Handreiking
WARC-validatie voor webarchiefbestanden

Waarover gaat deze handreiking?

Deze handreiking geeft voorlichting over beschikbare tools die gebruikt kunnen worden om de *technische kwaliteit* van webarchiefbestanden te controleren.

Voor het archiveren van websites is het Web-ARChive of WARC-formaat ontwikkeld. Dit WARC-formaat is gepubliceerd als de internationale standaard ISO 28500. De Richtlijn archiveren overheidswebsites van het Nationaal Archief schrijft voor dat geharveste overheidswebsites in dit bestandsformaat worden bewaard.

Om te controleren of een webarchiefbestand aan de standaard voldoet, kunnen validatietools gebruikt worden. Er zijn verschillende validatietools op de markt. Maar niet iedere tool is even handig, nuttig of geschikt. Om overheidsorganisaties op weg te helpen, heeft het Nationaal Archief een aantal tools onderzocht en met elkaar vergeleken. In deze handreiking delen we de beste oplossingen, gebaseerd op onze testervaringen.

Waarom deze handreiking?

Deze handreiking helpt overheidsorganisaties bij het uitvoeren van technische kwaliteitscontroles in het webarchiveringsproces.

Om de duurzame toegankelijkheid van webarchieven te waarborgen, moet een webarchiefbestand van goede kwaliteit zijn. Welke eisen essentieel zijn staat beschreven in de Richtlijn archiveren overheidswebsites. Een webarchief dat niet aan deze eisen voldoet, kan voor een deel of helemaal onbruikbaar zijn.

Achteraf zijn fouten vaak moeilijk of niet meer te herstellen. Daarom is het belangrijk fouten op tijd op te sporen. Validatietools zijn daarbij een onmisbaar hulpmiddel. Met behulp van deze tools kan geverifieerd worden of een webarchiefbestand aan de WARC-standaard voldoet en dus *technisch* van goede kwaliteit is.

Validatietools kunnen *niet* de inhoudelijke correctheid van webarchiefbestanden controleren. Daarvoor zijn aanvullende handmatige en visuele controles nodig. Ook deze inhoudelijke check is een vereist onderdeel van de periodieke kwaliteitscontrole die overheidsorganisaties moeten uitvoeren op gevormde webarchieven. Zie voor verdere informatie de paragrafen 4.8 en 5.15 van de Richtlijn archiveren overheidswebsites.

Hoe gebruik je deze handreiking?

De handreiking kan gebruikt worden als een praktische bijlage bij de Richtlijn archiveren overheidswebsites.

Om de validatietools in de praktijk toe te kunnen passen is kennis vereist van Linux, command-line interfaces (CLI) en het kunnen uitvoeren van Java en/of Python-applicaties. Om de validatie-uitkomsten goed te kunnen interpreteren is daarnaast ook enige kennis nodig van de Richtlijn archiveren overheidswebsites en de WARC-standaard ISO 28500.

Voor wie is deze informatie?

De handreiking is bedoeld voor overheidsorganisaties. Meer specifiek is de informatie gericht op:

- Archiefbeheerders en/of IT-specialisten die de uitvoerende verantwoordelijkheid dragen voor de kwaliteitscontrole.
- Adviseurs die verantwoordelijk zijn voor (het organiseren van) de archivering van websites. Zoals documentaire informatieadviseurs, informatiemanagers en -beheerders, proces- of informatieanalisten, en adviseurs digitale archivering.
- Projectleiders die de opdracht voor het archiveren van websites uitvoeren.
- De informatie op deze startpagina kan ook dienen als een managementsamenvatting. En relevant zijn voor managers die verantwoordelijk zijn voor de informatie in werkprocessen en de bijbehorende informatiesystemen. Zij nemen de besluiten over de eisen en ontwerpen en zien erop toe dat deze besluiten worden uitgevoerd. Bijvoorbeeld de directeur Bedrijfsvoering of CIO. Managers gebruiken de rest van deze handreiking niet zelf, maar kunnen wel de opdracht geven het toe te passen.

Toepassingskader

Deze handreiking is van toepassing op WARC-validatie als onderdeel van de periodieke kwaliteitscontrole van webarchieven. De informatie is relevant op het moment dat een webarchiveringstrategie wordt ingericht. Dan wordt bepaald welke tools ingezet gaan worden. Het is daarbij aan te raden om de selectie van tools ieder jaar aan de hand van de informatie op deze webpagina te herijken en zo nodig bij te stellen. De markt voor WARC-validatietools ontwikkelt zich. En zo mogelijk ook de adviezen in deze handreiking.

Vragen of suggesties?

Het toepassen van de informatie kan in de praktijk nog vragen oproepen, problemen opleveren en tot nieuwe inzichten leiden. Met die praktijkervaringen wil het Nationaal Archief de handreiking verbeteren en eventueel uitbreiden. Vragen of suggesties voor verbeteringen of ervaringen uit de praktijk ontvangen we dan ook graag. Je kunt deze sturen via info@nationaalarchief.nl.

Disclaimer
De inhoud van deze handreiking is met grote zorg samengesteld. Gezien de snelle ontwikkeling van software kan het desondanks voorkomen dat informatie verouderd, achterhaald of onvolledig is. De handreiking bevat soms verwijzingen naar andere websites die niet door het Nationaal Archief worden onderhouden. Hoewel we uiterst selectief zijn ten aanzien van de sites waarnaar verwezen wordt, aanvaardt het Nationaal Archief geen aansprakelijkheid voor de inhoud en het functioneren daarvan.

Aan de slag met WARC-validatie

Valideren betekent geldig-verklaren. WARC-validatie houdt in dat gecontroleerd wordt of het webarchiefbestand voldoet aan de WARC-standaard ISO-28500. Op dit moment bestaan er twee versies van deze standaard: versie 1.0 (ISO-28500:2009) en versie 1.1 (ISO-28500:2017). Door webarchiefbestanden te valideren kunnen fouten worden opgespoord en kan actie worden ondernomen om ze zo nodig te herstellen.

Wanneer valideren?

Het is aan te raden om de validatietools op regelmatige basis te gebruiken, maar ten minste bij de volgende gebeurtenissen.

Door de verantwoordelijke overheidsorganisatie:
- Direct na de jaarlijkse volledige harvest.
- Voordat de webarchieven naar een andere organisatie worden gebracht. Bijvoorbeeld bij overbrenging naar een archiefinstelling voor blijvende bewaring of naar een andere harvestingdienstverlener voor beheer van de webarchieven.

Door de archiefinstelling:
- Op het moment dat de webarchieven worden overgebracht voor blijvende bewaring.

Welke validatietools zijn geschikt?

Er bestaan verschillende applicaties die webarchiefbestanden op hun conformiteit met de WARC ISO 28500-standaard toetsen. Hieronder zetten we de tools uiteen die het Nationaal Archief adviseert om te gebruiken. Deze tools zijn geselecteerd na een uitgebreide inventarisatie van de markt, waarna ze zijn getest door materiedeskundigen van het Nationaal Archief.

[dropdown met links:]
- JHOVE
- JWAT
- WARCAT
- WARCIO

JHOVE (JSTOR/Harvard Object Validation Environment)

Versie getest: JHOVE-versie 1.22 (van april 2019), met daarin versie 1.0.3 van JWAT-WARC

Algemeen

JHOVE is Java-opensourcesoftware voor identificatie, validatie en karakterisatie van diverse soorten computerbestanden. De software is bruikbaar op UNIX-, Windows- of OS X-besturingssystemen met een passende versie van de *Java Runtime Environment*[1]. JHOVE kan worden gebruikt als *commandoregel-tool*[2], *grafische gebruikersomgeving*[3] of *API*[4].

Sinds versie 1.14 (van mei 2016) kan JHOVE ook WARC-bestanden valideren. Hiervoor is de tool JWAT-WARC (versie 1.0.3 van juni 2015) in JHOVE geïntegreerd. Zie voor een beschrijving van wat JHOVE valideert de beschrijving van JWAT.

Onderhoud en support

Goed - De Open Preservation Foundation ontwikkeld en onderhoudt JHOVE actief .

Ondersteuning WARC ISO 28500-standaard

Versie 1.0 (ISO 28500:2009) ☑

Versie 1.1 (ISO 28500:2017) ☐

Voor- en nadelen

JHOVE is beoordeeld als een nuttig hulpmiddel bij het beheer van WARC-bestanden. Een voordeel van JHOVE is dat het naast WARC's ook andere soorten computerbestanden kan valideren, zoals GIF, JPEG, PDF, TIFF en WAVE. De tool is dus breder inzetbaar dan alleen voor WARC-validatie. Door de actieve gebruikerscommunity en ondersteuning vanuit de Open Preservation Foundation is er bovendien een solide basis voor doorontwikkeling.

Een nadeel van JHOVE is dat de huidige versie niet de laatste versie van JWAT in zich heeft.

Let op: door een bug is het nodig om bij installatie expliciet de WARC-module te selecteren. De automatische herkenning van modules gaat niet altijd goed.

Meer informatie

- Productgebruik en installatie: https://openpreservation.org/products/jhove
- Code en issues: https://github.com/openpreserve/jhove
- Onderhoud: https://openpreservation.org

---

[1] Java Runtime Environment: software om computerprogramma's die zijn geschreven in de programmeertaal Java uit te voeren. Zie ook https://nl.wikipedia.org/wiki/Java_Runtime_Environment.

[2] Commandoregeltool: computerprogramma waarin de gebruiker met tekst opdrachten aan het systeem geeft, zoals de Opdrachtprompt in Windows. Zie ook https://nl.wikipedia.org/wiki/Command-line-interface.

[3] Grafische gebruikersomgeving: computerprogramma waarin de gebruiker in een grafische omgeving met het systeem interacteert, zoals Microsoft Windows. Zie ook https://nl.wikipedia.org/wiki/Grafische_gebruikersomgeving.

[4] Application Programming Interface": definities op basis waarvan computerprogramma's met elkaar kunnen communiceren. Zie ook https://nl.wikipedia.org/wiki/Application_programming_interface.

JWAT (Java Web Archive Toolkit)

Versie getest: JWAT-WARC-versie 1.1.1 (van maart 2018), ingebouwd in de commandoregel-tool JWAT-Tools 0.6.6 (van maart 2018).

Algemeen
JWAT is Java-opensourcesoftware voor het lezen, schrijven en valideren van WARC-, ARC- en GZIP-computerbestanden. De software is bruikbaar op UNIX-, Windows- of OS X-besturingssystemen met een passende versie van de Java Runtime Environment. JWAT kan worden gebruikt als *softwarebibliotheek*[5] of als onderdeel van een commandoregel-tool. Een consortium van (nationale en universiteits)bibliotheken onderhoudt JWAT als onderdeel van de NetarchiveSuite.

JWAT kan WARC-, ARC- en GZIP-computerbestanden valideren. De werking van het WARC-leesproces is hier gedocumenteerd: https://sbforge.org/display/JWAT/WARC+reader+process.

Onderhoud en support
Beperkt – Al enkele jaren is er weinig ontwikkeling en onderhoud aan de software.

Ondersteuning WARC ISO 28500 standaard
Versie 1.0 (ISO 28500:2009) ☑
Versie 1.1 (ISO 28500:2017) ☐

Voor- en nadelen
Een voordeel van JWAT is dat deze software naast WARC-computerbestanden ook ARC's (de voorloper van het WARC-formaat) en gecomprimeerde GZIP-computerbestanden aankan. En meer kan dan alleen valideren (zie Productgebruik en installatie). Een nadeel is dat de tool op dit moment niet heel actief onderhouden wordt.

Meer informatie
- Productgebruik en installatie: https://sbforge.org/display/JWAT/JWAT
- Code en issues: https://github.com/netarchivesuite/jwat en https://sbforge.org/jira/projects/JWAT/issues/
- Onderhoud: https://sbforge.org/display/NAS

[5] (Software)bibliotheek: verzameling code voor gebruik door computerprogramma's. Zie ook https://nl.wikipedia.org/wiki/Bibliotheek_(informatica).

WARCAT (Web ARChive Archiving Tool)

Versie getest: WARCAT 2.2.5 (van april 2017)

Algemeen

WARCAT is Python-opensourcesoftware voor het laagdrempelig en snel werken met WARC-bestanden. De software is bruikbaar op UNIX-, Windows- of OS X-besturingssystemen met een passende Pythonversie. WARCAT kan worden gebruikt als softwarebibliotheek of als commandoregel-tool. Het Archive Team heeft WARCAT ontwikkeld.

WARCAT kan WARC-bestanden valideren. Het verifieert de *digest*[6] en valideert de conformiteit van ISO 28500 (versie 1.0). Zie ook https://github.com/chfoo/warcat/blob/develop/warcat/tool.py.

WARCAT is Python-opensourcesoftware voor het laagdrempelig en snel werken met WARC-bestanden. De software is bruikbaar op UNIX-, Windows- of OS X-besturingssystemen met een passende Pythonversie. WARCAT kan worden gebruikt als softwarebibliotheek of als commandoregel-tool. Het Archive Team heeft WARCAT ontwikkeld.

Onderhoud en support

Beperkt – Al enkele jaren is er weinig ontwikkeling en onderhoud aan de software.

Ondersteuning WARC ISO 28500-standaard

Versie 1.0 (ISO 28500:2009) ☑

Versie 1.1 (ISO 28500:2017) ☐

Voor- en nadelen

Een voordeel van WARCAT is dat het meer kan dan alleen WARC-bestanden valideren. Een nadeel is dat de tool op dit moment niet heel actief onderhouden wordt. En er geen officiële beheerorganisatie achter de tool zit.

Meer informatie

- Productgebruik en installatie: https://warcat.readthedocs.io/en/latest/index.html
- Code en issues: https://github.com/chfoo/warcat
- Onderhoud: https://www.archiveteam.org/

[6] Digest: controlegetal, meestal in de vorm van een checksum, voor een WARC-record (WARC-Block-Digest) of de in dat record opgenomen data (WARC-Payload-Digest).

WARCIO (Streaming WARC/ARC library for fast web archive IO)

Versie getest: WARCIO 1.7.1 (van juli 2019)

Algemeen

WARCIO is Python-opensourcesoftware voor snel, standalone lezen van ARC's en WARC's, en het schrijven van WARC's. De software is bruikbaar op UNIX-, Windows- of OS X-besturingssystemen met een passende Pythonversie. WARCIO kan worden gebruikt als softwarebibliotheek of als commandoregel-tool en is onderdeel van de webrecordersoftware Conifer. Webrecorder Software onderhoudt WARCIO.

WARCIO kan WARC-computerbestanden controleren. Het check-commando controleert zo mogelijk de block- en payloaddigests van versie 1.0 en versie 1.1 van de WARC-standaard.

Onderhoud en support

Goed – WARCIO wordt actief ontwikkeld en onderhouden door Webrecorder Software.

Ondersteuning WARC ISO 28500-standaard

Versie 1.0 (ISO 28500:2009) ☑

Versie 1.1 (ISO 28500:2017) ☑

Voor- en nadelen

Een voordeel van WARCIO is dat het zowel versie 1.0 als versie 1.1 van de WARC-standaard ondersteunt, en actief onderhouden wordt. Een nadeel is dat de check beperkt is tot de controle van block- en payloaddigests.

Meer informatie

- Productgebruik en installatie: https://pypi.org/project/warcio/
- Code en issues: https://github.com/webrecorder/warcio
- Onderhoud: https://webrecorder.net/

Wat controleren de tools?

Een WARC-bestand bestaat volgens de ISO-standaard uit een of meer WARC-records. Deze WARC-records bevatten regels met een veldnaam en een waarde, die van elkaar gescheiden worden door een dubbele punt (veldnaam: waarde). Validatietools verifiëren of de bestandsopbouw en de veldwaarden aan de ISO-28500 normeisen voldoen. De tools geven signalen als er afwijkingen zijn. De output is een waarschuwingsrapport met het aantal invalide en ontbrekende gegevens, wat als trigger dient om actie te ondernemen.

Uit het onderzoek is gebleken dat niet alle geteste validatietools alle relevante velden toetsen aan de WARC-norm. WARCIO bijvoorbeeld is beperkt, maar wel heel gericht gespecialiseerd in het controleren van block- en payloaddigests. In de hieronder gepresenteerde tabel is inzichtelijk gemaakt wat elke tool precies controleert.

Relevante velden WARC NEN-ISO 28500	JHOVE	JWAT	Warcat	WARCIO
WARC-Record-ID	●	●	●	
Content-Length	●	●	●	
WARC-Date	●	●	●	
WARC-Type	●	●	●	
Content-Type	●	●	●	
WARC-Concurrent-To	●	●	●	
WARC-Block-Digest	●	●	●	●
WARC-Payload-Digest	●	●		●
WARC-IP-Address	●	●	●	
WARC-Refers-To	●	●	●	
WARC-Refers-To-Target-URI	●	●		
WARC-Refers-To-Date	●	●		
WARC-Target-URI	●	●	●	
WARC-Truncated	●	●		
WARC-Warcinfo-ID	●	●		
WARC-Filename	●	●	●	
WARC-Profile	●	●		
WARC-Identified-Payload-Type	●	●		
WARC-Segment-Number	●	●		
WARC-Segment-Origin-ID	●	●	●	
WARC-Segment-Total-Length	●	●	●	

NB: De velden WARC-Refers-To-Target-URI en WARC-Refers-To-Date zijn toegevoegd in de NEN ISO-28500:2017-versie van de WARC-standaard.

In een onderzoek uitgevoerd door het Nationaal Archief met een testbestand van 23 webarchieven van verschillende grootte, rapporteerden de verschillende validatietools diverse (fout)meldingen over de geteste WARC-bestanden. De onderstaande tabel bevat een overzicht van de meldingen die voorkwamen. En wat die meldingen (volgens ons) betekenen.

JHOVE

Getest: JHOVE-versie 1.22 (van april 2019), met daarin versie 1.0.3 van JWAT-WARC.
Gebruikte toepassing: commandoregel-tool.
Gebruikt commando: *jhove –m WARC-kb [bestandsnaam].warc*

Melding	Verklaring
Incorrect payload digest, 0F4E929DD5BB2564F7AB9C76338E04E292A42ACE, DA39A3EE5E6B4B0D3255BFEF95601890AFD80709	De in het WARC-record opgeslagen checksum van de payload komt niet overeen met de berekende waarde.
'WARC-Target-URI' value <http://www.website.nl/> Unexpected encapsulating '<' and '>' characters	De WARC-Target-URI bevat onverwachte < en > karakters. Conform de huidige toepassing van de WARC 1.1-specificatie mag dit niet. De 'defacto toepassing' van de WARC 1.0-specificatie maakt ook geen gebruik van deze zogenaamde 'angled brackets'. JHOVE valideert conform de 'defacto toepassing' van de WARC 1.0-specificatie en niet naar de letterlijke specificatie.

JWAT

Getest: JWAT-WARC-versie 1.1.1 (van maart 2018), ingebouwd in de commandoregel-tool JWAT-Tools 0.6.6 (van maart 2018).
Gebruikte toepassing: commandoregel-tool.
Gebruikt commando: *jwattools test -i -e [bestandsnaam].warc*

Melding	Verklaring
'WARC-Target-URI'	De WARC-Target-URI bevat onverwachte < en > karakters. Zie verder verklaring van dezelfde JHOVE-melding.
Incorrect payload digest	De in het WARC-record opgeslagen checksum van de payload komt niet overeen met de berekende waarde.

Warcat

Getest: Warcat 2.2.5 (van april 2017).
Gebruikte toepassing: commandoregel-tool.
Gebruikt commando: *warcat verify [bestandsnaam].warc*

Melding	Verklaring
Warcat.tool.VerifyProblem: ('Bad payload digest.', '5.9', True)	De in het WARC-record opgeslagen checksum, van de payload die aanwezig is of waar naar wordt verwezen, komt niet overeen met de berekende waarde.
Warcat.tool.VerifyProblem: ('Concurrent Record ID <urn:uuid:10399947-52fa-4b4d-bfac-ce1b01c2a22f> not seen yet', None, False)	Het record-ID waar naar verwezen wordt is nog niet voorgekomen in de WARC. Conform de standaard moet een 'concurrent record-ID' al voorgekomen zijn in een eerder WARC-record. Wel mag binnen een WARC-bestand naar een WARC-record vooruit verwezen worden.
Warcat.tool.VerifyProblem: ('Duplicate Record ID.', None, True)	Het record-ID is niet uniek (ten opzichte van alle andere record-ID's in de WARC).

Warcio

Getest: Warcio 1.7.1 (van juli 2019)
Gebruikte toepassing: commandoregel-tool
Gebruikt commando: *warcio check [bestandsnaam].warc*

Melding	Verklaring
Digest present but not checked (revisit)	Het WARC-record bevat een checksum (digest) van de payload maar deze is niet gecontroleerd, omdat het een 'revisit record' betreft. Het controleren is niet eenvoudig, omdat de inhoud van het record dat opnieuw werd bezocht (revisit) elders staat. Het 'revisit record' verwijst hiernaar.
No digest to check	Het WARC-record bevat geen checksum (digest) en kan niet gecontroleerd worden.
Digest present but not checked	Het WARC-record bevat een checksum van de payload, maar die is niet gecontroleerd.
Payload digest failed: sha1:22TRD4UTL6ARBYHUPHEO3BABNW56FYJY	De in het WARC-record opgeslagen checksum van de payload komt niet overeen met de berekende waarde.

Conclusie en advies

De geselecteerde validatietools vormen een goed vertrekpunt om vast te stellen of een webarchiefbestand technisch van goede kwaliteit is. Validatietools zijn echter niet feilloos. Er kunnen bugs in de software zitten. Ook kan het voorkomen dat verschillende tools of verschillende versies van een tool op hetzelfde onderdeel een ander resultaat teruggeven. Uit de testen is bovendien gebleken dat de tools kunnen achterlopen op de ontwikkeling van de WARC-standaard of alleen maar een beperkt aantal aspecten, zoals alleen block- en payloaddigests, controleren.

Concluderend kunnen we stellen dat er geen ultieme WARC-validatietool is die alles afvangt. Daarvoor is het vakgebied ook eigenlijk nog te jong en hebben de tools over het algemeen nog een te laag volwassenheidsniveau. Van de onderzochte tools biedt JHOVE de meeste voordelen. Wat deze tool onderscheidt is de bredere inzetbaarheid, de grafische gebruikersinterface en de actieve beheerorganisatie. Door ook andere tools mee te nemen, kan een nog betere en completere analyse worden gemaakt. Een gecombineerde inzet van tools is daarom voorlopig de beste strategie.

Ervaring opdoen

Ondanks dat WARC-validatie nog in de kinderschoenen staat, kan het nut van validatietools niet genegeerd worden. Het is voor de toekomst van webarchivering belangrijk om te kijken hoe we deze instrumenten van toegevoegde waarde kunnen maken. Ervaring opdoen met het gebruik van validatietools is daarom van groot belang.

Het Nationaal Archief roept organisaties op om ervaringen en opgedane kennis met elkaar uit te wisselen, om van elkaar te leren. Organisaties die hiertoe willen bijdragen kunnen met elkaar en onze experts in gesprek gaan op het kennisplatform Webarchivering van het Kennisnetwerk Informatie en Archief (KIA) of een e-mail sturen naar info@nationaalarchief.nl. Het Nationaal Archief actualiseert de handreiking als nieuwe ontwikkelingen en/of nieuwe inzichten daartoe aanleiding geven.