Nationaal Archief
Ministerie van Onderwijs, Cultuur en Wetenschap

Ervaringen met website-archivering in het Nationaal Archief

Jeroen van Luin, 8 april 2016

1 Inleiding

In de afgelopen maanden is een aantal websites opgenomen in het e-Depot en via de archiefinventarissen weer beschikbaar gemaakt voor het publiek. Bij het opnemen van deze websites zijn een aantal verschillende technieken gebruikt. In dit verslag wordt door het gebruik van een aantal casussen aangegeven welke ervaringen daarmee zijn opgedaan en wat daarvan geleerd is. Alle technieken maken gebruik van *harvesting*, waarbij een computerprogramma vanuit een startpagina probeert een hele website af te lopen en op te slaan.

1.1 Leeswijzer

In hoofdstuk 2 wordt de belangrijkste stap binnen website-archivering besproken: het doel van het archiveren van de website. De gebruikte harvest-technieken bieden mogelijkheden, maar kennen ook beperkingen. Voordat harvesting plaatsvindt moet eerst worden vast-gesteld of website-archivering wel de geschikte manier is voor de archiveringsbehoefte.

Hoofdstuk 3 gaat in op de gebruikte harvest-technieken. Deze technieken zijn:

1. Direct harvesten van een website die nog live is, via de Website ingest-workflow in het e-Depot. Hierbij wordt gebruik gemaakt van de in het e-Depot ingebakken versie van de tool Heritrix.
2. Harvesten via een live-website met een installatie van Heritrix buiten het e-Depot om.
3. Harvesten van een live-website met de tool Wget.
4. Harvesten van een website die niet meer live te benaderen is, maar waarvan de bronbestanden zijn aangeleverd.

In hoofdstuk 4 wordt beschreven wat er nodig is om de gearchiveerde website weer aan het publiek ter beschikking te stellen. Daarna wordt in hoofdstuk 5 een verslag gegeven van een aantal casussen waarbij website-archivering is toegepast.

2 Bepalen doel van website-archivering

Tijdens het uitvoeren van de verschillende archiveringstechnieken kwam al snel naar voren dat één van de belangrijkste vragen bij website-archivering niet is met welke techniek er gearchiveerd moet gaan worden, maar wat er met de archivering van de website bereikt moet worden.

2.1 Wat kan website-harvesting wel

Een harvester zal alleen die informatie kunnen opslaan die via het aanklikken van een linkje bereikbaar is. Alle informatie die alleen geraadpleegd kan worden na een ander soort handeling (bijvoorbeeld het invullen van een zoekformulier) zal niet worden opgenomen. Wat functionaliteit betreft zal ook alleen client-side functionaliteit beschikbaar blijven (bijvoorbeeld Javascript), en niet de functionaliteit die op de server wordt uitgevoerd (bijvoorbeeld zoeken in een database).

Een harvester ziet ook alleen die content die op het moment van harvesten aanwezig is. Alle data die vlak na een harvest online wordt geplaatst is niet meegenomen. En belangrijker: alle informatie die na een harvest is geplaatst, en vóór een volgende harvest weer is verwijderd, zal dus nooit in een gearchiveerde versie van de website terechtkomen. Op dit moment is er nog geen ervaring met korte, snel herhaalde harvests, waarbij steeds alleen de wijzigingen ten opzichte van de vorige harvest (incrementele harvest) worden opgeslagen.

Doordat harvesting geen server-side functionaliteit kan opnemen, en doordat het altijd een momentopname is, is website-archivering door middel van harvesting vooral geschikt om te dienen als digitaal tijdsbeeld. Bijvoorbeeld wanneer een website vervangen wordt door een nieuwe versie, waarbij de belangrijke inhoud wordt overgenomen in de nieuwe website, maar waarvan men wel wil bewaren hoe de oude website eruit zag.

2.2 Wat kan website-harvesting niet

Website-harvesting geeft wel goed weer hoe de informatie op dat moment aan de gebruikers getoond werd, maar het is niet geschikt om te dienen als vervanging voor reguliere overbrenging van te bewaren documenten. Alle functionaliteit die een website normaal gesproken biedt om de informatie in een website vindbaar te maken, ontbreekt.

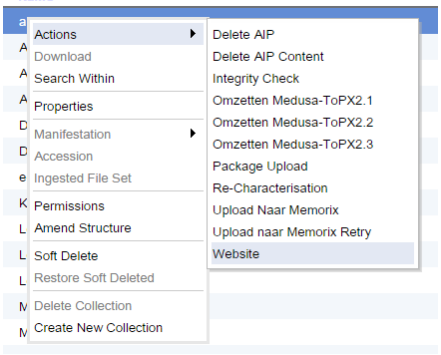
Wanneer een website het publicatiekanaal is waarmee besluiten aan een burger worden getoond, is website-archivering niet de manier om die besluiten te archiveren. De besluiten moeten dan ook als eigen archiefstukken worden overgedragen, zodat de gebruikers via de gebruikelijke weg de stukken kunnen vinden.

Nu kent elke regel een uitzondering, dus deze ook: in één casus is een gearchiveerde website wel geschikt gebleken om de plek te zijn om informatie aan de gebruiker te geven: bij het aantonen van de onveranderlijkheid van de informatie. In deze casus krijgt de Tweede Kamer op Prinsjesdag de miljoenennota aangeboden, waar onder andere ook websites als bijlage worden gebruikt. De kamer wil de zekerheid dat de informatie die op Prinsjesdag is aangeboden, tijdens de Algemene Politieke Beschouwingen niet ongemerkt gewijzigd wordt. In dat geval kan op Prinsjesdag een harvest van de bijlage-websites worden uitgevoerd, en kunnen de Kamerleden via de gearchiveerde website de zekerheid hebben dat de informatie niet is gewijzigd. Wel moeten de websites zo zijn ontworpen en gebouwd, dat ze zonder server-side functionaliteit goed te gebruiken zijn.

3 Gebruikte harvesting-technieken

3.1 Harvesten met Heritrix via de Website ingest-workflow

De makkelijkste manier van opnemen van een live-website is het gebruik van de Website ingest-workflow in het e-Depot. In de verkenner van het e-Depot wordt de collectie geselecteerd waarbinnen de website als nieuw dossier moet worden opgenomen, en wordt de Website ingest-workflow gestart.



3.1.1 Configureren van de Website ingest-workflow

Na het opstarten van de workflow kunnen een aantal configuratie-instellingen worden ingevuld, die hieronder worden toegelicht:

- Seed URL
- Title en Scope & Content
- Catalogue reference
- Security Tag
- Max Hops en Max Path Depth
- Max Download Size en Max Download Time
- Exclude extensions
- Honour robots.txt
- Output type

Alle andere configuratie-instellingen die bij Heritrix aanpasbaar zijn, krijgen een standaardwaarde van Preservica, en kunnen niet door de gebruiker per harvest worden aangepast.

3.1.1.1 Seed URL

Om een harvest te kunnen uitvoeren moet de harvester weten op welke pagina hij moet beginnen. Deze Seed URL wordt als eerste bezocht en gedownload. Vervolgens kijkt de harvester welke URL's op deze pagina voorkomen. Alle links binnen dezelfde website zullen daarbij worden aangemerkt als vervolgstappen. Links die niet wijzen naar dezelfde website worden genegeerd.

3.1.1.2 Title en Scope & content

Na het harvesten zal de website worden opgenomen als nieuw dossier binnen de collectie waarin de workflow was opgestart. Alle dossiers moeten een titel en een beschrijving hebben, en die kan hier worden ingevuld.

3.1.1.3 Catalogue reference

Het dossier dat na afloop van de harvest wordt aangemaakt krijgt de waarde die in dit veld wordt ingevuld als inventarisnummer.

3.1.1.4 Security Tag

De meeste websites die worden opgenomen zullen volledig openbaar zijn, maar voor het geval er een beperkt-openbare website moet worden geharvest, kan hier de bijhorende security-tag worden ingevuld.

3.1.1.5 Max Hops en Max Path Depth

De 'hops' is het aantal links dat gevolgd moet worden om vanaf de startpagina op een betreffende pagina te komen. Een pagina die bereikbaar is door vanaf de startpagina op een linkje te klikken, heeft een hopwaarde van 1. De pagina's die daar weer op gelinkt staan hebben een hopwaarde van 2. Een pagina die vanaf meerdere andere pagina's gelinkt wordt, neemt de kleinste waarde.
Door bij 'max hops' een waarde in te vullen, kan ervoor gezorgd worden dat maar een beperkt deel van de website wordt opgenomen. Een max hops waarde van 0 betekent dat het aantal hops geen criterium is bij het bepalen welke pagina's worden opgenomen.

De 'path depth' is het aantal URL-delen in het adres van een pagina volgend op de start-URL van de website. Bij bijvoorbeeld de start-URL 'www.nationaalarchief.nl' en een path depth van 1 zullen pagina's en files in

`http://www.nationaalarchief.nl/organisatie`

wel worden meegenomen, maar pagina's in

`http://www.nationaalarchief.nl/organisatie/architectuur-beleid`

niet, omdat die meer dan 1 path depth van de start-URL verwijderd zijn.

Wanneer de start-URL 'http://www.nationaalarchief.nl/organisatie' is, dan worden de pagina's en bestanden binnen

`http://www.nationaalarchief.nl/organisatie/architectuur-beleid`

wel meegenomen, omdat ze nu 1 diepte verwijderd zijn van de start-URL.

Door een waarde in te vullen kan er voor gezorgd worden dat pagina's met een path depth hoger dan de opgegeven waarde niet worden meegenomen. Op deze manier kan het deel van de website dat wordt opgenomen worden beperkt. Een max path depth waarde van 0 betekent dat de path depth geen criterium is bij het bepalen welke pagina's worden opgenomen.

3.1.1.6 Max Download Size en Max Download Time

Via de configuratie-instelling 'Max Download Size' kan worden aangegeven dat de harvest na een opgegeven aantal megabytes moet stoppen. Een max download waarde van 0 betekent dat de omvang van de gedownloade bestanden geen criterium is bij het bepalen welke pagina's worden opgenomen.

Via de configuratie-instelling 'Max Download Time' kan worden aangegeven na hoeveel doorlooptijd (in minuten) de harvest moet stoppen. Een max time waarde van 0 betekent dat de doorlooptijd van de harvest geen criterium is bij het bepalen welke pagina's worden opgenomen.

3.1.1.7 Exclude extensions

Via de exclude extensions configuratie-instelling kan worden aangegeven welke extensies moeten worden overgeslagen. Door hier bijvoorbeeld 'zip,exe' in te vullen, zullen alle in de website gelinkte zip-bestanden en executables worden overgeslagen. Een lege waarde voor deze instelling betekent dat alle soorten bestanden en pagina's zullen worden opgenomen.

3.1.1.8 Honour robots.txt

Alle websites kunnen via een bestandje met de naam 'robots.txt' aangeven hoe zij willen dat webcrawlers met de website omgaan. Hierin kan een website-eigenaar bijvoorbeeld

aangeven dat bepaalde onderdelen van de website niet door een zoekmachine moeten worden geïndexeerd. Het daadwerkelijk opvolgen van het verzoek in de robots.txt is niet verplicht, maar wordt door de website-eigenaar wel op prijs gesteld. Bij harvesting kan het wenselijk zijn om die website-onderdelen wel op te nemen. In dat geval kan via het uitzetten van het vinkje worden aangegeven dat de instructies in robots.txt moeten worden genegeerd.

3.1.1.9 Output type

De Heritrix-harvester kan drie formaten uitvoeren: het standaard WARC-formaat, het verouderde ARC-formaat, of een afspiegeling van de bestandsstructuur. Alleen WARC en ARC kunnen weer worden gepresenteerd als website in het e-Depot.

3.1.2 Uitvoeren van de Website ingest-workflow

Wanneer de Website ingest-workflow is geconfigureerd kan deze worden gestart. De Heritrix harvester is standaard zo ingesteld dat deze relatief langzaam pagina's binnenhaalt, om zo de druk op de webserver beperkt te houden. Hierdoor kan het dus enige tijd duren voordat de harvest van een website is afgerond.

Nadat een website-harvest klaar is, wordt de uitvoer eerst opgenomen in een XIP[1]-pakket, en volgt daarna alle stappen die een normale ingest ook zou doorlopen:

State	Name	Progress
	SelectUrl	
	ConfigureCrawl	
✓	Url Crawler	
✓	Create Website XIP	
✓	Virus Check	
✓	Metadata Integrity	
✓	Content Integrity	
✓	Fixity Check	
✓	SIP Validation	
✓	SIP Validation with Database Crosscheck	
✓	Characterise	
✓	Store Files	
✓	Store Metadata	
✓	Store Metadata File	
✓	Update Search Index	
✓	Thumbnail Creation	

Tijdens de ingest worden een aantal controlestappen uitgevoerd om ervoor te zorgen dat er geen virussen meekomen, en dat alle benodigde data en metadata in orde is, en dat de technische metadata (bijv. bestandsformaten) wordt toegevoegd. Daarna wordt de website daadwerkelijk opgenomen op het opslagsysteem, en wordt de metadata opgeslagen in de database. Na het registreren van het nieuwe dossier in de zoek-index en het maken van de thumbnail is de opname afgerond.

De meeste van deze controlestappen zullen nooit tot problemen leiden, omdat het e-Depot zelf net ter plekke de SIP heeft aangemaakt. Toch worden voor alle zekerheid de controles uitgevoerd.

---

[1] XIP is het interne metadata-formaat dat wordt gebruikt in het softwarepakket Preservica, dat de basis vormt van de e-Depot-applicatie. Het XIP-formaat kan gebruikt worden in alle drie informatiepakket-soorten uit het functionele model in de OAIS-standaard: SIP, AIP en DIP.

3.2 Harvesten met Heritrix buiten het e-Depot om

De harvest-tool Heritrix kent een enorme hoeveelheid configureerbare instellingen. Binnen het e-Depot is een versie geïnstalleerd waarbij voor de meeste van deze configuratie-instellingen standaardwaarden worden gebruikt. Alleen de 11 instellingen die in het vorige hoofdstuk zijn beschreven kunnen door de gebruiker van de ingest-workflow worden aangepast.

Wanneer de standaardwaarden van de andere configuratie-instellingen niet voldoet, of wanneer een externe organisatie verantwoordelijk is voor het harvesten van de website, kan ook buiten het e-Depot om gebruik worden gemaakt van Heritrix.

3.2.1 Configureren van Heritrix

In een stand-alone installatie van Heritrix kan per "job" de configuratie worden bepaald. Naast de 11 instellingen die in het e-Depot kunnen worden ingevuld, zijn er heel veel andere aanpassingen die kunnen worden gedaan door wijzigingen aan te brengen in de *Crawl job configuration file*.

save changes C:\Temp\heritrix-3.1.0\bin\jobs\Test\crawler-beans.cxml view

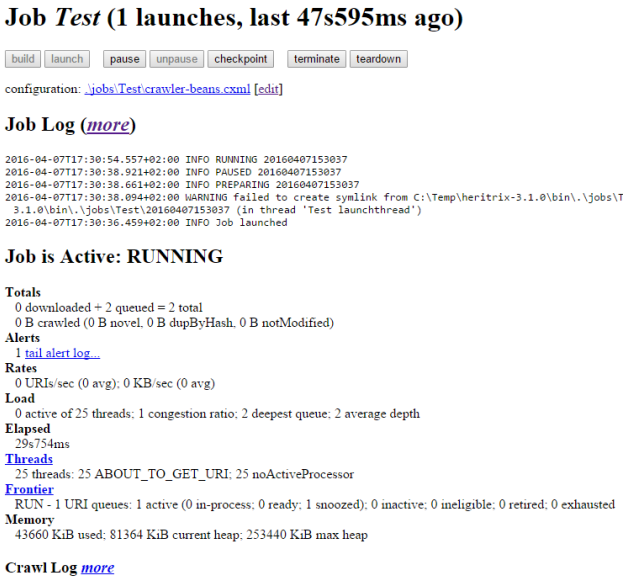
```
<?xml version="1.0" encoding="UTF-8"?>
<!--
  HERITRIX 3 CRAWL JOB CONFIGURATION FILE

  This is a relatively minimal configuration suitable for many crawls.

  Commented-out beans and properties are provided as an example; values
  shown in comments reflect the actual defaults which are in effect
  if not otherwise specified specification. (To change from the default
  behavior, uncomment AND alter the shown values.)
 -->
<beans xmlns="http://www.springframework.org/schema/beans"
            xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xmlns:context="http://www.springframework.org/schema/context"
            xmlns:aop="http://www.springframework.org/schema/aop"
            xmlns:tx="http://www.springframework.org/schema/tx"
            xsi:schemaLocation="http://www.springframework.org/schema/beans http://www.springframework.org/s
beans-3.0.xsd
          http://www.springframework.org/schema/aop http://www.springframework.org/schema/aop/spring-aop-3.0
          http://www.springframework.org/schema/tx http://www.springframework.org/schema/tx/spring-tx-3.0.xs
          http://www.springframework.org/schema/context http://www.springframework.org/schema/context/spring

 <context:annotation-config/>

<!--
  OVERRIDES
   Values elsewhere in the configuration may be replaced ('overridden')
   by a Properties map declared in a PropertiesOverrideConfigurer,
   using a dotted-bean-path to address individual bean properties.
   This allows us to collect a few of the most-often changed values
   in an easy-to-edit format here at the beginning of the model
   configuration.
 -->
 <!-- overrides from a text property list -->
 <bean id="simpleOverrides" class="org.springframework.beans.factory.config.PropertyOverrideConfigurer">
  <property name="properties">
   <value>
# This Properties map is specified in the Java 'property list' text format
# http://java.sun.com/javase/6/docs/api/java/util/Properties.html#load%28java.io.Reader%29

metadata.operatorContactUrl=ENTER_AN_URL_WITH_YOUR_CONTACT_INFO_HERE_FOR_WEBMASTERS_AFFECTED_BY_YOUR_CRAWL
metadata.jobName=basic
metadata.description=Basic crawl starting with useful defaults
```

Het kiezen van de juiste configuratie-instellingen, waaronder welke pagina's wel en niet worden geharvest, en hoe snel Heritrix pagina's mag binnenhalen, vereist wel een uitgebreide studie van de handleiding van Heritrix.

Nadat de configuratie naar wens is aangepast, kan Heritrix de opdracht worden gegeven om de harvest te starten.

3.2.2 Uitvoeren van de harvest

De harvest wordt gestart door het gebruik van achtereenvolgens de knoppen ‘build’, ‘launch’ en ‘unpause’. Tijdens de harvest kan op de console van Heritrix worden bekeken wat de voortgang van de harvest is.



Job *Test* (1 launches, last 47s595ms ago)

build | launch | pause | unpause | checkpoint | terminate | teardown

configuration: .\jobs\Test\crawler-beans.cxml [edit]

Job Log (*more*)

```
2016-04-07T17:30:54.557+02:00 INFO RUNNING 20160407153037
2016-04-07T17:30:38.921+02:00 INFO PAUSED 20160407153037
2016-04-07T17:30:38.661+02:00 INFO PREPARING 20160407153037
2016-04-07T17:30:38.094+02:00 WARNING failed to create symlink from C:\Temp\heritrix-3.1.0\bin\.\jobs\T
 3.1.0\bin\.\jobs\Test\20160407153037 (in thread 'Test launchthread')
2016-04-07T17:30:36.459+02:00 INFO Job launched
```

Job is Active: RUNNING

Totals
0 downloaded + 2 queued = 2 total
0 B crawled (0 B novel, 0 B dupByHash, 0 B notModified)
Alerts
1 tail alert log...
Rates
0 URIs/sec (0 avg); 0 KB/sec (0 avg)
Load
0 active of 25 threads; 1 congestion ratio; 2 deepest queue; 2 average depth
Elapsed
29s754ms
Threads
25 threads: 25 ABOUT_TO_GET_URI; 25 noActiveProcessor
Frontier
RUN - 1 URI queues: 1 active (0 in-process; 0 ready; 1 snoozed); 0 inactive; 0 ineligible; 0 retired; 0 exhausted
Memory
43660 KiB used; 81364 KiB current heap; 253440 KiB max heap

Crawl Log *more*

Nadat de harvest klaar is, kan het resulterende WARC-bestand worden gedownload en worden klaargemaakt voor opname in het e-Depot.

3.2.3 Maken van de SIP

Om de WARC te kunnen opnemen moet er een SIP van worden gemaakt. Hiervoor kan de SIPCreator worden gebruikt, die uit één of meer losse bestanden een XIP-pakket kan maken dat opgenomen kan worden. Tussen het aanmaken van de XIP en het starten van de ingest moet alleen handmatig één handeling worden uitgevoerd: het e-Depot moet verteld worden wat de start-URL van de website was, zodat hij bij het weergeven van de website weet op welke pagina hij moet starten.

Om dit voor elkaar te krijgen moet in het metadatabestand van de XIP een XML-fragment worden toegevoegd in de “accession”. Een voorbeeld van zo’n toe te voegen fragment is:

```
<AccessionEvent>
   <EventDate>2016-04-07T15:49:33.925+01:00</EventDate>
   <Process>http://www.ministervanboxtel.nl</Process>
   <Outcome>2016-04-07T15:49:33.925+01:00</Outcome>
</AccessionEvent>
```

Hieraan kan het e-Depot later zien dat voor het correct weergeven van deze WARC hij op zoek moet naar de pagina ‘http://www.ministervanboxtel.nl’ zoals deze was op 7 april 2016 om 15:49:33. Wanneer er meerdere versies van een website zijn opgenomen weet het e-Depot dus welke versie hier bedoeld wordt.

3.2.4 Ingesten van de SIP

Na het maken van de aangevulde SIP kan deze middels een standaard ingest-workflow worden opgenomen in het e-Depot.

3.3 Harvesten van een live-website met Wget

Als standaard-onderdeel van vrijwel elke Linux-installatie wordt de command-line tool Wget meegeleverd. In nieuwere versies van Wget bestaat de mogelijkheid om recursief een hele website te harvesten, en het resultaat als WARC-bestand uit te voeren.

3.3.1 Configureren van de Wget-tool

Doordat Wget een command-line tool is, is er geen andere configuratie nodig dan het bij het opstarten van de tool meegeven van de juiste opstart-parameters.

De gebruikte opstart-parameters zijn:

`-m / --mirror`	Maak een afspiegeling van de hele website (recursief downloaden van alle pagina's, ongeacht het aantal hops of link depth, -N --no-remove-listing
`-k / --convert-links`	Past links in gedownloade pagina's aan zodat ze geschikt zijn voor lokaal bekijken
`-p / --page-requisites`	Download alles wat nodig is voor het tonen van een pagina, inclusief alle gelinkte bestanden zoals afbeeldingen, geluidsfragmenten, stylesheets, etc.
`-E / --adjust-extension`	Voeg de extensie ".html" toe aan html-bestanden, en ".css" toe aan css bestanden, zodat de bestanden de juiste benaming krijgen.
`-w 1`	Wacht 1 seconde tussen downloads
`--warc-file="<naam>"`	Gebruik <naam> als bestandsnaam voor de WARC
`--no-warc-compression`	
`<URL>`	Gebruik <URL> als start-URL van de website

3.3.2 Uitvoeren van de harvest

De harvest is uitgevoerd door in Cygwin de wget-tool te gebruiken, met bijvoorbeeld de volgende command-line aanroep:

```
wget -m -k -p -E -w 1 --warc-file="ministervanboxtel"
http://www.ministervanboxtel.nl
```

Hierna zal in de console de lijst verschijnen met alle acties die door Wget worden uitgevoerd, en de pagina's die worden binnengehaald. Na afloop van de harvest is er een lokaal opgebouwde kopie van de website, en is er een WARC bestand met de naam "nationaalarchief.warc" waarin de website zit.

3.3.3 Maken van de SIP

Van dit WARC-bestand moet een SIP worden gemaakt, analoog aan de SIP die gemaakt moet worden bij gebruik van een externe Heritrix zoals beschreven in paragraaf 3.2.3.

3.3.4 Ingesten van de SIP

Na het maken van de aangevulde SIP kan deze naar de quarantaine-zone van het e-Depot worden geüpload, en daarna middels een standaard ingest-workflow worden opgenomen in het e-Depot.

3.4 Harvesten van losse bronbestanden

Websites die niet meer online te benaderen zijn, kunnen door de harvester niet worden opgevraagd, en dus niet worden gedownload. In dit geval zal de oude website eerst weer actief moeten worden gemaakt. Doordat in de meeste gevallen de URL weer hergebruikt is voor een nieuwe website, of doordat het ongewenst is dat de oude website weer voor het publiek benaderbaar is, moeten wat speciale technieken worden toegepast om ervoor te zorgen dat de website toch geharvest kan worden.

3.4.1 Inrichten van een lokale webserver

De harvester (Heritrix of Wget) zoekt tijdens het harvesten contact met een webserver om alle pagina's en andere website-onderdelen op te vragen. Zo'n webserver kan lokaal (dus op de eigen PC) worden geïnstalleerd om de lokaal beschikbare website-bestanden voor de harvester beschikbaar te maken. Welke webserver daarbij nodig is, hangt af van de techniek die bij de oude website werd gebruikt.

In het eenvoudigste geval bestaat de website alleen uit losse HTML-bestanden, met stylesheets, afbeeldingen en andere gelinkte inhoud. In dat geval is elke webserver geschikt om de website te serveren. Bijvoorbeeld:

- Python's SimpleHTTPServer (onderdeel van de standaard Python distributie)
- Apache's XAMPP (te downloaden bij ApacheFriends)
- Windows IIS (standaard in Windows 7, 8 en 10, maar moet aangezet worden)

Complexere websites kunnen server-side scripting hebben (bijvoorbeeld PHP of ASP), of gebruik maken van een database. In dat geval is een simpele webserver niet voldoende, en zal een geavanceerdere webserver als XAMPP of IIS gebruikt moeten worden. Hierbij is XAMPP meer geschikt voor PHP en MySQL-combinaties, en IIS meer geschikt voor ASP en MS-SQL-combinaties.

In het begin van de jaren 2000 was het populair om websites te bouwen in XML, die met XSLT's werden omgezet naar HTML. Voor dit soort websites is het vaak nodig om eerst de omzetting uit te voeren, zodat de HTML-bestanden gevormd zijn, en deze dan te behandelen als een 'eenvoudige' website.

3.4.2 Aanpassen van de lokale hosts-file

Wanneer de website niet meer online beschikbaar is, is het meestal ook niet mogelijk, of niet wenselijk, dat de oude website tijdelijk wel weer onder de oorspronkelijke naam beschikbaar komt. Daarom moet de PC waar de harvester op draait geïnstrueerd worden waar hij de oude website kan vinden. Dit kan door middel van het aanpassen van de hosts-file.

In de hosts file kan worden aangegeven welk IP-adres hoort bij een webadres. Dit bestand wordt altijd als eerste geraadpleegd wanneer een webadres wordt opgevraagd. Het aanpassen van deze hosts-file vereist beheerder-rechten (local admin op Windows, root op Linux).

Onder Windows is dit bestand te vinden op de locatie:

```
C:\Windows\System32\drivers\etc\hosts
```

Onder Linux is dit bestand te vinden op de locatie:

```
/etc/hosts
```

Met een tekst-editor kan worden aangegeven dat een bepaalde webserver te vinden is op de eigen machine, door toevoeging van de regel

```
127.0.0.1    <webadres>
```

bijvoorbeeld

```
127.0.0.1    www.ministervanboxtel.nl
```

Het IP-adres 127.0.0.1 wijst hierbij altijd naar de eigen machine, de 'localhost'. Via de webbrowser kan nu worden gekeken of het gelukt is. Door in de URL-balk de website-naam in te voeren (in dit geval 'www.ministervanboxtel.nl') zou de oude website zichtbaar moeten worden.

3.4.3 Uitvoeren van de harvest

Nu de website weer lokaal actief is, kan de harvest worden gestart op dezelfde manier als elke andere website. Doordat de harvester bij elke aanvraag van een URL te horen krijgt dat hij contact moet zoeken met de webserver op de eigen machine, zal steeds de lokale webserver bevraagd worden, en zal de oude website worden geharvest. Uiteindelijk resulteert dat in een WARC-bestand van de oude website.

3.4.4 Maken van de SIP

Van dit WARC-bestand moet een SIP worden gemaakt, analoog aan het maken van de SIP zoals beschreven in paragraaf 3.2.3.

3.4.5 Ingesten van de SIP

Na het maken van de aangevulde SIP kan deze naar de quarantaine-zone van het e-Depot worden geüpload, en daarna middels een standaard ingest-workflow worden opgenomen in het e-Depot.

4 Beschikbaar stellen van een gearchiveerde website

Nadat een website is opgenomen als digitaal object in het e-Depot, kan deze aan het publiek ter beschikking worden gesteld. In het e-Depot zit daarvoor de module Wayback-machine ingebouwd.

Voor het gebruik van de Wayback-machine zijn drie onderdelen nodig:

1. De basis-URL van de Wayback-machine voor een tenant
2. De UUID van het WARC-bestand in het e-Depot
3. Een publicatiekanaal waarop de URL aan het publiek wordt getoond

4.1 Tenant-specifieke basis-URL van de Wayback-machine

In het e-Depot worden meerdere tenants gebruikt. De naam van deze tenant is onderdeel van de basis-URL van de Wayback-machine. Deze tenant-naam is meestal gelijk aan de ISIL-code van de betreffende tenant-houder, zonder koppelteken tussen "NL" en de plaatsaanduiding:

Gelders Archief	NLAhGldA
Groninger Archieven	NLGrGRA
Het Utrechts Archief	NLUtHUA
Historisch Centrum Overijssel	NLZlHCO
Nationaal Archief	NA[2] (Productie) of NLHaNA (TED[3])
Nieuw Land Erfgoedcentrum	NLLlsNLE
Noord-Hollands Archief	NLHlmNHA
RHC Limburg	NLMtRHCL
Tresoar	NL040041000
Zeeuws Archief	NLMdbZA

Deze tenant-naam kan worden ingevuld in de basis-URL:

```
https://e-depot.nationaalarchief.nl/Render/render/external?
tenant=<tenantnaam>&entity=TypeFile&entityRef=
```

of voor de TED-omgeving:

```
https://e-depot-ted.nationaalarchief.nl/Render/render/external?
tenant=<tenantnaam>&entity=TypeFile&entityRef=
```

Aan het einde van deze URL moet de unieke identifier van het te tonen WARC-bestand worden geplakt.

Voor de productieomgeving van het Nationaal Archief is de basis-URL dus:

```
https://e-depot.nationaalarchief.nl/Render/render/external?
tenant=NA&entity=TypeFile&entityRef=
```

---

[2] De standaardisering van tenantnamen tot de ISIL-code is geïntroduceerd nadat de Nationaal Archief-tenant in de productieomgeving was aangemaakt. De functionaliteit voor het hernoemen van tenants wordt nog door de leverancier ontwikkeld.

[3] De TED-omgeving is de Training- en Demonstratie-omgeving. Deze is functioneel identiek aan de productieomgeving van het e-Depot, maar kan gebruikt worden voor test- en trainingsdata zonder dat de productieomgeving hiermee vervuild raakt. Voor de meeste tenants is de tenant-naam in Productie en TED identiek.

4.2 UUID van het WARC-bestand

De unieke identifier, of UUID van het te tonen WARC is tijdens het maken van de SIP gevormd. Zowel in de SIP-metadata als in het e-Depot zelf kan deze worden uitgelezen.

4.2.1 UUID ophalen uit de SIP

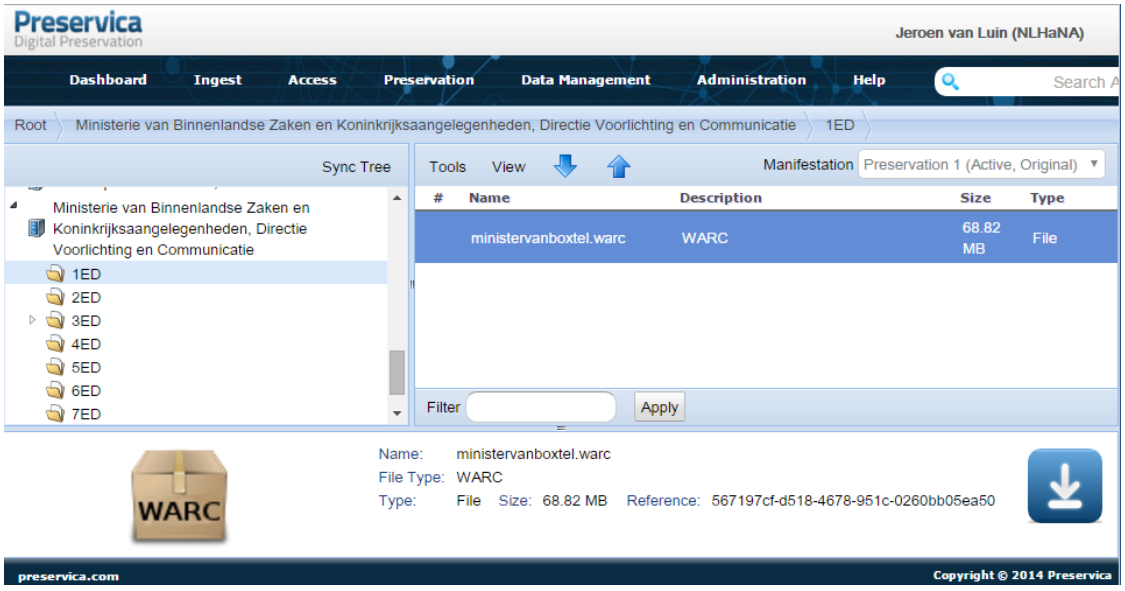
In de metadata.xml van de SIP die gebruikt is om de website op te nemen in het e-Depot, staan de collecties, deliverable units, manifestaties en bestanden onder elkaar beschreven. Voor het tonen van een website is alleen het stukje over bestanden nodig, beschreven tussen de XML-tags "`<Files>`" en "`</Files>`". Zoek in dat stukje het onderdeel dat gaat over het gewenste WARC-bestand. De UUID staat dan in de "`<FileRef>`".

```
<Files>
    <File status="new">
        <FileRef>567197cf-d518-4678-951c-0260bb05ea50</FileRef>
        <IngestedFileSetRef>13f53d3d-39c8-4720-911c-7e12ad6c2f96</IngestedFileSetRef>
        <FileName>ministervanboxtel.warc</FileName>
        <Extant>true</Extant>
        <Directory>false</Directory>
        <FileSize>72163207</FileSize>
        <LastModifiedDate>2016-01-12T17:58:38.721+01:00</LastModifiedDate>
        <FixityInfo>
            <FixityAlgorithmRef>2</FixityAlgorithmRef>
            <FixityValue>9fcc185e9185c922bdaa1b9473072319c45744e0</FixityValue>
        </FixityInfo>
        <WorkingPath>/</WorkingPath>
    </File>
</Files>
```

In dit voorbeeld wordt de UUID gezocht van de website van www.ministervanboxtel.nl. In de SIP staat bij die WARC de UUID `567197cf-d518-4678-951c-0260bb05ea50`.

4.2.2 UUID ophalen uit het e-Depot via de Explorer

In de Explorer van het e-Depot kan de WARC worden opgezocht. Na het selecteren van de WARC staat rechts onderin het beeld de gezochte UUID.



4.3 Samenstellen van de hele URL

Nu de basis-URL bepaald is, en bekend is wat de UUID van het te tonen WARC-bestand is, moeten deze twee gegevens achter elkaar worden gezet om te totale URL van de te tonen website te krijgen.

De basis-URL in het gebruikte voorbeeld is:

```
https://e-depot.nationaalarchief.nl/Render/render/external?
tenant=NA&entity=TypeFile&entityRef=
```

De UUID van de WARC uit het voorbeeld is:

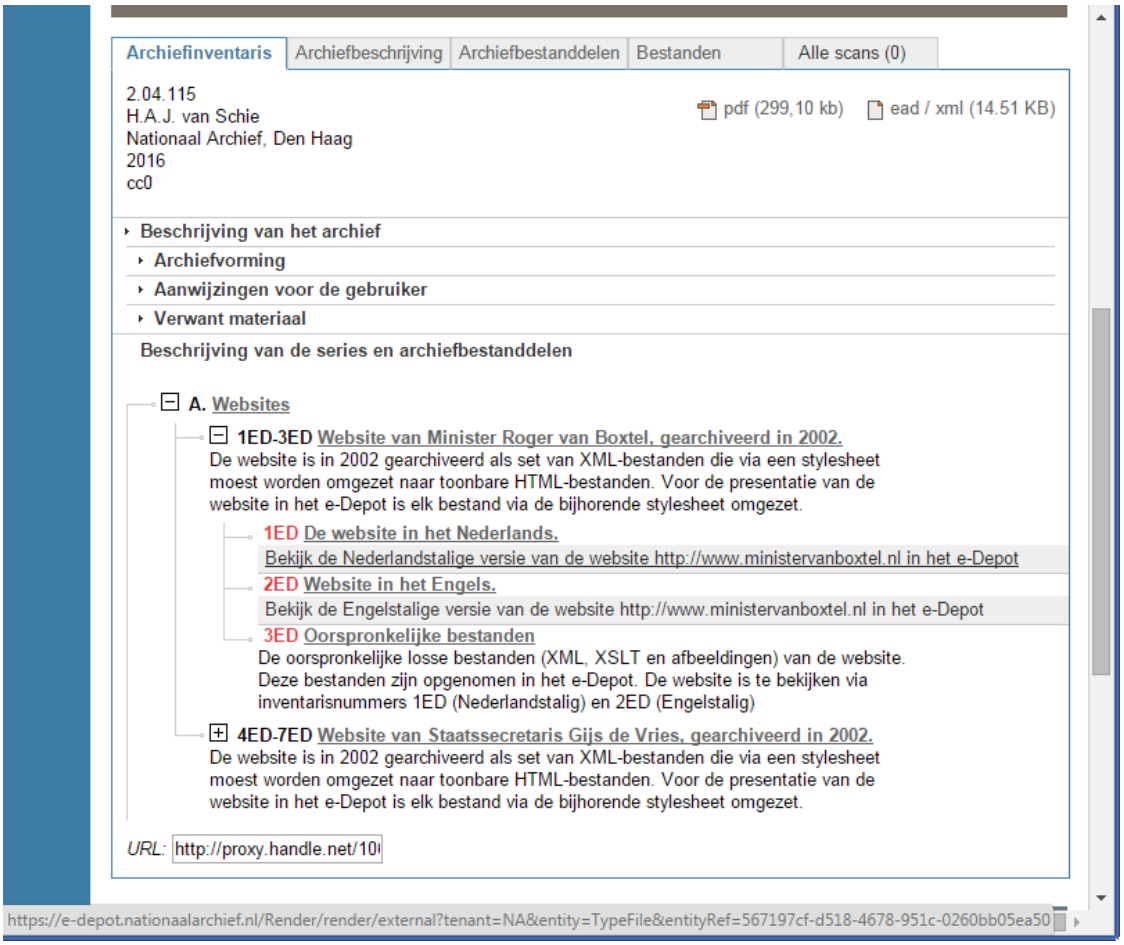
```
567197cf-d518-4678-951c-0260bb05ea50
```

Dus de totale URL van de te tonen website is:

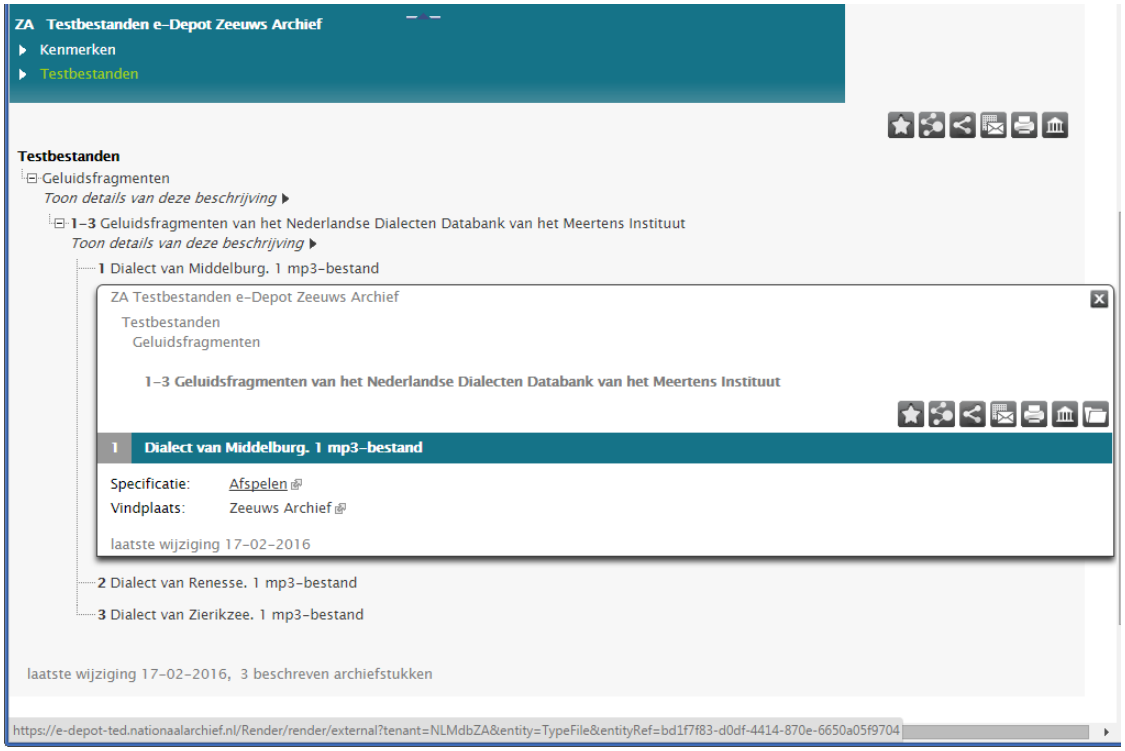
```
https://e-depot.nationaalarchief.nl/Render/render/external?
tenant=NA&entity=TypeFile&entityRef=567197cf-d518-4678-951c-
0260bb05ea50
```

4.4 Publiceren van de URL aan het publiek

Nu de hele URL voor het weergeven van de website bekend is, kan deze worden opgenomen in de bestaande publicatiekanalen, zoals het archievenoverzicht. Het Nationaal Archief gebruikt deze links in de archiefinventarissen op Gahetna.nl.



Het Zeeuws Archief heeft vanuit de TED-omgeving van hun e-Depot een aantal test-bestanden beschikbaar gesteld via Archieven.nl. Deze test-bestanden zijn weliswaar audiobestanden en geen websites, maar de achterliggende techniek van het opnemen van een URL naar een te tonen bestand is hetzelfde:



5 Ervaringen met het archiveren

De hiervoor beschreven technieken zijn een aantal keren uitgevoerd, met wisselend resultaten.

5.1 Casus 1: Rijksoverheid.nl

Eén van de eerste pogingen tot het opnemen van een website betrof de website van Rijksoverheid.nl. Hierbij is gebruik gemaakt van de ingebouwde versie van Heritrix.

5.1.1 Harvesten van de website

De harvest zelf verliep probleemloos, maar leidde helaas niet tot het gewenste resultaat: alle opmaak was verdwenen:

Dit probleem bleek te worden veroorzaakt door een specifiek stukje broncode in de website, bedoeld om de website te laten omgaan met hele oude versies van Internet Explorer:

```
<!--[if (gt IE 8)|!(IE)]><!-->
      <link rel="stylesheet" href="/presentation/responsive-
      2016.2.3.min.css" type="text/css" media="all"/>
<!--<![endif]-->
```

Dit deel zorgde ervoor dat Heritrix de aanwijzing voor het ophalen van de vormgevings-stylesheet als commentaar beschouwt, en dus overslaat. Latere pogingen met Wget leidden wel tot het gewenste resultaat, waarbij alle vormgeving was meegenomen.

5.1.2 Publicatie op GahetNA

Omdat deze website nog actief is, en geen overgedragen archiefstuk is, staat deze niet in de productieomgeving van het e-Depot, en is hij dus niet gepubliceerd op GahetNA.

5.2 Casus 2: Onderwijswebsite Nationaal Archief

Eind 2015 is de losse Onderwijswebsite van het Nationaal Archief opgenomen in de rest van de Gahetna.nl website. Voordat de losse website offline ging, is deze geharvest met de ingebouwde Heritrix-workflow.

5.2.1 Harvesten van de website

Doordat deze website niet de problematische code van Rijksoverheid.nl bevat, kwam deze in z'n geheel probleemloos in het e-Depot terecht:

Ook de Javascript-code waarmee de aankeiler-afbeelding wordt gewisseld, en waarmee de tekstvlakken onder de aankeiler omhoog komt wanneer de muis eroverheen beweegt, werken nog. De zoekmachine onderin het scherm doet het niet, dit is server-side functionaliteit die niet kon worden geharvest.

5.2.2 Publicatie op GahetNA

Dit voorbeeld is online te vinden via:
http://www.gahetna.nl/collectie/archief/ead/index/eadid/2.14.97

5.3 Casus 3: Website Minister van Boxtel

In hoofdstuk 3 is deze casus al een aantal keer als voorbeeld gebruikt: de website van Minister van Boxtel uit 2002, toen Minister van Grote Steden en Integratiebeleid.

De website werd aangeleverd op een CD, in de vorm van losse XML-bestanden met een aantal gekoppelde XSLT stylesheets die van de XML weer toonbare HTML konden maken. Om deze website te kunnen harvesten zijn twee voorbereidende stappen uitgevoerd: het omzetten van de XML naar HTML (inclusief intern aanpassen van de links) en het inrichten van een lokale webserver.

5.3.1 Omzetten van XML naar HTML

De bewerkingsslag voor het omzetten van XML naar HTML bevatte drie onderdelen:

1. Toepassen van de XSLT op het XML-bestand via de tool 'xsltproc'.
2. Interne links in de resulterende HTML-code aanpassen zodat ze niet eindigen op "`.xml`" maar op "`.xml.html`" via de tool 'sed'
3. Het gewijzigde HTML-bestand opslaan onder de oorspronkelijke bestandsnaam, aangevuld met "`.html`". Het oorspronkelijke bestand '`i000000.xml`' werd na de omzetting een HTML-bestand met de naam '`i000000.xml.html`'.

Alle XML-bestanden stonden in één map, en konden daardoor met één command-line opdracht in Cygwin worden omgezet:

```
for i in  `ls -1 *.xml`;
        do xsltproc.exe ${i} | sed 's/\.xml/\.xml.html/g' > ${i}.html;
done
```

5.3.2 Inrichten van de lokale webserver

Na de omzetting van XML naar HTML was de website een eenvoudige website met statische HTML-pagina's en gelinkte afbeeldingen geworden. Via de standaard in Python aanwezige SimpleHTTPServer kon deze in website-vorm worden getoond door in de command-line van Cygwin naar de hoofdmap van de website te gaan, en daar het commando uit te voeren:

```
python -m SimpleHTTPServer 80
```

Hiermee wordt lokaal een webserver geactiveerd die de opstartmap als startpunt gebruikt, en via de standaard website-poort 80 te benaderen is.

Door daarna in de hosts-file de volgende regel op te nemen:

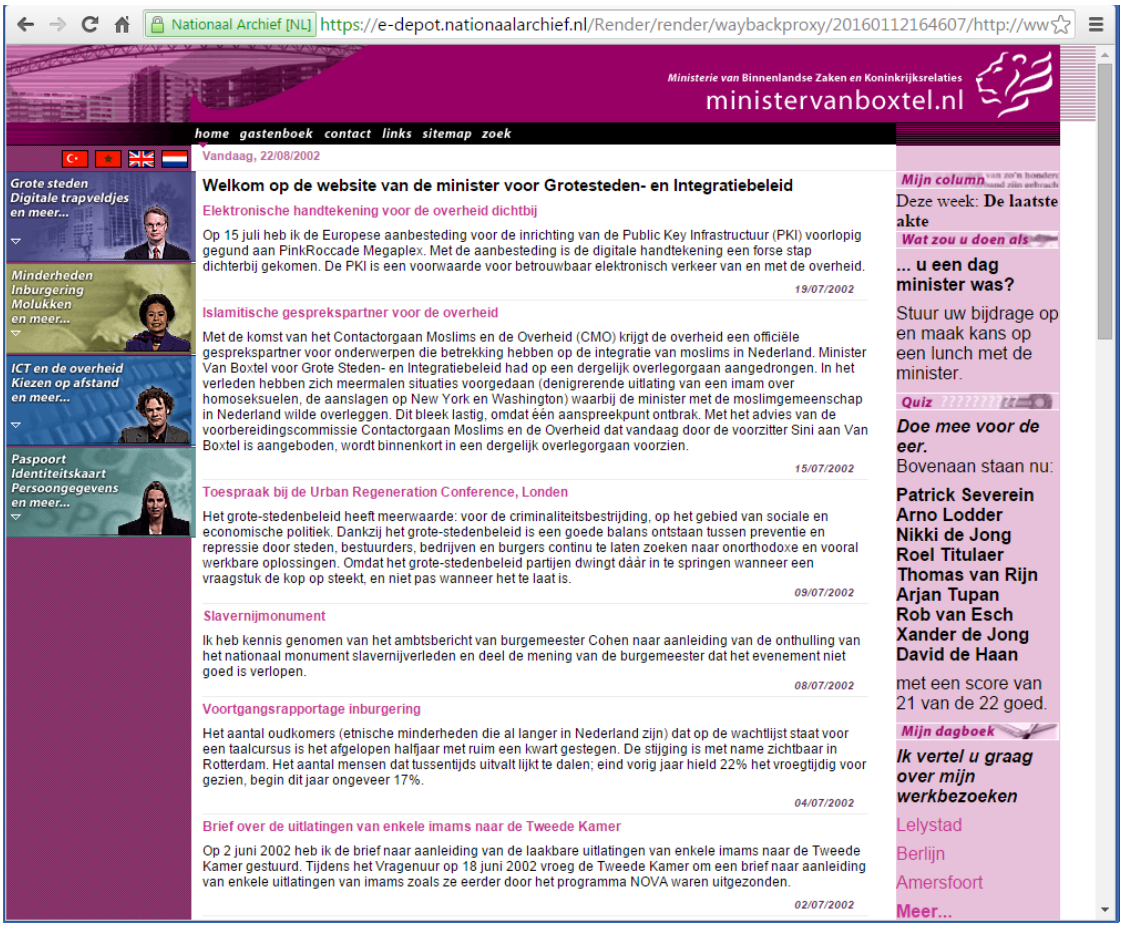
```
127.0.0.1    www.ministervanboxtel.nl
```

was het mogelijk om in de webbrowser naar http://www.ministervanboxtel.nl te gaan, en daarmee bij de oude website uit te komen.

5.3.3 Harvesten van de website

De website is vervolgens met Wget geharvest en daarna opgenomen in het e-Depot:

```
wget -m -k -p -E -w 1 --warc-file="ministervanboxtel"
http://www.ministervanboxtel.nl
```



5.3.4 Opname in het e-Depot

Via de SIPCreator is van deze WARC een SIP gemaakt. Na het maken van de SIP is in het metadata-bestand een stukje code toegevoegd om aan te geven wat de Seed URL van de website was, zoals beschreven in paragraaf 3.2.3.

Na opname van de WARC is een tweede SIP gemaakt waarin de oorspronkelijke, niet-omgezette losse bestanden zijn opgenomen. Mocht iemand de oorspronkelijke bronbestanden willen onderzoeken, of een betere techniek bedenken om de website weer tot leven te wekken, dan zijn de oorspronkelijke bestanden nog beschikbaar.

5.3.5 Publicatie op GahetNA

Dit voorbeeld is online te vinden via:
http://www.gahetna.nl/collectie/archief/ead/index/eadid/2.04.115

5.4 Casus 4: Jaarverslag 2000 Rijksdienst Wegverkeer

De Rijksdienst Wegverkeer (RDW) brengt sinds 2000 haar publieke jaarverslag uit in de vorm van een website. Als test voorafgaan aan de uiteindelijke overbrenging van dit jaarverslag zijn twee verschillende versies aangeleverd: de relatief simpele versie uit 2000 en de veel complexere versie uit 2014. In deze paragraaf worden de ervaringen met het Jaarverslag 2000 beschreven, in de volgende die over het Jaarverslag 2014.

5.4.1 Inrichten van de lokale webserver

De bestanden werden aangeleverd als losse HTML-bestanden met bijhorende stylesheets en afbeeldingen. Er was dus geen omzetting nodig, de website kon direct via een eenvoudige webserver worden getoond:

```
python -m SimpleHTTPServer 80
```

Bij het bekijken via een webbrowser bleek echter wel een probleem:

De afbeelding van het witte vlak met paarse balk die gebruikt is als achtergrond-afbeelding was 1600 pixels breed, niet meer breed genoeg voor de schermen van nu. Doordat de afbeelding nu aan de rechterkant van het scherm herhaald wordt, is de bovenliggende tekst niet meer leesbaar. De oplossing was een kleine aanpassing aan de stylesheet van de website, waarbij werd opgenomen dat de achtergrond-afbeelding alleen verticaal herhaald mag worden: `”background-repeat: repeat-y;”`

Na de aanpassing bleek de website probleemloos te kunnen worden geharvest.

5.4.2 Opname in het e-Depot

De website is na harvesting opgenomen in de TED-omgeving van het e-Depot en was daar zonder problemen te gebruiken. Wanneer deze website ooit in de productieomgeving van het e-Depot wordt opgenomen zal ook een SIP met daarin de ongewijzigde losse bestanden worden opgenomen.

5.4.3 Publicatie op GahetNA

Omdat deze website nog geen overgedragen archiefstuk is, staat deze niet in de productieomgeving van het e-Depot, en is hij dus niet gepubliceerd op GahetNA.

5.5 Casus 5: Jaarverslag 2014 Rijksdienst Wegverkeer

De 2014-versie van het jaarverslag is een nog actieve website, en dus direct te harvesten. Voor deze harvest is gebruik gemaakt van de ingebouwde Heritrix van het e-Depot.

Een complicatie bij deze website is dat er een filmpje van YouTube op de website staat. De harvester neem alleen gekoppelde bestanden op wanneer ze op dezelfde webserver staan als de hele website, en de Wayback-machine zal alle links vertalen naar gearchiveerde versies van de link. Bovendien is er geen garantie dat YouTube eeuwig blijft bestaan, en zou een externe koppeling naar YouTube dus niet duurzaam zijn.

Het gevolg is dat er op de gearchiveerde versie van de website alleen een alt-tekst staat:

In plaats van hoe het hoort te zijn, met filmpje:

Dit probleem is vooralsnog niet op te lossen met techniek: geen van de harvesters zal in staat zijn om hier een duurzame versie van te harvesten en op te nemen.

De tot dusver bedachte alternatieven zijn allen procedureel:

1. Harvesten zonder filmpje
2. Harvesten zonder filmpje, en in de archiefinventaris uitleggen dat er een YouTube-filmpje stond
3. Harvesten zonder filmpje, en in de archiefinventaris uitleggen dat er een YouTube-filmpje stond, en daarbij de inhoud van het filmpje beschrijven
4. Harvesten zonder filmpje, en het bronbestand van het filmpje als apart archiefstuk opnemen in het e-Depot
5. Harvesten zonder filmpje, het bronbestand van het filmpje als apart archiefstuk opnemen in het e-Depot, in de archiefinventaris uitleggen dat er een filmpje stond, en verwijzen naar het apart in het e-Depot opgenomen filmpje

De vijfde optie is daarbij het meeste werk, maar biedt wel de netste oplossing.

Ook voor deze versie van het jaarverslag geldt dat het nog geen overgedragen archiefstuk is, daardoor niet in de productieomgeving van het e-Depot staat, en dus ook niet gepubliceerd is op GahetNA.

5.6 Casus 6: Oude Nationaal Archief website

Bij de live-gang van de GahetNA-website in 2011 is de oude Nationaal Archief-website offline gegaan. De webserver waar de oude website op stond is gevirtualiseerd en op het interne netwerk van het Nationaal Archief nog beschikbaar.

5.6.1 Harvesten van de website

De eerste stap in het harvest-proces was het aanpassen van de lokale hosts-file, zodat voor het webadres 'www.nationaalarchief.nl' niet de huidige online versie, maar de interne oude versie wordt benaderd. Via de webbrowser was daarna de oude website te raadplegen:

Daarna kon via Wget de website worden geharvest:

```
wget -m -k -p -w 1.0 -E --warc-file="oude-nationaalarchief"
http://www.nationaalarchief.nl
```

Bij het harvesten worden alle dynamisch gegenereerde pagina's als individuele bestanden weggeschreven. Waar in de bronbestanden één ASP-pagina bestaat, die zijn gegevens uit een database met 100.000 records haalt, bestaat het harvest-resultaat uit 100.000 HTML-pagina's, voor elk record een resultaatpagina.

Alhoewel dit technisch op zich geen probleem is, zorgt het er wel voor dat het archievenoverzicht uit 235.000 pagina's bestaat, en het plaatsenoverzicht uit 155.000 pagina's. Zowel de karakterisatie- als de thumbnail creation-stappen uit de ingest-workflow worden ook op bestanden binnenin containers als ZIP en WARC-bestanden uitgevoerd, en kennen dus een aanzienlijke doorlooptijd.