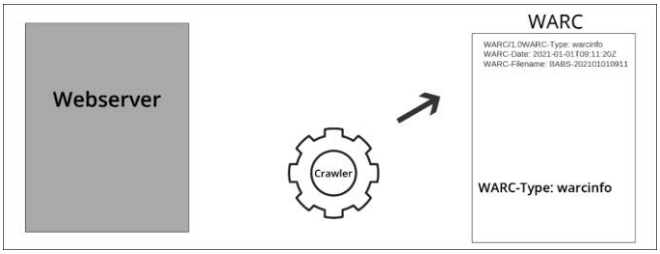
15 – Das Dateiformat WARC für die Webarchivierung

Ziel der Webarchivierung ist es, Ressourcen aus dem Internet - Websites, Blogs, Social Media etc. - zu speichern, dauerhaft zu bewahren und für die Nutzung zur Verfügung zu stellen. Dafür wird ein (Web-)Crawler bzw. Harvester eingesetzt, der eine Website bzw. einen definierten Ausschnitt des Web automatisch durchsucht. Komplexe dynamische Elemente und immer neue Internettechnologien können den Crawler dabei vor Herausforderungen stellen. Der Crawler lädt die gefundenen Inhalte herunter, z.B. HTML-Seiten, clientseitige Skripte (vor allem JavaScript Code), Stylesheets, Text-, Bild-, Audio- und Videodateien, und speichert diese zusammen mit Metainformationen zum Crawlingprozess ab. Für die Strukturierung und Speicherung dieser großen, heterogenen Datenmengen hat sich das Containerformat WARC (Web ARChive) etabliert.

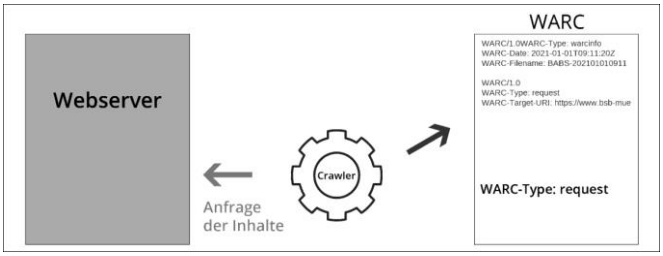
Aufbau einer WARC-Datei

Eine WARC-Datei enthält WARC-Records unterschiedlichen WARC-Typs, die aneinandergereiht werden. Vereinfacht dargestellt wird eine WARC-Datei während des Crawlprozesses wie folgt aufgebaut:



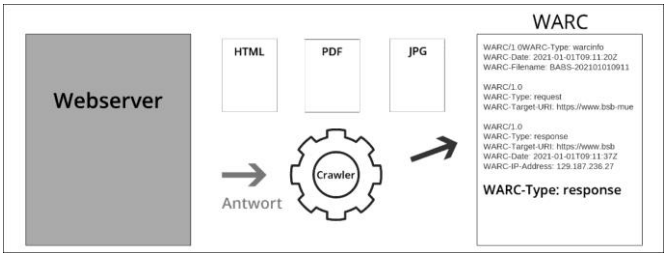
Schritt 1:

Der Crawler erzeugt eine neue WARC-Datei. Im Record vom Typ *warcinfo* speichert er Metadaten über den Crawl, wie Datum und Uhrzeit, die verwendete Software.



Schritt 2:

Der Crawler fordert Inhalte der zu archivierenden Website vom Webserver an und dokumentiert die Anfrage in dem Record vom Typ *request*.



Schritt 3:

Der Webserver antwortet, indem er die angeforderte Ressource, z.B. eine HTML-Datei, ein Bild oder ein PDF-Dokument, sendet. Der Crawler schreibt diese Datei in einen Record vom Typ *response* bzw. *resource*.

Abbildung 1: Aufbauprozess einer WARC-Datei

Gemäß seiner Prozesskette fordert der Crawler weitere Inhalte vom Webserver an und schreibt die Antworten jedes Mal sukzessive in die WARC-Datei. Schritt 2 und 3 werden so lange wiederholt, bis das Ziel des Crawlprozesses erreicht ist oder es zu einem Halt, z.B. wegen des Erreichens vordefinierter Limits, oder zu einem Abbruch kommt. Gegebenenfalls werden vom Crawler weitere WARC-Dateien angelegt.

Neben diesen WARC-Records, die die eigentlichen zu archivierenden Inhalte und die zur Dokumentation des Crawlingprozesses wesentlichen Metainformationen speichern, gibt es noch weitere Record-Types. Beispielsweise können zusätzliche Metadaten zur Beschreibung der einzelnen Inhalte in Records vom Typ *metadata* aufgenommen werden. Records vom Typ *continuation* geben an, dass Inhalte auf mehrere WARC-Dateien aufgeteilt wurden. Records vom Typ *revisit* enthalten Hinweise auf bereits archivierte Inhalte. Ferner können auch transformierte Inhalte, die im Zuge einer Langzeiterhaltungsmaßnahme erstellt werden, nachträglich in der WARC-Datei in Records vom Typ *conversion* gespeichert werden.

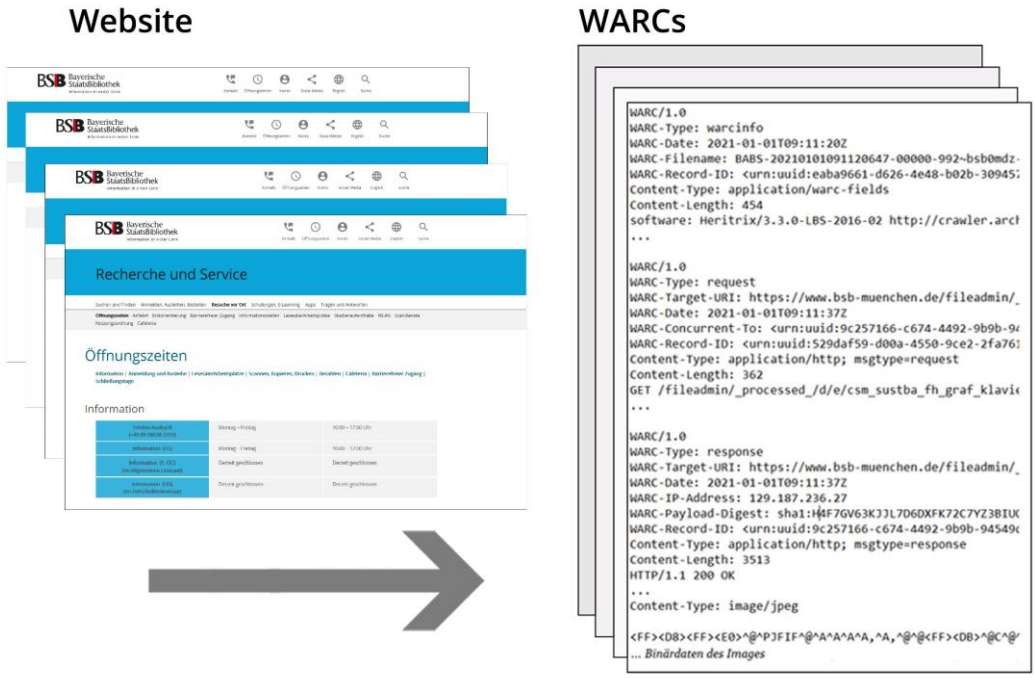


Abbildung 2: Das Archiv einer Website in einer oder mehreren WARC-Dateien, https://langzeitarchivierung.bib-bvb.de/wayback/*/https://www.bsb-muenchen.de/

Standardisierung

Das WARC-Format ist eine vom International Internet Preservation Consortium (IIPC) erarbeitete Erweiterung des ARC-Formats, das 1996 vom Internet Archive entwickelt worden war. Anders als das ARC-Format, das nur die Speicherung der Inhaltsinformationen ermöglichte, erlaubt WARC auch die Aufzeichnung von Metainformationen. Die ISO-Standardisierung 28500:2009 des WARC-Formats Version 1.0 eröffnete den Weg zum Aufbau und der breiten Nutzung von Webarchiven sowie dem Austausch von Ressourcen. Zuletzt wurde 2017 die WARC Version 1.1 als ISO-Standard 28500:2017 publiziert. Die Überarbeitung führte nur kleinere Veränderungen zur Vorversion ein. Derzeit wird eine zweite Revision vorbereitet.

Inzwischen hat sich WARC als das Format für die Webarchivierung durchgesetzt. Neben dem Internet Archive, dem weltweit größten Webarchiv, nutzen auch alle anderen großen nationalen, regionalen oder fachlichen Webarchive das Dateiformat WARC.

Nutzung von WARC-Dateien, heute und zukünftig

Die Standardisierung fördert die Entwicklung einer Vielzahl von interoperablen Open Source Anwendungen: Crawler zur Erstellung der WARCs (z.B. Heritrix), Indexer und Viewer für Suche und Wiedergabe (z.B. Open Wayback, Pywb, SolrWayback) sowie Tools für die Steuerung der Crawls und Verwaltung der Daten (z.B. WebCuratorTool oder der Dienst Archive-It).

Für die wissenschaftliche Nutzung der Webarchive bietet das reichhaltige WARC-Format nach der Extraktion von Metadaten-, Text- und Linkderivaten mittels etablierter Open Source Tools vielfältige Möglichkeiten. Neben der intellektuellen Auswertung einzelner Websites in einem Viewer sind damit auch automatische Text- und Linkanalysen größerer Datenmengen möglich. So können beispielsweise mittels Named Entity Recognition automatisch Eigennamen von Orten oder Personen erkannt, durch Topic Modeling Themen identifiziert sowie Trend- und Netzwerkanalysen durchgeführt werden.
Jedoch bleiben nach wie vor Fragen der Langzeitarchivierung offen, da das Containerformat unterschiedlichste von Obsoleszenz bedrohte Dateiformate enthalten kann. Neben der Migration einzelner im WARC-Container enthaltener Dateiformate wird das Verfahren der Emulation früherer Browsertypen als Langzeitarchivierungsstrategie erprobt.

Referenzen

ISO 28500:2017, Information and documentation — WARC file format
(online unter: https://www.iso.org/standard/68004.html - letzter Zugriff: 08.04.2021)

IIPC: The WARC Format 1.1
(online unter: https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/
- letzter Zugriff: 08.04.2021)

IIPC: Tools & Software (Übersicht des International Internet Preservation Consortiums)
(online unter: https://netpreserve.org/web-archiving/tools-and-software/#tools--software
- letzter Zugriff: 08.04.2021)

Konstanze Weimer, Astrid Schoger

Bayerische Staatsbibliothek
Digitale Bibliothek - Münchener Digitalisierungszentrum - Langzeitarchivierung
langzeitarchivierung@bsb-muenchen.de