Crash course webarchivering

NDE Webinar Webarchivering

Jesse de Vos, juni 2020

netwerk digitaal erfgoed

Wat gaan we behandelen?

- Wat is webarchivering
- Het waarom van webarchivering
- Vier fases in webarchivering
- Quickstart guide for het archiveren van websites

Je kunt vandaag nog beginnen met het archiveren van het web!

netwerk digitaal erfgoed

Wat is webarchivering?

"**Web archivering** is het proces van het verzamelen van onderdelen van het Wereld Wijde Web, het preserveren van deze collecties in een archiveerbaar formaat, en deze archieven aanbieden voor toegang en gebruik."

Waarom webarchivering?

Wettelijke redenen:

Het vastleggen van overheidsinformatie of zakelijke communicatie voor eventueel gebruik bij geschillen (resource: Richtlijn webarchivering NA

Cultuur-historische redenen:

Het webcontent als drager van creativiteit, communicatie, entertainment, technologische ontwikkeling, etc.

Waarom webarchivering?

It's Time to Archive the Internet Archive

Publishers are suing the Internet Archive for its emergency library, putting the whole project in danger.

By **Samantha Cole** and **Jason Koebler**

Jun 9 2020, 2:00pm Share Tweet Snap

HEADQUARTERS OF INTERNET ARCHIVE, LOCATED IN RICHMOND DISTRICT, SAN FRANCISCO, CALIFORNIA.
CREDIT: GIRL2K/WIKIMEDIA COMMONS

Related Articles

You Can Now Access 1.4 Million Books for Free Thanks to the Internet Archive

MADDIE BENDER

Viacom Forced Internet Archive to Remove Hundreds of Hours of MTV Broadcasts

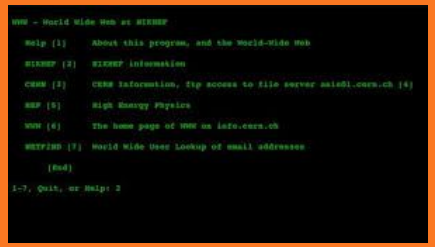
SAMANTHA COLE

Please Don't Sue LeVar Burton for Reading Soothing Stories to Scared Children

SAMANTHA COLE

Het web: een korte geschiedenis

Eerste NL website - 1992 Nikhef

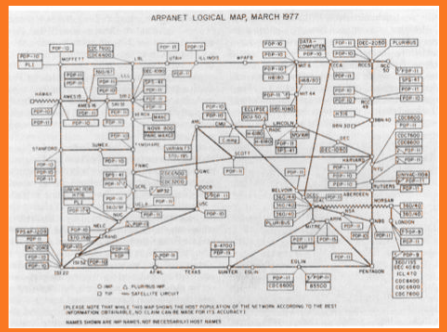


Google - 1998

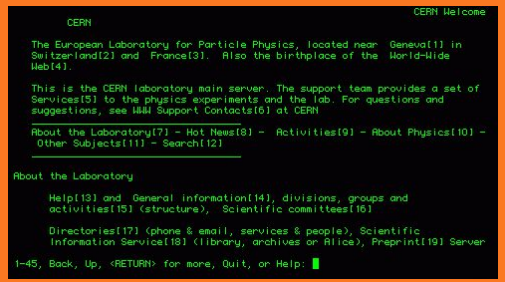
“QWERTYIOP”

Eerste e-mail - 1971

ARPANET - 1969



Eerste website - 1991 CERN



De Digitale Stad - 1995

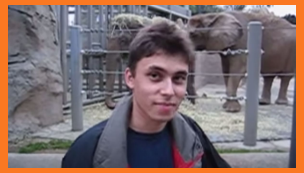
Het web: een korte geschiedenis

Wikipedia - 2001

MySpace - 2003

myspace.com

YouTube - 2005



Blogger - 1999

Blogger

Hyves - 2004

Facebook - 2004

facebook

“just setting up my twttr”

Twitter - 2006

Het web: de statistieken

4.4 miljard gebruikers

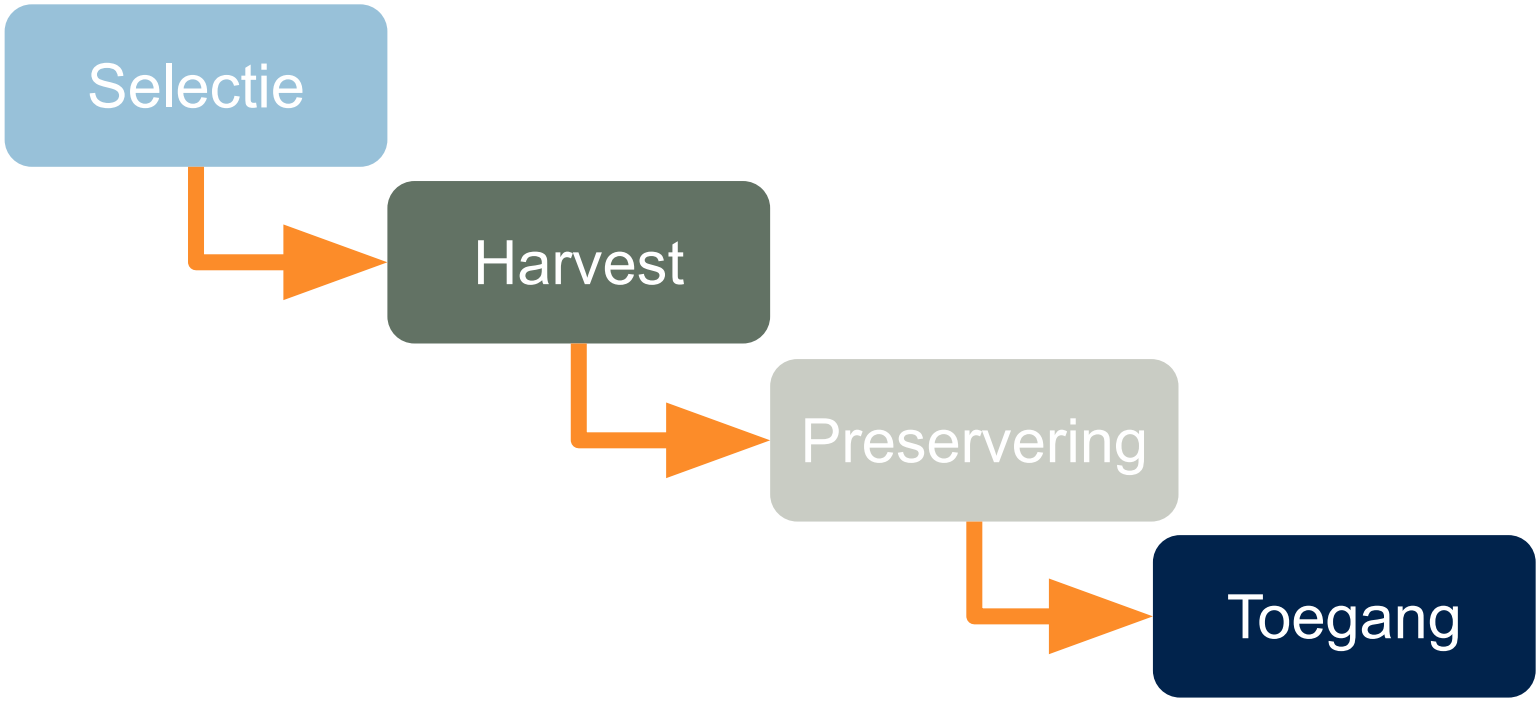
3.5 miljard op social media

52% van het webverkeer is door bots

6 miljoen .nl domeinen

1.7 miljard websites
(200 miljoen actief)

Vier fases in webarchivering



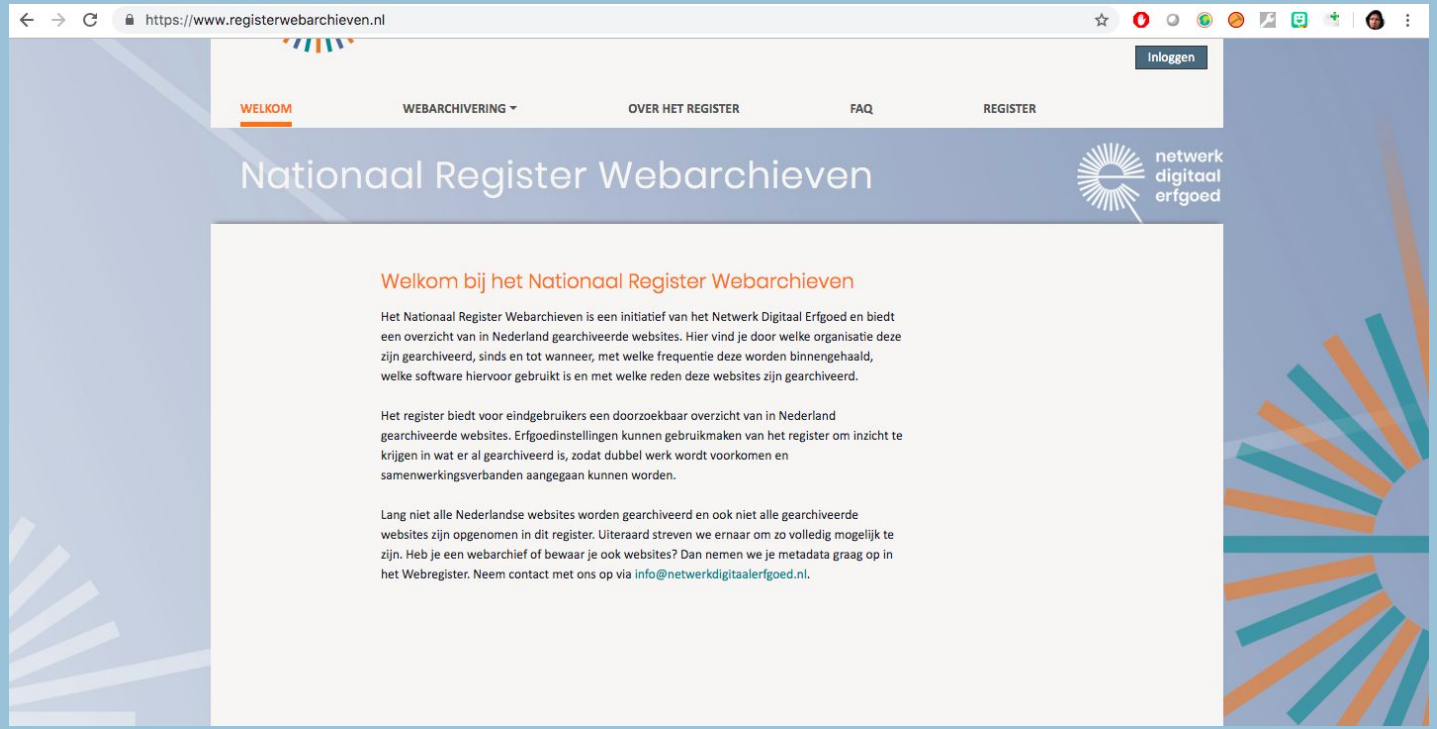
Vier fases in webarchivering

Fase 1: Selectie

- Kan gekoppeld zijn aan wettelijke taak
- Kan gebaseerd zijn op afwegingen over de cultuur-historische waarde van webcontent
- Denk na over de doelgroep
- Afstemming met andere webarchieven in Nederland kan via het Nationaal Register Webarchieven

Fase 1: Selectie

Fase 1: Selectie

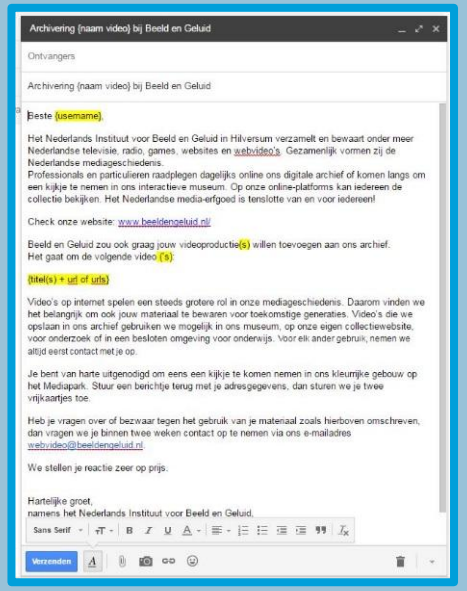


lin

Fase 1: Juridisch

- Opt-out, maar is niet juridisch houdbaar
- Meerdere rechthebbenden
- Ethische zaken m.n. Privacy

Maar: hoop op verruiming van regelgeving voor crawlen van webcontent.



Archivering (naam video) bij Beeld en Geluid

Ontvangers

Archivering (naam video) bij Beeld en Geluid

Beste (username),

Het Nederlands Instituut voor Beeld en Geluid in Hilversum verzamelt en bewaart onder meer Nederlandse televisie, radio, games, websites en webvideo's. Gezamenlijk vormen zij de Nederlandse mediageschiedenis.
Professionals en particulieren raadplegen dagelijks online ons digitale archief of komen langs om een kijkje te nemen in ons interactieve museum. Op onze online-platforms kan iedereen de collectie bekijken. Het Nederlandse media-erfgoed is tenslotte van en voor iedereen!

Check onze website: www.beeldengeluid.nl/

Beeld en Geluid zou ook graag jouw videoproductie(s) willen toevoegen aan ons archief.
Het gaat om de volgende video ('s):

(titel(s) + url of urls)

Video's op internet spelen een steeds grotere rol in onze mediageschiedenis. Daarom vinden we het belangrijk om ook jouw materiaal te bewaren voor toekomstige generaties. Video's die we opslaan in ons archief gebruiken we mogelijk in ons museum, op onze eigen collectiewebsite, voor onderzoek of in een besloten omgeving voor onderwijs. Voor elk ander gebruik, nemen we altijd eerst contact met je op.

Je bent van harte uitgenodigd om eens een kijkje te komen nemen in ons kleurrijke gebouw op het Mediapark. Stuur een berichtje terug met je adresgegevens, dan sturen we je twee vrijkaartjes toe.

Heb je vragen over of bezwaar tegen het gebruik van je materiaal zoals hierboven omschreven, dan vragen we je binnen twee weken contact op te nemen via ons e-mailadres webvideo@beeldengeluid.nl.

We stellen je reactie zeer op prijs.

Hartelijke groet,
namens het Nederlands Instituut voor Beeld en Geluid,

Sans Serif

Verzenden

Fase 2: Harvest

Ook wel ‘crawling’ of ‘capture’

- Het gebruik van een tool (“crawler” of “spider”) om systematisch door een website te browsen
- Configureer de crawler, definieer de scope
- De crawler download code (html, css), afbeeldingen, documenten, en andere bestanden
- Uitsluitend toegestaan met toestemming van rechthebbenden (opt-out is nu gangbaar)

Fase 2: Harvest

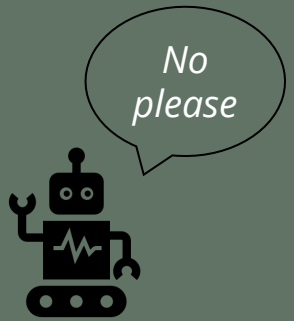
Domain crawl
Een beetje van alles

Select crawl
Alles van een beetje

Fase 2: Harvest

Seed URLs of **URIs**: startpunt(en) voor webcrawlers; de crawler volgt de links vanuit deze eerste URL of set van URLs, een zgn. “Seed List”

Robots.txt: een bestand besloten in een website die een crawler instrueert bepaalde content niet vast te leggen



Fase 2: Harvest

Crawl diepte of “**Hops**”: aantal links vanaf de Seed URL die de crawler crawlt

Crawl-Frequentie: hoe vaak dezelfde webcontent wordt gecrawld

Fase 2: Harvest

Crawler tools, o.a.:

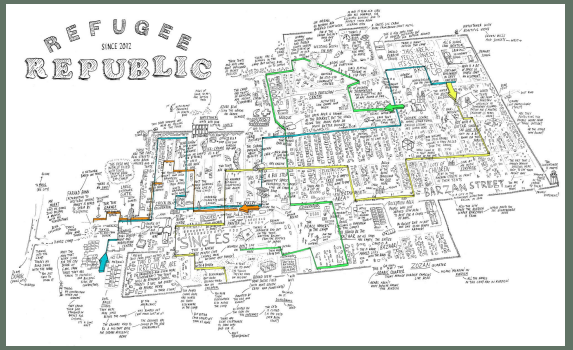
- wget
- Heritrix
- Brozzler
- Web Curator Tool (WCT)

Crawler diensten, o.a.:

- NetarchiveSuite
- Archive-It
- Archiefweb

Dynamische webcontent

Deferred representations: “client-side technologies (e.g., JavaScript) to change the client-side state of a representation after it has initially loaded.” (Brunelle et al 2015)



Dynamische webcontent

Mogelijkheden voor het vastleggen van dynamische webcontent:

- Conifer (was Webrecorder): door mens aangestuurde tool ipv automatische crawler.
- Server-side webarchivering: kopieer de bestanden op de server en maak een virtuele server.

Conifer
by RHIZOME.org

ReproZip

Webarcheologie

Link naar Factsheet Webarcheologie

Fase 3: Preservering

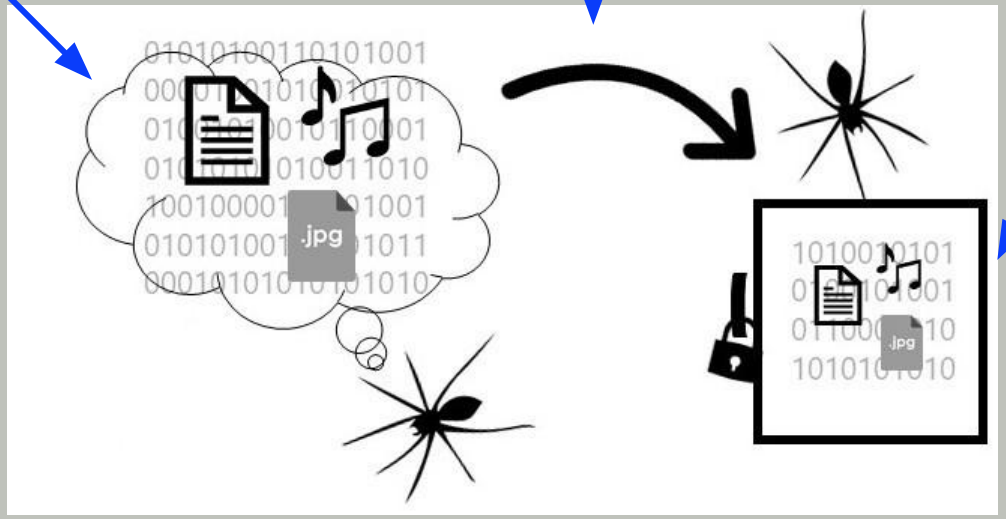
Gedownloade files worden gechecked op kwaliteit, geconverteerd naar een stabiel bestandstype indien noodzakelijk en er wordt op termijn zorg voor gedragen.

Fase 3: Kwaliteitscontrole

Gecrawlde webcontent
SIP

Kwaliteitscontrole!
Blogpost met tips

Webarchief
duurzaam
opgeslagen
AIP



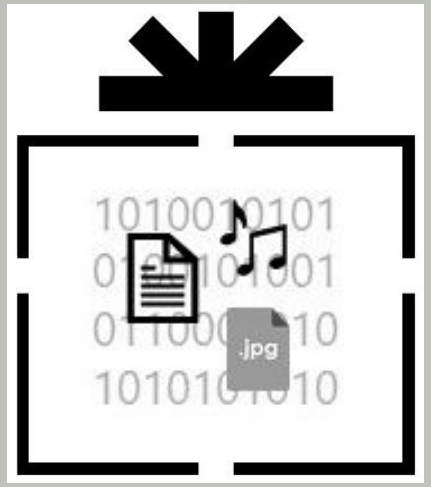
Fase 3: Kwaliteitscontrole

Fouten in de crawl? Onderzoek de broncode van de online website (dmv ‘inspecteren’ in Chrome Browser). Voeg ontbrekende URLs toe aan de ‘seedlist’ in de instellingen van de crawler, of pas het aantal ‘hops’ aan en laat de crawler opnieuw draaien.

Fase 3: Preservering

WARC (Web ARChive)

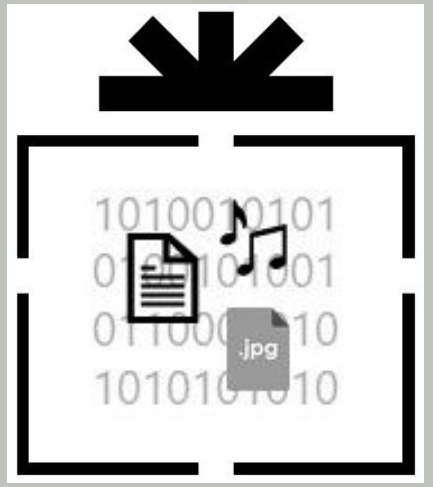
- Een WARC wordt aangemaakt door alle standaard crawling tools
- WARC is een wrapper voor gearchiveerde webobjecten ontwikkeld door de IIPC.
- ISO 28500:2017 (formerly ISO 28500:2009)
- Een WARC-bestand kan instromen in een digitaal preserveringssysteem.
- Voor het WARC-formaat was er het ARC (.arc) formaat



Fase 3: Preservering

WARC (Web ARChive)

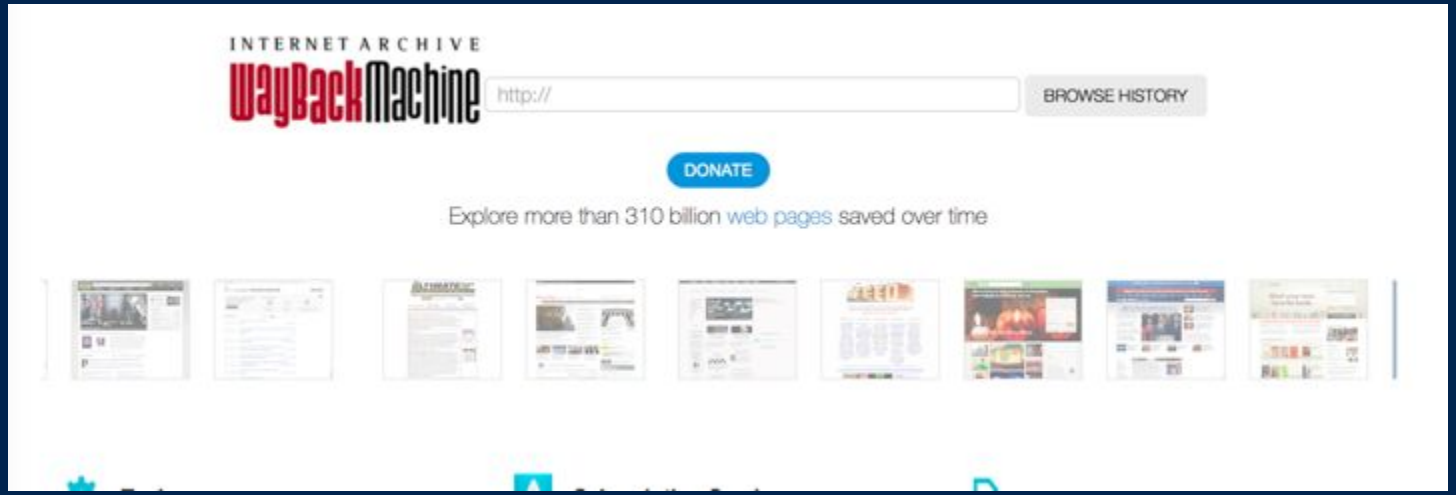
- Een bestand dat verschillende bestandstypen bevat die verkregen zijn door het crawlen
- Behoudt de relatie tussen webpagina's en gerelateerde content
- Bevat beperkte metadata in de header
- Gebruikt specifieke toegangstools



Fase 4: Toegang

- Het on-site of online beschikbaarstellen van het webarchief
- Uitsluitend mogelijk met toestemming van rechthebbenden

Fase 4: Toegang



Fase 4: Toegang

Tools en diensten voor beschikbaarstelling

- Webrecorder Player
- Web Archieven Dashboard (Archiefweb)
- Web Curator Tool (WCT)
- ArchiveIt
- Archives Unleashed Toolset (voor onderzoekers)

Fase 4: Toegang

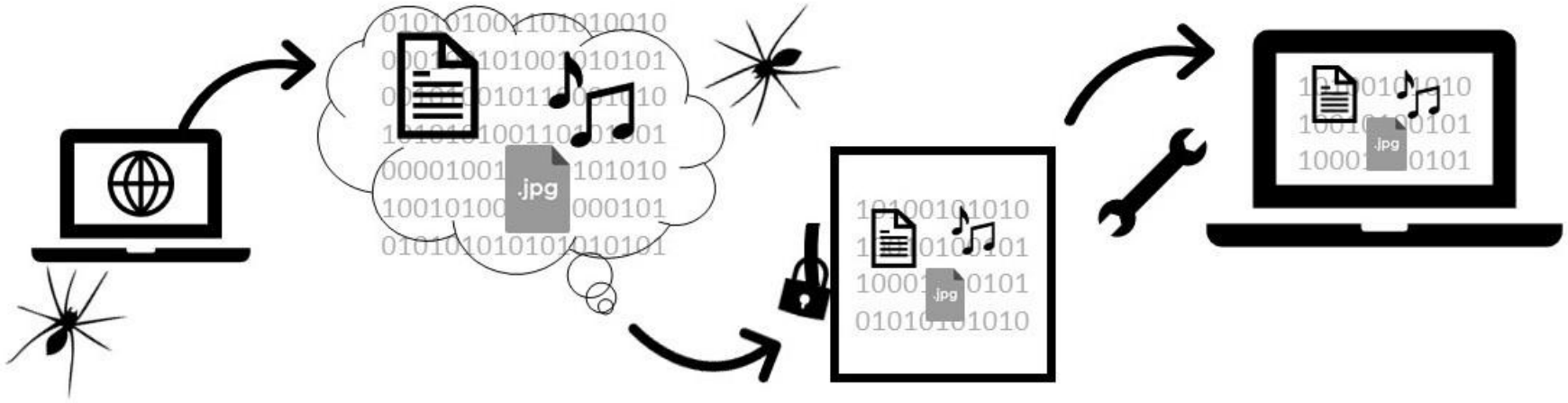
Browseremulatie voor toegang

Vier fases in webarchivering

Crawler

Code en bestanden nodig voor reproductie van website

Playback Tool (bijv. Wayback)



Webarchivering quickstart guide

- Ga naar https://conifer.rhizome.org/
- Crawl de gewenste website
- Download de WARC vanuit je Conifer archief
- Check de kwaliteit van de WARC
- Neem de WARC op in je digitale archief (e-depot)
- Gebruik Webrecorder Player om de gearchiveerde WARC weer te openen

Of doe een suggestie aan een ander Nederlands webarchief via het Nationaal Register Webarchieven

Bedankt!

Jesse de Vos
jdvos@beeldengeluid.nl