UAZ innerhalb der zwei verschiedenen Collections verknüpft. Derzeit ist aber noch kein Zugriff auf einzelne Seiten möglich via Seeds. Es ist geplant, dass die Seeds automatisiert durch Archive-It erstellt werden. Der Auftrag ist derzeit noch pendent und die Erledigung wird ein paar Wochen in Anspruch nehmen (Stand: 19.8.2024).

Wird der Initialaufwand bei der Nutzung von Archive-It mit demjenigen bei der damaligen Einführung der Nutzung vom Offline Explorer miteinander verglichen, dann deckt sich dieser ungefähr. Der Archivinformatiker Markus Kandlbinder erklärt, dass der Initialaufwand beim Offline Explorer sehr gross war. Dies sei nun aber auch bei der Einführung der Webarchivierung über Archive-It der Fall. Er musste damals das Offline-Explorer-Tool zunächst gründlich testen und auf die Bedürfnisse und Anforderungen des UAZ einstellen. Mit der Zeit konnte vieles automatisiert werden. Um die Qualität der abgezogenen Seiten zu prüfen, erfolgte bei der Nutzung vom Offline Explorer jeweils eine stichprobenartige Sichtung. Dafür hatte Kandlbinder jeweils einen Tag im Monat eingesetzt. Er hatte kontrolliert, ob die Qualität in Ordnung ist und falls nicht, einen neuen optimierten Abzug von der Website gemacht. Auch bei Archive-It ist der Initialaufwand gross. Die Seeds müssen zunächst erfasst werden (ca. 10 Stück pro Monat). Die Crawls erfolgen monatlich. Mit einem Zweijahresplan ist vorgesehen, sämtliche Seeds zu erfassen. Derzeit erfolgen jeweils die Masterabzüge und diese müssen genau auf die Qualität hin überprüft werden. So muss überprüft werden, ob alles Wesentliche mitgecrawlt wurde, die Reports müssen studiert und allenfalls die Crawls nochmals neu angesteuert werden. Die nächsten zwei Jahre bis alle Seeds erfasst sind und jeweils einmalig ein Mastercrawl erfolgt ist, bleibt der Kontrollaufwand entsprechend hoch. Nach diesen beiden Jahren sollte der Aufwand dann kleiner werden. Während der Phase, in welcher der Initialaufwand anfällt, investieren Markus Kandlbinder (Archivinformatiker UAZ) und Inge Moser (stv. Leiterin UAZ) jeweils einen Tag pro Monat (also gesamthaft zwei Tage) für die Webarchivierung. Es wird vermutet, dass dann später noch gesamthaft ein Tag zu investieren sein wird, der wiederum auf zwei Personen verteilt werden soll. Wieviel Zeit dereinst für die Überführung der WARC Files in das digitale Langzeitarchiv des UAZ aufgewendet werden muss, ist derzeit noch unklar. Auch dies wird noch Ressourcen benötigen. Ein Umstand, der im Vergleich zum Vorgehen vor der Neuorganisation für etwas Entlastung sorgt, ist, dass sich der UAZ-Archivinformatiker nicht länger darum kümmern muss, dass die notwendige Infrastruktur läuft, da diese Verantwortung an den Dienstleister Archive-It ausgelagert wurde.[80] Auch die Archivbenutzung verursacht keinen Personalaufwand mehr, da sich die Archivnutzenden die archivierten Seiten direkt via Wayback Machine anschauen können und die Mitarbeitenden des UAZ die Zeitschnitte nicht mehr aus dem digitalen Langzeitarchiv herausholen und für die Benutzung bereitstellen müssen.

[80] Austausch mit Markus Kandlbinder (Archivinformatiker UAZ) vom 13. Juni 2024.

Es sollten folglich genügend personelle Ressourcen vorhanden sein. Aufgrund der zeitnahen Pensionierung des Archivinformatikers, ist seit August neu an seiner Stelle die Autorin der vorliegenden Arbeit in die Webarchivierung involviert und darf seine Aufgaben in Zusammenhang mit der Webarchivierung übernehmen (Durchführung Crawls, Qualitätskontrolle).

In Bezug auf die finanziellen Ressourcen musste für die Neuorganisation des Webarchivs etwas aufgestockt werden. Die Nutzung eines Archive-It Pro Accounts mit einem Data Budget von 256 GB verursacht jährliche Kosten von 4’000 Dollar. Da durch die Nutzung von Archive-It jährliche Mehrkosten auf das UAZ zukamen, wurde im Jahr 2023 abgeklärt, ob das Budget ab 2024 jährlich um 5'000 Fr. erhöht werden kann. Dieser Bitte konnte nachgekommen werden. Zusätzlich zu den wiederkehrenden jährlichen Kosten ist zudem auf einmalige Kosten von 7'500 Dollar aufmerksam zu machen, die in Bezug auf den Waybackfill Service auf das UAZ zukamen. Diese Kosten konnten bereits im Jahr 2023 beglichen werden, da noch genügend finanzielle Ressourcen im laufenden Budget vorhanden waren.[81]

Vom verfügbaren Data Budget von 256 GB auf Archive-It für das Subscription-Year 2024 wurden bis anhin 133.5 GB (Stand: 4. September 2024) genutzt. Derzeit wird darüber diskutiert, ob allenfalls mehr Websites als gemäss Zweijahresplan terminiert, gecrawlt werden sollen, da noch verfügbares Data Budget vorhanden ist. Auf diese Weise könnte das Data Budget vollumfänglich ausgeschöpft werden.

Access/Use/Reuse

Wie bereits erwähnt, wird durch die Nutzung von Archive-It die Benutzerfreundlichkeit stark optimiert. Die über Archive-It vom UAZ archivierten Websites sind öffentlich zugänglich auf der Access Page auf Archive-It und können über die *Wayback Machine* (Beispiel: Website UAZ) betrachtet werden.

Bei der *Wayback Machine* handelt es sich um ein Open Source-Anzeigetool. Dieses ermöglicht es, dass Archivnutzende die archivierten Webseiten auf eine Art und Weise durchsuchen können, wie es einer Live-Website entspricht.[82] Archivnutzende können ortsungebunden und unabhängig von den Archivarinnen und Archivaren des UAZ auf die archivierten Websites zugreifen. Einschränkungen zur Einsicht sind keine festgelegt.

[81] Interne Dokumente (CMI G 2022-137).

[82] vgl. Archive-It Help Center. Glossary: Stichwort «Wayback Machine». Verfügbar unter: https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms [2.8.2024].

Das Aufgabenbiet des UAZ umfasst gemäss § 12 der Archivverordnung unter anderem auch, dass die Bestände durch Öffentlichkeitsarbeit bekannt gemacht werden.[83] Die Collections wurden bis anhin nicht aktiv beworben, da dies zum aktuellen Zeitpunkt noch zu früh wäre. Die Collections sind noch unvollständig. Das UAZ arbeitet mit einem Zweijahresplan, was das Crawling anbelangt. Es lohnt sich nicht, allzu früh die Öffentlichkeit auf die Sammlung aufmerksam zu machen, wenn erst wenige Inhalte vorhanden sind. Zudem müssen auch zuerst noch die vom UAZ unabhängigen Abzüge vom Internet Archive, die bis 2023 entstanden sind, in den Archive-It Account des UAZ überführt und nach Seeds recherchierbar gemacht werden.[84] So wird es dann dereinst möglich sein, dass den Archivnutzenden sämtliche Zeitschnitte zum Vergleich zur Verfügung stehen werden. Der Hinweis auf das Webarchiv soll dann auf der Website des UAZ aufgeschaltet werden.[85] Zusätzlich wäre es auch denkbar im online verfügbaren Archivkatalog des UAZ bei der entsprechenden Verzeichnungseinheit via den Metadaten auf den jeweiligen Zeitschnitt in der *Wayback-Machine* direkt zu verlinken.

Preservation

Die vom Internet Archive archivierten Daten werden redundant in verschiedenen Datenzentren gesichert. Das Internet Archive besitzt diese Datenzentren und betreibt sie unabhängig. Es gibt mehrere geografisch verteilte primäre Datenzentren in den USA sowie zusätzliche Datenzentren in Kanada und Europa. Der Zugang zu sämtlichen Datenzentren ist nur eingeschränkt möglich und wird mit Ausfallsicherheitsverfahren überwacht.[86]

Das UAZ möchte sich nicht allein auf eine Drittpartei zur dauerhaften Sicherung der WARC Files und der dazugehörigen Metadaten verlassen, auch wenn es sich beim Internet Archive um eine angesehene Institution mit strengen Sicherheitsvorkehrungen handelt. Das UAZ plant deshalb, die Daten von Archive-It herunterzuladen und zusätzlich in das eigene digitale Langzeitarchiv zu überführen. Wie genau und in welchen zeitlichen Abständen der Download erfolgen soll, ist derzeit noch nicht definiert.

Es gibt Institutionen, die bereits Erfahrung mit dem Download der Daten von Archive-It sammeln konnten. Die University of Georgia Libraries empfiehlt viermal jährlich die Daten herunterzuladen. Durch diese Regelmässigkeit wird verhindert, dass allzu viele Daten auf einmal heruntergeladen werden müssen. Ebenfalls wird empfohlen, dass zusätzlich

[83] vgl. Archivverordnung vom 9. Dezember 1998: §12, Buchstabe e, S. 3.

[84] Der aktuelle Stand der Dinge zum Waybackfill Service ist in der Fussnote 79 festgehalten.

[85] Austausch mit Inge Moser (stv. Leiterin UAZ) vom 23. Mai 2024.

[86] vgl. Internet Archive, Archiving & Data Services Division 2023: Seite 2.

zu den WARC Files sechs Metadatenberichte mit heruntergeladen werden sollen: Berichte zu Seed, Seed Scope, Collection, Collection Scope, Crawl Job und Crawl Definition. Durch diese Reports wird der Kontext zu den WARC Files mitgesichert. In den Reports wird der Erfolg des Crawls dokumentiert und zudem sind die Metadaten enthalten, welche durch die Partnerorganisation von Archive-It erfasst wurden.[87]

Noch unklar ist, welche Preservation Planning-Strategie bezüglich der langfristigen Erhaltung der WARC Files verfolgt werden soll. An dieser Stelle sei auf den Abschnitt 6 in der vorliegenden Arbeit verwiesen, worin auf diese Thematik und mögliche Lösungsansätze zu Migration und Emulation eingegangen wird.

Risk Management

Die UZH betreibt die Website selbst. Das Copyright ist bei der UZH. Bei den Websites handelt es sich um Publikationen öffentlichen Charakters, die das UAZ übernehmen darf und auch übernehmen muss. Die Websites gehören zum Archivsprengel. Aus diesem Grund wurde nicht eine offizielle Erlaubnis eingeholt, als 2012 die ersten Websites gesichert wurden. Selbstverständlich wurden die relevanten Personen und Stellen aber über das Vorgehen informiert (unter anderem das Generalsekretariat).[88] Bis anhin wurde ausschliesslich derjenige Teil des öffentlich zugänglichen Webbereichs der UZH archiviert. Websites mit internen Bereichen werden nicht gecrawlt. Wenn nun dereinst an der UZH ein Intranet eingeführt wird, dann müsste überprüft werden ob diese Inhalte erst nach einer Schutzfrist veröffentlicht werden sollen.[89] Vermutlich würde zur Beurteilung der Archivwürdigkeit der Inhalte des Intranets eine separate Bewertung erfolgen, bei welcher das konkrete Vorgehen definiert wird.

Beim Crawling können robots.txt Files ein Problem darstellen. Es handelt sich dabei um Dateien, die ein Website-Besitzer zu seiner Website hinzufügen kann, um Crawlern den Zugriff auf die gesamte Website oder Teilen davon zu verwehren.[90] Das UAZ darf aufgrund seines Auftrages robots.txt ignorieren, damit die archivierte Website vollständig gesichert und wiedergegeben werden kann. Die ersten Erfahrungen mit Archive-It haben gezeigt, dass manchmal aus Sicherheitsgründen der Zugriff auf eine Seite verweigert wird. Wird robots.txt dann ignoriert, dann funktioniert in der Regel das Crawling. Bis jetzt

[87] vgl. Hanson, Archive-It Blog 2021. Verfügbar unter: https://archive-it.org/blog/post/automated-preservation-workflow/ [24.7.2024].

[88] Austausch mit Inge Moser (stv. Leiterin UAZ) vom 23. Mai 2024.

[89] Austausch mit Inge Moser (stv. Leiterin UAZ) vom 23. Mai 2024.

[90] vgl. Archive-It Help Center. Glossary: Stichwort «robots.txt». Verfügbar unter: https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms [14.7.2024].

wurde der Zugriff aber nur ganz selten verweigert. Inge Moser (stv. Leiterin UAZ) schätzt, dass dies auf nur ca. 5% aller Fälle zutrifft.[91]

Innerer oranger Rahmen

Appraisal and Selection

Als Grundlage welche Websites gesichert werden sollen dient der Bewertungsbericht vom 6. September 2021. Die berücksichtigten Websites sind:

- Hauptseite der UZH
- Seiten der zentralen Dienste und Abteilungen
- Seiten der Fakultäten, der Kompetenzzentren und der universitären Forschungsschwerpunkte
- Seiten von Instituten, Seminaren und Kliniken
- Websites, die für besondere Anlässe wie beispielsweise Ausstellungen oder Jubiläen erstellt wurden
- Seiten von universitären Organisationen.

In einer Excelliste sind sämtliche URLs der zu archivierenden Webseiten aufgeführt.[92]

Ob dereinst auch spezifische Collections zu bestimmten Themen angelegt werden sollen, ist wie bereits unter dem Abschnitt «Vision and Objectives» erwähnt derzeit noch unklar.

Scoping

Das UAZ arbeitet mit einer selektiven Webarchivierung. Dabei werden ausgewählte Websites bei der Archivierung berücksichtigt. Diese Websites sollen möglichst in ihrer Gesamtheit gecrawlt werden. Bei den ausgewählten Seiten muss also nicht exakt definiert werden, welche Bestandteile gecrawlt werden sollen. Das UAZ verwendet standardmässig den Crawler *Heritrix*, der einen umfassenden Crawl ermöglicht. Bestandteile einer Website, die nicht übernommen werden sind zum Beispiel, wenn in einem Blog mit verschiedenen Artikeln diverse Links vorhanden sind, beispielsweise auch auf Youtube-Videos oder weitere externe Inhalte. Für eine Übernahme dieser Inhalte müssten theoretisch sämtliche Urheberrechte abgeklärt werden.[93]

[91] Austausch mit Inge Moser (stv. Leiterin UAZ) vom 23. Mai 2024.

[92] Interne Dokumente (CMI G 2012-240).

[93] Austausch mit Inge Moser (stv. Leiterin UAZ) vom 23. Mai 2024.

Anhand der Excel-Liste «_Webpages-zum-Archivieren»[94] werden die Seeds für den entsprechenden Monat erstellt. Als Seed Type wird der Standard drin gelassen. Müssen spezifische Scoping Rules bei einem Seed gesetzt werden, dann ist dies in der entsprechenden Collection unter Seed zu erfassen (beispielsweise ignore robots.txt). Es wird zunächst ein Test Crawl durchgeführt. Beim Crawl an sich werden keine Limitationen festgelegt. Es wird lediglich ein Zeitlimit von drei Tagen angesetzt. Die Qualitätskontrolle erfolgt sobald der Test Crawl erfolgreich abgeschlossen wurde.[95] Auf die Verwendung eines Tools, welches limitiert, wie viel von einer Seite gecrawlt werden soll, wird verzichtet. Auch wird der Crawl nicht ausschliesslich auf ein bestimmtes Dateiformat (beispielsweise PDF) limitiert. Wenn nach einem Test Crawl Probleme mit einer spezifischen URL ersichtlich sind, dann wird erneut ein Test Crawl durchgeführt, wobei ein Exclude der entsprechenden URL in Archive-It im «Seed Scope» erfolgt.

Generell ist darauf zu achten, dass nach dem ausgelösten Crawl ein gründliches Monitoring stattfindet. Der Crawl kann über die Registerkarte «Crawls», «Current Crawls» verfolgt werden. Durch die Überwachung des laufenden Crawls kann beispielsweise das Zeitlimit verlängert werden, wenn während des Crawlprozesses festgestellt wird, dass die als Zeitlimit definierten drei Tage nicht ausreichend sind. Wird der Crawl aufgrund der Überschreitung des Zeitlimits gestoppt, dann kann nach dem Abbruch keine Verlängerung erfolgen, sondern es müsste ein neuer Test Crawl gestartet werden mit einem grösseren Zeitfenster.

Data Capture

Das UAZ zieht monatlich Seiten ab. Dies wurde auch bereits vor der Neuorganisation auf diese Weise durchgeführt. Für unterschiedliche URLs werden unterschiedliche Crawl-Intervalle angesetzt. Die meisten Seiten werden in der Regel alle zwei Jahre abgezogen. Es gibt auch einige Seiten, die jedes halbe Jahr gesichert werden. Die Hauptdomäne wird zweimal jährlich abgezogen, da sie auch Informationen über spezielle Ereignisse enthalten kann (beispielsweise zu Covid19). Die Websites der Fakultäten werden einmal pro Jahr gesichert, um die Studienprogramme und Studienordnungen zu dokumentieren. Für die Website der UZH Alumni und das längerfristige Projekt Stadtuniversität wird ein Abzug im Zweijahresrhythmus durchgeführt. Websites von zeitlich begrenzten Anlässen (beispielsweise Jubiläen, Ausstellungen etc.) werden einmalig abgezogen. Es wird dabei

[94] Internes Dokument (CMI G 2012-240).

[95] Austausch mit Inge Moser (stv. Leiterin UAZ) vom 7. Juni 2024.

mit der Excel-Liste «_Webpages-zum-Archivieren» gearbeitet, die einen Gesamtüberblick bietet und beim Erfassen der Seeds zum Einsatz kommt.[96]

Es wird jeweils ein Test Crawl durchgeführt. Wenn das Resultat des Test Crawls den Qualitätsanforderungen entspricht, dann wird der Crawl gesichert. Falls nicht, dann müssen Anpassungen erfolgen und ein erneuter Test Crawl ausgelöst werden. Vor allem zum aktuellen Zeitpunkt, bei welchem die Seeds neu erstellt und Mastercrawls durchgeführt werden, ist es wichtig, dass mit Test Crawls gearbeitet wird.

Archive-It ermöglicht neun wiederkehrende Crawl-Frequenzen, die von zweimal täglich bis jährlich variieren, sowie einen einmaligen Crawl, der sich nicht wiederholt.[97] Das UAZ plant nach Abschluss der zweijährigen Testphase, in welcher sämtliche Seeds erfasst werden sowie die Masterabzüge erfolgen, das Intervall zum Crawl der Seiten auf einmal jährlich festzusetzen und die Crawls allenfalls zu automatisieren.

Storage and Organization

Wie unter dem Abschnitt «Preservation» bereits erwähnt, möchte sich das UAZ nicht allein auf das Internet Archive zur dauerhaften Sicherung der WARC Files und der dazugehörigen Metadaten verlassen. Die Daten sollen von Archive-It heruntergeladen und zusätzlich in das eigene digitale Langzeitarchiv überführt werden. Um WARC Files von Archive-It herunterzuladen, kann die Schnittstelle *WASAPI* in einem Webbrowser verwendet werden. Auf diese Weise können WARC Files manuell über die mit Hyperlinks versehenen URLs heruntergeladen werden. Jeder Link muss einzeln angeklickt werden, um das jeweilige WARC File herunterzuladen. Je nach Anzahl der WARC Files kann dies sehr zeitaufwendig sein, so dass es sinnvoll sein kann, ein externes Tool für den Massendownload in Betracht zu ziehen.[98] Auf Anfrage stellt Archive-It auch Festplatten mit den WARC Files für seine Partnerorganisationen zur Verfügung.[99] Zusätzlich zu den WARC Files ist es auch sinnvoll die Liste mit sämtlichen Metadaten zu einem Crawl auf Archive-It herunterzuladen. In diesem File sind sämtliche Dokumente aufgelistet, die in den WARC-Containern enthalten sind.

[96] Interne Dokumente (CMI G 2012-240).

[97] vgl. Bragg/Hanna 2013: Seite 25.

[98] vgl. Blumenthal, Archive-It Help Center 2024. Verfügbar unter: https://support.archive-it.org/hc/en-us/articles/360015225051-Find-and-download-your-WARC-files-with-WASAPI [24.6.2024].

[99] Korrespondenz mit Kody Willis (Archive-It) vom 13. Mai 2024.

Quality Assurance and Analysis

Bezüglich der Qualitätskontrolle soll zunächst auf ein paar Darstellungsprobleme hingewiesen werden, die derzeit bestehen. Unter anderem gibt es Probleme beim Abzug der offiziellen Teamseite der UZH (Beispiel Teamseite vom UAZ). Die Fotos der einzelnen Teammitglieder werden als Dateien zwar heruntergeladen aber nicht auf der archivierten Seite über die *Wayback Machine* angezeigt. Ebenso verhält es sich mit Galerieansichten. Ein weiteres Problem stellen die Dropdown-Menüs dar, die ebenfalls nicht korrekt dargestellt werden können. Zudem stellte sich heraus, dass teilweise Videos beim Crawl nicht mitgesichert werden. Es handelt sich dabei um Videos, die auf der Plattform SWITCHcast MediaSpace eingebunden sind. Diese Quellen sind eigentlich nicht extern. Sie werden aber vom Crawler als extern beurteilt und nicht gesichert, da sie über den Player abgespielt werden und nicht als Asset direkt in der Website hinterlegt wurden. Diese Problematik bezüglich den Videos kann aber vernachlässigt werden, da das UAZ wie bereits erwähnt Videos von MELS direkt übernimmt. Ein weiteres bekanntes Problem ist Java.[100] Bei Flash und Javascript handelt es sich um zwei Dateiformate, die schwer zu erfassen und wiederzugeben sind.[101] Generell problematisch für den Crawler sind Flash-Dateien, JavaScript, eingebettete Video-Streams sowie Datenbankinhalte.[102] Bei der Qualitätskontrolle muss bewusst auf diese Komponenten geachtet werden.

Die Qualitätskontrolle im UAZ wird einerseits durch das Studium der im Report enthaltenen Informationen in der «Crawling History» und andererseits durch einen Abgleich zwischen der durch den Test Crawl gesicherten Seite und der Seite auf dem Live-Web durchgeführt.[103] Für die Qualitätskontrolle ist die Rubrik «Crawling History» zentral. Nur an dieser Stelle ist ersichtlich, wie der aktuelle Status des Crawls ist und auch wie oft die Seite in der Vergangenheit bereits gecrawlt wurde. Auch der Grund für einen allfälligen Abbruch des Crawls ist darin festgehalten. Über die «Crawling History» kann auch auf den Report zugegriffen werden. Im «Hosts Report» sind sämtliche mit dem Crawl mitgesicherten Sites ersichtlich. Hier ist es wichtig in Zusammenhang mit der Hauptseite zu prüfen ob und weshalb bestimmte Inhalte unter «Blocked» oder «Queued» von einer Sicherung ausgeschlossen wurden. Im Report unter «Hosts» kann zudem nach spezifischen Seiten gesucht werden, was ebenfalls sehr hilfreich sein kann. Vor allem da wie erwähnt derzeit ein Problem besteht mit der Sicherung des Dropdown-Menüs. Wenn nun

[100] Austausch mit Inge Moser (stv. Leiterin UAZ) vom 7. Juni 2024.

[101] vgl. Bragg/Hanna 2013: Seite 24.

[102] vgl. Schrimpf/Beinert/Wolf 2011b: Seite 2.

[103] Austausch mit Inge Moser (stv. Leiterin UAZ) vom 23. Mai 2024.

Unsicherheit darüber besteht, ob diese spezifischen via Dropdown-Menü verlinkten Seiten mitgesichert wurden oder nicht, dann kann nach diesen Seiten unter «Hosts» gesucht werden. Bisherige Erfahrungen haben gezeigt, dass es sich lediglich um ein Darstellungsproblem handelt, die Seiten aber in der Regel mit dem Crawl abgezogen wurden. Im Report unter dem Register «Seed» kann auf den Testserver zugegriffen werden, auf welchem die mit dem Test Crawl vorübergehend gesicherten Seiten dargestellt werden können. Auf der produktiven *Wayback Machine* sind die Seiten nicht ersichtlich, da es sich erst um einen Test Crawl handelt. Wichtig ist es immer einen visuellen Abgleich zu machen zwischen dem Test Crawl und der Website auf dem Live-Web. Durch das Anklicken von Links auf der Website wird die archivierte Website auf ihre Vollständigkeit und Funktionalität (Stichproben) überprüft. Im Report unter dem Register «File Types» kann überprüft werden, wieviele verschiedene Filetypen in welchem Umfang gesichert worden sind. Der «File Types Report» organisiert sämtliche gecrawlten URLs nach Dateityp und stellt den Zugriff auf diese zur Verfügung. Es kann nach spezifischen Inhalten gesucht werden, beispielsweise nach Bildern. Wenn nun Unsicherheit darüber besteht, ob die Fotografien, die es bei der Darstellung des Test Crawls auf der Teamseite nicht anzeigt, gesichert worden sind, dann kann dies über den Report «File Types» überprüft werden. Dazu wird das entsprechende Dateiformat ausgewählt und ein Suchbegriff eingegeben. Genau gleich kann vorgegangen werden, wenn Darstellungsprobleme bezüglich Galerien vorkommen und überprüft werden soll, ob die entsprechenden Bilder mit dem Crawl gesichert wurden.

Es ist wichtig Qualitätskriterien zu entwickeln, damit eine Einheitlichkeit in den Collections erzielt werden kann.[104] Das International Internet Preservation Consortium (IIPC) plant, ein technisches Instrument zu entwickeln, dass eine zuverlässige Qualitätssicherung durchführen kann. Allenfalls wäre die Nutzung dieses Instruments auch für das UAZ dereinst interessant. Bis dieses Instrument zur Verfügung steht, wird die Qualitätskontrolle weiterhin manuell und anhand von Stichproben durchgeführt werden. Bei der manuellen Überprüfung soll auf verschiedene Aspekte geachtet werden. Der Gesamteindruck der archivierten Website ist mit der Originalwebsite zu vergleichen. Bezüglich Darstellung kann beispielsweise darauf geachtet werden, ob eine Fotogalerie korrekt angezeigt wird. Die Vollständigkeit kann stichprobenartig überprüft werden, indem Links verfolgt werden und die Website auf verschiedene Sprachversionen überprüft wird.[105] Wenn die Qualität nicht zufriedenstellend ist (beispielsweise aufgrund von fehlenden Inhalten, falscher

[104] vgl. Schrimpf/Beinert/Wolf 2011b: Seite 2.

[105] vgl. Schweizerische Nationalbibliothek, Webarchiv Schweiz: Merkblatt Archivieren 2024: Seite 13.

Darstellung oder Übernahme von Daten, die nicht in das Sammelgebiet gehören), dann kann es sich lohnen, nochmals einen neuen Test Crawl anzusteuern.[106] Dieser Crawl kann dann versuchsweise auch mit dem Crawler *Brozzler* erfolgen. *Brozzler* hat den Vorteil, dass er mit interaktiven Inhalten besser zurechtkommt als der Crawler *Heritrix*, der vom UAZ standardmässig genutzt wird.[107] Ein Test Crawl mit *Brozzler* kann Abhilfe schaffen, wenn eine Website, die mit *Heritrix* gecrawlt wurde, Darstellungsprobleme aufweist. Dies ist oft bei neueren oder speziellen Layouteinstellungen der Fall. Aus den bis anhin gewonnenen Erfahrungen zeigte es sich allerdings, dass es besser ist, den Crawler *Brozzler* zurückhaltend zu verwenden, da mit dem *Brozzler* teilweise Inhalte in mehrfacher Ausführung übernommen werden.

Zum Abschluss der Qualitätsprüfung kann der Test Crawl – wenn er den Qualitätsanforderungen genügt – via dem Button «Save Crawl Data» gesichert werden. Soll ein Test Crawl nicht gesichert werden, dann erfolgt die Löschung via den Button «Delete Crawl Data». Wird der Test Crawl weder gespeichert noch aktiv gelöscht, dann wird er automatisch nach 60 Tagen entfernt, da es sich nur um eine temporäre Sicherung handelt. Zuletzt ist die Excelliste «_Webpages-zum-Archivieren» zu aktualisieren. Der letzte Abzug der Website sowie der nächste geplante Abzug sind zu vermerken. Auch Bemerkungen sind zu erfassen, beispielsweise wenn eine Seite nicht mehr existiert und deshalb beim nächsten Abzug nicht mehr berücksichtigt werden muss.

4.5 Weiterführende Gedanken und offene Fragen

Wie bereits im Prozessschritt «Preservation» erwähnt, wird auf das Thema Preservation Planning in Abschnitt 6 der vorliegenden Arbeit eingegangen. Sobald der Entschluss für eine Preservation Planning-Strategie feststeht, können die entsprechenden Angaben dazu in die Prozessbeschreibung integriert werden.

Gesamthaft kann ausgesagt werden, dass diese erste Fassung einer Prozessbeschreibung im vorangehenden Abschnitt 4.4 als eine Grundlage dienen soll. Die Beschreibung kann überarbeitet und ergänzt werden, sobald sich der Prozess mit Archive-It gefestigt hat, derzeit noch offene Fragen geklärt werden konnten und sich eine Routine in Bezug auf den Workflow bewährt hat.

[106] vgl. Schweizerische Nationalbibliothek, Webarchiv Schweiz: Merkblatt Archivieren 2024: Seite 14.

[107] Austausch mit Inge Moser (stv. Leiterin UAZ) vom 7. Juni 2024.

5 Datenmodell

Der Abschnitt 5 befasst sich mit grundlegenden Fragen in Zusammenhang mit der Überführung der über Archive-It gesicherten Daten in das digitale Langzeitarchiv des UAZ. Der Abschnitt resultiert in Empfehlungen, welche dem UAZ für die spätere Erarbeitung eines Datenmodells dienen können.

Untergliedert ist der Abschnitt wie folgt: Der Abschnitt 5.1 befasst sich mit dem für die Webarchivierung zentralen Dateiformat WARC und seinen Herausforderungen und Spezifikationen. In Abschnitt 5.2 wird anschliessend aufgezeigt, wie die drei für die Webarchivierung relevanten Komponenten Archive-It, digitales Langzeitarchiv und AIS zusammenspielen. In Abschnitt 5.3 sind einige Fragen zusammengetragen, die in Zusammenhang mit der Erarbeitung eines Datenmodells zentral sind. Eine Auseinandersetzung mit diesen Fragen erfolgt im selbigen Abschnitt. In Abschnitt 5.4 werden aufgrund der aus den Abschnitten 5.1 bis 5.3 gewonnenen Erkenntnisse erste Empfehlungen ausformuliert, welche Aspekte bezüglich der Erarbeitung eines Datenmodells beachtet werden müssen. Zuletzt werden in Abschnitt 5.5 weiterführende Gedanken sowie offene Fragen thematisiert.

5.1 Das Dateiformat WARC und seine Herausforderungen und Spezifikationen

Allgemeine Informationen zum Dateiformat WARC

Beim Crawlprozess wird eine Vielfalt an unterschiedlichen Inhalten von einem definierten Ausschnitt des Webs heruntergeladen: unter anderem HTML-Seiten sowie Text-, Bild-, Audio- und Videodateien. Diese Files werden zusammen mit Metadaten zum Prozess gesichert. Zur Sicherung dieser Daten hat sich das Dateiformat WARC (Web ARChive) durchgesetzt. Es handelt sich dabei um ein Containerformat.[108] Die Nutzung des Dateiformates WARC ist weit verbreitet. Sowohl das Internet Archive als auch grosse nationale, regionale oder fachliche Webarchive setzen für die Webarchivierung auf dieses Format.[109] Das Dateiformat WARC ist auf Basis des ARC-Formates geschaffen worden. Das 1996 vom Internet Archive entwickelte ARC-Format wurde vom IIPC zum WARC-Format erweitert. Während es mit dem ARC-Format ausschliesslich möglich war Inhaltsinformation zu sichern, ermöglichte das WARC-Format zusätzlich auch eine

[108] vgl. Weimer/Schoger 2021: Seite 1.

[109] vgl. Weimer/Schoger 2021: Seite 3.

Sicherung von Metainformation.[110] Das WARC-Format bietet die Möglichkeit Informationen zu Duplikaterkennung, zu Migrationen sowie zur Aufteilung grosser Ressourcen auf unterschiedliche WARC Files zu dokumentieren. Im Vergleich mit dem ARC-Format bietet das WARC-Format Optimierungen in den Bereichen Harvesting, Zugriff und Austausch.[111] Das Dateiformat WARC wurde im Jahr 2009 zu einem ISO-Standard ernannt. Eine erste Revision führte zur WARC Version 1.1, die 2017 publiziert wurde. Weimer/Schoger weisen darauf hin, dass eine weitere Revision in Planung ist.[112] Das Dateiformat WARC ermöglicht die strukturierte Darstellung in der *Wayback Machine*.[113] WARC Files liegen meist in gezippter Form vor, erkennbar durch die Dateiendung GZIP. Es handelt sich dabei um eine verlustfreie Kompression, die durch die WARC-Spezifikationen erlaubt ist.[114] Eine WARC-Datei besteht aus der Verknüpfung von einem oder mehreren WARC-Datensätzen. Ein WARC-Datensatz enthält einen Header, gefolgt von einem Content Block und zwei Zeilenumbrüchen. Der Header enthält die folgenden Angaben: Datum, Typ und Länge des Datensatzes. Es können acht verschiedene Typen von WARC-Records unterschieden werden: warcinfo, response, resource, request, metadata, revisit, conversion und continuation. In den Content Blocks einer WARC-Datei können Ressourcen in jedem beliebigen Format enthalten sein.[115]

Der Aufbau einer WARC-Datei erfolgt während dem Crawlprozess. Eine WARC-Datei besteht aus WARC-Records unterschiedlichen WARC-Typs. Diese werden aneinandergereiht.

Die WARC-Datei entsteht in mehreren Schritten, die nachfolgend erläutert werden:

- Die WARC-Datei wird durch den Crawler erstellt.
- Im WARC-Record des Typs **warcinfo** (steht am Anfang der Datei) speichert der Crawler Metadaten, die den Crawl dokumentieren: unter anderem Datum, Zeitpunkt des Crawls sowie die verwendete Software.
- In einem nächsten Schritt stellt der Crawler eine Anfrage nach Inhalten der zu archivierenden Website an den Webserver. Die Anfrage wird dokumentiert im Record des Typs **request**.

---

[110] vgl. Weimer/Schoger 2021: Seite 2.

[111] vgl. Library of Congress: Sustainability of Digital Formats. Verfügbar unter: https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml [17.7.2024].

[112] vgl. Weimer/Schoger 2021: Seite 2.

[113] vgl. Schoger/Beinert/Schmid/Donig/Eckl 2021: Folie 11.

[114] vgl. Frappart 2023: Folie 16.

[115] vgl. Library of Congress: Sustainability of Digital Formats. Verfügbar unter: https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml [17.7.2024].

- Der Webserver stellt daraufhin die gewünschte Ressource zur Verfügung. In einen Record vom Typ **response** beziehungsweise **resource** hält der Crawler diese Datei und die dazugehörigen Metadaten fest. Dieser und der vorherige Schritt werden so oft wiederholt, bis das Ziel des Crawlprozesses erreicht ist oder der Prozess aus anderen Gründen (beispielsweise Erreichen eines Limits) unterbrochen wird.
- Daneben gibt es auch noch den Typ **metadata**. In Records von diesem Typ können zusätzliche Metadaten zur Beschreibung der einzelnen Inhalte festgehalten werden.
- In Records vom Typ **continuation** ist vermerkt, wenn Inhalte auf verschiedene WARC-Dateien aufgeteilt wurden.
- In Records vom Typ **revisit** sind Hinweise auf bereits archivierte Inhalte festgehalten.
- In Records vom Typ **conversion** können im Rahmen von Langzeiterhaltungsmassnahmen migrierte Inhalte nachträglich in der WARC-Datei gesichert werden.[116]

Herausforderungen

Das Dateiformat WARC ist zwar zentral für die Webarchivierung, doch es bringt einige Herausforderungen mit sich. Ein Problem ist die Abhängigkeit von zusätzlicher Software zur Anzeige und Darstellung des Inhaltes. Es sind zwar sämtliche Ressourcen, die zur Darstellung eines Webinhaltes notwendig sind in einer WARC-Datei enthalten, doch es ist nicht möglich das WARC autark zu nutzen. Für die Betrachtung der archivierten Website ist immer eine zusätzliche Software (Viewer) notwendig.[117] Desweitern problematisch ist es, dass WARC Files und ihr Inhalt nur aufwendig bearbeitet werden können. Von anderen Containerformaten wie ZIP ist sich der User gewohnt, dass sich diese auf sehr einfache Art und Weise entzippen und die Inhalte bearbeiten lassen. Auf den WARC-Container trifft dies nicht zu.[118] In langfristiger Perspektive ist zudem zu problematisieren, dass in einem WARC-Container eine Vielzahl sehr unterschiedlicher Dateiformate vereint sein kann. Einzelne dieser Dateiformate könnten dereinst obsolet werden. Die grosse Vielfalt innerhalb des WARC-Containers stellt das Preservation Planning vor Herausforderungen. Weimer/Schoger erläutern, dass sowohl Ansätze zur Migration als auch Ansätze zur Emulation getestet werden, um bezüglich der Gefahr vor Verlusten Abhilfe leisten zu können. Bei der Migration werden einzelne im WARC-Container enthaltene Dateiformate migriert. Bei der Emulation werden frühere Browsertypen

[116] vgl. Weimer/Schoger 2021: Seite 1–2.

[117] vgl. KOST, Katalog archivischer Dateiformate (KaD), WARC. Verfügbar unter: https://kost-ceco.ch/cms/kad_warc_de.html [17.7.2024].

[118] Austausch mit Angela Gastl (Webarchiv ETH) vom 25. März 2024.

emuliert.[119] Auf die Thematik Preservation Planning wird eingehend in Abschnitt 6 eingegangen.

5.2 Die drei verschiedenen in den Prozess involvierten Komponenten (Archive-It, digitales Langzeitarchiv, CMI AIS) mit ihren jeweiligen Elementen und die Suche nach Konkordanz

Um ein übergeordnetes Datenmodell für die Webarchivierung im UAZ erarbeiten zu können, ist es zentral, eine Übersicht über alle in den Prozess involvierten Komponenten zu gewinnen. Von Angela Gastl (Webarchiv ETH) erfolgte der zentrale Hinweis, dass im Prinzip gleichzeitig drei verschiedene Datenmodelle vorliegen mit den jeweils eigenen Elementen, die aufeinander abgestimmt werden müssen: das Datenmodell vom Webarchiv-Service Archive-It, das Datenmodell vom digitalen Langzeitarchiv und das Datenmodell vom AIS.[120] Diese grundlegende Überlegung soll als Startpunkt dienen bei der Erarbeitung eines übergeordneten Datenmodells spezifisch für die Webarchivierung im UAZ. Nachfolgend wird zunächst jeweils einzeln auf die verschiedenen in den Prozess involvierten Komponenten und ihre Elemente eingegangen, zuletzt erfolgt der Versuch einer Zusammenführung und Abstimmung aufeinander.

[119] vgl. Weimer/Schoger 2021: Seite 3.

[120] Austausch mit Angela Gastl (Webarchiv ETH) vom 25. März 2024.

Archive-It

Der Webarchiv-Service Archive-It gliedert seine Elemente in **Collection**, **Seed** und **Document**.

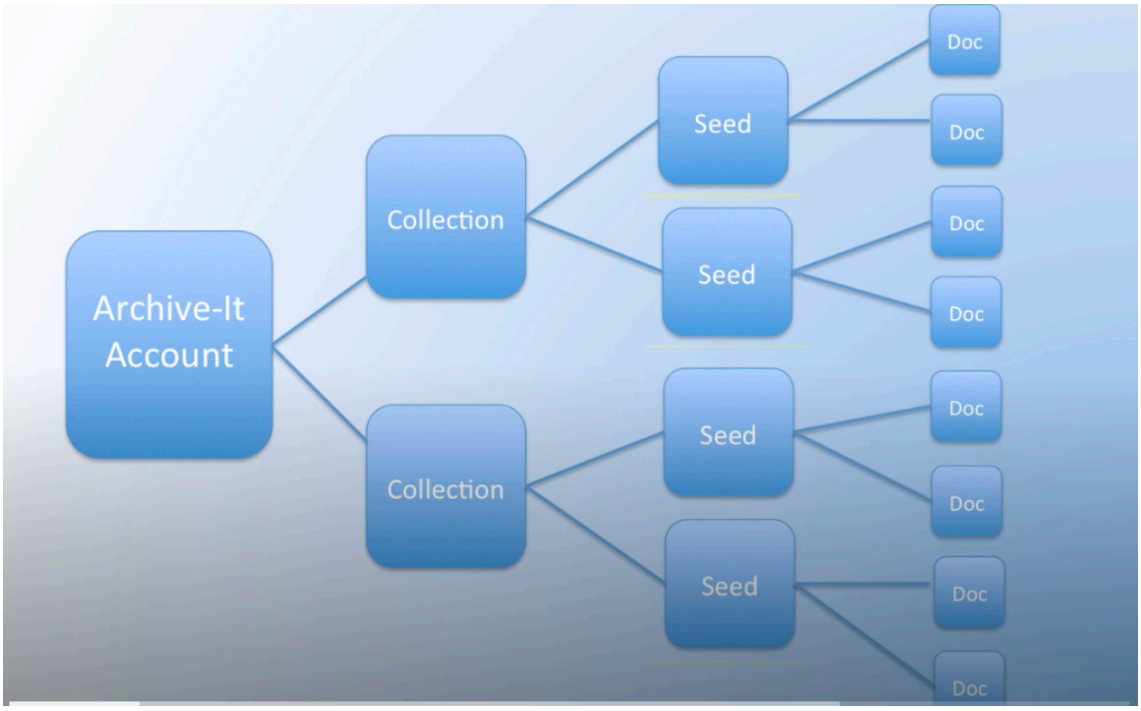

Abbildung 3: Archive-It Help Center. Navigating Archive-It (Video, Screenshot 00:48). URL.: https://support.archive-it.org/hc/en-us/articles/216489103-Archive-It-Video-Curriculum [4.8.2024].

Die Struktur der Archive-It-Elemente wird in Abbildung 3 dargestellt. Jede Partnerorganisation von Archive-It verfügt über einen eigenen Archive-It Account. In der Abbildung ist ersichtlich, dass ein Archive-It Account aus verschiedenen Collections bestehen kann. Eine Collection kann als eine Gruppe von archivierten Webdokumenten verstanden werden. Diese sind in der Collection zu einer bestimmten Thematik oder Domain kuratiert.[121] Die URLs, welche die Collection bilden, werden Seeds genannt. Beim Seed handelt es sich um ein Element in Archive-It mit einer eindeutigen ID-Nummer. Die Seed-URL teilt dem Crawler mit, an welche Stelle er sich im Live-Web bewegen muss, und dient als Zugangspunkt zu archivierten Inhalten.[122] Jeder Seed umfasst Documents. Es handelt sich dabei um Files mit einer eindeutigen URL (beispielsweise ein PDF, HTML, Video).[123] Archive-It arbeitet mit dem Dateityp WARC. Der Crawlprozess resultiert in WARC Files.

[121] vgl. Archive-It Help Center. Glossary: Stichwort «Collection». Verfügbar unter: https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms [14.7.2024].

[122] vgl. Archive-It Help Center. Glossary: Stichwort «Seed». Verfügbar unter: https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms [14.7.2024].

[123] vgl. Archive-It Help Center. Glossary: Stichwort «Document». Verfügbar unter: https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms [14.7.2024].

Die Partnerorganisationen haben die Möglichkeit ihre archivierten Websites als WARC-Container von ihrem Archive-It Account herunterzuladen.[124]

CMI AIS

Das UAZ arbeitet mit der Software *CMI AIS* der Firma CMI. Das AIS basiert auf dem Standard ISAD(G). Gearbeitet wird mit Verzeichnungseinheiten auf den Stufen **Archiv**, **Abteilung**, **Bestand**, **Dossier** und **Einzelstück**. Bestellbare Einheiten stellen Verzeichnungseinheiten der Stufen Dossier und Einzelstück dar. [125] Die Eingabemaske Serie wird im UAZ nicht verwendet und ist in *CMI AIS* ausgeblendet.[126] Zusätzlich kann für eine Optimierung der visuellen Gliederung eines Bestandes mit dem Element **Klassifikation** gearbeitet werden. Dieses Element befindet sich ausserhalb der Logik von ISAD(G). Informationen dieses Elements werden nicht nach unten vererbt und sollen ausdrücklich redundant sein.[127] Auf den Stufen Bestand, Dossier und Einzelstück sind Pflichtelemente definiert, die ausgefüllt werden müssen (unter anderem: Signatur, Titel etc.). Bei welchen Feldern es sich für welche Stufe um ein Pflichtelement handelt, kann dem Erschliessungshandbuch des UAZ im Detail entnommen werden.[128]

Digitales Langzeitarchiv

Das UAZ verwendet für die Speicherung das Open Source Repository Fedora Commons auf NetApp.[129] Als Grundlage für die digitale Archivierung im UAZ dient das Referenzmodell OAIS (Open Archival Information System). Gemäss dem OAIS Modell werden sämtliche zu archivierenden Informationen zu Paketen zusammengefasst. Im OAIS Modell sind drei verschiedene Typen solcher Pakete definiert. Diese werden entlang des Archivierungsprozesses mit unterschiedlichen Informationen angereichert. Bei einem Paket vom Typ **SIP (Submission Information Package)** handelt es sich um ein für die Ablieferung und Archivierung bereitgestelltes Paket. Ein Paket vom Typ **AIP (Archival Information Package)** wird im Archiv gespeichert. Ein Paket des Typs **DIP (Dissemination Information Package)** wird den Archivnutzenden zur Verfügung gestellt. Diese Informationspakete bestehen aus Primär- und Metadaten. Bei den Primärdaten handelt es sich um die Informationen, die bei der aktenbildenden Stelle aus ihrer Tätigkeit resultiert

[124] vgl. Blumenthal, Archive-It Help Center 2024. Verfügbar unter: https://support.archive-it.org/hc/en-us/articles/360015225051-Find-and-download-your-WARC-files-with-WASAPI [24.6.2024].

[125] vgl. Erschliessungshandbuch UZH Archiv (UAZ) V1.1 vom 22. Februar 2021: Seite 4.

[126] vgl. Erschliessungshandbuch UZH Archiv (UAZ) V1.1 vom 22. Februar 2021: Seite 3.

[127] vgl. Erschliessungshandbuch UZH Archiv (UAZ) V1.1 vom 22. Februar 2021: Seite 4.

[128] vgl. Erschliessungshandbuch UZH Archiv (UAZ) V1.1 vom 22. Februar 2021: ab Seite 7.

[129] Internes Dokument (CMI G 2018-312).

sind. Diese Primärdaten werden durch die Metadaten beschrieben. Bei den Metadaten ist zu unterscheiden zwischen technischen Metadaten, die automatisch vom System vergeben werden (beispielsweise Zeitpunkt einer Fotoaufnahme) und deskriptiven Metadaten, die durch das UAZ bewusst zur Beschreibung des Inhaltes ergänzt werden (beispielsweise Titel, Signatur). Da sich das Informationspaket auf diese Weise selbst beschreibt, kann es bis zu einem gewissen Grad ohne zusätzliche Konsultation des AIS selbständig verstanden und interpretiert werden.[130] Die Erschliessung des SIPs wird im UAZ mit dem Tool *docuteam packer* vorgenommen. Das UAZ arbeitet mit fünf Grundstrukturen, die im SIP erzeugt werden können (Undefiniert, Bestand, Klassifikation, Dossier, Einzelstück). Die oberste Ordnerebene entspricht einem Bestand oder einem Dossier. Auf der zweiten Ebene wird dann mit der Rubrik Dossier oder Einzelstück gearbeitet. Die einzelnen Dokumente, die den Dossiers oder Einzelstücken auf einer dritten Ebene zugeordnet sind, werden als «Undefiniert» markiert und es wird in diesem Fall auf eine Erfassung von Metadaten verzichtet. Wenn ein Dossier die oberste Ordnerebene bildet, dann ist für die darin enthaltenen Dokumente auf der zweiten Ebene ebenfalls die Struktur «Undefiniert» auszuwählen. Mit der Web-Applikation *docuteam feeder* durchläuft das SIP einen bestimmten Workflow, sodass es in das digitale Langzeitarchiv ingestiert werden kann. Beim DIP handelt es sich um eine ZIP-Datei, in welcher die digitalen Archivalien sowie eine XML-Datei mit den Objektmetadaten nach METS enthalten sind.[131]

Abstimmung der drei in den Prozess involvierten Komponenten aufeinander

Das Zusammenspiel zwischen dem digitalen Langzeitarchiv und *CMI AIS* hat sich aufgrund der bereits langjährigen Erfahrung mit der digitalen Langzeitarchivierung im UAZ bereits etabliert. Die beiden Komponenten mit ihren jeweiligen Elementen wurden ideal aufeinander abgestimmt. Der Grossteil der Metadaten, die als Pflichtfelder dereinst im AIS zur Verfügung stehen müssen, kann bereits bei der Vorbereitung zum Ingest in *docuteam packer* erfasst werden. Nach der Überführung der Daten in das digitale Langzeitarchiv erfolgt ein Import aus Ingest in *CMI AIS* mittels der EAD-Datei. Durch den Import werden in *CMI AIS* automatisiert die entsprechenden Verzeichnungseinheiten angelegt sowie auch die Metadaten übernommen.[132] Anschliessend erfolgt im AIS eine Nacherschliessung von einigen wenigen Metadaten, die nicht direkt vom *docuteam packer* übernommen werden konnten.

[130] vgl. UZH Archiv Policy Digitale Langzeitarchivierung Version 1.0 vom 1. Oktober 2022: Seite 7–8.

[131] Internes Dokument (CMI G 2018-312).

[132] Internes Dokument (CMI G 2018-312).

Die Aufgabe besteht nun darin, den Webarchiv-Service Archive-It mit den beiden anderen bereits zusammenspielenden Komponenten sinnvoll zu verknüpfen.

Eine Möglichkeit bestünde darin, dass jeweils ein einzelner Crawl eines Seeds in Archive-It (also ein bestimmter Zeitschnitt) einem Dossier in *CMI AIS* entsprechen könnte. Im digitalen Langzeitarchiv bestünde ein AIP dann aus einem Crawl eines Seeds. Dieser Crawl kann mehrere WARC Files umfassen. Dies würde an das im UAZ bereits etablierte Vorgehen anschliessen. Bis anhin wurde im AIS in PUB.010 jeweils die URL als eine Einheit und jeder Zeitschnitt als ein Dossier erfasst. Beispielsweise enthält das Dossier mit der Signatur PUB.010.022 einen Zeitschnitt vom 22.6.2012 von der URL www.archiv.uzh.ch.

Im nachfolgenden Abschnitt 5.3 wird eingehender auf diese Thematik eingegangen und eine weitere Option angedacht.

5.3 Diskussion zentraler Fragen

Bevor auf mögliche Lösungsansätze bezüglich der nachstehenden Fragen eingegangen wird, soll an dieser Stelle betont werden, dass das Dateiformat WARC ein zentraler Faktor darstellt, wenn es darum geht, sich Gedanken bezüglich einem Datenmodell für die Webarchivierung zu machen. Wie bereits in der Prozessbeschreibung unter «Storage and Organization» erwähnt, können die erfolgten Crawls via Schnittstelle *WASAPI* von Archive-It heruntergeladen werden. Resultat des Downloads sind die WARC Files. Wichtig ist es, sich bewusst zu sein, dass ein Crawl eines Seeds mehrere WARC Files erzeugen kann, da ein von Archive-It erstelltes WARC File eine Maximalgrösse von einem GB aufweist.[133] Das Dateiformat WARC bildet die Grundlage für alle weiterführenden Überlegungen.

[133] vgl. Blumenthal, Archive-It Help Center 2024. Verfügbar unter: https://support.archive-it.org/hc/en-us/articles/360015225051-Find-and-download-your-WARC-files-with-WASAPI [24.6.2024].

Nachfolgend ist eine Übersicht an Fragen zusammengetragen. Im vorliegenden Abschnitt werden diese Fragen thematisiert und mit möglichen Lösungsansätzen beantwortet.

- Von Archive-It lassen sich die erfolgten Crawls im WARC-Format herunterladen. Wie sollen diese einzelnen Crawls in das digitale Langzeitarchiv des UAZ überführt werden?
- Manchmal besteht ein Crawl aus mehreren WARC Files. Welche Einheit bildet das AIP im Repository?
- Welche Struktur soll ein SIP, AIP, DIP aufweisen?
- Wie gross dürfen die zu ingestierenden Dateien maximal sein? Wo liegt die Kapazitätsgrenze für ein Paket?
- Wie kann das in Abschnitt 5.2 angesprochene Zusammenspiel zwischen den Datenmodellen von Archive-It, dem digitalen Langzeitarchiv und *CMI AIS* umgesetzt werden?
- Für wie lange soll der Mastercrawl (Erstabzug) als Basis für die anschliessenden ergänzenden Deltacrawls aufrecht erhalten bleiben bis von einer neuen Basis ausgegangen wird?
- Das Dateiformat WARC kann ohne zusätzliche Hilfsmittel/Software nicht benutzt werden. Auf welche Hilfsmittel soll zurückgegriffen werden, damit mit den WARC Files gearbeitet werden kann?

Informationspakete SIP/AIP/DIP

Wenn die WARC Files von Archive-It heruntergeladen wurden und in das digitale Langzeitarchiv des UAZ überführt werden sollen, dann ist eine zentrale Frage, die thematisiert werden muss, was ein Informationspaket genau umfassen soll.

Corinne Frappart (EU Publications Office) macht in Zusammenhang mit der Modellierung des AIPs darauf aufmerksam, dass es sich um eine komplexe Angelegenheit handelt, da aus einem Crawl mehrere WARC Files resultieren können. Zudem besteht bei einer Website, die regelmässig erneut gecrawlt wird, eine zentrale Überlegung darin, wie die verschiedenen Crawls miteinander in Verbindung gebracht werden sollen. Die Frage ist dann, ob es der einzelne Crawl ist oder sämtliche Crawls einer URL, die als eine einzelne Entität in einem AIP archiviert werden soll.[134]

[134] vgl. Frappart 2023: Folie 19.

Hier an dieser Stelle sollen zwei mögliche Ansätze zur Bildung eines Informationspaketes angedacht werden.

- **Ansatz 1**

  Der Master – und die nachfolgenden Deltacrawls werden jeweils als individuelle Pakete in das digitale Langzeitarchiv ingestiert. Es wird mit einem Verweis gearbeitet, sodass bei Bedarf problemlos eruiert werden kann, welche Pakete zusammengehören. In diesem Fall würde ein Crawl eines Seeds in Archive-It (also ein bestimmter Zeitschnitt) einem einzelnen Dossier in *CMI AIS* entsprechen. Wenn aus dem Crawl mehrere WARC Files resultiert sind, dann können diese zusammen in ein AIP gepackt werden. Jeder einzelne gesicherte Zeitschnitt würde dann über eine separate Signatur verfügen und wäre in einem separaten Dossier erfasst, ganz unabhängig davon, ob es sich um den Master- oder einen nachfolgenden Deltacrawl derselben URL handelt.

  Verschiedene Fachleute empfehlen es, den Ansatz 1 zu verfolgen. Unter anderem spricht sich Youssef Eldakar (Bibliotheca Alexandrina in Ägypten, Mitglied IIPC) dafür aus, dass sämtliche WARC Files von einem Crawl zu einem AIP verbunden werden sollen.[135] Auch Corinne Frappart (EU Publications Office) spricht die Empfehlung aus, dass ein Crawl einem AIP entsprechen soll. Das AIP soll sämtliche WARC Files vom selben Crawl enthalten.[136] Die ZB betreibt zwar selbstständig keine Webarchivierung, aber es wird ein dem hier erwähnten Ansatz 1 ähnliches Vorgehen bezüglich Master- und Deltakapseln bei archivierten Daten von e-rara und e-manuscripta verfolgt. Die Master- und Deltakapseln werden jeweils als individuelle AIPs archiviert. Als Masterkapseln werden die Daten der Erstpublikation auf den Plattformen in das digitale Langzeitarchiv überführt. Treten zu einem späteren Zeitpunkt Änderungen auf (beispielsweise eine neue Transkription, OCR-Erkennung, aktualisierte Metadaten), dann wird eine Deltakapsel erstellt, die ebenfalls in das digitale Langzeitarchiv überführt wird. Welche AIPs zusammengehören bleibt nachvollziehbar über den Namen des SIPs sowie weiteren Metadaten, die im mets.xml festgehalten sind (insbesondere der Digital Object Identifier DOI als Identifikator). Folglich müssen Daten aus mehreren AIPs gemergt werden, wenn die aktuellen Daten zu einem Objekt aus mehreren Paketen aus dem

[135] Korrespondenz mit Youssef Eldakar (Bibliotheca Alexandrina, Ägypen) vom 2. Mai 2024.

[136] vgl. Frappart 2023: Folie 19.

digitalen Langzeitarchiv hergestellt werden sollen.[137] Auch die NB verfolgt diesen Ansatz für das Webarchiv Schweiz. Ein Crawl stellt die Einheit für ein AIP dar. Die einzelnen AIPs sind via Verweis miteinander verknüpft und beziehen sich so auf eine gemeinsame Domain.[138]

- **Ansatz 2**

  Der Ansatz 2 wird verfolgt, wenn es das Ziel ist, dass das Paket aus sich heraus die ganze Website herausspielen kann. Bei diesem Ansatz werden der Master- und sämtliche nachfolgenden Deltacrawls in einem einzigen Paket ingestiert.

  Diese Möglichkeit verfolgt das Hochschularchiv der ETH. Eine Verzeichnungseinheit – ein Dossier – im AIS umfasst alles, was zu einer Seed-URL gehört, also den Master- sowie sämtliche Deltacrawls. Sämtliche WARC Files von einer Seed-URL kommen in einem AIP zusammen. Die Laufzeit des Dossiers umfasst den gesamten Zeitraum bis ein neuer Mastercrawl durchgeführt wird.[139]

Eine weitere zentrale Frage ist, was zusätzlich neben den WARC Files in das Informationspaket gehört und ob die WARC Files lose oder in gezippter Form eingelagert werden sollen. Die von Archive-It heruntergeladenen Daten liegen im Dateiformat warc.gz vor, also in gezippter Form. Corinne Frappart (EU Publications Office) empfiehlt, dass die WARC Files in der gezippten Form belassen und auf diese Weise ingestiert werden sollen.[140] Für Frappart ist es zentral, dass das AIP sämtliche WARC Files vom selben Crawl enthält. Was dann noch dazukommt sei von der jeweiligen Institution abhängig, diesbezüglich besteht kein Konsens. Es kann sich dabei um verschiedene Dokumente handeln: CDX, seed list, crawl log, error log etc.[141] Wie in Abschnitt 4.4 unter dem Prozessschritt «Storage and Organization» erwähnt, ist es sinnvoll die Liste mit sämtlichen Metadaten zu einem Crawl auf Archive-It mit herunterzuladen, da in diesem File sämtliche Dokumente aufgelistet sind, welche die WARC-Container beinhalten. So wären als Basis für das Informationspaket mindestens die gezippten WARC Files sowie die Index-Liste des Crawls vorhanden.

Was die Metadaten anbelangt, ist zu beachten, dass einige wesentliche Metadaten bereits in den WARC Files selbst enthalten sind (beispielsweise die Primary Seed URL, das

[137] Korrespondenz mit Lea Fuhrer (ZB) vom 29. April 2024.

[138] Austausch mit Barbara Signori (Webarchiv Schweiz) vom 19. April 2024.

[139] Austausch mit Angela Gastl (Webarchiv ETH) vom 25. März 2024.

[140] vgl. Frappart 2023: Folie 16.

[141] vgl. Frappart 2023: Folie 19.

Harvest-Datum, Informationen zum Harvest etc.).[142] David Giaretta (Chair of Consultative Committee for Space Data Systems / Data Archive Interoperability Working Group (CCSDS-DAI)) weist darauf hin, dass die WARC Files alleine nicht das AIP bilden können, sondern dass diese noch mit zusätzlichen Informationen angereichert werden müssen. Beispielsweise kann dies eine Angabe zum GZIP-Standard sein, wenn die WARC Files in komprimierter Form vorliegen sowie Angaben zum geltenden WARC-Standard. Auch Informationen zur *Wayback*-Software und zum Webbrowser mit seiner Konfiguration seien notwendig. Ein OAIS-taugliches AIP muss eine Anzahl an Elementen aufweisen und bestimmte Bedingungen erfüllen. Das WARC File alleine erfüllt diese Anforderungen nicht. So sind im WARC File selbst beispielsweise nur wenige Informationen zu Provenienz, Kontext und Zugriffsrechten enthalten.[143] Um die WARC Files als AIP archivieren zu können, müssten also diese Angaben noch ergänzt werden.

Um aber ein tatsächlich OAIS-konformes Informationspaket zu erhalten, müsste das Informationspaket sämtliche Informationen enthalten, um autark verstanden und benutzt werden zu können. Da es aber zur Benutzung der WARC Files mindestens noch eine externe zusätzliche Software (einen Viewer) braucht, gestaltet es sich als sehr schwierig ein tatsächlich OAIS-konformes Informationspaket zu modellieren. Aus diesem Grund könnte sogar ausgesagt werden, dass es sich bei der Webarchivierung in diesem Sinne gar nicht um digitale Langzeitarchivierung handelt, wie Angela Gast (Webarchiv ETH) kritisch thematisiert.[144]

Kapazitätsgrenzen von Informationspaketen

Werden die WARC Files in komprimierter Form ingestiert (GZIP) – eine verlustfreie Kompression, die durch die WARC-Spezifikationen erlaubt ist –, dann kann Speicherplatz eingespart werden.[145] Trotzdem handelt es sich in der Regel um umfangreiche SIPs, die eingelagert werden müssen. Die Grösse der Informationspakete muss folglich im Auge behalten werden, vor allem wenn der zweite Ansatz angestrebt wird (ein Paket mit dem Master- und sämtlichen Deltacrawls in einem AIP vereint). Die Grösse von Informationspaketen spielt generell bei der digitalen Langzeitarchivierung – unabhängig von der Webarchivierung – eine entscheidende Rolle. In den Richtlinien zur Archivischen

[142] vgl. Frappart 2023: Folie 21.

[143] Consultative Committee for Space Data Systems (CCSDS), Data Archive Interoperability (DAI) Working Group, Giaretta 2020. Verfügbar unter: (https://www.youtube.com/watch?v=vdEaz109uAo [19.7.2024].

[144] Austausch mit Angela Gastl (Webarchiv ETH) vom 25. März 2024.

[145] vgl. Frappart 2023: Folie 16.

Ablieferungsschnittstelle eCH-0160 ist festgehalten, dass ein SIP maximal 8 GB gross sein sollte. Empfohlen wird aber – aus Gründen der schnelleren Übertragung und Vermittlung –, dass die Grösse eines einzelnen SIP für die Ablieferung unter 2 GB gehalten wird.[146] Wird der oben erwähnte zweite Ansatz verfolgt, dann würde diese Limite vermutlich schnell überschritten werden. In Gesprächen mit Fachleuten stellte sich heraus, dass bezüglich der gewählten Grösse von Informationspaketen sehr starke Unterschiede bestehen. Franziska Geisser (docuteam) gibt bezüglich der Grösse von SIPs/AIPs/DIPs an, dass diese beispielsweise in Zusammenhang mit e-rara/e-manuscripta stark variieren kann. Die Grösse kann sich zwischen wenigen KB (für eine Deltakapsel mit lediglich einem einzelnen Metadatenfile) bis hin zu einer Grösse von 200–300 GB (für eine Masterkapsel eines digitalisierten Werks mit mehreren 1000 Seiten) bewegen.[147] Bei der Schweizerischen Nationalbibliothek gibt es keine Kapazitätsgrenze für ein Informationspaket.[148] Anna Vögeli (Spezialistin digitale Langzeitarchivierung an der Universitätsbibliothek Bern) wiederum bezieht sich bezüglich der maximalen Grösse von Informationspaketen auf einen Hinweis, den sie von docuteam erhalten hat. Es wurde ihr geraten, dass sie jeweils vorzugsweise mehrere kleinere als ein grosses Paket einlagern soll. Je grösser das Paket ist, desto grösser ist das Risiko, dass es irgendwo beim Ingest hängen bleibt. Für die Paketierung von SIPs ist Vögeli dringlich empfohlen worden, unter einem Wert von 50 GB pro Paket zu bleiben.[149] Grundsätzlich ist es auch so, dass die Frage nach der Kapazitätsgrenze sehr stark mit dem System zusammenhängt, welches die jeweilige Institution verwendet.[150] Zu beachten ist, dass Archive-It WARC Container mit einer Maximalgrösse von einem GB bildet. Aus diesem Grund können aus einem Crawl mehrere WARC Files resultieren.[151] Im UAZ spielt die Grösse des Informationspaketes für den Ingest sicher eine grosse Rolle (also für das SIP), damit dieser problemlos erfolgen kann. Für das DIP im UAZ ist die Grösse vermutlich weniger relevant. Die Archivbenutzung erfolgt über die *Wayback Machine*, sodass die DIPs des Webarchivs vermutlich nicht sehr häufig aus dem digitalen Langzeitarchiv geholt werden müssen und ein Download nicht innert kurzer Zeit abgeschlossen sein muss. In Zusammenhang mit der Grösse – und damit verbunden der Kapazitätsgrenze – von einem SIP, ist entscheidend, was genau im SIP alles enthalten sein soll. Wenn jeder Crawl separat in einem SIP ingestiert

[146] vgl. eCH E-Government Standards (2022), eCH-0160: Seite 29.

[147] Korrespondenz mit Franziska Geisser (docuteam) vom 2. April 2024.

[148] Austausch mit Barbara Signori (Webarchiv Schweiz) vom 19. April 2024.

[149] Austausch mit Anna Vögeli (Universitätsbibliothek Bern) vom 8. Mai 2024.

[150] vgl. Frappart 2023: Folie 19.

[151] Korrespondenz mit Kody Willis (Archive-It) vom 13. Mai 2024.

wird und einem AIP entspricht, dann können die einzulagernden Pakete klein gehalten werden. Dies hätte aber zur Folge, dass sehr viele SIPs gebildet werden müssten. Personelle Ressourcen könnten geschont werden, wenn das SIP aus mehreren Crawls gebildet und dann in verschiedene AIPs aufgeteilt werden könnte und somit nicht jeder Crawl einzeln als separates SIP eingelagert werden müsste. Denkbar wäre es dann, dass monatliche Pakete für den Ingest gebildet werden könnten. Sollten diese Pakete nicht zu gross werden für den Ingest, dann wäre dies ein interessanter Lösungsansatz.

Umsetzung Zusammenspiel zwischen dem Webarchiv-Service Archive-It, dem digitalen Langzeitarchiv und *CMI AIS*: Problemfall Metadaten

Beim Zusammenspiel zwischen Archive-It, dem digitalen Langzeitarchiv und *CMI AIS* ist es wichtig, sich Gedanken zu den Metadaten zu machen. Angela Gastl (Webarchiv ETH) weist darauf hin, dass die Metadaten in den drei verschiedenen Systemen (Archive-It, digitales Langzeitarchiv und AIS) unterschiedlich sein können und teilweise sogar unterschiedlich sein müssen. Bei der ETH steht im AIS beispielsweise im Titel «Webseite: […]», dies macht auf Archive-It keinen Sinn, da es sich auf dieser Plattform bei sämtlichen gesicherten Daten um Websites handelt. Es kann folglich sein, dass Anpassungen bezüglich der Metadaten notwendig sind, wenn diese von einem System in das andere übertragen werden. Somit sind die Metadaten nicht auf allen drei in den Prozess involvierten Komponenten – Archive-It, digitales Langzeitarchiv, AIS – identisch. Wertvoll ist es, dass durch die Metadaten von einer Komponente auf die andere verwiesen werden kann. Die ETH nutzt diese Möglichkeit, indem sie in ihrem AIS direkt auf die *Wayback*-Ansicht über das Feld «Link auf digitales Original» (beispielsweise Signatur EZ-INF1.12/001) verlinkt. So können Archivnutzende direkt via Link auf die gewünschte archivierte Website in *Wayback* zugreifen.[152]

Master- und Deltacrawls

Bezüglich der Frage für wie lange der Mastercrawl als Basis für die anschliessenden ergänzenden Deltacrawls aufrecht erhalten bleiben soll bis von einer neuen Basis ausgegangen wird, besteht unter den Fachleuten kein Konsens. Angela Gastl (Webarchiv ETH) plant eine zeitliche Limite zu setzen. Ein neuer Mastercrawl soll jeweils nach einem Zeitintervall von fünf Jahren erstellt werden. Im Jahr 2026 wäre dies bei der ETH soweit. Ob im Jahr 2026 aber tatsächlich ein neuer Gesamtcrawl der Seed-URLs durchgeführt wird, ist aktuell noch unklar. Das Problem ist, dass im Viewer im Zeitstrahl jeweils nach

[152] Austausch mit Angela Gastl (Webarchiv ETH) vom 25. März 2024.

einem neuen Master die alten Ansichten nicht mehr direkt ersichtlich sind, sondern lediglich der neue Master und die auf dessen Basis nachfolgend erstellten Deltas. Für Archivnutzende ist es nicht ideal, wenn der Zeitstrahl nicht komplett abgebildet ist, sondern nur alles angezeigt wird, das seit dem letzten Gesamtcrawl gesichert wurde. Gastl erklärt, dass dieser Umstand an der ETH aber nicht so problematisch ist, da im AIS jeweils direkt eine Verknüpfung hinterlegt ist auf den Viewer zum entsprechenden Zeitschnitt (via Feld «Link auf das digitale Original»). Gastl weist darauf hin, dass es auch denkbar sei einen neuen Gesamtcrawl unabhängig von einer festgelegten Frist zu erstellen. Dies könnte beispielsweise dann sinnvoll sein, wenn das CMS wechselt oder allgemein, wenn es zu einem Medienbruch kommt.[153] Barbara Signori (Webarchiv Schweiz) betont, dass es wichtig sein könnte mit der Zeit einen neuen Mastercrawl zu erstellen, vor allem wenn der Master- zusammen mit den Deltacrawls im digitalen Langzeitarchiv gesichert werden soll. Würde in diesem Fall lange Zeit auf einen neuen Gesamtcrawl verzichtet werden, dann würde das Paket dereinst zu gross werden, um es ohne Schwierigkeiten einzulagern.[154] Annabel Walz (Friedrich Ebert Stiftung) schlägt vor, dass allenfalls anhand von Statistiken der geeignete Zeitpunkt für einen neuen Mastercrawl ausgemacht werden könnte. Im Voraus wäre zu definieren, was als Grenzwert festgelegt werden soll. Sobald dieser Grenzwert an Veränderung erreicht oder überschritten wurde, kann dann ein neuer Gesamtabzug erstellt werden.[155]

Tools für das Dateiformat WARC

Mit einem WARC File alleine kann nicht viel gemacht werden. Der Container lässt sich nicht einfach entpacken, sodass die Inhalte bearbeitet werden können, wie es etwa bei einem ZIP File der Fall ist. Auch braucht es einen spezifischen Viewer, damit der Inhalt des WARC Files – die Website – betrachtet werden kann. Es gibt verschiedene Tools, die in Zusammenhang mit WARC Files verwendet werden können. Eine Auswahl ist auf der Webseite des IIPC unter Tools & Software ersichtlich. Für die Analyse von WARC Files bieten sich unter anderem die Tools *Hadoop* und *JHOVE2* an.[156] Ein für die Benutzung sehr wichtiges Tool stellt die *Wayback Machine* dar. Die Kalenderansicht ermöglicht es, dass Benutzende verschiedene Zeitschnitte einer Website anschauen und auch miteinander vergleichen können. Als interessante Wiedergabewerkzeuge stehen *Open*

[153] Austausch mit Angela Gastl (Webarchiv ETH) vom 25. März 2024.

[154] Austausch mit Barbara Signori (Webarchiv Schweiz) vom 19. April 2024.

[155] Austausch mit Annabel Walz (Friedrich Ebert Stiftung) vom 5. Juli 2024.

[156] vgl. Lazorchak, Library of Congress Blogs 2011. Verfügbar unter: https://blogs.loc.gov/thesignal/2011/08/web-archive-preservation-planning/ [10.8.2024].

Wayback, *Python Wayback (Pywb)* und *SolrWayback* als Open Source Tools zur Verfügung.[157] Angela Gastl (Webarchiv ETH) verwendet beim Ingest in das digitale Langzeitarchiv Rosetta *JHOVE* und *DROID* für die Formaterkennung und -validierung.[158] Auf einige weitere spezifische Tools wird in Abschnitt 6 zum Preservation Planning eingegangen.

5.4 Empfehlungen bezüglich der Erarbeitung eines Datenmodells für die Webarchivierung im UZH Archiv

Die Ausführungen im voranstehenden Abschnitt 5.3 sollen als eine Anregung dazu dienen auf deren Basis Entscheidungen getroffen werden können. Im erwähnten Abschnitt wurde auf zwei mögliche Ansätze zur Bildung des Informationspaketes eingegangen. Es gibt verschiedene Argumente, die für eine Realisierung von Ansatz 1 (1 Crawl = 1 AIP) sprechen. Wird pro Crawl ein AIP gebildet und eingelagert, dann sind die einzulagernden Pakete kleiner, als wenn der Master- und sämtliche nachfolgenden Deltacrawls allesamt gemeinsam in einem Paket ingestiert werden (Ansatz 2). Wenn mit einem Verweis der einzelnen Pakete aufeinander gearbeitet wird, dann bleibt ersichtlich, welche Pakete zusammengehören. Da Archivnutzende des UAZ die archivierten Websites über die *Wayback Machine* anschauen können, ist die Wahrscheinlichkeit eher gering, dass Mitarbeitende des UAZ auf die eingelagerten AIPs (den Master- und die Deltacrawls) zugreifen und diese für eine Archivbenutzung aufwendig zusammen- und bereitstellen müssten. Daher ist mit diesem Aufwand – der bei der Verfolgung von Ansatz 2 nicht entstünde – konkret für das UAZ nicht zu rechnen und dieser Mehraufwand muss somit im Falle des UAZ nicht als Nachteil von Ansatz 1 beurteilt werden. Für den Ansatz 1 spricht ebenfalls, dass dieser der bisherigen Vorgehensweise im UAZ entspricht, wonach jeder Zeitschnitt separat in das digitale Langzeitarchiv überführt wurde.

Empfohlen wird, dass die WARC Files für den Ingest in der gezippten Form belassen werden. Da die WARC Files alleine noch kein OAIS-konformes AIP bilden, lautet eine entscheidende Frage, was zusätzlich in einem Informationspaket enthalten sein soll, d.h. welche zusätzlichen Informationen und Files zu den WARC Files miteingelagert werden sollen. Es ist eine Anreicherung von Metadaten notwendig, da im WARC File selbst nur wenige Informationen zu Provenienz, Kontext und Zugriffsrechten enthalten sind. Zusätzlich zu den WARC Files kann die über Archive-It downloadbare Liste mit sämtlichen

[157] vgl. Weimer/Schoger 2021: Seite 3.

[158] Korrespondenz mit Angela Gastl (Webarchiv ETH) vom 27. Juni 2024.

Metadaten zu einem Crawl im Informationspaket miteingelagert werden. Im AIP würde diese Liste dann als Nachweis dienen. Zusätzlich ist es sinnvoll, diese Liste auch in *CMI AIS* unter «Dateien» als Findmittel zu hinterlegen. Mit dieser Liste ist es dann möglich, nach spezifischen Inhalten zu suchen.

Bezüglich dem Zusammenspiel zwischen den Systemen Archive-It, dem digitalen Langzeitarchiv und *CMI AIS* ist es zentral, sich Gedanken über die Metadaten zu machen. Nicht in allen Systemen sind die Metadaten identisch, beispielsweise kann der Titel variieren (in *CMI AIS* ist von «Website» die Rede, bei Archive-It ist dies nicht notwendig, da es sich allesamt um archivierte Websites handelt). Damit Archivnutzende über den Online-Recherchekatalog direkt auf die gewünschte Ressource in der *Wayback*-Ansicht zugreifen können, kann in *CMI AIS* der entsprechende Link für den Zeitschnitt erfasst werden.

Bezüglich der Frage wie lange ein Mastercrawl als Basis für die anschliessenden ergänzenden Deltacrawls aufrecht erhalten bleiben soll bis von einer neuen Basis ausgegangen wird, gehen die Meinungen auseinander. Wenn für ein AIP der Master- und sämtliche Deltacrawls gemeinsam ingestiert werden sollen (Ansatz 2), dann ist es ratsam, nicht allzu lange mit einen neuen Gesamtabzug zu warten, da das zu ingestierende Paket ansonsten sehr umfangreich wird. Entspricht ein AIP einem Crawl (Ansatz 1), dann kann zugewartet werden mit der Durchführung eines neuen Gesamtcrawls bis grundlegende Veränderungen der Website vorhanden sind (beispielsweise ein neues CMS), da die Pakete separat (aber mit Verweis aufeinander) in das digitale Langzeitarchiv überführt werden.

Damit mit den WARC Files gearbeitet werden kann, stehen verschiedene Tools zur Verfügung. Das IIPC stellt auf seiner Website eine Übersicht mit verschiedenen Tools & Software zusammen.[159] Sehr zentral für die Benutzung ist die *Wayback Machine*.

5.5 Weiterführende Gedanken und offene Fragen

Wenn es um die Erstellung eines Datenmodells geht, sind noch einige Fragen offen. Die Empfehlungen in Abschnitt 5.4 sollen als erste Grundlage dienen. Im Zentrum steht vor allem die Modellierung der Informationspakete. Sind dereinst Entscheidungen getroffen worden wie vorzugehen ist, dann kann entsprechend ein Datenmodell konzipiert werden.

[159] IIPC: Tools and Software. Verfügbar unter: https://netpreserve.org/web-archiving/tools-and-software/ [10.8.2024].

Auch wäre es denkbar zu gegebener Zeit, die Elemente des Datenmodells direkt in die Prozessbeschreibung zu integrieren.

Ein interessanter Punkt ist, dass die Grundsatzfrage unter Fachleuten diskutiert wird, ob die Webarchivierung überhaupt die Anforderungen an die digitale Langzeitarchivierung erfüllen kann und ob es in Zusammenhang mit dem Dateiformat WARC überhaupt möglich ist, ein OAIS-konformes AIP zu erstellen. Problematisiert wird, dass die WARC Files immer auf externe Software (einen Viewer) angewiesen sind, damit sie benutzt werden können. Ein OAIS-konformes Informationspaket müsste aber autark funktionieren.

6 Preservation Planning

Abschnitt 6 widmet sich verschiedenen möglichen Preservation Planning-Strategien. In Abschnitt 6.1 erfolgen zunächst ein paar allgemeine Bemerkungen zum Thema Preservation Planning. Anschliessend wird in Abschnitt 6.2 auf den Lösungsansatz Migration und in Abschnitt 6.3 auf den Lösungsansatz Emulation eingegangen. In Abschnitt 6.4 folgen Empfehlungen bezüglich einer möglichen Strategie für das UAZ. In Abschnitt 6.5 schliesslich wird auf weiterführende Gedanken und offene Fragen eingegangen.

6.1 Allgemeine Bemerkungen zum Preservation Planning

Preservation Planning hat zum Ziel, dass Inhalte langfristig erhalten bleiben, trotz Veränderungen der Hard- und Software sowie allfälliger technologischer Obsoleszenz[160]. Wenn es um digitale Erhaltungsstrategien geht, dann kommen grundsätzlich zwei verschiedene Ansätze in Frage: Migration oder Emulation.

Ganz unabhängig von der favorisierten Strategie gilt es bezüglich der digitalen Langzeitarchivierung von Webinhalten zu beachten, dass das Dateiformat WARC zentral ist. Einerseits geht es dabei um den WARC-Container als Ganzes, andererseits aber auch um die diversen Files verschiedenen Formattyps, welche im WARC-Container enthalten sein können. Die möglichen Preservation Planning-Strategien haben an den WARC Files als Basis anzusetzen.

Migration

Bei der Migration werden digitale Objekte an ein neues Umfeld angepasst.[161] Eine Gefahr bezüglich Migration besteht darin, dass es zu einem Informationsverlust kommen kann. Die Migration ist verlustfrei, wenn sowohl das Original- wie auch das Ziel-Format eindeutig spezifiziert sind.[162] Bei der Migration ist es zentral, dass man sich Gedanken darüber macht, welches die signifikanten Eigenschaften sind, also welche Eigenschaften des Objekts wesentlich sind und erhalten bleiben müssen.[163] Die signifikanten Eigenschaften umfassen sowohl den Inhalt, das Layout, die Struktur sowie auch die Funktionalität einer Website. Felix Lange (Deutsches Bundesarchiv) nennt als Erhaltungsziele sowohl inhaltliche als auch strukturelle und funktionelle Aspekte. Beim Inhalt geht es um die Erhaltung

[160] vgl. Pennock 2013: Seite 15.

[161] vgl. Funk 2016: Folie 12.

[162] vgl. Neuroth/Osswald/Scheffel/Strathmann/Huth 2010: Kapitel 8.3, Seite 11.

[163] vgl. Funk 2016: Folie 13.

aller Webpages einer Website, inklusive des textuellen und audiovisuellen Inhalts. Bei der Struktur geht es um eine Abbildung der Binnenreferenzen und des HTML. Bezüglich Funktionalität geht es darum, ein Abbild der ursprünglichen Funktionalität zu erzielen sowie möglichst auch das dynamische Verhalten der Website zu erhalten.[164]

Emulation

Bei der Emulation wird das Umfeld an die digitalen Objekte angepasst.[165] Die Emulation simuliert das originäre Umfeld der digitalen Objekte. An den Dateien selbst wird nichts verändert.[166] Da an den Originaldateien nichts verändert wird, kann mit der Emulation die Gefahr eines möglichen Informationsverlustes, der bei einer Migration besteht, vermieden werden.[167]

Bisherige Preservation Planning-Strategie im UAZ

Bis anhin hat das UAZ vollumfänglich auf den Lösungsansatz der Migration gesetzt. In der Policy Digitale Langzeitarchivierung des UAZ ist festgehalten, dass das UAZ digitale Unterlagen basierend auf dem Migrationsprinzip archiviert. Eine Archivierung nach dem Migrationsprinzip stellt sicher, dass Dateien auf lange Zeit lesbar bleiben, unabhängig von ihrem Umfeld. Das UAZ kümmert sich um die Konvertierung von Dateien, sobald gewisse Dateiformate von Obsoleszenz bedroht sind. Die Konvertierung in ein neues archivtaugliches Format muss möglichst verlustfrei erfolgen und dokumentiert werden, damit die Nachvollziehbarkeit gewährleistet ist. Das UAZ steht im Austausch mit anderen Archiven und orientiert sich an den Vorgaben der KOST. Im UAZ kommt – bis anhin – weder das Emulationsprinzip noch das Technologiekonservierungsprinzip zur Anwendung.[168] Obwohl bis anhin im UAZ ausschliesslich auf eine Migrationsstrategie fokussiert wurde, soll in der vorliegenden Arbeit auch die Emulation als ein möglicher Lösungsansatz thematisiert werden.

[164] vgl. Lange 2024: Folie 4.

[165] vgl. Funk 2016: Folie 15.

[166] vgl. Neuroth/Osswald/Scheffel/Strathmann/Huth 2010: Kapitel 8.3, Seite 10.

[167] vgl. Neuroth/Osswald/Scheffel/Strathmann/Huth 2010: Kapitel 8.4, Seite 16.

[168] vgl. UZH Archiv Policy Digitale Langzeitarchivierung Version 1.0 vom 1. Oktober 2022: Seite 6–7.

Übersicht zentraler Fragen

Wenn es darum geht einen Lösungsansatz zu favorisieren, dann sind unter anderem die folgenden Fragen zentral:

- Welche Strategie ist mit weniger Aufwand realisierbar: Migration oder Emulation? Mit welcher Strategie werden weniger Ressourcen verbraucht: weniger Aufwand durch Personal und/oder auch weniger Speicherplatz?
- Welche Strategie ist sinnvoller umzusetzen? Ist es sinnvoller auf Migration zu setzen, da dies der bisherigen Tradition des UAZ entspricht?
- Für welche Strategie gibt es in der Community der Gedächtnisinstitutionen den grösseren Erfahrungsschatz?
- Hat sich eine Strategie allenfalls bereits etabliert?

6.2 Lösungsansatz Migration

Allgemeine Hinweise

Bezüglich Migration sind zwei mögliche Szenarien denkbar.

Szenario 1: Migration WARC-Container

Wäre dereinst das Dateiformat WARC von Obsoleszenz bedroht, dann müsste der WARC-Container als Ganzer in ein archivtaugliches Nachfolgeformat migriert werden. Dies sollte kein Problem darstellen. WARC gilt als Standard in der Webarchivierung. Zahlreiche Institutionen verwenden das Dateiformat WARC, darunter auch das Internet Archive. Droht das Dateiformat WARC dereinst tatsächlich obsolet zur werden, dann würde sicherlich in der Community der Gedächtnisinstitutionen rechtzeitig eine Anschlusslösung gefunden werden.

Szenario 2: Migration eines Dateiformates innerhalb des WARC-Containers

Das zweite Szenario betrifft die Files innerhalb des WARC-Containers. Ein WARC-Container enthält Files ganz unterschiedlichen Formattyps. Wenn nun ein einzelnes Format (oder auch mehrere Formate) innerhalb des WARC-Containers obsolet werden würde, dann könnte ebenfalls die Migration als mögliche Preservation Planning-Strategie in Frage kommen.

Eine Migration innerhalb eines WARC-Containers, wobei eine Preservation Action für ein Dateiformat durchgeführt wird, ist laut Andreas Rauber (Universität Wien) zum jetzigen

Zeitpunkt bereits realisierbar.[169] Eine solche gestaltet sich jedoch als aufwendig. Um eine Migration durchführen zu können, müssten zunächst alle Dateien, die das von der Obsoleszenz bedrohte Dateiformat aufweisen, ausgemacht, herausgefiltert und extrahiert werden.[170] Eine Problematik diesbezüglich stellt es allerdings dar, dass aufgrund des Containerformates die Dateien innerhalb des WARC Files nicht über das Cockpit des digitalen Langzeitarchivs überwacht werden können.

Konnten nun aber trotz der erwähnten Problematik die entsprechenden Daten erfolgreich aus dem WARC-Container extrahiert werden, dann fände in einem nächsten Schritt die Migration statt. Nach der Preservation Action wird ein neues WARC erstellt.[171] Zudem sind auch im HTML die Änderungen, die aufgrund der Preservation Action resultiert sind, zu erfassen. Dieser Punkt wird von vielen Fachleuten problematisiert, da die Anpassung des HTML einen Aufwand darstellt, der ohne Automatisierung nicht geleistet werden kann.[172]

Festzuhalten ist, dass (noch) kein «Super-Tool» existiert, das dies alles auf Knopfdruck durchführen kann. Allenfalls wäre es denkbar mit einem Workflow zu arbeiten, wobei mehrere Tools zusammenspielen. Was es bereits gibt, ist das Konzept eines solchen Workflows wie die nachfolgende Grafik zeigt. Auch wurden diesbezüglich bereits erste Versuche durchgeführt.

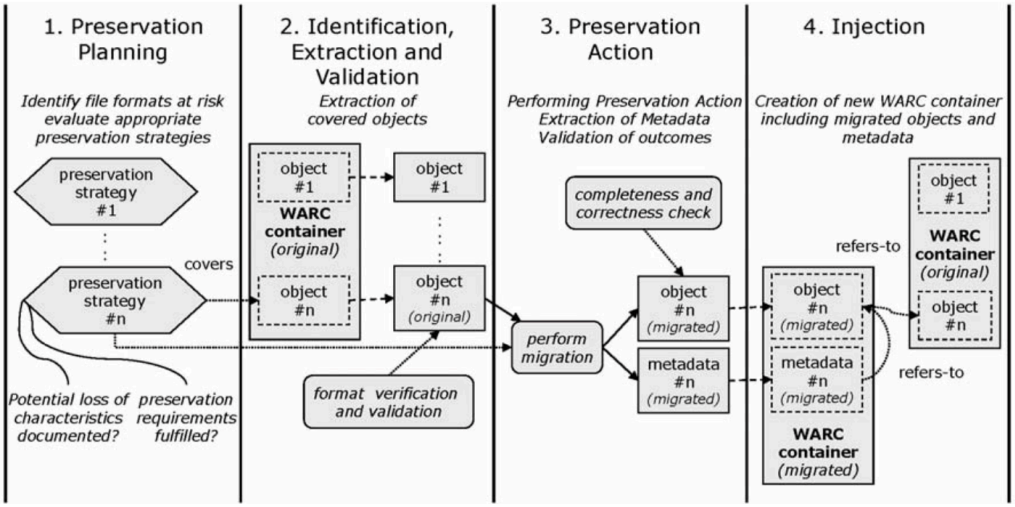

Abbildung 4: Workflow for migration of WARC records (Strodl/Beran/Rauber 2009: Seite 46)

[169] Korrespondenz mit Andreas Rauber (Universität Wien) vom 16. Juli 2024.

[170] vgl. Strodl/Beran/Rauber 2009: Seite 45.

[171] vgl. Strodl/Beran/Rauber 2009: Seite 45.

[172] Die Problematik bezüglich der Anpassung des HTML nach einer Migration wurde unter anderem im Austausch mit Barbara Signori (Webarchiv Schweiz) vom 19. April 2024 (sowie Korrespondenz mit Barbara Signori (Webarchiv Schweiz) vom 24. Juli 2024) und im Austausch mit Markus Kandlbinder (Archivinformatiker UAZ) vom 12. April 2024 thematisiert.

In Schritt 1 «Preservation Planning» in der voranstehenden Abbildung 4 wird ausgemacht, welche Dateiformate innerhalb des WARC-Containers von Obsoleszenz bedroht sind und migriert werden müssen. In Schritt 2 «Identification, Extraction and Validation» geht es dann darum, die entsprechenden Dateien zu identifizieren und zu extrahieren. In Schritt 3 «Preservation Action» findet dann die eigentliche Migration der Dateien des betroffenen Dateiformates statt. In Schritt 4 «Injection» wird ein neuer WARC-Container erstellt, in welchem die aus der Preservation Action entstandenen Objekte mit Verweisen auf die ursprünglichen Objekte sowie die extrahierten Metadaten enthalten sind. Sämtliche migrierten Objekte werden als WARC-Records des Typs conversion gespeichert. Das WARC-Feld «WARC-Refers-To» dient dabei als Verweis auf den Datensatz, der das ursprüngliche Objekt enthält (siehe dazu die nachfolgende Abbildung 5).[173]

```
WARC/0.18
WARC-Type: conversion
WARC-Target-URI: http://www.tu-sofia.bg/Bul/
norm-dok/prav-UEP.pdf
WARC-Date:
2009-07-22T11:08:24Z
WARC-Refers-To: <urn:uuid:1b5d742f-2f6c-4f03-
8a79-3dd9551b9570>
WARC-Record-ID: <urn:uuid:cf7e6b9a-4f26-447b-
9a2a-25c0cc5e419e>
Content-Type: txt/pdf
Content-Length: 64799
...
```

Abbildung 5: WARC-Record des Typs conversion: Die Abbildung zeigt den Header eines migrierten Objektes. In diesem Fall wurde ein Word File in ein PDF umgewandelt (Strodl/Beran/Rauber 2009: Seite 46)

Der Beitrag «Migrating Content in WARC Files»[174] stammt aus dem Jahr 2009. Im Beitrag werden erste kleinere Migrationen präsentiert. Seit der Publikation des Beitrages ist einige Zeit vergangen. Unter den Fachleuten sind keine weiteren Bemühungen oder Anschlusspublikationen in diesem Zusammenhang bekannt.[175]

Migration on Request Strategy

In Bezug auf eine Preservation Planning-Strategie sind auch Ansätze im Gespräch, bei welchen nicht vollumfänglich migriert werden muss. Dies trägt dazu bei, den Aufwand

[173] vgl. Strodl/Beran/Rauber 2009: Seite 45–46.

[174] Strodl/Beran/Rauber 2009.

[175] Die Anfrage nach einer Anschlusspublikation zum Beitrag von 2009 und weiteren Erfahrungen bezüglich Migration wurde an diverse Fachleute gestellt. Unter anderem wurde die Thematik im Austausch mit Annabel Walz (Friedrich Ebert Stiftung) vom 5. Juli 2024 und in der Korrespondenz mit Andreas Rauber (Universität Wien) vom 16. Juli 2024 angesprochen.

etwas überschaubarer zu halten. Pennock verweist auf eine «Migration on Request Strategy», bei welcher ein speziell entwickelter Webbrowser auf Anfrage des Nutzers automatisch Daten migriert.[176] Auch Stephanie Kortyla (Sächsisches Staatsarchiv) kann sich vorstellen, eine Strategie zu verfolgen, bei welcher nur bei konkretem Bedarf migriert wird. Wird für eine Benutzung ein konkreter Zeitschnitt benötigt und stellt das Archivpersonal bei der Aufbereitung dieser Daten fest, dass die Inhalte nicht mehr korrekt angezeigt werden, dann würde bei diesem konkreten Zeitschnitt die notwendige Migration durchgeführt.[177]

Weitere Ansätze

Ein weiterer Ansatz besteht darin, zusätzlich zum WARC File auch einen ZIP-Container miteinzulagern, der sämtliche Files enthält. Mit diesen Dateien könnten dann auf einfache Art und Weise Migrationen durchgeführt werden. So müsste der WARC-Container nicht mithilfe von Tools für die Durchführung einer Preservation Action geöffnet werden.[178] Das ingestierte ZIP File würde dann erst bei Bedarf (beispielsweise für eine Archivbenutzung) in das Dateiformat WARC umgewandelt.

Die ETH hat noch keine definitive Strategie bezüglich Preservation Planning entwickelt. Angela Gastl (Webarchiv ETH) weist darauf hin, dass es auch eine Option darstellen könnte, den Ansatz zu verfolgen, konkrete archivwürdige Dateien auf Websites doppelt zu überliefern – einmal über das Webarchiv (also in die Website integriert) und einmal als separates File. Ein PDF würde beispielsweise zusätzlich zur Webarchivierung auch noch als einzelnes PDF heruntergeladen und auf diese Weise langzeitarchiviert werden. Es könnte sich dabei beispielsweise um die Rechtssammlung, einen Geschäftsbericht oder andere graue Literatur handeln.[179]

Vorteile Migration

Angenommen es gibt dereinst ein Tool, das auf relativ einfache Art und Weise eine Migration durchführen kann, dann müsste nicht extra ein spezifischer Emulator gebaut werden. Dies ist von Vorteil, da das Erstellen von Emulatoren sehr aufwendig sein kann. Ein weiterer Vorteil besteht darin, dass das Dateiformat WARC Migrationsbestrebungen unterstützt. So besteht die Möglichkeit Migrationen in WARC-Records vom Typ conversion

[176] vgl. Pennock 2013: Seite 15.

[177] Austausch mit Stephanie Kortyla (Sächsisches Staatsarchiv) am 20. Juni 2024.

[178] Austausch mit Felix Lange (Deutsches Bundesarchiv) vom 7. Juni 2024.

[179] Korrespondenz mit Angela Gastl (Webarchiv ETH) vom 27. Juni 2024.

zu dokumentieren. Diese Möglichkeit ist allerdings noch nirgends implementiert.[180] Im Austausch mit Fachleuten stellte sich ausserdem heraus, dass dieser WARC-Typ noch nicht sehr bekannt ist. Ein weiterer Vorteil bezüglich Migration besteht darin, dass die Mitarbeitenden des UAZ sich bereits eingehend mit dieser Preservation Planning-Strategie auseinandergesetzt haben und – wie bereits erwähnt – in der Policy Digitale Langzeitarchivierung des UAZ verschriftlicht wurde, dass diese Strategie verfolgt werden soll.

Nachteile Migration

Ein Nachteil der Migration ist es, dass damit unter Umständen ein hoher Speicherbedarf verbunden sein kann. Zusätzlich zu den migrierten Inhalten kann es sinnvoll sein, auch die originären Objekte zu behalten, um bei Bedarf zu einem späteren Zeitpunkt darauf zurückgreifen zu können. Alternativ kann ein Archiv sich aber auch darauf beschränken, lediglich die aktuellste Version zu erhalten sowie eine Dokumentation der Migrationsschritte. Das UAZ plant nach der zweiten Variante zu verfahren, das heisst, dass die Erhaltungsmassnahme dokumentiert und nur jeweils die aktuelle Version des AIP behalten wird.[181] Ein weiterer Nachteil besteht darin, dass über die Zeit mehrere Migrationen notwendig sein können. Dies ist nicht vorauszusehen, sondern die Dateiformate sind stetig zu überwachen. Die Überwachung von Dateiformaten ist als kontinuierlicher Prozess zu verstehen, der nie abgeschlossen ist. Weiterhin problematisch ist auch die Gefahr von Informationsverlusten. Es ist zentral, dass darauf geachtet wird, dass es bei Migrationen nicht zu einem Informationsverlust kommt. Ein weiterer Nachteil besteht darin, dass ohne effizientes Tool die Migration derzeit aufgrund des grossen Aufwandes kaum durchführbar ist. Neben den einzelnen zu konvertierenden Files müsste auch das HTML mit den darin enthaltenden Verknüpfungen (Links auf Dateien) aktualisiert werden. Vor allem auch für die Anpassung des HTML wäre eine Automatisierung mittels eines Tools notwendig. Werden die Filenamen der migrierten Dateien im HTML nicht angepasst, dann kann die Website im Viewer nicht mehr korrekt angezeigt werden. Zwischen den Files besteht eine komplexe Beziehung. Diese Struktur und die Links zwischen den Files müssen erhalten bleiben.[182] Ein weiterer Nachteil ist, dass es bezüglich der Darstellung von Flash keine migrationsbasierten Lösungen gibt.[183]

[180] vgl. Blumenthal, Archive-It Blog 2021. Verfügbar unter: https://archive-it.org/post/the-stack-warc-file/ [24.7.2024].

[181] vgl. UZH Archiv Policy Digitale Langzeitarchivierung Version 1.0 vom 1. Oktober 2022: Seite 14.

[182] Austausch mit Markus Kandlbinder (Archivinformatiker UAZ) vom 12. April 2024.

[183] vgl. Veenendaal/Takema/Wijsman/Rappard 2023: Seite 5.

Tools

Ein einzelnes Tool, welches die Migration gesamthaft durchführen kann, gibt es nicht. Es existieren aber einige interessante Tools, die einen Teil des Migrations-Workflows durchführen können. Das IIPC weist auf seiner Website auf verschiedene geeignete Tools hin.[184] Auch GitHub listet eine Anzahl unterschiedlicher Tools auf.[185] Stephan Strodl, Peter Paul Beran und Andreas Rauber (Universität Wien) stellen in ihrem gemeinsamen Beitrag «Migrating Content in WARC Files» aus dem Jahr 2009 verschiedene Tools vor, welche die Extraktion, Migration, Validierung und Injektion von Objekten in WARC Files unterstützen.[186] Nachfolgend sind einige interessante Tools erwähnt:

Um eine Migration durchführen zu können müsste in einem ersten Schritt das komprimierte warc.gz entpackt werden. Dies ist mit dem Tool *7-Zip* möglich.[187] Um in einen WARC-Container hineinschauen und herausfinden zu können, was für Dateiformate darin vorhanden sind, ist eine Indexierung notwendig. Mit dem vom UK Web Archive entwickelten Tool *WARC-Indexer* ist es möglich die Indexierung durchzuführen.[188] Wird das WARC File mit dem Texteditor *Notepad++* geöffnet, dann ist es möglich, dass gesamte WARC und die darin enthaltenen Bausteine zu betrachten. Mit der Tastenkombination Ctrl + F kann nach Fileendungen gesucht werden.[189] Bevor die Extraktion stattfindet, ist es wichtig sicherzustellen, dass auch die richtigen Objekte extrahiert werden. Da die Dateiendung nicht als zuverlässiger Indikator für das Format dienen kann, ist es notwendig, dass die Objekte mit Hilfe von Formaterkennungsdiensten identifiziert werden. Dazu kann *DROID* von The National Archives verwendet werden.[190] Mit der quelloffenen Software *Archives Unleashed Toolkit* kann die Extraktion von Daten aus WARC-Dateien erfolgen.[191] Als ein weiteres Tool für die Extraktion von Files eignet sich *webrecorder/warcio*. Das Tool *webrecorder/warcit* bietet die Möglichkeit WARC-Records vom Typ conversion anzulegen.[192] Migrationstools gibt es verschiedene. Als Tool zur Migration von Bildern (konkret von GIF respektive JPG in PNG) verwenden Strodl/Beran/Rauber in

[184] IIPC: Tools and Software. Verfügbar unter: https://netpreserve.org/web-archiving/tools-and-software/ [10.8.2024].

[185] GitHub: Awesome Web Archiving. Verfügbar unter: https://github.com/iipc/awesome-web-archiving [4.8.2024].

[186] Strodl/Beran/Rauber 2009.

[187] Das Tool ist verfügbar unter: https://www.7-zip.org/ [10.8.2024].

[188] vgl. Walz, FEShistory Blog 2023. Verfügbar unter: https://www.fes.de/feshistory/blog/archivierte-webseiten-mit-pywb-und-solrwayback-nutzen [24.7.2024].

[189] vgl. Veenendaal/Takema/Wijsman/Rappard 2023: Seite 3.

[190] vgl. Strodl/Beran/Rauber 2009: Seite 45.

[191] vgl. Schoger/Beinert/Schmid/Donig/Eckl 2021: Folie 18.

[192] Austausch mit Annabel Walz (Friedrich Ebert Stiftung) vom 5. Juli 2024.

«Migrating Content in WARC Files» das Tool *ImageMagick*. Für die Migration von Word Files in PDF kommt bei Strodl/Beran/Rauber der *JavaOpenDocument Converter* zum Einsatz.[193]

6.3 Lösungsansatz Emulation

Allgemeine Hinweise

Eine grundsätzliche Frage bezüglich Emulation, die es zunächst zu klären gilt, lautet: Was, respektive welche Komponente(n) soll(en) überhaupt emuliert werden?

Browseremulation

Masanès weist darauf hin, dass sich die Webstandards und die von den Browsern unterstützten Technologien im Laufe der Zeit verändern, sodass es wichtig sein kann, mehrere Versionen der Browsersoftware aufzubewahren, um sicherzustellen, dass die archivierten Websites in geeigneter Weise wiedergegeben werden können.[194] Auch die KOST macht darauf aufmerksam, dass die Authentizität bei der Darstellung nur mit Browseremulation zu erreichen ist. Zudem weist die KOST in ihrer Studie aus dem Jahr 2016 auf den raschen Technologiewandel und in diesem Zusammenhang auf immer wieder neue daraus resultierende Browsergenerationen hin.[195] Tony Franzky (Projektleiter Digitales Archiv und Magazin am Erzbischöflichen Archiv Freiburg) hingegen vermutet, dass es bezüglich Browseremulation ausreichend wäre, wenn alle paar Jahre ein neuer Emulator erstellt wird. Der Generationenwechsel habe sich unterdessen etwas verlangsamt.[196]

Es wäre folglich eine Option, dass bei jedem Generationenwechsel des Browsers ein neuer Emulator entwickelt würde. Allenfalls wäre es aber auch denkbar unabhängig von einem konkreten Generationenwechsel Emulatoren in einem definierten zeitlichen Abstand zu bauen, beispielsweise alle fünf Jahre. Je nachdem, welche Ressourcen sich Archivnutzende anschauen möchten, müsste für das WARC File dann der entsprechende Browseremulator angesteuert werden. Auf jeden Fall wäre die Browseremulation ein möglicher Ansatz für das Preservation Planning.

[193] vgl. Strodl/Beran/Rauber 2009: Seite 48.

[194] vgl. Masanès 2006: Seite 186.

[195] vgl. KOST 2016: Seite 1.

[196] Austausch mit Tony Franzky (Erzbischöfliches Archiv Freiburg) vom 24. Mai 2024.

Emulation der Browserumgebung

Umfassender als die Browseremulation wäre es aber auch denkbar, dass die gesamte Browserumgebung mitemuliert würde. In diesem Fall wird das komplette Ökosystem, das auf dem Server läuft, mitgesichert. Dieses Vorhaben ist allerdings mit sehr grossem Aufwand verbunden. Tony Franzky (Projektleiter Digitales Archiv und Magazin am Erzbischöflichen Archiv in Freiburg) hat diesbezüglich mit anderen Fachleuten erste Versuche durchgeführt.[197] Auch aus der Sicht von Tobias Wildi ist es sinnvoller, die Emulation nicht nur auf den Browser einzuschränken, sondern die Browserumgebung mitzuberücksichtigen. Vor allem die Plug-Ins und die Flashwebsites sind entscheidend, nicht der Browser allein. Ohne die Emulation von Flash ist es nicht möglich, sich archivierte Websites anzeigen zu lassen, die zwischen 2000 und 2010 online verfügbar waren.[198] Flash stellt bezüglich dem Webauftritt eine wichtige Komponente dar. Bei *Adobe Flash* handelt es sich um eine Softwareplattform, die es ermöglicht Animationen, Webvideos und Webanwendungen zu erstellen. *Adobe Flash* wurde in erster Linie zu Gestaltungszwecken eingesetzt. Der *Adobe Flash Player* war der Viewer, mit dem die mit *Adobe Flash* erstellten Inhalte angezeigt werden konnten. Flash verfügte über eine große Nutzerbasis für die Erstellung interaktiver Websites. Mit der Einführung von HTML5 nahm diese jedoch ab. Außerdem wurden Sicherheitsprobleme festgestellt, die Adobe dazu veranlassten, auf die Adobe Air-Plattform umzusteigen. Der *Flash Player* ist schliesslich im Jahr 2017 veraltet und wurde 2020 für die Nutzer ausserhalb Chinas und für sämtliche Privatnutzer eingestellt.[199] Felix Lange (Deutsches Bundesarchiv) macht ebenfalls darauf aufmerksam, dass der Flashplayer ein grosses Problem darstellt. Eine Darstellung solcher Ressourcen sei für das Internet Archive gar nicht mehr möglich. Lange weist darauf hin, dass derzeit bezüglich Preservation Planning-Bestrebungen Java-Script gänzlich ausgeklammert werden muss. Irgendwann müsse zwingend ein Java-Script-Emulator gebaut werden.[200]

Vorteile Emulation

Ein grosser Vorteil der Emulation besteht darin, dass die Originalobjekte unverändert bleiben. Sie verbleiben über die Zeit so wie sie ursprünglich waren, die Authentizität ist gewährleistet. Eine Konvertierung der Objekte ist nicht notwendig.

---

[197] Austausch mit Tony Franzky (Erzbischöfliches Archiv Freiburg) vom 24. Mai 2024.

[198] Zwischengespräch zur Masterarbeit mit Tobias Wildi und Ana Petrus vom 10. Juni 2024.

[199] vgl. Veenendaal/Takema/Wijsman/Rappard 2023: Seite 2.

[200] Austausch mit Felix Lange (Deutsches Bundesarchiv) am 7. Juni 2024.

Im Austausch mit Fachleuten stellte sich heraus, dass die Emulation von vielen gegenüber der Migration als der wahrscheinlichere Lösungsansatz bewertet wird. Dabei geht es vor allem um die Flash-Inhalte in den WARC Files. Annabel Walz (Friedrich Ebert Stiftung) betont, dass gerade für Flash eine Browseremulation unverzichtbar ist.[201] Bjarne Andersen (Royal Danish Library) spricht sich klar für eine Preservaton Planning-Strategie mittels Emulation aus. Er vermutet, dass dies langfristig (in den nächsten 50–100 Jahren) die beste Strategie sein wird, da bis dahin sehr viele Dateiformate obsolet sein werden. Auch gäbe es bereits jetzt zuverlässige Emulatoren für alte Betriebssysteme wie Windows 3, Windows 3.11 etc.[202]

Nachteile Emulation

Ein Nachteil der Emulation besteht darin, dass für komplexe Objekte/Systeme (wie Betriebssysteme oder Anwendungsprogramme) die Emulatoren technisch schwer zu implementieren sind. Zudem entsteht ein hoher Aufwand pro Hardware-Generationenwechsel. Für jede Plattform ist es notwendig, dass neue Emulatoren entwickelt werden. Eine weitere Herausforderung besteht darin, dass die Spezifikationen für die Objekte/Systeme, die emuliert werden sollen, nicht immer hinreichend bekannt sind.[203] Emulationsbestrebungen können sehr komplex und umfassend sein. Es kann erforderlich sein, dass für die Emulation bestimmter Software zur Darstellung von Daten sowohl die Erhaltung des Betriebssystems, als auch der Anwendungssoftware und der Daten gleichzeitig sichergestellt werden muss. Die Daten wären verloren und könnten nicht wiederhergestellt werden, wenn auch nur eine dieser Komponenten ausfallen würde.[204] Dies alles erfordert sehr viel Know-How und Personalkosten für die Fachleute, die sich dieser Arbeit widmen.

Ein weiterer Aspekt, den es zu bedenken gilt, ist, dass auch Emulatoren als Software über die Zeit erhalten bleiben müssen. Es ist also auch bezüglich der Emulatoren ein Preservation Planning zu betreiben. Dies ist eine wichtige Erkenntnis, wenn man sich für eine Strategie entscheiden möchte. Mit einer Emulation ist das Problem nicht für immer gelöst, sondern es ist entscheidend, wachsam zu bleiben und bei Bedarf die notwendigen Anpassungen vorzunehmen, damit der Emulator auch in Zukunft zur Verfügung steht.

Tony Franzky (Projektleiter Digitales Archiv und Magazin am Erzbischöflichen Archiv Freiburg) macht auf einen weiteren Nachteil der Emulation aufmerksam. Bei der

[201] Austausch mit Annabel Walz (Friedrich Ebert Stiftung) vom 5. Juli 2024.

[202] Korrespondenz mit Bjarne Andersen (Royal Danish Library) vom 3. Mai 2024.

[203] vgl. Neuroth/Osswald/Scheffel/Strathmann/Huth 2010: Kapitel 8, Seite 23.

[204] vgl. Strodl/Beran/Rauber 2009: Seite 45.

Emulation kann es sich als ein Problem herausstellen, dass sich die Nutzergewohnheiten mit der Zeit verändern. So kann es sein, dass ein Eingabeinterface eine bestimmte Tastenkombination erfordert, die nicht mehr bekannt oder überliefert ist, sodass mit einem Emulator dereinst vielleicht gar nichts mehr angefangen werden kann. Bei Emulatoren braucht es also auch immer eine Medienkompetenz, die es zu erhalten gilt. Bei der Migration ist dies nicht der Fall.[205]

Eine weitere Herausforderung besteht darin, dass die Softwarelizenzierung häufig Rechtsstreitigkeiten auslöst. Die Rechtsverhältnisse müssen eindeutig ausgemacht werden, bevor emuliert wird.[206]

Tools

Emulationslösungen sind verfügbar und im Open Source-Bereich erhältlich. Auch ist es vorstellbar, dass in Zukunft Dienstleister die Emulation der Umgebung übernehmen könnten, im Sinne von «Emulation as a Service».[207] Nachfolgend werden einige interessante Tools erwähnt.

Bei Webrecorder handelt es sich um eine Plattform, die verschiedene Open Source-Tools zur Verfügung stellt, unter anderem auch zur Softwareemulation: beispielsweise *oldweb.today* und *pywb-remote-browsers*.[208] Pennock weist auf das Projekt Keeping Emulation Environments Portable (KEEP) hin. In Zusammenarbeit mit der Niederländischen Nationalbibliothek sollte aus dem Projekt ein Emulations-Framework resultieren, welches es ermöglicht, archivierte Websites im Laufe der Zeit zu erhalten und korrekt wiederzugeben.[209] Das Projekt fand 2012 seinen Abschluss und die Software steht über die Emulation Framework (EF)-Website zur Verfügung.[210] Bezüglich einer Lösung der Flash-Problematik hat die Niederländische Nationalbibliothek Erfahrungen mit dem Tool *Ruffle* gesammelt. Über die *Ruffle*-Website kann ein Flash-Emulator heruntergeladen werden. Dieser ermöglicht es dem Benutzer, lose Flash-Objekte zu rendern, während das Browser-Plugin die Objekte auf den Websites selbst anzeigt.[211]

---

[205] Austausch mit Tony Franzky (Erzbischöfliches Archiv Freiburg) vom 24. Mai 2024.

[206] vgl. Digital Preservation Coalition. Digital Preservation Handbook 2015, Preservation action. Verfügbar unter: https://www.dpconline.org/handbook/organisational-activities/preservation-action [24.7.2024].

[207] Austausch mit Annabel Walz (Friedrich Ebert Stiftung) vom 5. Juli 2024.

[208] Die Tools sind verfügbar unter: https://webrecorder.net/tools [24.7.2024].

[209] vgl. Pennock 2013: Seite 16.

[210] vgl. KEEP project: Emulation Framework (EF). Verfügbar unter: https://emuframework.sourceforge.net/ [11.8.2024].

[211] vgl. Veenendaal/Takema/Wijsman/Rappard 2023: Seite 5.

6.4 Empfehlungen bezüglich der Erarbeitung einer Preservation Planning-Strategie für die Webarchivierung im UZH Archiv

Da keine Gedächtnisinstitution bekannt ist, die bereits eine konkrete Preservation Action durchgeführt hat, ist es schwierig an dieser Stelle eine Empfehlung auszuformulieren, da es an der Erfahrung in der Praxis fehlt. Es gibt derzeit sowohl Fachleute, die sich eher für Migration, als auch solche, die sich für Emulation aussprechen. Wobei Emulation von vielen als die wahrscheinlichere Lösungsstrategie genannt wird. Zum jetzigen Zeitpunkt scheint Emulation realisierbarer zu sein und mit weniger Aufwand verbunden.

Da sowohl Migrations- als auch Emulationsvorhaben möglich sind, wird die Entscheidung für eine Preservation Planning-Strategie letztlich auch von den verfügbaren Ressourcen (Personal, Budget) sowie dem mit der Preservation Action verbundenen Aufwand (Personalaufwand, Speicherplatz) abhängig sein. Aufwand und Ertrag müssen sorgfältig gegeneinander abgewogen werden. Der Bedarf an personellen Ressourcen bezüglich der Migration hängt auch stark damit zusammen, welche Tools verwendet werden sollen und auch in welche Richtung sich diese Tools in Zukunft entwickeln werden. Bezüglich der verfügbaren Ressourcen ist es auch zentral festzustellen, wie oft es einen neuen Emulator brauchen würde und wie aufwendig es ist, diesen zu bauen. Bezüglich Speicherplatz steht die Migration eher im Nachteil, da es sinnvoll sein kann, das originäre WARC File zusätzlich zum migrierten WARC File zu behalten, ein Vorgehen, das eine hohe Speicherplatzbelegung mit sich bringt. Alternativ könnte hierbei aber auch lediglich die aktuellste Version gesichert werden sowie eine Dokumentation, die sämtliche Migrationsschritte schriftlich festhält – ein Vorgehen, welches das UAZ anstrebt. Der Speicherplatzbedarf darf aber auch bezüglich Emulatoren nicht unterschätzt werden. Wenn der Browser und seine Umgebung in einer gewissen Regelmässigkeit emuliert werden müssen, dann braucht es ebenfalls verfügbaren Speicherplatz, um die Software sichern zu können. Da es noch an der praktischen Erfahrung fehlt, kann nicht klar ausgesagt werden, welche Option ressourcensparender ist. Bezüglich Effizienz ist es zentral, dass die weitere Entwicklung der verfügbaren Tools im Auge behalten wird. Felix Lange (Deutsches Bundesarchiv) weist darauf hin, dass auch die Entwicklung der KI in diesem Bereich von Interesse sein kann. Es ist denkbar, dass die KI allenfalls dereinst spezifische Aufgaben übernehmen könnte. [212]

[212] Austausch mit Felix Lange (Deutsches Bundesarchiv) vom 24. Mai 2024.

Es gibt Fachleute, die sich einen Lösungsansatz vorstellen könnten, der Migration und Emulation miteinander kombiniert. Annabel Walz (Friedrich Ebert Stiftung) bringt die Überlegung an, ob sich dereinst allenfalls auch eine Mischform zwischen Emulation und Migration durchsetzen könnte, dies vor allem auch, da zum Abspielen der WARC Files immer ein Browser benötigt werden wird.[213] Auch die in Abschnitt 6.2 erwähnte «Migration on Request Strategy» ist ein Ansatz, der in der Community diskutiert wird. Wenn die Zeitschnitte für Archivnutzende aus dem digitalen Langzeitarchiv herausgeholt und von Archivmitarbeitenden geprüft werden würden, dann wäre dies ein interessanter Ansatz, da bei Bedarf Migrationen durchgeführt werden könnten. Doch für das UAZ ist dieser Ansatz nicht geeignet, da sich Archivnutzende die archivierten Websites über die *Wayback Machine* anzeigen lassen können. So gelangt die Information, welche Bestandteile einer Seite nicht mehr korrekt angezeigt werden können, gar nicht an das UAZ.

Viele Fachleute möchten sich noch nicht auf eine konkrete Strategie festlegen und visieren eher ein wachsames Abwarten und Beobachten an. Da es sich beim Dateiformat WARC um einen Standard handelt, der weit verbreitet ist, ist die Überzeugung unter den Fachleuten gross, dass in der Community der Gedächtnisinstitutionen zu gegebener Zeit eine gute Lösung gefunden werden wird. Felix Lange (Deutsches Bundesarchiv) schlägt vor, derzeit möglichst vollständige und gut dokumentierte AIPs zu bilden, an denen möglichst über Jahrzehnte keine Preservation Action durchgeführt werden muss. So ist es realistisch, sich über die nächsten Jahrzehnte zu «retten» und Zeit zu gewinnen, um durchdachte Lösungsansätze zu erarbeiten. Auch empfiehlt er, dass eine Arbeitsgruppe zur Webarchivierung mit Schwerpunkt Preservation Planning gebildet wird. Dazu müsste das richtige Format gefunden werden, dies kann beispielsweise nestor oder auch der VSA sein. Diese Arbeitsgruppe könnte gemeinsam allgemeine Strategien erarbeiten, sodass nicht jede Institution auf sich allein gestellt ist und individuelle Lösungen erarbeiten muss.[214] Von zahlreichen Fachleuten wurde der Wunsch nach einem intensiveren Austausch geäussert, sodass es realistisch ist, dass dereinst eine solche Arbeitsgruppe gebildet werden kann, die sich intensiv mit der Thematik auseinandersetzt.

Auf der Suche nach konkreten Lösungen für das UAZ ist wichtig zu beachten, dass die zu überwachenden Daten, die von Obsoleszenz bedroht sein können, an zwei unterschiedlichen Stellen archiviert werden: einerseits auf den Servern vom Internet Archive, andererseits im digitalen Langzeitarchiv des UAZ. Das UAZ kann ausschliesslich in

[213] Austausch mit Annabel Walz (Friedrich Ebert Stiftung) vom 5. Juli 2024.

[214] Austausch mit Felix Lange (Deutsches Bundesarchiv) vom 24. Mai 2024.

Bezug auf die Daten im eigenen digitalen Langzeitarchiv aktiv Preservation Planning betreiben. Im Rahmen der vorliegenden Arbeit wurde eine Anfrage an Archive-It gerichtet bezüglich allfällig geplanten Preservation Planning-Strategien des Internet Archive. Eine Rückmeldung ist derzeit noch ausstehend.[215] Einige konkrete Erhaltungsmassnahmen des Internet Archive konnten über einen Beitrag im Helpcenter ausgemacht werden. So führt das Internet Archive für sämtliche über Archive-It gesicherten Webarchive regelmässig integritiy checks durch, um die Unveränderlichkeit der archivierten Daten sicherstellen zu können. Zudem werden die archivierten Daten regelmässig auf neue physische Medien migriert, um die Zuverlässigkeit der physischen Medien proaktiv zu gewährleisten. Treten Probleme bei der Hardware auf, dann wird dank Überwachungs-, Protokollierungs- und Benachrichtigungssysteme eine entsprechende Meldung an ein Bereitschaftsteam, welches für die Wartung der Infrastruktur zuständig ist, ausgelöst.[216] Dies sind die verfügbaren Informationen in Zusammenhang mit Erhaltungsmassnahmen durch das Internet Archive. Informationen bezüglich einer intendierten Preservation Planning-Strategie sowie Angaben zu diesbezüglichen Migrations- oder Emulationsvorhaben konnten nicht ausfindig gemacht werden.

Angenommen das Internet Archive führt dereinst selbst Migrationen durch, dann müsste das UAZ an den im eigenen digitalen Langzeitarchiv gesicherten Daten keine Preservation Action durchführen, sondern könnte die migrierten Daten von Archive-It herunterladen und die alten Daten im digitalen Langzeitarchiv mit den migrierten Daten ersetzen. Archive-It müsste in diesem Fall eine Dokumentation mitliefern, die aufführt, welche Migrationsschritte konkret bei der Preservation Action durchgeführt wurden.

Setzt das Internet Archive nicht auf Migration sondern auf Emulation, dann ist es allenfalls vorstellbar, dass Emulatoren direkt in die *Wayback Machine* integriert werden können, sofern dies technisch möglich ist. Wenn sich Archivnutzende dann einen spezifischen Zeitschnitt anschauen möchten und diesen anwählen, dann könnte im Viewer im Hintergrund allenfalls der entsprechende Emulator angesteuert werden, der die WARC-Daten spezifisch aus diesem Zeitraum anzeigen kann. Allenfalls könnte das UAZ – sowie auch andere Gedächtnisinstitutionen – diese Emulatoren über das Internet Archive beziehen, damit auch im eigenen Archiv darüber verfügt werden kann und Daten aus dem eigenen digitalen Langzeitarchiv mithilfe dieser Emulatoren bei Bedarf angesehen

[215] Die Anfrage wurde am 22. Juni 2024 an Kody Willis (Archive-It) gesandt. Eine Antwort ist noch ausstehend. Ein Reminder wurde am 1. September 2024 versandt.

[216] vgl. Blumenthal, Archive-It Help Center 2023. Verfügbar unter: https://support.archive-it.org/hc/en-us/articles/208117536-Archive-It-Storage-and-Preservation-Policy [4.9.2024].

werden können. Eine Emulation muss und kann in der Regel nicht von einem einzelnen Archiv bewerkstelligt werden. Dies ist aufgrund der verfügbaren Ressourcen zumeist gar nicht möglich. Für das Internet Archive ist das Webarchiv das Hauptgeschäft, auf die einzelnen Partnerorganisationen von Archive-It trifft dies nicht zu.

Je nachdem wie das Internet Archive vorzugehen plant, müssen die Partnerorganisationen von Archive-It vielleicht selbst gar nicht tätig werden bezüglich einer Preservation Planning-Strategie. Die Entscheidung für das UAZ, welche Preservation Planning-Strategie verfolgt werden soll, ist daher vermutlich auch stark davon abhängig wie das Internet Archive zukünftig plant, bezüglich der Erhaltung der Daten vorzugehen. Vielleicht kann darauf vertraut werden, dass die *Wayback Machine* angepasst und weiterentwickelt wird und möglicherweise auch Emulatoren direkt darin integriert werden können, damit die korrekte Darstellung von älteren Seiten sichergestellt bleibt. Das Internet Archive aktualisiert den Viewer stetig. Dieser muss mit einem breiten Spektrum an Seiten umgehen können.

Als äusserst positiv zu bewerten ist, dass Archive-It und damit das Internet Archive gut auf Probleme aus Nutzersicht reagieren. Treten bei den Partnerorganisationen von Archive-It Probleme auf (beispielsweise die bereits erwähnte Problematik mit den Dropdown-Menüs im UAZ), versucht Archive-It jeweils Lösungen zu finden. Dies stimmt zuversichtlich, dass auch bezüglich Preservation Planning Lösungen gefunden werden können, ohne dass jede einzelne Organisation Individuallösungen erarbeiten muss.

Wenn nun aber vom Szenario ausgegangen wird, dass das Internet Archive auf eine Preservation Planning-Strategie gänzlich verzichtet[217], dann könnten die nachfolgenden Überlegungen für das UAZ interessant sein:

Wenn das UAZ sich für eine Lösung mit Emulation entscheidet, dann müsste zunächst genau durchdacht werden, wie die Emulation in die bestehenden Strukturen im UAZ integriert werden kann. Auch müsste abgeklärt werden, ob es allenfalls ausreichend wäre mit Browseremulatoren zu arbeiten oder ob auch die Browserumgebung mitemuliert werden muss. Ebenfalls müsste eruiert werden, in welchen zeitlichen Abständen wieder ein neuer Emulator gebaut werden muss. Denkbar wäre es beispielsweise, dass in einem

[217] Dass dieses Szenario Realität wird, ist unwahrscheinlich. Die Entwicklung einer Preservation Planning-Strategie ist für das Internet Archive eigentlich unverzichtbar, ansonsten könnten sich die Archivnutzenden die vom Internet Archive gesicherten Inhalte mit der Zeit gar nicht mehr in der *Wayback Machine* ansehen. Es handelt sich aber um eine ausserordentlich umfangreiche Aufgabe, da das Internet Archive über einen immensen Umfang an Daten verfügt. Die Preservation Planning-Strategie müsste automatisiert durchführbar sein, ansonsten besteht keine Chance für sämtliche Daten Preservation Planning zu betreiben und gegebenenfalls eine Preservation Action durchführen zu können. Es ist nicht bekannt wie intensiv sich das Internet Archive bereits mit solchen Überlegungen auseinandergesetzt hat. Allenfalls ist der Zeitpunkt für solche Überlegungen noch zu früh und das Internet Archive wird sich erst späterhin auf eine Preservation Planning-Strategie festlegen.

Intervall von 5 Jahren immer wieder ein neuer Browseremulator erstellt würde. Soll dann ein spezifischer gesicherter Zeitschnitt einer Website betrachtet werden, dann kann das entsprechende DIP mit dem archivierten WARC File aus dem digitalen Langzeitarchiv herausgeholt werden. Das WARC File könnte dann mit dem entsprechenden Browseremulator angesteuert werden, der spezifisch für WARC Files aus diesen Zeitraum geschaffen wurde. Die Frage ist dann auch, ob die Emulatoren direkt in das digitale Langzeitarchiv integriert und gesichert werden können und sollen oder ob diese separat und unabhängig vom digitalen Langzeitarchiv gesichert werden. Wie bereits erwähnt, ist es denkbar, dass allenfalls dereinst Dienstleister die Emulation übernehmen könnten im Sinne von «Emulation as a Service». Da die Mitarbeitenden des UAZ sich bis anhin in Zusammenhang mit der digitalen Langzeitarchivierung auf eine Migrationsstrategie fokussiert haben, wäre es vielleicht sinnvoll, sich für eine Beratung und einen allfälligen späteren Auftrag an einen solchen Dienstleister zu wenden, wenn die Möglichkeit dazu bestünde.

Soll ein Migrationsansatz verfolgt werden, dann muss unterschieden werden, ob der WARC-Container als Ganzer migriert werden muss oder ob es um ein spezifisches Dateiformat innerhalb des WARC-Containers geht, welches einer Preservation Action unterzogen werden muss. Geht es lediglich darum, den WARC-Container zu migrieren und braucht es keine Migration der Files innerhalb des Containers, dann wäre Migration ein guter Weg und fügt sich nahtlos ein in die Tradition des UAZ. Droht das Dateiformat WARC dereinst obsolet zu werden, dann wäre für ein Nachfolgeformat sicher gesorgt, da es sich um den Standard in der Webarchivierung handelt, der von vielen Gedächtnisinstitutionen genutzt wird. Müssten aber einzelne Dateiformate im WARC-Container migriert werden, dann sieht die Situation anders aus. Aktuell wäre der Aufwand zu gross, um einzelne Dateien in einem WARC File zu migrieren und das HTML anzupassen. Gäbe es aber dereinst ein «Super-Tool», das dies kann oder einen nicht zu komplexen Workflow, wobei mehrere Tools zusammenspielen, um die Migration durchzuführen, dann wäre dies sicherlich eine interessante Option für das UAZ. Dies auch da das WARC File mit dem WARC-Record des Typs conversion Migrationsbestrebungen unterstützt. Dies ist allerdings noch nicht erprobt und müsste vermutlich bei einer nächsten Version erst noch weiter spezifiziert werden, bevor diese Möglichkeit tatsächlich zur Anwendung kommen kann. Es kommt in der Zukunft darauf an, ob es Tools gibt, die eine effiziente und einfache Migration ermöglichen. Sicherlich ist es zentral, wachsam zu bleiben und im Austausch mit anderen Gedächtnisinstitutionen. Interessant wäre zudem die Teilnahme in einem Austauschgremium, wenn dieses zustande kommt. So könnten gemeinsame Lösungsstrategien entwickelt werden.

6.5 Weiterführende Gedanken und offene Fragen

Die im Abschnitt 6 genannten Tools wurden im Rahmen der vorliegenden Arbeit noch nicht getestet. Dies könnte im Rahmen von Versuchen nach Abschluss der Arbeit erfolgen. Das Testen der Tools würde dann eine erste Einschätzung ermöglichen, wie aufwendig sich die Nutzung der Tools tatsächlich gestaltet und wie effektiv diese funktionieren. Diese Erfahrungen könnten dann auch wesentlich zur Entscheidungsfindung beitragen ob bezüglich Preservation Planning der archivierten Websites auf Migration oder Emulation gesetzt werden soll.

7 Schlussbetrachtung

In Abschnitt 7.1 erfolgt ein abschliessendes Fazit mit einem Ausblick bezüglich dem weiteren Vorgehen im UAZ. In Abschnitt 7.2 ist der Fokus nicht mehr auf das UAZ gelegt, sondern wechselt auf eine allgemeinere Ebene. Für Gedächtnisinstitutionen, die sich ebenfalls mit der Webarchivierung befassen, wurden die zentralsten Erkenntnisse, Empfehlungen und Leitfragen zusammengetragen.

7.1 Abschliessendes Fazit und Schlusswort

In der vorliegenden Arbeit wurden verschiedene Themen behandelt, die in Zusammenhang mit der Neuorganisation des Webarchivs im UAZ stehen.

Aus der Arbeit sind im Wesentlichen drei Komponenten resultiert:

- eine erste Fassung einer Prozessbeschreibung auf der Grundlage des WALCM
- Empfehlungen für die Erarbeitung eines Datenmodells
- Empfehlungen bezüglich der Wahl einer Preservation Planning-Strategie

Die in Abschnitt 4 erarbeitete Prozessbeschreibung hält den Prozess der Webarchivierung im UAZ erstmals schriftlich fest. Sie soll als Grundlage dienen, auf deren Basis weitergearbeitet werden kann. Allenfalls lässt sich durch die Verschriftlichung Optimierungspotenzial im Prozess ausfindig machen und umsetzen. Bezüglich der Erstellung eines Datenmodells wurden in der vorliegenden Arbeit einige zentrale Fragen aufgeworfen und diskutiert. Als Grundlage für die Konzipierung eines Datenmodells wurden erste Empfehlungen ausformuliert (Abschnitt 5). Die Prozessbeschreibung und die Empfehlungen für das Datenmodell wurden als zwei separate und voneinander unabhängige Bestandteile erarbeitet. In der Praxis verhält es sich allerdings so, dass die Prozessbeschreibung und das Datenmodell in einem Abhängigkeitsverhältnis zueinanderstehen, indem sie sich gegenseitig ergänzen oder teilweise auch bedingen. In der vorliegenden Arbeit wurde von einer Vermischung abgesehen, damit eine distinkte Erarbeitung möglich war. Allenfalls könnte es aber dereinst interessant sein, das Datenmodell in die Prozessbeschreibung direkt zu integrieren, indem die einzelnen Teilschritte der Prozessbeschreibung mit den jeweils zentralen Informationen aus dem Datenmodell angereichert werden. Diese Integration macht selbstverständlich aber erst Sinn, wenn das Datenmodell konzipiert wurde und auch die Prozessbeschreibung komplett ist (also auch der Teil zum Preservation Planning mit der damit zusammenhängenden Migrations- respektive Emulationsstrategie in die Prozessbeschreibung eingefügt werden konnte). Wenn die Prozessbeschreibung dann dereinst in finalisierter Form vorliegt und mit den Informationen zum

Datenmodell angereichert werden konnte, dann könnte allenfalls ein Leitfaden daraus entstehen. Denkbar ist es dann auch, dass die Abteilung Visuelle Gestaltung an der UZH für eine individuelle Grafik kontaktiert werden könnte, sodass nicht mehr die Grafik des WALCM dazu verwendet werden müsste. Dies würde dem Dokument öffentlichen Charakter und mehr Gewicht verleihen. Das Dokument könnte dann für Interessierte online auf der Website des UAZ zur Verfügung gestellt werden.

Neben der Prozessbeschreibung und dem Datenmodell wurde in der vorliegenden Arbeit auch auf das Thema Preservation Planning eingegangen (Abschnitt 6). Das Thema Preservation Planning wird die Gedächtnisinstitutionen in Zukunft noch stark beschäftigen. Noch ist offen, welcher Ansatz – Migration oder Emulation – sich durchsetzen wird. Es ist aber heute schon zentral, sich damit auseinanderzusetzen. Es ist zwar noch keine Institution bekannt, die bereits eine konkrete Preservation Action durchgeführt hat, aber das Thema Preservation Planning beschäftigt die verschiedenen Gedächtnisorganisationen in einem starken Ausmass. Die Community zeigt sich offen für einen Austausch. Der Wunsch nach einem spezifischen Austauschgremium wurde geäussert, damit gemeinsam nach Lösungen gesucht werden kann. Entscheidend wird es auch sein, welche Tools in Zukunft für das Preservation Planning zur Verfügung stehen und wie sich diese mit der Zeit weiterentwickeln werden. Allenfalls stehen in Zukunft effiziente Tools zur Verfügung, die eine Migrationsstrategie unterstützen, die sich dann in die bestehende Migrationstradition des UAZ integrieren liesse. In Bezug auf das Preservation Planning ist es sinnvoll, wachsam zu bleiben, aktiv zur Diskussion beizutragen und auf diese Weise in Zusammenarbeit mit anderen Gedächtnisinstitutionen gemeinsame Lösungen zu entwickeln. Sehr zentral ist es auch, sich zu informieren, was das Internet Archive in Zukunft selbst für einen Lösungsansatz anstrebt. Falls beispielsweise Emulatoren direkt in die *Wayback Machine* integriert werden können, dann müssten die Partnerorganisationen von Archive-It gar nicht selbst in Bezug auf das Preservation Planning aktiv werden.

Die intensive Auseinandersetzung mit dem Thema Webarchivierung im Rahmen der Erarbeitung der vorliegenden Arbeit hat aufgezeigt, dass der Webarchivierung in der Community einen hohen Stellenwert zugesprochen wird und bereits grosse Bemühungen unternommen werden und in der Vergangenheit bereits unternommen worden sind, um gute Lösungsansätze zur Sicherung des Webauftrittes erarbeiten und realisieren zu können. Wir befinden uns in einer zugleich herausfordernden sowie auch spannenden Zeit, um sich mit dieser Thematik auseinanderzusetzen, da noch viele Fragen offen sind, insbesondere was das Preservation Planning anbelangt. Das grosse Engagement und Interesse der Fachpersonen stimmt zuversichtlich, sodass darauf vertraut werden darf, dass diese derzeit noch offenen Fragen und Herausforderungen gemeinsam

angegangen und dereinst auch gelöst werden können. Denn es besteht in der Community Einigkeit darüber, dass das Web ein wichtiges Zeitdokument darstellt. Es ist zentral, dass etwas, das die Gesellschaft in einem solch beträchtlichem Ausmass prägt und die Art und Weise der Verbreitung und Beschaffung von Information sowie die Form der Kommunikation beeinflusst durch Gedächtnisinstitutionen für zukünftige Forschende, Historiker*innen und die interessierte Öffentlichkeit langfristig gesichert und zur Verfügung gestellt werden kann.

7.2 Zusammenstellung Erkenntnisse, Empfehlungen und zentrale Leitfragen für andere Gedächtnisinstitutionen

Aus der Arbeit sind einige Erkenntnisse resultiert, die auch für andere Gedächtnisinstitutionen, die sich mit der Webarchivierung befassen, von Interesse sein können. Nachfolgend sind einige der zentralsten Erkenntnisse, Empfehlungen und Leitfragen zusammengetragen. Die Gliederung der Hinweise orientiert sich dabei an den drei Hauptteilen der Arbeit: Prozessbeschreibung, Datenmodell und Preservation Planning.

Prozessbeschreibung

- Um Schwachstellen zu erkennen und so den Prozess zu optimieren, hilft eine detaillierte Beschreibung des Webarchivierungs-Prozesses. Die Prozessbeschreibung schafft einen Überblick über den Gesamtprozess und dient für neue Mitarbeitende als Einführung in die Webarchivierung.
- Bei der Verschriftlichung des Prozesses ist es hilfreich, sich an bestehenden Modellen zu orientieren. Beim Web Archiving Life Cycle Model (WALCM) vom Archive-It Team handelt es sich um ein Modell, das sich hervorragend als Vorlage eignet.[218] Es schafft einen umfassenden Überblick über sämtliche Teilschritte des Prozesses und hilft dabei, keinen Prozessschritt auszulassen. Vorteilhaft ist es, dass es sich beim WALCM um ein übergeordnetes Konzept handelt, welches sich nicht auf konkrete Software bezieht. Dadurch kann das Modell individuell auf die Gegebenheiten und Bedürfnisse der jeweiligen Gedächtnisinstitution angepasst werden. Archive-It hat das WALCM aus der Erfahrung im Austausch mit verschiedenen Gedächtnisinstitutionen weltweit entwickelt, sodass das Modell eine grosse Nähe zur Praxis aufweist.
- Der Webarchivierungs-Prozess ist in seiner Gesamtheit zu betrachten (von den grundlegenden Policy-Entscheidungen bis zum Preservation Planning). Zwischen den verschiedenen Teilschritten kommt es immer wieder zu Überschneidungen.

[218] Bragg/Hanna 2013.

Beispielsweise kann die Metadatierung nicht auf einen bestimmten Teilschritt reduziert werden, sondern spielt während des ganzen Prozesses eine entscheidende Rolle (Erstellung, Import, Export von Metadaten). Die kreisförmig dargestellte Grafik des WALCM visualisiert den sich wiederholenden Charakter der einzelnen Arbeitsschritte im Lebenszyklus.[219]

- Für die Verschriftlichung des Webarchivierungsprozesses können die für jeden Teilschritt des WALCM zusammengetragenen Fragen in Abschnitt 4.3 durchgegangen werden.[220]

Datenmodell

- Zentral ist das Bewusstsein, dass gleichzeitig drei verschiedene Datenmodelle vorliegen, die aufeinander abgestimmt werden müssen. Die Logiken des AIS (ISAD(G), hierarchiebasiert), des Webarchiv-Services Archive-It (sammlungsbasiert) und des digitalen Langzeitarchivs (paketbasiert) lassen sich nicht ohne vorausgehende gründliche Reflexion zusammenbringen. Insbesondere gilt es diesbezüglich die Metadatierung zu beachten, da diese nicht in allen drei Komponenten identisch sein muss. Es macht Sinn, sich genau zu überlegen, wo und in welcher Ausführlichkeit Metadaten erfasst werden sollen.
- Die über Archive-It gesicherten Daten liegen auf den Servern des Internet Archive. Eine Partnerorganisation von Archive-It muss für sich entscheiden, ob diese Sicherung den eigenen Anforderungen genügt oder ob sie die Daten (WARC Files) von Archive-It herunterladen und zusätzlich im eigenen digitalen Langzeitarchiv sichern möchte. In diesem Zusammenhang stellen sich Fragen wie: Welche Ziele werden mit der Sicherung im eigenen digitalen Langzeitarchiv verfolgt? Handelt es sich primär um ein Backup? Sollen die ingestierten Daten auch für Benutzungen zur Verfügung stehen? Es ist zentral, das angestrebte Ziel zu definieren, da der Download der Daten sowie die Verarbeitung für den Ingest und die Wartung im digitalen Langzeitarchiv Ressourcen benötigen. Je nachdem welches Ziel mit den im eigenen digitalen Langzeitarchiv gesicherten Daten verfolgt wird, können als Konsequenz verschiedene Vorgehensweisen resultieren (z.B. müssen kleinere AIPs gebildet werden, wenn diese für die Benutzenden zeitgerecht aus dem digitalen Langzeitarchiv geholt werden müssen).

[219] vgl. Abbildungen 1 und 2 auf den Seiten 16 und 21.
[220] Der Fragenkatalog ist auf den Seiten 17 bis 20 ersichtlich.

- Grundlegende Fragen in Bezug auf die Bildung von Informationspaketen für den Ingest in das eigene digitale Langzeitarchiv:
    - Sollen der Master- und die nachfolgenden Deltacrawls jeweils als individuelle Pakete ingestiert werden? Kann in diesem Fall mit einem Verweis gearbeitet werden, sodass ersichtlich bleibt, welche Pakete zusammengehören? Soll der Crawl eines Seeds einem Dossier im AIS entsprechen? Soll jeder einzelne gesicherte Zeitschnitt über eine separate Signatur verfügen?
    - Ist es das Ziel, dass das Informationspaket aus sich heraus die ganze Website herausspielen kann? Falls ja: Sollen der Master- und sämtliche Deltacrawls in einem einzigen Paket ingestiert werden? Wird das Paket auf diese Weise zu umfangreich oder bleibt eine gute Handhabung für den Ingest und die Bereitstellung für eine allfällige Benutzung gewährleistet?
    - Welche zusätzlichen Dokumente sollen mitingestiert werden? Mit welchen Metadaten muss das Informationspaket angereichert werden?

Preservation Planning

- Ganz allgemein stellen sich für eine Gedächtnisinstitution bezüglich Preservation Planning zunächst die folgenden Fragen:
    - Was entspricht der bisherigen Tradition einer Institution? Wurden bereits erste Erfahrungen gesammelt? Ist bereits Know-how vorhanden, das ausgebaut werden kann?
    - Welche Möglichkeiten stehen zur Verfügung (Ressourcen wie Finanzen, Personal, Infrastruktur)?
- Eine Institution, die Webarchivierung betreibt, sollte sich eingehend mit dem Dateiformat WARC auseinandersetzen, dem Standardformat der Webarchivierung. In einem WARC-Container kann eine Vielzahl sehr unterschiedlicher Dateiformate enthalten sein. Diese grosse Vielfalt innerhalb des Containers stellt insbesondere das Preservation Planning vor Herausforderungen.
- Bezüglich Preservation Planning kommen sowohl Migration als auch Emulation als mögliche Strategie in Frage. Bei der Migration muss unterschieden werden, ob der WARC-Container als Ganzer oder die Migration eines Dateiformates innerhalb des WARC-Containers erfolgen soll. Bezüglich des zweiten Falles ist die Lektüre des Beitrags «Migrating Content in WARC Files» von Strodl/Beran/Rauber zu empfehlen.

[221] Im Beitrag wird ein Workflow für die Migration vorgestellt. Bei der Emulation werden frühere Browsertypen (oder die ganze Browserumgebung) emuliert. Emulation ist insbesondere bezüglich der Darstellung von Flashinhalten zentral, da es diesbezüglich keine migrationsbasierten Lösungen gibt. Denkbar ist es, dass sich dereinst eine Mischform zwischen Migration und Emulation durchsetzen wird.

- Bis anhin ist noch keine Institution bekannt, die tatsächlich bereits eine Preservation Action durchgeführt hat. Das Thema Preservation Planning ist aber sehr präsent und es besteht ein grosses Bedürfnis nach einem intensiveren Austausch zwischen den Fachleuten. Es ist wichtig, informiert zu bleiben und insbesondere auch mitzuverfolgen, welche Massnahmen das Internet Archive selbst bezüglich Preservation Planning ergreifen wird. Je nachdem kann es sein, dass es gar nicht notwendig sein wird, dass individuelle Institutionen selbständig Preservation Actions durchführen müssen.

[221] Strodl/Beran/Rauber 2009.

8 Fachliteratur und weitere Informationsquellen

Archive-It Helpcenter: Glossary of Archive-It and Web Archiving Terms. Verfügbar unter: https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms [9.7.2024].

Beinert, Tobias/Schrimpf, Sabine/Wolf, Stefan: Collect now – Ask later why?! nestor-Expertengespräch zur Archivierung von Websites im deutschsprachigen Raum. 2011a. Verfügbar unter: https://www.langzeitarchivierung.de/Webs/nestor/SharedDocs/Downloads/DE/berichte/collectNowAskLaterWhy.pdf?__blob=publicationFile&v=1 [25.7.2024].

Beinert, Tobias/Schrimpf, Sabine/Wolf, Stefan: Wie gut ist gut genug? Qualität in der Webarchivierung. 2. nestor-Expertengespräch zur Archivierung von Websites. 2011b. Verfügbar unter: https://www.langzeitarchivierung.de/Webs/nestor/SharedDocs/Downloads/DE/berichte/qualit%C3%A4tWebarchivierung.pdf?__blob=publicationFile&v=1 [25.7.2024].

Blumenthal, Karl-Rainer (2021): Archive-It Blog: *The stack: An introduction to the WARC file*. Beitrag vom 1. April 2021. Verfügbar unter: https://archive-it.org/post/the-stack-warc-file/ [24.7.2024].

Blumenthal, Karl-Rainer (2023): Archive-It Help Center: *Archive-It Storage and Preservation Policy*. Verfügbar unter: https://support.archive-it.org/hc/en-us/articles/208117536-Archive-It-Storage-and-Preservation-Policy [4.9.2024].

Blumenthal, Karl-Rainer (2024): Archive-It Help Center: *Find and download your WARC files with WASAPI*. Verfügbar unter: https://support.archive-it.org/hc/en-us/articles/360015225051-Find-and-download-your-WARC-files-with-WASAPI [24.6.2024].

Bragg Molly/Hanna, Kristine (The Archive-It-Team, Internet Archive): The Web Archiving Life Cycle Model. March 2013.

Brunner, Kerstin/Debenath, Olivier: Arbeitsbericht zur Archivierung von Netzressourcen im Staatsarchiv des Kantons Basel-Stadt. In: Informationswissenschaft: Theorie, Methode und Praxis. Band 5. 2018, Nr. 1, S. 111–118.

Digital Preservation Coalition: Digital Preservation Handbook, 2nd Edition, 2015. Verfügbar unter: https://www.dpconline.org/handbook [24.7.2024].

eCH E-Government Standards. eCH-0160 – Archivische Ablieferungsschnittstelle. 2022.

Franzky, Tony (Erzbischöfliches Archiv Freiburg): Multimodale Ansätze der Webarchivierung: Einblick in das Konzept des Erzbischöflichen Archivs Freiburg. Präsentation im Rahmen der 27. Tagung des Arbeitskreises AUdS in Zürich am 5.3.2024. Verfügbar unter: https://www.sg.ch/kultur/staatsarchiv/Spezialthemen-/auds/2024/_jcr_content/Par/sgch_downloadlist_1831161160/DownloadListPar/sgch_download.ocFile/01_BC_Zwingli_Franzky%20-%20Multimodale%20W.pdf [26.7.2024].

Frappart, Corinne (EU Publications Office): Towards an effective long-term preservation of the web: the case of the EU Publications Office. Präsentation im Rahmen der IIPC web archiving conference 2023. Die Präsentation von Corinne Frappart kann als Video online eingesehen werden. Verfügbar unter: https://www.youtube.com/watch?v=f_672x9s5xk [17.7.2024].

Funk, Stefan E.: Migration und Emulation – Angewandte Magie? Präsentation im Rahmen vom nestor-Praktikertag vom 14. Juni 2016 in Dresden. Verfügbar unter: https://www.langzeitarchivierung.de/Webs/nestor/SharedDocs/Downloads/DE/praesentationen/2016PraktikertagFunk.pdf?__blob=publicationFile&v=1 [29.3.2024].

Gastl, Angela: ETH-Webarchiv: Ein attraktives Produkt des Hochschularchivs. Präsentation im Rahmen des Treffens der ArchivarInnen der Schweizer Universitäten und Hochschulen in Mendrisio am 5.5.2022. Verfügbar unter: https://www.research-collection.ethz.ch/handle/20.500.11850/545817 [14.7.2024].

Geisler, Felix/Dannehl, Wiebke/Keitel, Christian/Wolf, Stefan: Zum Stand der Webarchivierung in Baden-Württemberg. In: Bibliotheksdienst 2017. Vol. 51 (6), S. 481–489.

GitHub: Awesome Web Archiving. Verfügbar unter: https://github.com/iipc/awesome-web-archiving [4.8.2024].

Hanson, Adriane (2021): Archive-It Blog: *For safekeeping: An automated preservation workflow for Archive-It content*. Beitrag vom 12. Januar 2021. Verfügbar unter: https://archive-it.org/blog/post/automated-preservation-workflow/ [24.7.2024].

International Internet Preservation Consortium (IIPC): Tools and Software. Verfügbar unter: https://netpreserve.org/web-archiving/tools-and-software/ [10.8.2024].

Internet Archive, Archiving & Data Services Division: Data & Data Center Security & Procedures. January 2023.

Kahle, Brewster (2021): Internet Archive Blogs: *Reflections as the Internet Archive turns 25*. Beitrag vom 21. Juli 2021. Verfügbar unter: https://blog.archive.org/2021/07/21/reflections-as-the-internet-archive-turns-25/ [2.7.2024].

KEEP (Keeping Emulation Environments Portable) project. Emulation Framework (EF). Verfügbar unter: https://emuframework.sourceforge.net/ [11.8.2024].

Keitel, Christian: Digitale Archivierung beim Landesarchiv Baden-Württemberg. In: Archivar, 63. Jahrgang, Heft 01, Februar 2010, S. 19–26.

Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST): Webarchivierung. Eine Studie der KOST. Version 0.4, 2016. Verfügbar unter: https://kost-ceco.ch/cms/dl/60298dec33a27e685be7b956d37eecf1/Studie_Webarchivierung_v0.4.pdf?target=1 [9.7.2024].

Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST): Katalog archivischer Dateiformate (KaD). Verfügbar unter: https://kost-ceco.ch/cms/kad_main_de.html [17.7.2024].

Kortyla, Stephanie (Sächsisches Staatsarchiv): Archivierung von Netzressourcen. Ist-Stand und Ausblick – Best Practices und Baustellen. Präsentation vom 8.5.2024.

Lange, Felix (Deutsches Bundesarchiv Koblenz): Perspektiven für eine semiautomatische Qualitätssicherung bei der Archivierung von Webseiten. Präsentation im Rahmen der 27. Tagung des Arbeitskreises AUdS in Zürich am 5.3.2024. Verfügbar unter: https://www.sg.ch/kultur/staatsarchiv/Spezialthemen-/auds/2024/_jcr_content/Par/sgch_downloadlist_1831161160/DownloadListPar/sgch_download_1839531972.ocFile/02_BC_Zwingli_Lange_Webarch.pdf [26.7.2024].

Lazorchak, Butch (2011): Library of Congress Blogs: *Web Archive Preservation Planning*. Beitrag vom 18. August 2011. Verfügbar unter: https://blogs.loc.gov/thesignal/2011/08/web-archive-preservation-planning/ [10.8.2024].

Le Follic, Annick/Stirling, Peter/Wendland, Bert: Putting it all together: creating a unified web harvesting workflow at the Bibliothèque nationale de France. 2012.

Library of Congress: Sustainability of Digital Formats: WARC, Web ARChive file format. Verfügbar unter: https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml [17.7.2024].

Locher, Hansueli: Archivierung von Internetseiten: eine Standortbestimmung. In:

Geschichte und Informatik. Band 13–14. 2002–2003, S. 111–119.

Loewenich, Maria von (Deutsches Bundesarchiv Berlin): Archivische Bewertung im digitalen Zeitalter. Präsentation im Rahmen der 27. Tagung des Arbeitskreises AUdS in Zürich am 5.3.2024. Verfügbar unter: https://www.sg.ch/kultur/staatsarchiv/Spezialthemen-/auds/2024/_jcr_content/Par/sgch_downloadlist_1834052559/DownloadListPar/sgch_download_1835963853.ocFile/02_PLDI_Loewenich_Archivische_Bewertung.pdf [8.9.2024].

Loewenich, Maria von (Deutsches Bundesarchiv Berlin): Archivische Bewertung im digitalen Zeitalter. Präsentation im Rahmen der 27. Tagung des Arbeitskreises AUdS in Zürich am 5.3.2024. Script zur Präsentation.

Masanès, Julien: Web Archiving. Berlin Heidelberg 2006.

Mayr, Michaela: Bedeutung der Webarchivierung am Beispiel von Web@rchiv Österreich. Präsentation im Rahmen des Workshops Webarchivierung DNB am 13. April 2011. Verfügbar unter: https://de.slideshare.net/slideshow/bedeutung-der-webarchivierung-nestordnb/7624633 [27.7.2024].

Messner, Philipp (2015): UZH Archiv Vitrine: *Die Dokumentation universitärer Publikationstätigkeit im Medienwandel*. Beitrag vom 3. November 2015. Verfügbar unter: https://www.archiv.uzh.ch/de/vitrine/aeltere_beitraege.html#Die_Dokumentation_universit%C3%A4rer_Publikationst%C3%A4tigkeit_im_Medienwandel [31.8.2024].

Neuroth, Heike/Osswald, Achim/Scheffel, Regine/Strathmann, Stefan/Huth, Karsten (Hgg.): nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.3, 2010. Verfügbar unter: https://nestor.sub.uni-goettingen.de/handbuch/nestor-handbuch_23.pdf [29.3.2024].

Pennock, Maureen: Web-Archiving. DPC Technology Watch Report 13-01 March 2013. Verfügbar unter: https://www.dpconline.org/docs/technology-watch-reports/865-dpctw13-01-pdf/file [29.3.2024].

Schoger, Astrid/Beinert, Tobias/Schmid, Katharina/Donig, Simon/Eckl, Markus: WWW Webarchivierung. Wieso? Weshalb? Warum? Präsentation im Rahmen von nestor virtuell am 22.7.2021. Verfügbar unter: https://www.langzeitarchivierung.de/Webs/nestor/SharedDocs/Downloads/DE/praesentationen/2021wwwWebarchivierung.pdf?__blob=publicationFile&v=1 [9.7.2024].

Schweizerische Nationalbibliothek, Webarchiv Schweiz. Glossar, Version 1.7 vom 19.

April 2024.Verfügbar unter: https://www.nb.admin.ch/snl/de/home/fachinformationen/e-helvetica/webarchiv-schweiz.html#993259453 [27.7.2024].

Schweizerische Nationalbibliothek, Webarchiv Schweiz. Grundlagenpapier vom 3. Juli 2024.Verfügbar unter: https://www.nb.admin.ch/snl/de/home/fachinformationen/e-helvetica/webarchiv-schweiz.html#993259453 [27.7.2024].

Schweizerische Nationalbibliothek, Webarchiv Schweiz. Merkblatt Archivieren, Version 1.7 vom 19. April 2024.Verfügbar unter: https://www.nb.admin.ch/snl/de/home/fachinformationen/e-helvetica/webarchiv-schweiz.html#993259453 [27.7.2024].

Strodl, Stephan/Beran, Peter Paul/Rauber, Andreas: Migrating Content in WARC Files. In: Proceedings of the 9th International Web Archiving Workshop (IWAW) 2009, S. 43–49.

UZH News vom 23.3.2023: «Ein frischer Look für die UZH-Webseiten». Verfügbar unter: https://www.news.uzh.ch/de/articles/news/2023/neuer-webauftritt.html [9.7.2024].

Veenendaal, Remco van/Takema, Jacob/Wijsman, Lotte/Rappard, Marin: Around for decades, gone in a Flash. How we dealt with Flash objects at the National Archives of the Netherlands. In: iPRES 2023: The 19th international Conference on Digital Preservation, Champaign-Urbana, IL, US. 19. – 23. September 2023. Verfügbar unter: https://www.ideals.illinois.edu/items/128299 [24.7.2024]

Walter Nagel GmbH & Co. KG. Webarchivierung. Verfügbar unter: https://www.walternagel.de/webarchivierung#:~:text=Die%20durchschnittliche%20%E2%80%9ELebensdauer%E2%80%9C%20von%20Webseiten,sichern%20damit%20alle%20relevanten%20Inhalte. [9.7.2024].

Walz, Annabel (2023): FEShistory Blog: *Archivierte Webseiten mit pywb und Solr-Wayback nutzen*. Beitrag vom 18. August 2023. Verfügbar unter: https://www.fes.de/feshistory/blog/archivierte-webseiten-mit-pywb-und-solrwayback-nutzen [24.7.2024].

Weimer, Konstanze/Schoger, Astrid: Das Dateiformat WARC für die Webarchivierung. In: Reihe «nestor Thema». 2021. Verfügbar unter: https://files.dnb.de/nestor/kurzartikel/thema_15-WARC.pdf [29.3.2024].

Video:

Consultative Committee for Space Data Systems (CCSDS), Data Archive Interoperability (DAI) Working Group; Michael W. Kearney III; David Giaretta; John Garrett; Steve Hughes: What's missing from WARC? To preserve Web Pages. Engaging with Web Archives, Opportunities, Challenges and Potentialities (#EWAVirtual) 21.–22. September 2020. Verfügbar unter: https://www.youtube.com/watch?v=vdEaz109uAo [19.7.2024].

Abbildungen

Abbildungen 1 und 2: Bragg Molly/Hanna, Kristine (The Archive-It-Team, Internet Archive): The Web Archiving Life Cycle Model. March 2013.

Abbildung 3: Archive-It Help Center. Archive-It Video Curriculum. Getting Started. Navigating Archive-It (Screenshot aus Video). URL.: https://support.archive-it.org/hc/en-us/articles/216489103-Archive-It-Video-Curriculum [4.8.2024].

Abbildungen 4 und 5: Strodl, Stephan/Beran, Peter Paul/Rauber, Andreas: Migrating Content in WARC Files. In: Proceedings of the 9th International Web Archiving Workshop (IWAW) 2009, S. 43–49.

9 Anhang

Rechtliche Grundlagen

Archivgesetz vom 24. September 1995 (170.6). Verfügbar unter: https://www.notes.zh.ch/appl/zhlex_r.nsf/WebView/4921593D1D43C0CAC1257C3F003AA606/$File/170.6_24.9.95_83.pdf [16.8.2024].

Archivverordnung vom 9. Dezember 1998 (170.61). Verfügbar unter: https://www.notes.zh.ch/appl/zhlex_r.nsf/WebView/176075E25907D4AEC12584B800295226/$File/170.61_9.12.98_107.pdf [16.8.2024].

UZH Archiv

Publizierte Dokumente (sämtliche Dokumente verfügbar unter «Dokumente zum Download» auf der UZH Archiv Webseite) [16.8.2024].

- UZH Archiv Policy Digitale Langzeitarchivierung, Version 1.0 vom 1. Oktober 2022
- Archivtaugliche Dateiformate, aktualisiert am 18. Oktober 2022
- Erschliessungshandbuch UZH Archiv (UAZ), Version 1.1 vom 22. Februar 2021

Online-Recherchekatalog (Webclient) UAZ. Verfügbar unter: https://mobile.cmistar.ch/webclients-r22/uzh/#/ [16.8.2024].

Interne Dokumente: *CMI* Benutzungsverwaltung (GEVER)

Besteht Interesse an einer Auskunft oder an der Einsicht in ein in der Arbeit erwähntes internes Dokument (Berichte, Handbuch, Korrespondenz etc.), dann können die Mitarbeitenden des UAZ diesbezüglich kontaktiert werden.

Übersichtstabelle Austausch mit Fachleuten

Datum	Austauschpartner	Thema
25. März 2024	Angela Gastl (Archivarin, Hochschularchiv der ETH, Verantwortliche Webarchiv ETH)	Webarchivierung mit Archive-It (Austausch via MS Teams)
2. April 2024	Franziska Geisser (docuteam)	Sicherung ZIP-Kapseln von e-rara und e-manuscripta / Master- und Deltakonzept (Korrespondenz per E-Mail)
9. April 2024	Franziska Geisser (docuteam)	Dateiformat WARC (Korrespondenz per E-Mail)

12. April 2024	Markus Kandlbinder (Archivinformatiker UAZ)	Webarchivierung mit Offline Explorer (Austausch via MS Teams)
19. April 2024	Markus Kandlbinder (Archivinformatiker UAZ)	Dateiformat WARC (Austausch via MS Teams)
19. April 2024	Barbara Signori (Webarchiv Schweiz)	Austausch Webarchiv Schweiz (Telefonat)
29. April 2024	Lea Fuhrer (Wissenschaftliche Mitarbeiterin, Plattformen und Daten, Bibliotheksinformatik, Zentralbibliothek Zürich)	AIP / Master- und Deltakapseln (Korrespondenz per E-Mail)
29. April 2024	Mark Philips (University of North Texas Libraries)	Tools für Dateiformat WARC (Korrespondenz per E-Mail)
29. April 2024	Bjarne Andersen (Head of Data, Royal Danish Library)	SIP/AIP/DIP / Preservation Planning (Korrespondenz per E-Mail)
30. April 2024	Anders Klindt Myrvoll (Programme Manager, Danish Web Archive)	Download WARC from Archive-It (Korrespondenz per E-Mail)
30. April 2024	Alex Osborne (Assistant Director, Web Archive Systems, National Library of Australia)	Emulation (Korrespondenz per E-Mail)
2. Mai 2024	Youssef Eldakar (Head of Department, International School of Information Science, Bibliotheca Alexandrina, Ägypten)	Überführung WARC Files in eigenes digitales Langzeitarchiv / AIP / Emulation (Korrespondenz per E-Mail)
3. Mai 2024	Bjarne Andersen (Head of Data, Royal Danish Library)	Tools in Zusammenhang mit Dateiformat WARC / Migration (Korrespondenz per E-Mail)
8. Mai 2024	Anna Vögeli (Spezialistin digitale Langzeitarchivierung, Universitätsbibliothek Bern)	SIP und AIP / Master- und Deltakonzept (Korrespondenz per E-Mail)
13. Mai 2024	Kody Willis (Product Operations Manager, Archiving & Data Services, Internet Archive / Ansprechpartner vom UAZ bei Archive-It)	Dateiformat WARC (Korrespondenz per E-Mail)

23. Mai 2024	Inge Moser (stv. Leiterin UAZ)	Stand der Dinge, Zwischenbericht / Prozessbeschreibung (Austausch via MS Teams)
24. Mai 2024	Tony Franzky (Projektleiter digitale Langzeitarchivierung, Erzbischöfliches Archiv Freiburg)	Fragen zu Präsentation AUdS in Zürich 2024 / Dateiformat WARC / Preservation Planning (Austausch via MS Teams)
4. Juni 2024	Kody Willis (Product Operations Manager, Archiving & Data Services, Internet Archive / Ansprechpartner vom UAZ bei Archive-It)	Aktualität WALCM (Korrespondenz per E-Mail)
7. Juni 2024	Felix Lange (Deutsches Bundesarchiv Koblenz)	Preservation Planning / AIP (Austausch via MS Teams)
7. Juni 2024	Inge Moser (stv. Leiterin UAZ)	Vorgehen Archive-It (Seeds erfassen, Crawls durchführen und Qualitätsprüfung) (Besprechung im UAZ)
10. Juni 2024	Tobias Wildi, Ana Petrus (GutachterIn)	Zwischengespräch zur Masterarbeit mit Zwischenbericht (Austausch via Webex)
13. Juni 2024	Markus Kandlbinder (Archivinformatiker UAZ)	Klärung einiger Fragen (Austausch via MS Teams)
20. Juni 2024	Stephanie Kortyla (Sachbearbeiterin, Sächsisches Staatsarchiv in Dresden)	Preservation Planning / Dateiformat WARC / Offline Explorer (Austausch via MS Teams)
22. Juni 2024	Kody Willis (Product Operations Manager, Archiving & Data Services, Internet Archive / Ansprechpartner vom UAZ bei Archive-It)	Preservation Planning Strategy von Archive-It (Korrespondenz per E-Mail) → Rückmeldung ausstehend (Reminder am 1. September 2024 versandt)
27. Juni 2024	Angela Gastl (Archivarin, Hochschularchiv der ETH, Verantwortliche Webarchiv ETH)	Dateiformat WARC / Preservation Planing / Überführung WARC in digitales Langzeitarchiv (Korrespondenz per E-Mail)

5. Juli 2024	Annabel Walz (Archiv der sozialen Demokratie, Friedrich Ebert Stiftung)	Dateiformat WARC / SIP, AIP, DIP / Master- und Deltakonzept / Preservation Planning (Austausch via MS Teams)
5. Juli 2024	Inge Moser (stv. Leiterin UAZ)	Qualitätsprüfung / Klärung einiger Fragen (Austausch via MS Teams)
16. Juli 2024	Andreas Rauber (Universität Wien, Mitverfasser Publikation *Migrating Content in WARC Files (2009)*)	Preservation Planning (Korrespondenz per E-Mail)
24. Juli 2024	Barbara Signori (Webarchiv Schweiz)	Tool um WARC zu öffnen für Virencheck (Korrespondenz per E-Mail)

Zu sämtlichen Gesprächen, die mündlich stattfanden, wurden Aktennotizen verfasst. Diese Dokumente sind der Arbeit nicht angegliedert. Besteht ein Interesse in die Einsicht einer Notiz, dann kann die Autorin der Arbeit diesbezüglich kontaktiert werden.

Bisher erschienene Schriften

Ergebnisse von Forschungsprojekten erscheinen jeweils in Form von Arbeitsberichten in Reihen. Sonstige Publikationen erscheinen in Form von alleinstehenden Schriften.

Derzeit gibt es in den Churer Schriften zur Informationswissenschaft folgende Reihen:
Reihe Berufsmarktforschung

Weitere Publikationen

Churer Schriften zur Informationswissenschaft – Schrift 164
Herausgegeben von Wolfgang Semar
Flurin Böni
Das verborgene Gold am Ende des Rainbow-Washing
Eine Analyse der Vereinbarkeit sozialen Engagements mit unternehmerischen Zielen
Chur, 2023
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 165
Herausgegeben von Wolfgang Semar
Alina Viert
Herausforderungen in der Aufbewahrung von Videospielen und ihrer Peripherie
Fragen und Antworten insbesondere zur Peripherie und zur Emulation als Lösungsansatz
Chur 2023
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 166
Herausgegeben von Wolfgang Semar
Susanne Knöpfel
Wissenslandkarten als Grundlage für Visualisierungen im Wissensmanagement
Chur, 2023
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 167
Herausgegeben von Wolfgang Semar
Lorena Staiger
Deep Web und Bibliotheken: Stand der Dinge
Chur, 2023
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 168
Herausgegeben von Wolfgang Semar
Karin Mattmann
Positive Darstellungen archivarischer Tätigkeiten in Fiktion
Wie das Abbild von fiktionalem Archivpersonal in der Öffentlichkeit positiv und realistisch dargestellt werden kann
Chur, 2023
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 169
Herausgegeben von Wolfgang Semar
Stefan Banzer
Codemigration mit ChatGPT
Evaluation von ChatGPT als Tool zur teilautomatisierten Codeübersetzung von COBOL Code zu Python Code
Chur, 2023
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 170
Herausgegeben von Wolfgang Semar
Marion Spitz
Digitale Nudges zwischen Moral und Manipulation
Eine quantitative Inhaltsanalyse zu den Auswirkungen ethischer Aspekte auf die erforschte Wirksamkeit von digitalen Nudges
Chur, 2024
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 171
Herausgegeben von Wolfgang Semar
Joy Walser
Erschliessungsmöglichkeiten einer Sammlung mit Records in Contexts
Entwicklung und Anwendung eines konzeptionellen Modells für die Sammlung
«Pfarrer F. Tschugmell, Siegel- und Stempelsammlung»
Chur, 2024
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 172
Herausgegeben von Wolfgang Semar
Alessio Monte
Potenzialanalyse zur Anwendung von KI-basierten
Chur, 2024
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 173
Herausgegeben von Wolfgang Semar
Lisa Köllner
Der Familienbezug und seine Bedeutung für die Nutzung von Firmenarchiven durch Familienunternehmen am Beispiel aktuell tätiger Unternehmen
Chur, 2024
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 174
Herausgegeben von Wolfgang Semar
Silvia Rutz
Psychologische Sicherheit in virtuellen agilen Teams
Eine explanative Analyse der Einflussfaktoren auf die psycholo-gische Sicherheit in virtuellen agilen Software-Teams
Chur, 2024
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 175
Herausgegeben von Wolfgang Semar
Jérôme Gander
Information Governance und öffentliche Verwaltung
Definitionen, Nutzen und die Rolle der Verwaltungsarchive.
Chur, 2024
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 176
Herausgegeben von Wolfgang Semar
Rade Jevdenic
Governance von Social-Media-Algorithmen im Digital Services Act
Analyse der Aufsicht und Regulation von
ML-basierten Empfehlungssystemen
Chur, 2024
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 177
Herausgegeben von Wolfgang Semar
Ramona Kälin
Verantwortungs- & respektvoller Umgang im Metaverse
Eine Untersuchung, welche Rolle die Medienkompetenz spielt, wenn Jugendliche Hatespeech im Metaverse erfahren.
Chur, 2024
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 178
Herausgegeben von Wolfgang Semar
Felicia Perrucci
Eine Erhebung des Status Quo der Therapiehun-de in Deutschschweizer Hochschulbibliotheken
Chur, 2024
ISSN 1660-945X

Über die Informationswissenschaft der Fachhochschule Graubünden

Die Informationswissenschaft ist in der Schweiz noch ein relativ junger Lehr- und Forschungsbereich. International weist diese Disziplin aber vor allem im anglo-ameri-kanischen Bereich eine jahrzehntelange Tradition auf. Die klassischen Bezeichnungen dort sind Information Science, Library Science oder Information Studies. Die Grundfragestellung der Informationswissenschaft liegt in der Betrachtung der Rolle und des Umgangs mit Information in allen ihren Ausprägungen und Medien sowohl in Wirtschaft und Gesellschaft. Die Informationswissenschaft wird in Chur integriert betrachtet.

Diese Sicht umfasst nicht nur die Teildisziplinen Bibliothekswissenschaft, Archiv-wissenschaft und Dokumentationswissenschaft. Auch neue Entwicklungen im Bereich Medienwirtschaft, Informations- und Wissensmanagement und Big Data werden gezielt aufgegriffen und im Lehr- und Forschungsprogramm berücksichtigt.

Der Studiengang Informationswissenschaft wird seit 1998 als Vollzeitstudiengang in Chur angeboten und seit 2002 als Teilzeit-Studiengang in Zürich. Seit 2010 rundet der Master of Science in Business Administration das Lehrangebot ab.

Der Arbeitsbereich Informationswissenschaft vereinigt Cluster von Forschungs-, Entwicklungs- und Dienstleistungspotenzialen in unterschiedlichen Kompetenzzentren:

- Information Management & Competitive Intelligence
- Collaborative Knowledge Management
- Information and Data Management
- Records Management
- Library Consulting
- Information Laboratory
- Digital Education

Diese Kompetenzzentren werden im Swiss Institute for Information Science (SII) zusammengefasst.

Impressum

Impressum

FHGR - Fachhochschule
Graubünden
Information Science
Pulvermühlestrasse 57
CH-7000 Chur

www.informationscience.ch
www.fhgr.ch
ISSN 1660-945X

Institutsleitung

Prof. Dr. Ingo Barkow
Telefon: +41 81 286 24 61
Email: ingo.barkow@fhgr.ch

Sekretariat

Telefon: +41 81 286 24 24
Fax: +41 81 286 24 00
Email: clarita.decurtins@fhgr.ch