Nationaal Archief
Ministerie van Onderwijs, Cultuur en Wetenschap

Experiences with web archiving in the Dutch National Archives

Jeroen van Luin, May 25th, 2016

1 Introduction

In the past few months, several attempts have been made to archive several types of websites, with varying degrees of success. This report will describe the approach used to archive the websites, show the results of the attempts, and addresses some of the issues encountered.

In all attempts, the technique of *website harvesting* is used: a computer program that takes a start page of a website as input, and continues to download that page, and all the pages linked. The resulting website archive is then ingested into either the production environment or the training environment of the *e-Depot*, the digital repository used by the Dutch National Archives.

1.1 Outline

Chapter 2 describes one of the primary questions in website archiving that needs to be addressed: the goal of website archiving, given the possibilities and limitations of the current website archiving techniques. Depending on the goal, website archiving may or may not be the best answer.

Chapter 3 explains the different techniques used in the website harvesting:

1. Direct harvesting of a live website, using the "Website ingest"-workflow in the e-Depot. This workflow uses a build-in version of the Heritrix webcrawler.
2. Harvesting of a live website using a Heritrix installation outside of the e-Depot.
3. Harvesting a live website using the Linux Wget tool.
4. Harvesting an offline website whose source files have been submitted to the Dutch National Archives on a portable storage device, or internet transfer.

In chapter 4 we'll go into some of the procedure to make the archived website accessible to the general public. Chapter 5 will end with a description of six cases of website archiving and the issues we encountered.

Both chapter 3 and 4 are quite technical in nature, explaining the configuration options and command-line parameters used in the harvests. When the technical details of web archiving aren't part of your field of interest, these two chapters can safely be skipped.

1.2 Terminology

This document assumes that the reader is familiar with the terminology of digital archiving and the functional model of the Open Archival Information System-standard (OAIS). Terms that are used, include SIP (Submission Information Package), DIP (Dissemination Information Package), AIP (Archival Information Package), Ingest (the process to add new content to a digital repository), web harvest and web crawling (recursively downloading webpages by following hyperlinks), UUID (a Universally Unique Identifier) and Tenant (a logically divided part of a digital repository containing the digital collection of one institute).

2 Determining the goal of web archiving

While using the web archiving techniques, it became clear that one of the most important questions in web archiving has to answered before starting the harvesting itself. This question is not about which technique to use, but is about what the goal is of the archival procedure.

2.1 What web archiving can do

A harvester will download and save the information that is reachable by following a path of links from the starting page. A harvester will see all the content that is present on the website while it is harvesting, provided this content is hosts on the website-domain it was told to harvest. Any content added after the harvest session will not be in the archive, and any content deleted from the website will not be deleted from the archived version.

If a website harvest session takes a long time (it is very possible that big websites takes several days or weeks to be fully harvested) the website itself can change during the harvest. This can result in a web archive that is in part the version before the change, and in part the version after the change.

A technique that should be possible, but has not yet been experimented with at the Dutch National Archives, is incremental harvests. With incremental harvests, a website is harvested multiple times over a length of time, and each time only the changes between the current harvest and the previous harvest are stored. This could give the user of the website archive an idea of the changes during the archived time frame. However, content that was added and removed again between two consecutive harvests is still missed so the frequency of harvests should be high enough.

2.2 What web archiving can't do

A website harvester will only follow links present in the source-code of the webpages that it downloads. Any information that cannot be found by clicking on links in the code will not be present in the web archive. Most notably: search forms and login forms, both server-side on-demand functionality, will not work in an archived version of a website. Likewise, links that are generated in a browser using JavaScript will not be followed, as the harvester that we used will not interpret or execute the JavaScript code.

When a webpage contains content that is hosted outside of the webserver itself, or contain links that point to other websites, most harvesters will automatically skip those links and that content. Webpages containing YouTube-clips or links to Wikipedia will not cause the harvester to harvest the entirety of YouTube or Wikipedia. When examining a website for suitability for web archiving, these cases should be inspected carefully.

As mentioned above, the on-demand interactive functionality of a website is mostly absent in the archived version. This reduces the possible uses of the archived version of the website. When a website is mostly (or even worse, exclusively) used by filling in forms and using search result pages and filters to find information, users will struggle to find the same information in the archived version. In most cases, this will make the archived version of the website unusable as replacement for the real website. When access to the information or documents on the website is necessary, this information should be part of the regular digital archive, accessible through the regular finding aid.

2.3 When to use web archiving

Web archiving has several very good uses and limitations. From what we concluded so far in our experiments, web archiving is good for:

1. Showing the public what a website looked like, to give them a general idea of what information could be found on the website, and what types of interaction were possible on that website
2. Showing a guaranteed fixed, immutable version of the website, as additions and deletions to the live website will not influence the archived version.

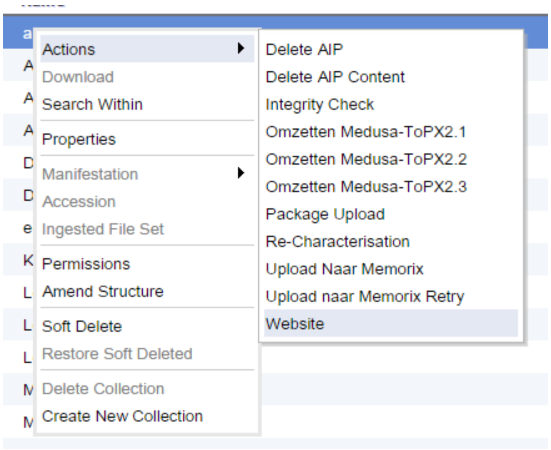
Web archiving is less usable for:

1. Being the only replacement for a highly interactive website where server-side functionality is used to find specific information.
2. Websites that have very frequent additions and removals of content
3. Websites that depend on a lot of content hosted on other websites

3 Harvesting techniques

3.1 Harvesting with Heritrix using the Website ingest workflow

The by far easiest way to harvest a website is using the built-in version of the webcrawler Heritrix that is part of our e-Depot. In the e-Depot Explorer, we can select the Collection that the new archived website should be part of, and start the Website ingest workflow.



3.1.1 Configuring the Website ingest workflow

After starting the workflow, we need to fill in some configuration fields:

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

The open source tool Heritrix itself has many more configuration options, but these are filled in with default values by the e-Depot and don’t need to be supplied by the user.

3.1.1.1 Seed URL

To start the harvest, the harvester needs to know what webpage to start on. The Seed URL will be the first URL to visit and download. The downloaded webpage source file will then be examined for links to content, stylesheets and other webpages. Links to pages or content on different websites than the Seed URL will be ignored.

One special case is Seed URLs that are redirected to a different URL, for example, unsecure webpages (using HTTP) that are redirected to secure versions (using HTTPS). The redirect will cause the harvester to only download the secure version and move on from there. However, the configuration still has the unsecure version als Seed URL. When the e-Depot later tries to render the website, it will find that the configured start page (the unsecure one) is missing from the archive, and will fail to render the website.

3.1.1.2 Title and Scope & content

After harvesting, the website will be added to the collection in which the workflow was started, as a new record. In these configuration options, you can give the values that should be used as Title and Scope/content fields in the metadata of the record.

3.1.1.3 Catalogue reference

The record that is created when the archived website is ingested will get the value of this configuration option as catalogue reference.

3.1.1.4 Security Tag

Most websites that are harvested will be open to the public. In case an internal, restricted website is harvested, the corresponding security tag can be filled in.

3.1.1.5 Max Hops en Max Path Depth

The number of 'Hops' is the smallest number of links that have to be followed to get to a page from the Seed URL. A page that is linked from the start page will have a hop count of 1, a page that is linked from a page with hop count 1 will have hop count 2, etc. When a page is linked from several other pages, the smallest hop count is taken. By filling in a max hop count value, you can limit set of pages that are harvested. A max hop value of 0 indicates that the hop count is not used.

The 'Path Depth' is the number of URL parts (separated by forward-slashes '/') in the website address. For example, with Seed URL 'http://en.nationaalarchief.nl' and a Max Path Depth of 1, pages and files in

```
       http://en.nationaalarchief.nl/news
```

will be harvested, but pages and files in

```
       http://en.nationaalarchief.nl/news/press
```

will not, as they have a path depth of 2.

A Max Path Depth of 0 indicates that the path depth is not used.

3.1.1.6 Max Download Size en Max Download Time

The options 'Max Download Size' and 'Max Download Time' don't influence which pages are harvested, but control how much is harvested and how long a harvest can take. Using a positive value for Max Download Size will tell Heritrix to stop harvesting after downloading that amount of megabytes of web content. A positive Max Download Time will tell Heritrix to stop harvesting after that amount of minutes has passed since the start of the harvest.

In both cases, a value of 0 will mean the criterion is not used to determine when to stop harvesting. And obviously, if a harvest is completed before any of the limits are reached, the harvest will finish normally.

3.1.1.7 Exclude extensions

With the 'Exclude extensions' option, Heritrix can be told to ignore one or more types of files linked on webpages. By using, for example, the value 'zip,exe', all zip-containers and executables will be excluded from the webarchive. An empty value for this configuration option will allow Heritrix to download all types of files.

3.1.1.8 Honour robots.txt

Websites can use a special file 'robots.txt' to tell robots what parts of the website should not be processed by crawlers and search indexers. Whether or not a webcrawler or search indexer actually honours this request is up to the user of the crawler or indexer. With this option, Heritrix can be configured to honour, or ignore the robots.txt directions.

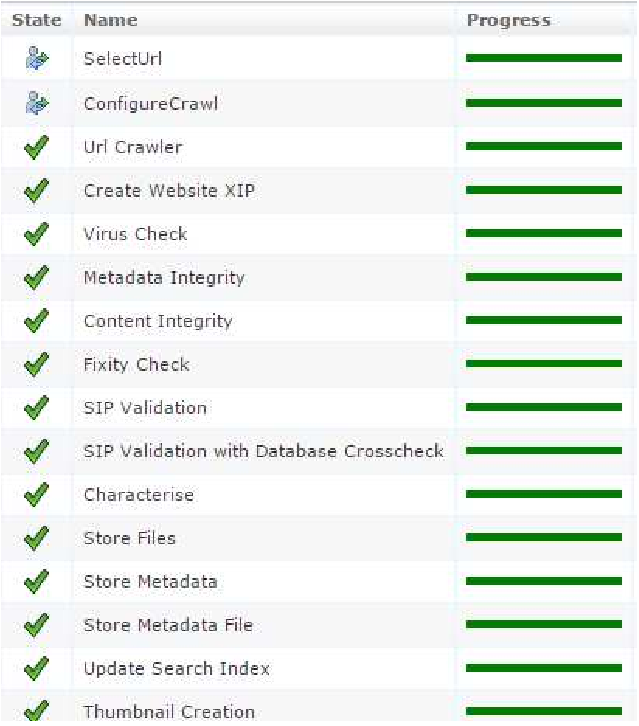
3.1.1.9 *Output type*

Heritrix is capable of using three different formats for output: the standard WARC-format, the obsolete ARC-format and a mirror of the file-structure of the website. Of these formats, only WARC and ARC can be rendered in the WayBack-machine that is part of the e-Depot.

3.1.2 Running the Website ingest workflow

After configuring the website ingest workflow, the Workflow will start Heritrix to do the actual URL crawling. One of the default configurations that cannot be changed by the user, is the 'aggressiveness' of the harvest. Being a robot, Heritrix would be capable of requesting the pages very fast, putting a strain on the webserver hosting the website. To avoid this, the configuration of the e-Depot forces Heritrix to be slow in harvesting.

When the harvest is done, the resulting WARC file is turned into a XIP[1]-package, and then follows the normal ingest workflow steps:



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	▬

During any ingest, several checks are performed to make sure that the archival package is in a good shape, free of viruses, and with consistent metadata. As the e-Depot created the archival package itself, problems in this phase are rare, but it's better to be safe. After the checks and validations, the package is characterized, and then stored in the repository. Finally the search index is updated and thumbnails are created.

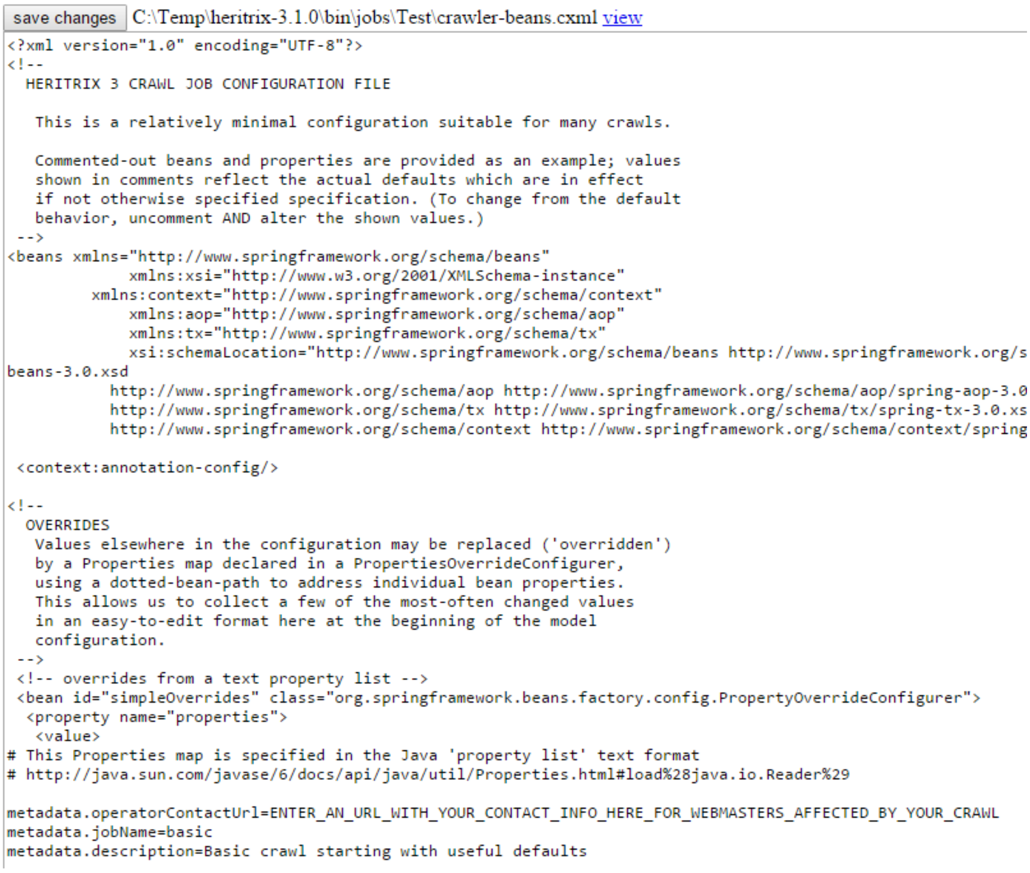
[1] XIP is the internal metadata-format used by the Preservica-software that is the base of the Dutch e-Depot system. The XIP format is used for all three information package types in the functional model of the OAIS standard: SIP, AIP and DIP.

3.2 Harvesting with an external Heritrix instance

As mentioned in the previous section, the harvest tool Heritrix can be configured with a huge amount of configuration options. The Website ingest workflow built into the e-Depot allows the configuration of only 8 of them (Seed URL, max hops, max path depth, max download size, max download time, excluded extensions, honouring robots.txt and output type). When one or more of the other configuration options needs to be changed, or when we want to harvest a website in a location that is inaccessible for the e-Depot, we can use a Heritrix instance outside of the e-Depot.

3.2.1 Configuring Heritrix

A stand-alone version of Heritrix can have several ‘jobs’ that each have their own configuration file:

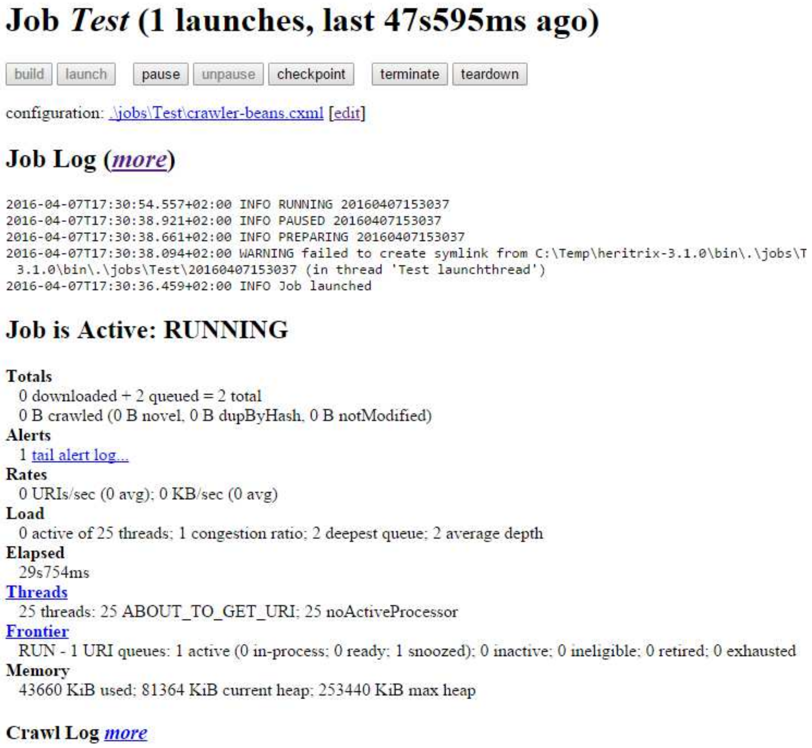

save changes
C:\Temp\heritrix-3.1.0\bin\jobs\Test\crawler-beans.cxml view

```
<?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
```

Choosing the right configuration options and the right values for each configuration option does require an extensive study of the Heritrix manual. After completing the configuration, the harvest job can be started.

3.2.2 Running the harvest

The actual harvest job is started by using the buttons ‘build’, ‘launch’ and ‘unpause’ in that order. During the harvest, the Heritrix job dashboard can be used to monitor the progress of the harvest:



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

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*

3.2.3 Creating the SIP

When the harvest is finished, the resulting WARC-file can be downloaded and prepared for ingest into the e-Depot using the SIP Creator. After creating the SIP, one piece of metadata has to be manually added: the Seed URL. This is needed by the render tool of the e-Depot to know which start page should be shown when showing the archived website to the user.

To add this metadata, we need to edit the XIP metadata file, and add an ‘AccessionEvent’ fragment:

```
<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>
```

This fragment tells the e-Depot that, when rendering this package, it should use the URL ‘http://www.ministervanboxtel.nl’ as it was on April 7th, 2016 at 15:49:33. When multiple versions of a website have been ingested, the e-Depot will now know which version it should show.

3.2.4 Ingesting the SIP

When the SIP is created and the metadata for the AccessionEvent has been added, the archival package can be ingested using a standard ingest workflow.

3.3 Harvesting a live website using Wget

A standard part of almost every Linux installation is the command-line tool Wget. Recent versions of this tool allow the recursive download of a website and creation of WARC-files.

3.3.1 Configuring Wget

Being a command line tool, Wget needs to be supplied with the configuration options as parameters during the execution of the tool. The parameters that we use for archiving a website are:

`-m / --mirror`	Recursively download an entire website, regardless of number of hops or link depth.Maak een afspiegeling van de hele website (recursief downloaden van alle pagina's, ongeacht het aantal hops of link depth,
`-k / --convert-links`	Convert absolute links to relative links, to allow local viewing of the website content
`-p / --page-requisites`	Download all the files that are necessary to completely render a page, including all style sheets, images, audio files, etc.
`-E / --adjust-extension`	Add the extension ".html" to all HTML-files, and ".css" to all stylesheet files.
`-w 1`	Wait 1 second between page downloads
`--warc-file="<name>"`	Use <name> as filename for the WARC
<URL>	Use <URL> as start page

3.3.2 Running the harvest

The harvest is executed using Cygwin (Linux for Windows) by runnig the following command in the Cygwin terminal to harvest the website for Minister Van Boxtel:

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

During the harvest, the terminal is flooded with all the actions performed by Wget and the pages and files that are downloaded. At the end of the harvest, there will be a file called 'ministervanboxtel.warc' containing the archived website.

3.3.3 Creating the SIP

The next step in the process is turning the WARC-file into a SIP and adding the Seed URL metadata as AccessionEvent, as described in section 3.2.3.

3.3.4 Ingesting the SIP

After the SIP has been created and the AccessionEvent metadata has been added, the archival package can be ingested using a standard ingest workflow.

3.4 Harvesting a website using source code

Websites that are off-line, but whose source code is still available, cannot be accessed, and thus cannot be harvested. In this case, we need to ‘revive’ the old website using the original code. As the original URL is often reused for a newer version of the website, we need to apply some special techniques to make sure the website can be harvested.

3.4.1 Installing a local webserver

A harvester (Heritrix or Wget) will contact a webserver to download the webpages and other files that make up the website. When a website is no longer available online, we can install a local webserver on our workstation to simulate the original webserver. Several types of local webservers can be used, depending on the type of workstation used, and on the technology that is used in the offline website.

Simple, static websites using HTML-files, stylesheets and images can easily be served using any of the following webservers:

- Pythons SimpleHTTPServer (part of any standard Python distribution)
- Apaches XAMPP (available at the ApacheFriends website)
- Windows IIS (standard in Windows 7, 8 and 10, but generally not switched on)

More complex websites can use server-side scripting (for example, PHP or ASP) or use a database. In this case, the simple HTTP server will not be sufficient, and XAMP or IIS needs to be used. In general, XAMPP is easier to use for websites that used PHP and MySQL, and IIS is easier to use for websites that used ASP and MS-SQL.

At the start of the 2000’s, it became very popular to create websites using XML, that needed XSLT-stylesheets to be transformed into HTML. For these websites, it can be necessary to first run the transformation (generating static HTML-files) and then treat the resulting files as a ‘simple’ website.

3.4.2 Editing the local hosts-file

When a website is offline, the original URL has been reused, or it is undesirable to have an old website back online using its original URL. In these cases we need a way to trick the harvester into contacting the local webserver every time the original URL is requested. We can accomplish this by editing the local ‘hosts’ file of the workstation running the harvest.

In a hosts file, a list of IP-addresses and hostnames can be entered. When contacting a host, the operating system will always first check the hosts file to find the IP address of the webserver. By adding a line that maps the website hostname to the IP address for the local host (`127.0.0.1`) or the IP address of the machine running the local webserver, the harvester will always contact that IP address when requesting a URL with that hostname.

Editing the hosts file requires administrative privileges under both Windows and Linux.

For Windows, this file can be found in:

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

In Linux, the file is located in:

```
	/etc/hosts
```

Using a text editor, we can add the line

```
      <IP address> <web address>
```

For example:

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

3.4.3 Running the harvest

When the website has been made available on a local webserver and the harvester has been tricked into contacting that webserver any time it requests the old website, we can used any of the previously described harvesting techniques. At the end of the harvest, we'll have a WARC file containing the archived website.

3.4.4 Creating the SIP

The next step in the process is turning the WARC-file into a SIP and adding the Seed URL metadata as AccessionEvent, as described in section 3.2.3.

3.4.5 Ingesting the SIP

After the SIP has been created and the AccessionEvent metadata has been added, the archival package can be ingested using a standard ingest workflow.

4 Publishing an archived website

When a website has been archived and ingested into the e-Depot, it can be made available online for the public to use. The e-Depot has a built-in version of the Wayback Machine to render archived websites.

In using the Wayback Machine, we need three pieces of information:

1. The base URL of the Wayback Machine for a tenant of the e-Depot.
2. The UUID of the WARC-file in the e-Depot
3. An online publishing channel to announce the URL to the public.

4.1 Tenant specific base URL of the Wayback-machine

The Dutch e-Depot is a multi-tenant installation. Every tenant has its own logically separated part in the e-Depot, designated by the tenant name. This tenant name is part of the base URL used in rendering digital object stored in the e-Depot. For the national and provincial archives that are a tenant in the e-Depot, the tenant names are based on the ISIL[2] code of the institute:

Gelders Archief	NLAhGldA
Groninger Archieven	NLGrGRA
Het Utrechts Archief	NLUtHUA
Historisch Centrum Overijssel	NLZlHCO
Nationaal Archief	NA[3](Production) or NLHaNA (TED[4])
Nieuw Land Erfgoedcentrum	NLLlsNLE
Noord-Hollands Archief	NLHlmNHA
RHC Limburg	NLMtRHCL
Tresoar	NL040041000
Zeeuws Archief	NLMdbZA

The tenant name can be entered into the URL:

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

or in case of the TED system:

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

For the Production tenant of the Dutch National Archives, the base URL is:

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

In the final version of the URL, the UUID of the rendered object is appended to this base URL

[2] International Standard Identifier for Libraries and related organizations

[3] Standardizing tenant names to the ISIL code was introduced after the tenant of the National Archives was created on the Production system. A standard database maintenance script to rename tenants is being developed by the supplier of the e-Depot software.

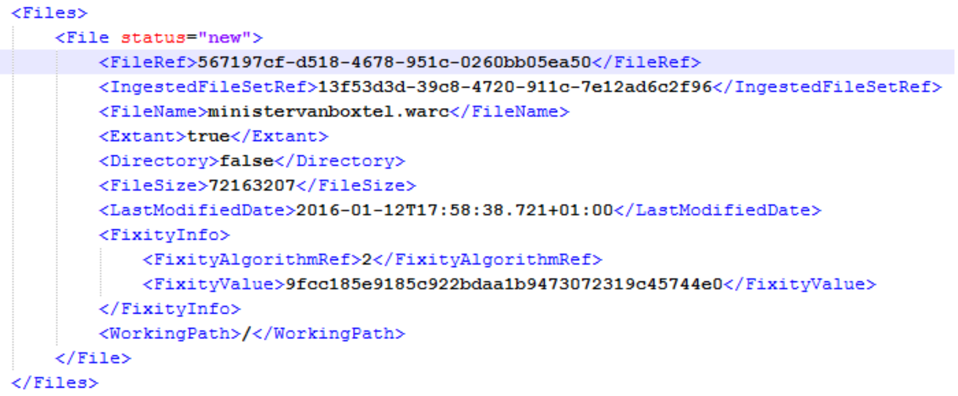
[4] The Training and Demonstration system of the e-Depot. Functionally, this system is identical to the Production system, but can be used with training data, and can be used for demonstration purposes.

4.2 UUID of the WARC file

The Unique Identifier, or UUID, of the WARC file is generated when the SIP is created, and stored into the e-Depot database during ingest. Both the SIP and the database can be used to find this UUID.

4.2.1 Finding the UUID in the SIP

In the same metadata.xml file of the SIP that we edited to add the Seed URL (see section 3.2.3) we can find the UUID of the WARC file. This metadata file contains the information for all collections, deliverable units, manifestations and files in an XML-structure. To find the UUID, we turn to the `<Files>` section of the metadata. In the XML for the WARC file, we'll find the `<FileName>` of the WARC, and just above it the `<FileRef>`. The value of this `<FileRef>` is the UUID that the WARC will get during ingest.

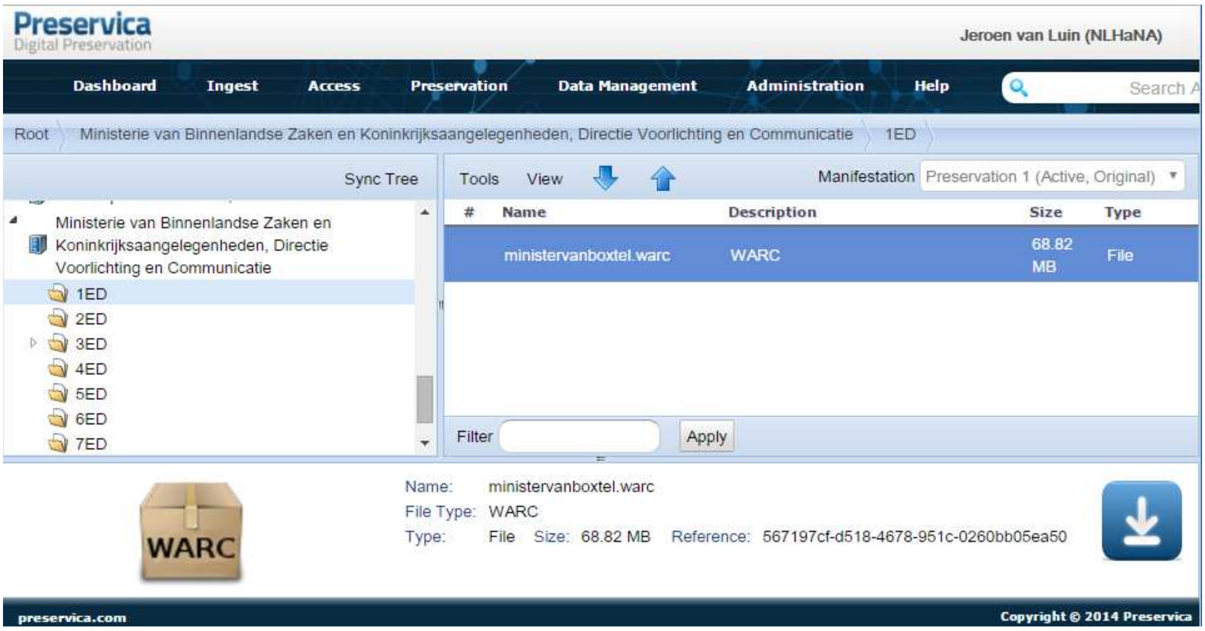


```
<Files>
    <File status="new">
        <FileRef>567197cf-d518-4678-951c-0260bb05ea50</FileRef>
        <IngestedFileSetRef>13f53d3d-39c8-4720-911c-7e12ad6c2f96</IngestedFileSetRef>
        <FileName>ministervanboxtel.warc</FileName>
        <Extant>true</Extant>
        <Directory>false</Directory>
        <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 the example above, the UUID of the WARC that contains the archived version of the website www.ministervanboxtel.nl is `567197cf-d518-4678-951c-0260bb05ea50`.

4.2.2 Finding the UUID in the e-Depot Explorer

In de Explorer of the e-Depot, we can browse to the location of the WARC. After selecting the file, the UUID will be displayed in the right part of the bottom section.



Preservica
Digital Preservation
Jeroen van Luin (NLHaNA)
Dashboard
Ingest
Access
Preservation
Data Management
Administration
Help
Search A
Root
Ministerie van Binnenlandse Zaken en Koninkrijksaangelegenheden, Directie Voorlichting en Communicatie
1ED
Sync Tree
Tools
View
Manifestation
Preservation 1 (Active, Original)
Ministerie van Binnenlandse Zaken en Koninkrijksaangelegenheden, Directie Voorlichting en Communicatie
1ED
2ED
3ED
4ED
5ED
6ED
7ED
Filter
Apply
WARC
preservica.com
Copyright © 2014 Preservica

#	Name	Description	Size	Type
	ministervanboxtel.warc	WARC	68.82 MB	File

Name: ministervanboxtel.warc
File Type: WARC
Type: File Size: 68.82 MB Reference: 567197cf-d518-4678-951c-0260bb05ea50

4.3 Constructing the complete URL

When the base URL is generated, and the UUID of the WARC file is found, we can construct the complete URL by adding the UUID to the end of the base URL.

If the base URL is:

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

and the UUID of the WARC file is:

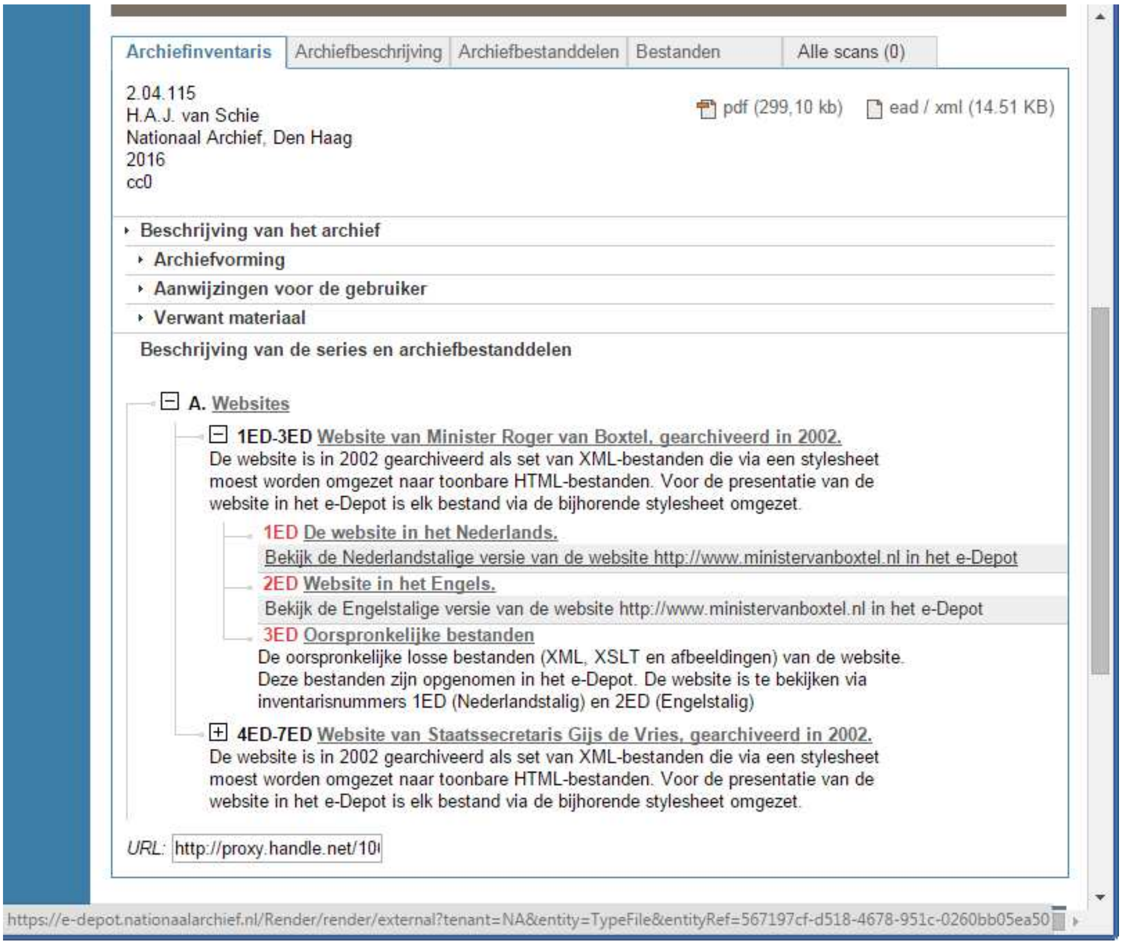
```
567197cf-d518-4678-951c-0260bb05ea50
```

then the complete URL for rendering the website is:

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

4.4 Publishing the URL to the public

After the complete URL has been constructed, it can be added to the existing publishing methods, like website content, finding aids and on social media. The Dutch National Archives uses their website Gahetna.nl to display the finding aids. The finding aid for the archive of the Communications Department of the Ministry that Minister Van Boxtel was head of, contains the link to the archived version of the website (in Dutch):



2.04.115
H.A.J. van Schie
Nationaal Archief, Den Haag
2016
cc0

- Beschrijving van het archief
  - Archiefvorming
  - Aanwijzingen voor de gebruiker
  - Verwant materiaal

Beschrijving van de series en archiefbestanddelen

⊟ A. Websites

⊟ 1ED-3ED Website van Minister Roger van Boxtel, gearchiveerd in 2002.
De website is in 2002 gearchiveerd als set van XML-bestanden die via een stylesheet moest worden omgezet naar toonbare HTML-bestanden. Voor de presentatie van de website in het e-Depot is elk bestand via de bijhorende stylesheet omgezet.

1ED De website in het Nederlands.
Bekijk de Nederlandstalige versie van de website http://www.ministervanboxtel.nl in het e-Depot

2ED Website in het Engels.
Bekijk de Engelstalige versie van de website http://www.ministervanboxtel.nl in het e-Depot

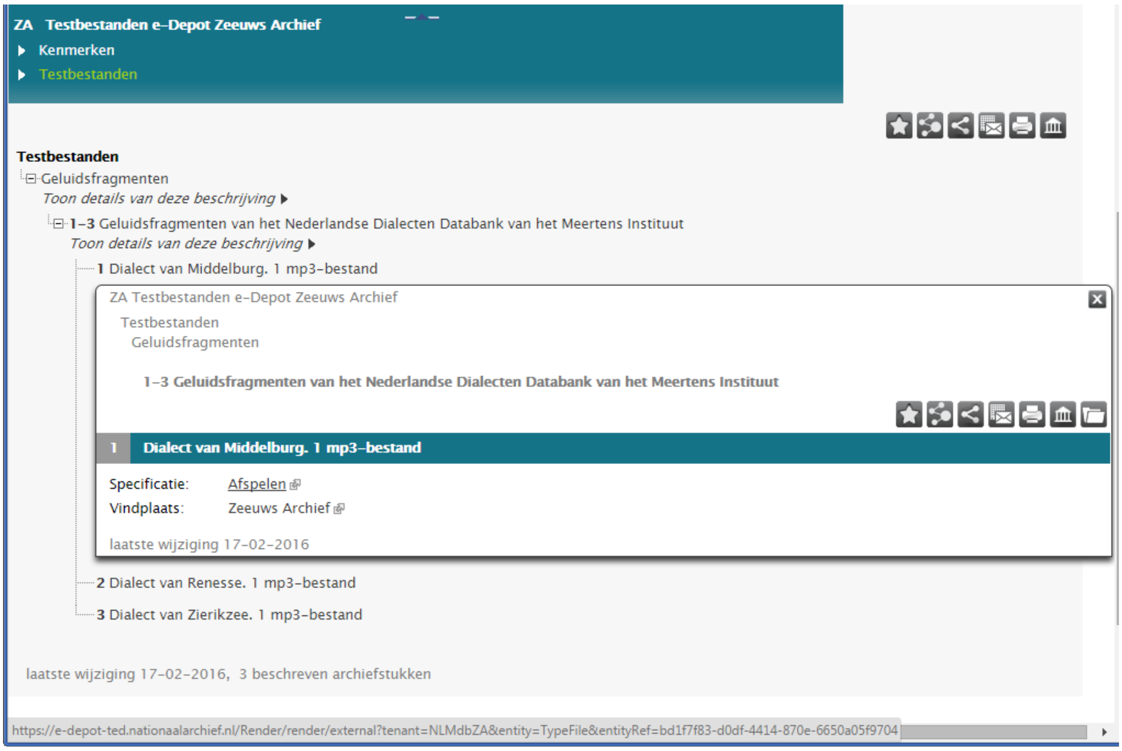
3ED Oorspronkelijke bestanden
De oorspronkelijke losse bestanden (XML, XSLT en afbeeldingen) van de website. Deze bestanden zijn opgenomen in het e-Depot. De website is te bekijken via inventarisnummers 1ED (Nederlandstalig) en 2ED (Engelstalig)

⊞ 4ED-7ED Website van Staatssecretaris Gijs de Vries, gearchiveerd in 2002.
De website is in 2002 gearchiveerd als set van XML-bestanden die via een stylesheet moest worden omgezet naar toonbare HTML-bestanden. Voor de presentatie van de website in het e-Depot is elk bestand via de bijhorende stylesheet omgezet.

URL: http://proxy.handle.net/10

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

The archive of the province Zeeland, the Zeeuws Archief, demonstrated publishing objects to the public using the central website for provincial archives 'Archieven.nl'. Though these files are audio files and not websites, the technique of generating the link and publishing it remains the same.



ZA Testbestanden e-Depot Zeeuws Archief

- Kenmerken
- Testbestanden

Testbestanden

⊟ Geluidsfragmenten

Toon details van deze beschrijving ▶

⊟ **1-3** Geluidsfragmenten van het Nederlandse Dialecten Databank van het Meertens Instituut

Toon details van deze beschrijving ▶

1 Dialect van Middelburg. 1 mp3-bestand

ZA Testbestanden e-Depot Zeeuws Archief

Testbestanden

Geluidsfragmenten

1-3 Geluidsfragmenten van het Nederlandse Dialecten Databank van het Meertens Instituut

1 Dialect van Middelburg. 1 mp3-bestand

Specificatie:	Afspelen
Vindplaats:	Zeeuws Archief

laatste wijziging 17-02-2016

2 Dialect van Renesse. 1 mp3-bestand

3 Dialect van Zierikzee. 1 mp3-bestand

laatste wijziging 17-02-2016, 3 beschreven archiefstukken

https://e-depot-ted.nationaalarchief.nl/Render/render/external?tenant=NLMdbZA&entity=TypeFile&entityRef=bd1f7f83-d0df-4414-870e-6650a05f9704

5 Experiences in archiving websites

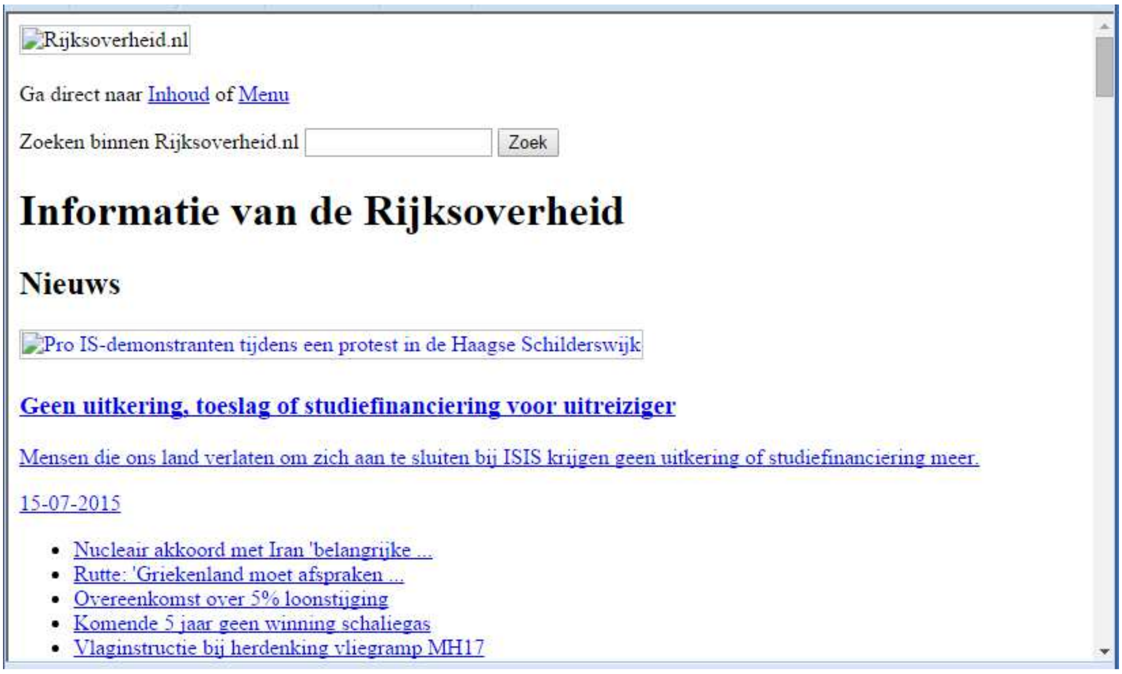
The techniques described in the previous chapters have been used in several cases, with varying degrees of success.

5.1 Case 1: Rijksoverheid.nl

One of the first attempts to webarchive used the website of the national government, Rijksoverheid.nl. In this case, we used the built-in version of Heritrix and the Website ingest workflow of the e-Depot.

5.1.1 Harvesting the website

The harvest workflow ran without problems, but the end result was not what we hoped it would be: all layout was missing.



Rijksoverheid.nl

Ga direct naar Inhoud of Menu

Zoeken binnen Rijksoverheid.nl Zoek

Informatie van de Rijksoverheid

Nieuws

Pro IS-demonstranten tijdens een protest in de Haagse Schilderswijk

Geen uitkering, toeslag of studiefinanciering voor uitreiziger

Mensen die ons land verlaten om zich aan te sluiten bij ISIS krijgen geen uitkering of studiefinanciering meer.

15-07-2015

- Nucleair akkoord met Iran 'belangrijke ...
- Rutte: 'Griekenland moet afspraken ...
- Overeenkomst over 5% loonstijging
- Komende 5 jaar geen winning schaliegas
- Vlaginstructie bij herdenking vliegramp MH17

The missing layout was caused by a small part in the website source code, meant to work around an issue with old versions of Internet Explorer:

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

Heritrix interprets this code as comments, instead of an actual link to the stylesheet, and skips the stylesheet. In later attempts using Wget, we did manage to get a successful ingest of the website including the layout.

5.1.2 Publishing the archived website

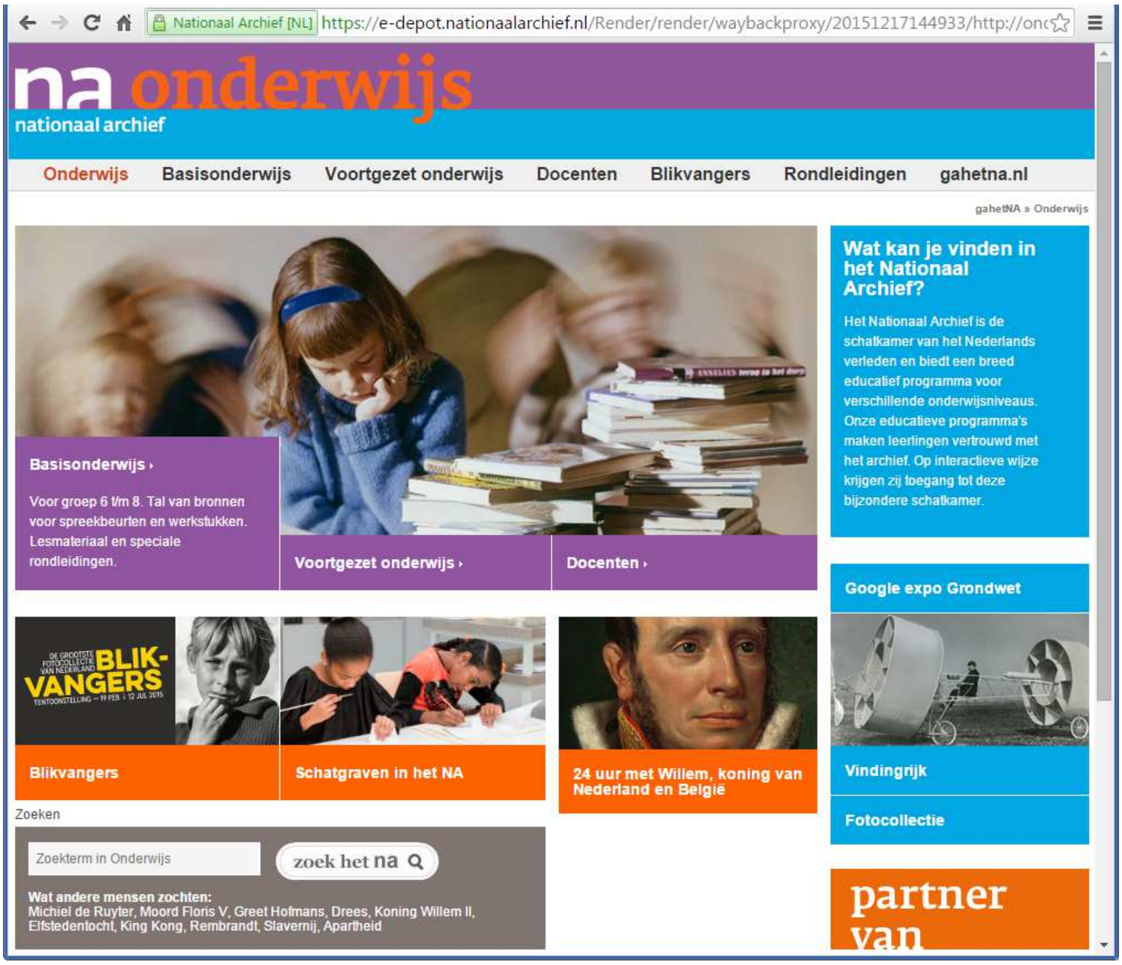
This website is still live and is not part of an archive, so it isn’t available in our collection.

5.2 Case 2: The education website of the National Archives

Towards the end of 2015, the separate Education Website of the Dutch National Archives was integrated into the main website. Just before the separate website was taken offline, we harvested it using the build-in version of Heritrix and the Website ingest workflow.

5.2.1 Harvesting the website

The website was successfully harvested and ingested using the Website ingest workflow:



Wat kan je vinden in het Nationaal Archief?

Het Nationaal Archief is de schatkamer van het Nederlands verleden en biedt een breed educatief programma voor verschillende onderwijsniveaus. Onze educatieve programma's maken leerlingen vertrouwd met het archief. Op interactieve wijze krijgen zij toegang tot deze bijzondere schatkamer.

Basisonderwijs ›

Voor groep 6 t/m 8. Tal van bronnen voor spreekbeurten en werkstukken. Lesmateriaal en speciale rondleidingen.

Wat andere mensen zochten:
Michiel de Ruyter, Moord Floris V, Greet Hofmans, Drees, Koning Willem II, Elfstedentocht, King Kong, Rembrandt, Slavernij, Apartheid

Client-side interaction using JavaScript still works, the image carousel works, and the text blocks that should move up on mouse-over still move. The search engine at the bottom of the screen and the embedded YouTube clips don't work.

5.2.2 Publishing the archived website

This website is available in the online collection:
http://www.gahetna.nl/collectie/archief/ead/index/eadid/2.14.97

5.3 Case 3: Website Minister Van Boxtel

This case has been used as an example in several of the previous chapters: the website of Minister Van Boxtel that went offline in 2002.

The website was transferred on CD-ROM, as XML documents with several linked XSLT stylesheets that needed to be applied to make presentable HTML files. To harvest this website, the technique described in section 3.4 were used: the source code was transformed from XML to HTML, and a local webserver was installed to serve the website to the harvester.

5.3.1 Transforming XML to HTML

Transforming the XML to HTML involved two small steps:

1. Applying the XSLT to the XML-files, using a tool 'xsltproc' and saving the new HTML files under the original name, appended with ".html". For example, the transformed version of the original page '`i000000.xml`' was saved as '`i000000.xml.html`'
2. Editing the resulting HTML files so that all internal links point to the new HTML files instead of the old XML files, using a tool 'sed'.

The two steps were executed in one Cygwin command:

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

5.3.2 Installing the local webserver

After the transformation from XML to HTML, the resulting website could be treated as a simple HTML website with static content. Using the Python SimpleHTTPServer, this website could be made available to a local harvester:

```
python -m SimpleHTTPServer 80
```

By adding the following line to the local hosts file:

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

we could use a normal web browser or a harvester to visit 'www.ministervanboxtel.nl' as served on our own workstation.

5.3.3 Harvesting the website

The website is then harvested using Wget:

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

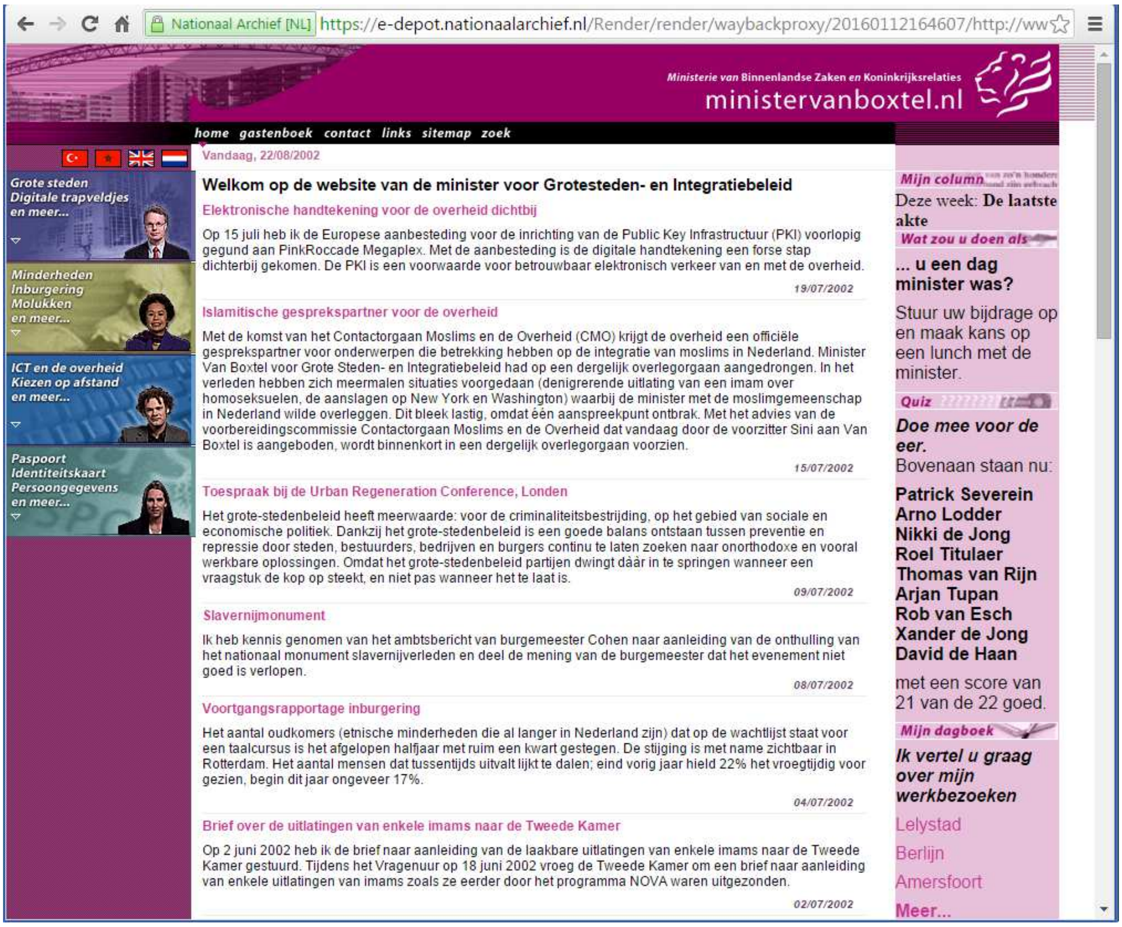
5.3.4 Ingesting the SIP

Using the SIP Creator, we created a SIP that contained the archived website. The AccessionEvent was added to the metadata file as described in section 3.2.3. and the SIP was ingested using a normal ingest workflow.

After ingesting the SIP with the WARC of the website, a second SIP was created containing the original, unchanged files that we received on the CD-ROM. If, at some point in the future, someone else wants to see what the original files were, or if someone finds a better technique to make the website available, the original files can be used.

5.3.5 Publishing the website

This website is available in the online collection:
http://www.gahetna.nl/collectie/archief/ead/index/eadid/2.04.115



Vandaag, 22/08/2002

Welkom op de website van de minister voor Grotesteden- en Integratiebeleid

Elektronische handtekening voor de overheid dichtbij

Op 15 juli heb ik de Europese aanbesteding voor de inrichting van de Public Key Infrastructuur (PKI) voorlopig gegund aan PinkRoccade Megaplex. Met de aanbesteding is de digitale handtekening een forse stap dichterbij gekomen. De PKI is een voorwaarde voor betrouwbaar elektronisch verkeer van en met de overheid.

19/07/2002

Islamitische gesprekspartner voor de overheid

Met de komst van het Contactorgaan Moslims en de Overheid (CMO) krijgt de overheid een officiële gesprekspartner voor onderwerpen die betrekking hebben op de integratie van moslims in Nederland. Minister Van Boxtel voor Grote Steden- en Integratiebeleid had op een dergelijk overlegorgaan aangedrongen. In het verleden hebben zich meermalen situaties voorgedaan (denigrerende uitlating van een imam over homoseksuelen, de aanslagen op New York en Washington) waarbij de minister met de moslimgemeenschap in Nederland wilde overleggen. Dit bleek lastig, omdat één aanspreekpunt ontbrak. Met het advies van de voorbereidingscommissie Contactorgaan Moslims en de Overheid dat vandaag door de voorzitter Sini aan Van Boxtel is aangeboden, wordt binnenkort in een dergelijk overlegorgaan voorzien.

15/07/2002

Toespraak bij de Urban Regeneration Conference, Londen

Het grote-stedenbeleid heeft meerwaarde: voor de criminaliteitsbestrijding, op het gebied van sociale en economische politiek. Dankzij het grote-stedenbeleid is een goede balans ontstaan tussen preventie en repressie door steden, bestuurders, bedrijven en burgers continu te laten zoeken naar onorthodoxe en vooral werkbare oplossingen. Omdat het grote-stedenbeleid partijen dwingt dáár in te springen wanneer een vraagstuk de kop op steekt, en niet pas wanneer het te laat is.

09/07/2002

Slavernijmonument

Ik heb kennis genomen van het ambtsbericht van burgemeester Cohen naar aanleiding van de onthulling van het nationaal monument slavernijverleden en deel de mening van de burgemeester dat het evenement niet goed is verlopen.

08/07/2002

Voortgangsrapportage inburgering

Het aantal oudkomers (etnische minderheden die al langer in Nederland zijn) dat op de wachtlijst staat voor een taalcursus is het afgelopen halfjaar met ruim een kwart gestegen. De stijging is met name zichtbaar in Rotterdam. Het aantal mensen dat tussentijds uitvalt lijkt te dalen; eind vorig jaar hield 22% het vroegtijdig voor gezien, begin dit jaar ongeveer 17%.

04/07/2002

Brief over de uitlatingen van enkele imams naar de Tweede Kamer

Op 2 juni 2002 heb ik de brief naar aanleiding van de laakbare uitlatingen van enkele imams naar de Tweede Kamer gestuurd. Tijdens het Vragenuur op 18 juni 2002 vroeg de Tweede Kamer om een brief naar aanleiding van enkele uitlatingen van imams zoals ze eerder door het programma NOVA waren uitgezonden.

02/07/2002

Deze week: **De laatste akte**

... u een dag minister was?

Stuur uw bijdrage op en maak kans op een lunch met de minister.

Doe mee voor de eer.

Bovenaan staan nu:

Patrick Severein
Arno Lodder
Nikki de Jong
Roel Titulaer
Thomas van Rijn
Arjan Tupan
Rob van Esch
Xander de Jong
David de Haan

met een score van 21 van de 22 goed.

Ik vertel u graag over mijn werkbezoeken

Lelystad
Berlijn
Amersfoort
Meer...

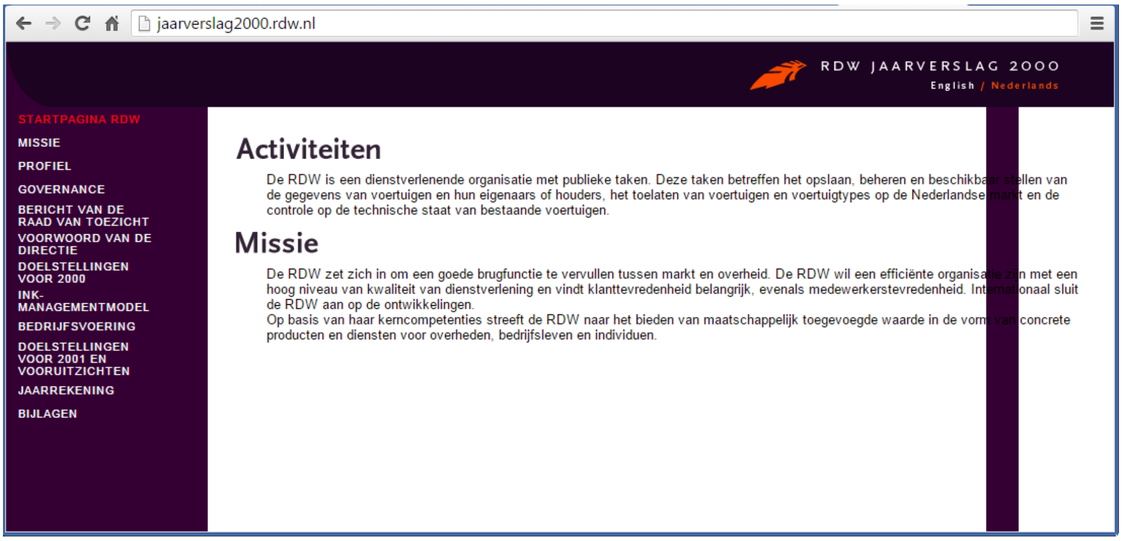
5.4 Case 4: Dutch Vehicle Authority annual report over the year 2000

The Dutch Vehicle Authority (RDW) publishes its public annual report in a web form since 2000. At their request, we tested two versions: a relatively simple version from 2000 (this case) and a more complex version from 2014 (described in section 5.5)

The 2000 version was transferred to us as individual HTML files, stylesheets, and images.

5.4.1 Installing the local webserver

The source files were already in a presentable form, with no transformation needed. We could simply start the Python webserver, edit the hosts file, and view the website:



Activiteiten

De RDW is een dienstverlenende organisatie met publieke taken. Deze taken betreffen het opslaan, beheren en beschikbaar stellen van de gegevens van voertuigen en hun eigenaars of houders, het toelaten van voertuigen en voertuigtypes op de Nederlandse markt en de controle op de technische staat van bestaande voertuigen.

Missie

De RDW zet zich in om een goede brugfunctie te vervullen tussen markt en overheid. De RDW wil een efficiënte organisatie zijn met een hoog niveau van kwaliteit van dienstverlening en vindt klanttevredenheid belangrijk, evenals medewerkerstevredenheid. Internationaal sluit de RDW aan op de ontwikkelingen.
Op basis van haar kerncompetenties streeft de RDW naar het bieden van maatschappelijk toegevoegde waarde in de vorm van concrete producten en diensten voor overheden, bedrijfsleven en individuen.

Most of the website is shown correctly, but one problem became clear: the background image of the lower right section (a purple vertical bar with a white field) was 1600 pixels wide. Wide enough for computer screens that were common in 2000, but not wide enough for the computer screens that we use today. This caused the purple bar to be repeated on the rights side of the section, blocking parts of the text. The solution to this problem required a small change in the website stylesheet. A CSS entry was added for the website body, telling the browser to only repeat the background image vertically, not horizontally:

```
”background-repeat: repeat-y;”
```

5.4.2 Harvesting the website

After the change, the website could be harvested using Wget.

5.4.3 Ingesting the website

Since the website was harvested as a test, and not as part of an actual archival process, we created a SIP using the SIP Creator and ingested it into the TED system. If at some point in the future the website will be ingested into the Production system, we will ingest a second SIP containing the unchanged source files, as we did with Minister van Boxtel website.

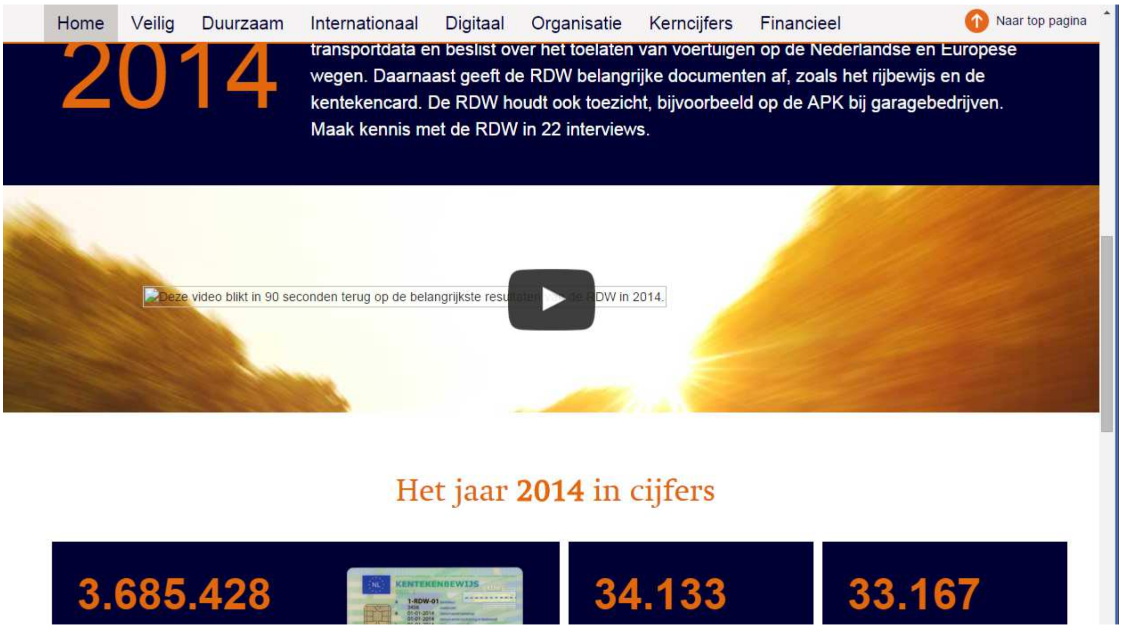
5.4.4 Publishing the website

The website is not part of the National Archives collection yet, so it is not available online.

5.5 Case 5: Dutch Vehicle Authority annual report over the year 2014

The 2014 version of the annual report by the Dutch Vehicle Authority is still online and could be harvested using the built-in version of Heritrix in the Website ingest workflow.

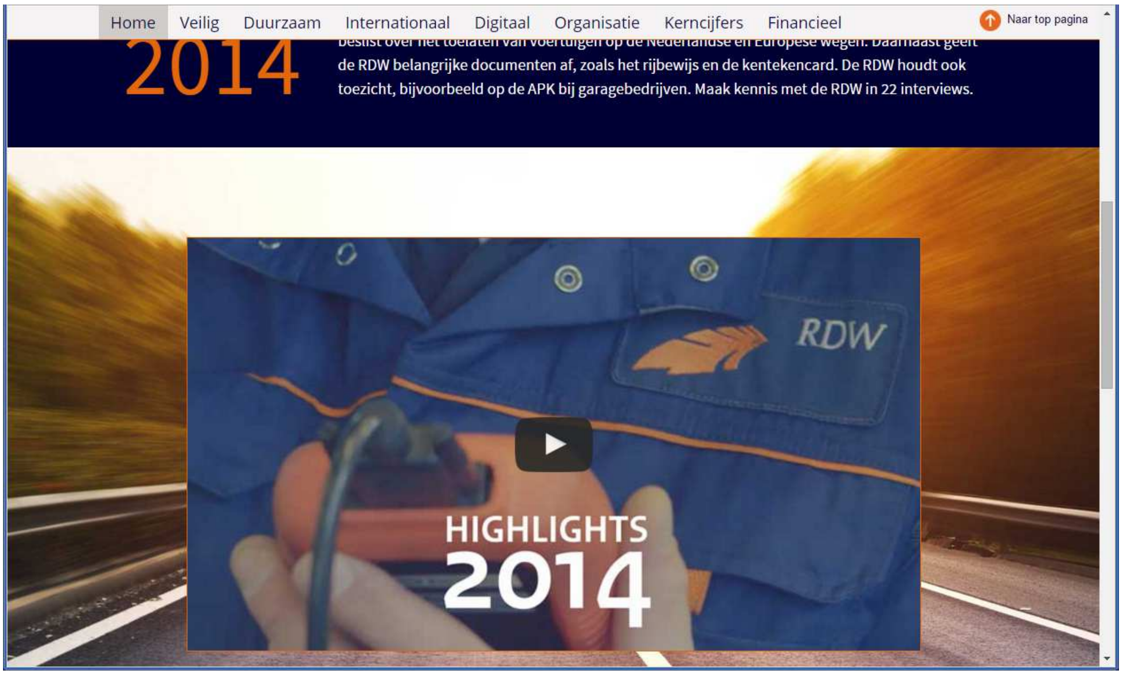
One complication in this website is the use of a YouTube video. The harvester will not harvest YouTube, resulting in a missing part of the web experience:



transportdata en beslist over het toelaten van voertuigen op de Nederlandse en Europese wegen. Daarnaast geeft de RDW belangrijke documenten af, zoals het rijbewijs en de kentekencard. De RDW houdt ook toezicht, bijvoorbeeld op de APK bij garagebedrijven. Maak kennis met de RDW in 22 interviews.

Het jaar **2014** in cijfers

Instead of the online version:



beslist over het toelaten van voertuigen op de Nederlandse en Europese wegen. Daarnaast geeft de RDW belangrijke documenten af, zoals het rijbewijs en de kentekencard. De RDW houdt ook toezicht, bijvoorbeeld op de APK bij garagebedrijven. Maak kennis met de RDW in 22 interviews.

This problem cannot be solved using any of the current harvesting techniques, so we need to find a different way of dealing with this problem.

The solutions that have been identified are procedural, not technical:

1. Harvest without video
2. Harvest without video, and explain in the finding aid that a YouTube video was shown on the website
3. Harvest without video, explain in the finding aid that a video was shown, and describe the content of the video
4. Harvest without video, and ingest the source video separately into the e-Depot
5. Harvest without video, ingest the source video separately, explain in the finding aid that a video was shown on the website, point to the separately ingested video, and add a screenshot of how the video was shown on the webpage.

The last option is probably the best solution, but requires the most work.

5.5.1 Publishing the website

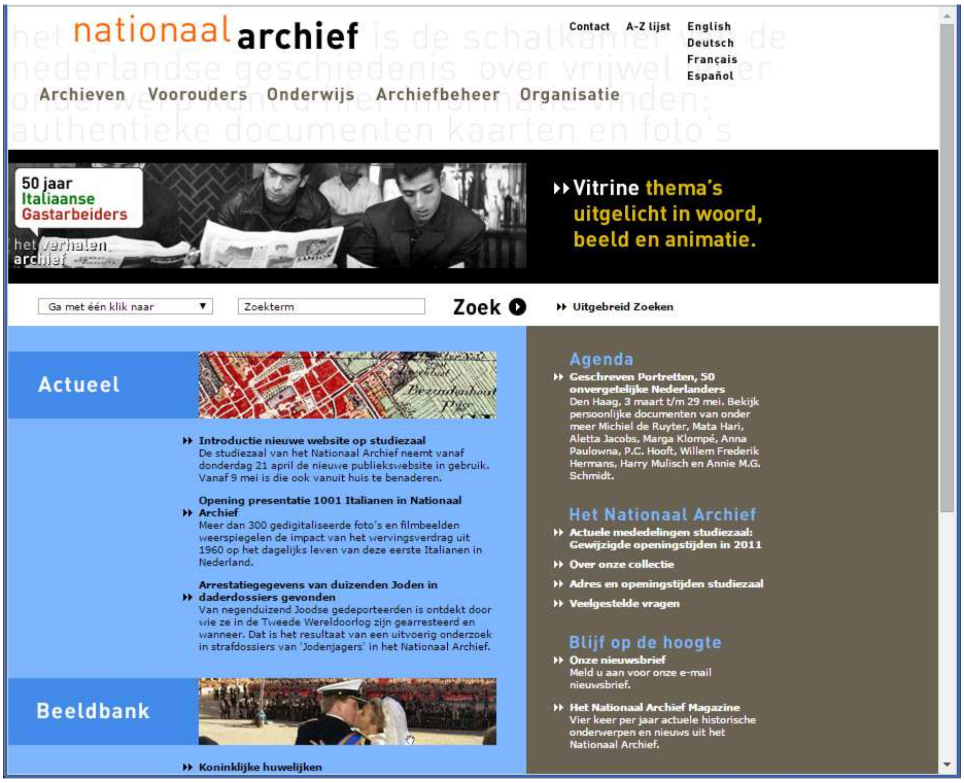
The website is not part of the National Archives collection yet, so it is not available online.

5.6 Case 6: The old website of the Dutch National Archives

With the introduction of the new website Gahetna.nl, the old website of the Dutch National Archives was taken offline. The IT department kept a virtual copy of the old webserver and were able to run the website on the internal network.

5.6.1 Harvesting the website

The first step in the harvest process was editing the local hosts file, to add the IP address of the virtual webserver and link it to the old hostname. The website was then available for viewing in a browser and crawling by a harvester:



Introductie nieuwe website op studiezaal
De studiezaal van het Nationaal Archief neemt vanaf donderdag 21 april de nieuwe publiekswebsite in gebruik. Vanaf 9 mei is die ook vanuit huis te benaderen.

Opening presentatie 1001 Italianen in Nationaal Archief
Meer dan 300 gedigitaliseerde foto's en filmbeelden weerspiegelen de impact van het wervingsverdrag uit 1960 op het dagelijks leven van deze eerste Italianen in Nederland.

Arrestatiegegevens van duizenden Joden in daderdossiers gevonden
Van negenduizend Joodse gedeporteerden is ontdekt door wie ze in de Tweede Wereldoorlog zijn gearresteerd en wanneer. Dat is het resultaat van een uitvoerig onderzoek in strafdossiers van 'Jodenjagers' in het Nationaal Archief.

Koninklijke huwelijken

Agenda

Geschreven Portretten, 50 onvergetelijke Nederlanders
Den Haag, 3 maart t/m 29 mei. Bekijk persoonlijke documenten van onder meer Michiel de Ruyter, Mata Hari, Aletta Jacobs, Marga Klompé, Anna Paulowna, P.C. Hooft, Willem Frederik Hermans, Harry Mulisch en Annie M.G. Schmidt.

Het Nationaal Archief

Actuele mededelingen studiezaal: Gewijzigde openingstijden in 2011

Over onze collectie

Adres en openingstijden studiezaal

Veelgestelde vragen

Blijf op de hoogte

Onze nieuwsbrief
Meld u aan voor onze e-mail nieuwsbrief.

Het Nationaal Archief Magazine
Vier keer per jaar actuele historische onderwerpen en nieuws uit het Nationaal Archief.

We then used Wget to perform the actual harvest:

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

This case had a new challenge: part of the website contained the old finding aids in a database. One ASP script dynamically generates webpages based on the requested finding aid and archival record ID. This way, the script could generate several 100000's of HTML pages containing the records described in the finding aids.

Although technically not a problem in itself, WARC files are considered to be container files (just like ZIP and PST files) and the contents of the container files are individually characterized and have a thumbnail created. This website contained more than 400,000 files, taking a considerable time to ingest.