Hoe archiveer ik websites?

De meeste organisaties hebben al een of meer websites versleten. Bij de overgang naar een nieuwe website stellen organisaties zich wel eens de vraag hoe ze de oude kunnen archiveren. Vaak bevat zo'n oude website interessante gegevens die niet meer relevant zijn voor de nieuwe, maar die wel een historische waarde hebben voor de organisatie. Wat is dan de eenvoudigste manier om die informatie te archiveren?

Nog niet zo heel lang geleden bestonden websites enkel uit statische html-pagina's.[1] Dit zijn eenvoudige tekstpagina's met een opmaak die de webbrowser kan omvormen tot een webpagina. Om deze websites te archiveren, volstond het om het mapje met de bestanden naar je eigen computer te kopiëren. Recente websites maken echter gebruik van een Content Management Systeem (CMS).[2] Dit is een databank waarin de website-informatie wordt beheerd en waarin webpagina's samengesteld worden op het ogenblik dat ze geopend worden. Dit maakt de website dynamisch, maar ook veel moeilijker om te archiveren.

In dit artikel bespreken we hoe zo'n (dynamische) website op een eenvoudige manier digitaal gearchiveerd kan worden. De website zal terug statisch gemaakt worden en offline opgeslagen worden in een vorm waarin ze op lange termijn bewaard kan worden. Net zoals bij e-mails is het digitale karakter bij websites een essentiële eigenschap die bewaard moet worden. Zonder digitale bewaring zou je de *look & feel* en de ervaring om door de website te surfen missen.[3]

Te ondernemen stappen

Analyseer je website

Maak eerst een analyse van je website. De keuze voor een archiveringsmethode is afhankelijk van het type, de inhoud en de elementen van je website.

Er bestaan grofweg drie types van websites: statische websites met vaste inhoud, dynamische websites waarbij de inhoud gehaald wordt uit het deep web en een tussenvorm van die twee.[4] Statische websites bestaan uit een aantal aan elkaar gekoppelde pagina's en zijn meestal in html opgemaakt. Er kunnen zich links met afbeeldingen of links naar andere websites in bevinden. Alle bestanden zijn in een hiërarchische mappenstructuur op de webserver gestockeerd. Een dynamische website is een website die samengesteld wordt op het moment dat ze geopend wordt. Hierbij hebben de pagina's zelf geen inhoud, maar worden ze opgevuld met inhoud die zich in een achterliggende databank bevindt, zoals bij een CMS. Door middel van cookies wordt specifieke gebruikersinformatie op de computer van de gebruiker bewaard waarmee de browser de inhoud van een webpagina kan aanpassen aan de persoonlijke voorkeuren van de gebruiker. De meeste websites zijn een tussenvorm van statisch en dynamisch.[5]

Bekijk daarnaast uit welke inhoud en elementen je website bestaat. Bevat je website veel links naar andere websites? Maakt je website gebruik van externe diensten, zoals kaarten van Google Maps, filmpjes op YouTube of foto's die op een online fotoservice staan? Ook geanimeerde of interactieve beelden en knoppen zorgen voor een extra uitdaging bij het archiveren. Deze elementen maken het archiveren van websites complex en zijn vaak moeilijk te bewaren. Bepaalde functionaliteiten kun je verliezen, zoals het afspelen van Flash-animaties[6] of elementen waarvoor plug-ins[7] geïnstalleerd moeten worden. Interactieve elementen kunnen in gearchiveerde websites niet meer werken, net zoals bestanden die van een andere website opgehaald worden.

Leg doelstellingen vast

Daarnaast is het belangrijk om een aantal duidelijke doelstellingen te formuleren alvorens een archiveringsmethode te kiezen. Het bepalen van een archiveringsmethode hangt namelijk samen met een aantal keuzes. Een eerste keuze betreft wat van de website vastgelegd moet worden bij archivering: de volledige website, inclusief de externe webpagina's waarnaar je website verwijst, of enkel het domein van je eigen website? Een tweede keuze betreft de frequentie waarmee de onderdelen gearchiveerd moeten worden.[8]

Het vastleggen van webpagina's houdt een aantal uitdagingen in die voortvloeien uit hun speciale karakter. Websites hebben een erg vluchtig karakter omdat ze regelmatig geactualiseerd en aangepast worden. Bovendien is de presentatie van een webpagina op het scherm afhankelijk van de interactie met de gebruiker (onder andere webbrowser, persoonlijke instellingen en voorkeuren). Webpagina's zijn tevens sterk met elkaar verweven: ze zijn aan elkaar gekoppeld, worden soms op meerdere servers gehost of halen informatie uit externe services of websites op.[9]

Je zal dus moeten bepalen wanneer je je website gaat archiveren en hoe je de te archiveren website afbakent. Ga je enkel de website capteren als hij offline gehaald wordt, jaarlijks, of bij iedere update? Wordt enkel de website van je eigen domein of ook alle pagina's waarnaar verwezen wordt gearchiveerd? Bij het archiveren van websites zal je moeten accepteren dat er steeds leemten zullen zijn.

Bewaar de essentiële kenmerken van je website

Door de vluchtigheid van het medium en de persoonlijke ervaring bij webpagina's is authenticiteit een moeilijk begrip bij het archiveren van websites. Toch kunnen een aantal essentiële eigenschappen gedefinieerd worden.[10]

1 **Context:** dit zijn gegevens die aanduiden wat de relatie van de website tot de archiefvormer is. Je kan dit onder meer bewaren door beschrijvende metadata over je website vast te leggen.[11]
2 De **inhoud** waaruit je website bestaat: tekst, foto's, video's, kaarten, enzovoort. Sommige elementen, zoals informatie die van externe diensten opgehaald worden (bijvoorbeeld YouTube, Google Maps en Flickr), zijn moeilijk te archiveren. Documenteer daarom de externe diensten die je website gebruikt.
3 **Structuur:** dit geeft de relatie weer tussen de website en zijn onderdelen. De meeste websites hebben een sitemap die de structuur van de website toont.[12] Je kan deze eigenschap bewaren door de originele structuur van je website (de originele structuur van de webpagina's van je website op de webserver) te bewaren en de relaties tussen de verschillende webpagina's te behouden.
4 **Look & feel:** Bij een website is niet enkel de inhoud, structuur en context belangrijk, maar ook de *look & feel* is een essentiële component die bewaard moet worden. Documenteer daarom steeds de technische omgeving waarin je website gemaakt is: bijvoorbeeld de CMS-software die je gebruikt, de plug-ins die je website nodig heeft om bepaalde componenten weer te geven en de serverconfiguratie. Registreer ook de periode waarin je website online was. Dit geeft een beeld van de gebruikte html-versie, de software en de versies van browsers waarin de website getoond kan worden. Op basis van die informatie kan een reconstructie van de website gemaakt worden.
5 Websites kunnen ook **specifiek gedrag en functionaliteiten** hebben, zoals animaties, interactieve elementen en hyperlinks. Daarvoor registreer je ook de technische omgeving van je website. Functionaliteiten kun je verliezen bij het kiezen van een bepaalde archiveringsmethode.

Essentiële kenmerken worden bewaard zodat een getrouwe reconstructie van de website mogelijk is en de website binnen zijn context gearchiveerd wordt. Op de website van eDAVID kan je een document vinden met een lijst van alle metadata die bewaard dienen te worden.[13] Sla dit document op als een gestructureerd tekstbestand (bijvoorbeeld als XML-, CSV- of Excel-bestand) en bewaar dit samen met de gearchiveerde website in het digitale archief. Hou ook alle bijkomende documentatie over je website bij. Dit kan van pas komen indien emulatie in de toekomst nodig zou zijn.[14]

Door te documenteren welke plug-ins de website gebruikt kun je de website met bijvoorbeeld emulatie reconstrueren en vermijd je dat bepaalde elementen niet meer geopend kunnen worden.

Archiveer een website steeds alvorens hem offline te halen en van de webserver te verwijderen. Dit geeft je de mogelijkheid om na het archiveren kwaliteitscontrole uit te voeren en te controleren of alle essentiële eigenschappen bewaard zijn.

Bewaar de website duurzaam

Voor de preservering van websites gelden de algemene regels met betrekking tot duurzame bewaring.[15] Zorg steeds dat je goede back-upprocedures gebruikt en dat je van je bestanden verschillende back-ups hebt die op verschillende (geografische) locaties bewaard worden. Bewaak de integriteit van je gearchiveerde website door checksums te gebruiken en de bestanden periodiek te controleren.[16]

Een uitdaging voor de langetermijnbewaring van websites is de grote hoeveelheid aan bestandsformaten die op websites geplaatst kunnen worden. Het is complex om deze te migreren naar duurzame bestandsformaten omdat de relatie tussen webpagina en bestand op deze manier verbroken kan worden. Onderzoek wijst echter uit dat websites hoofdzakelijk gestandaardiseerde formaten gebruiken, zoals html, jpeg, mp3, enzovoort, waardoor dit probleem te relativeren valt. Een oplossing voor deze uitdaging is om websites te archiveren in het WARC-formaat. Dit is een open standaardformaat om verschillende digitale bronnen met metadata in één archiefbestand op te slaan. Het archiveren van websites in het WARC-formaat is echter complex en wordt in dit artikel niet behandeld.[17]

Archiveringsmethodes

In dit deel worden drie archiveringsmethodes besproken:

1 je website laten archiveren door een organisatie die dit als missie heeft;
2 zelf een offline kopie maken;
3 zelf een video van een surfsessie maken.

Elke methode heeft haar gebreken. Je kunt daarom een aantal methodes combineren om ieder aspect van je website te bewaren.

Laat je website archiveren door een organisatie die dit als missie heeft

The Internet Archive heeft als doel om alle kennis van het web te verzamelen en te bewaren.[18] De Wayback Machine.[19] van The Internet Archive is de grootste externe (gratis) webarchiveringsdienst. Op deze manier werden al meer dan 40 miljard pagina's gearchiveerd. De meeste websites worden gearchiveerd zonder dit te melden. De Wayback Machine maakt op meerdere tijdstippen een momentopname van websites. Neem dus zeker eens een kijkje om te zien of er al opnames van jouw website gearchiveerd werden. Je kunt je website invoeren en de opdracht geven om die te archiveren als dit nog niet gebeurd is.

Het voordeel van deze methode is dat je website gearchiveerd wordt zonder dat je er zelf tijd of kennis voor nodig hebt. Een nadeel is dat je afhankelijk bent van een externe dienst en zelf geen controle hebt over wanneer de momentopnames gearchiveerd worden. Ook moet je steeds naar de Wayback Machine gaan om je gearchiveerde website te raadplegen en heb je de gearchiveerde website niet in eigen bezit.

Controleer of je website al gearchiveerd werd

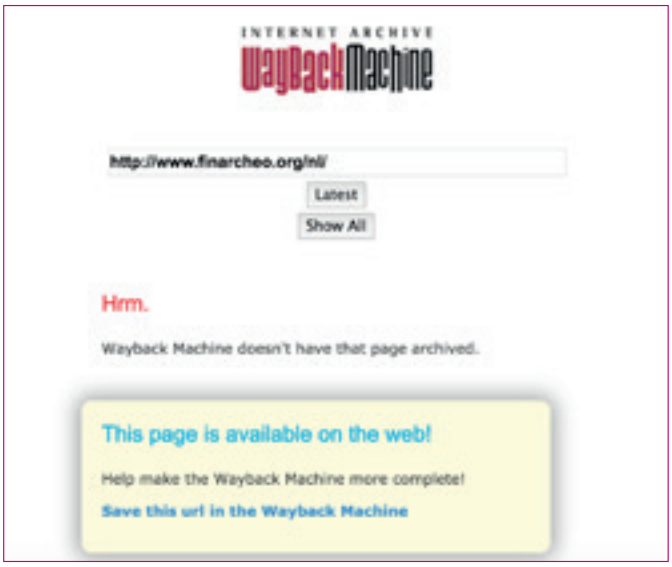
1. Ga naar de website **https://archive.org/web/**
2. Typ de url van je website in de tekstbalk en klik op 'browse history'.

3. Je kunt zien dat de website van PACKED vzw al 93 keer werd gecapteerd tussen 15 februari 2004 en 9 oktober 2016.

4. Door op een datum te klikken kom je op een oude versie van de gearchiveerde website. Dit is de website van PACKED vzw op 15 februari 2004.

Meld je website aan in Wayback Machine

1. Typ je website in de tekstbalk en klik op 'browse history'.
2. Je krijgt een bericht dat je website nog niet gearchiveerd werd. Klik op 'Save this url in the Wayback Machine'.



3. Je website wordt gearchiveerd.

4. Je website is gearchiveerd. Je krijgt een url naar de gearchiveerde versie van je website.

Maak een offline kopie

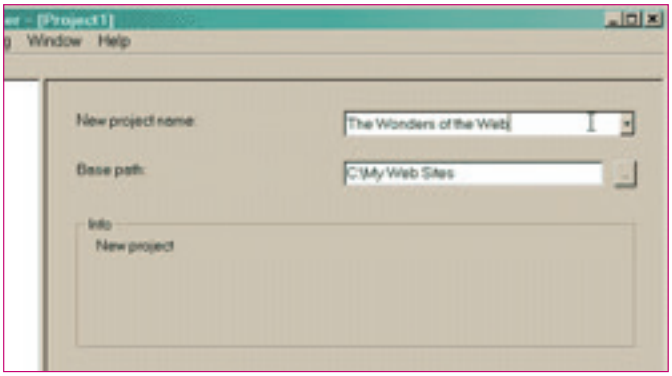
Dit is de meest toegepaste vorm van websitearchivering. Een *crawler* of offline browser maakt een snapshot van je website en slaat alle bestanden op als een html-bestand. Dit is mogelijk doordat de crawler zich als een browser voordoet die iedere pagina van de website bezoekt. In een browser wordt iedere webpagina als een html-pagina weergegeven, en daarom wordt iedere pagina als een html-bestand opgeslagen. Absolute padaanduidingen worden hierbij omgezet naar relatieve padaanduidingen, zodat de website offline geopend kan worden zoals de oorspronkelijke website.[20]

Deze methode kun je toepassen wanneer je alle pagina's en bestanden waaruit je website bestaat, wil bewaren. Het houdt de oorspronkelijke structuur van je website relatief intact en maakt het mogelijk om je website offline te openen en erin te navigeren zoals bij de oorspronkelijke website. Er bestaan eenvoudige tools om een snapshot van je website te maken.

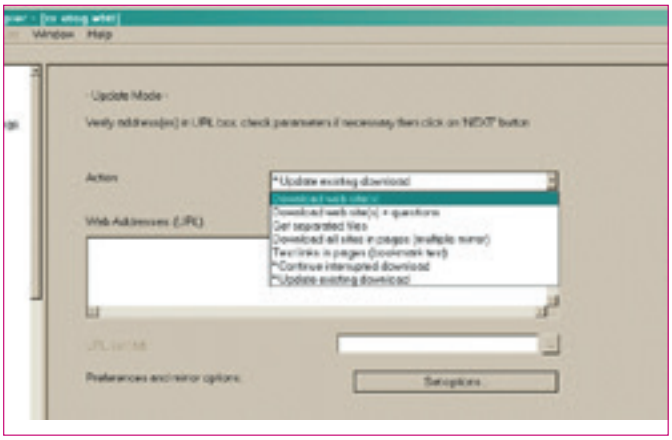
Hou er rekening mee dat crawlers beperkingen hebben. Dynamische webpagina's waarbij de inhoud gevormd wordt op basis van gegevens die een gebruiker invoert, kunnen niet gearchiveerd worden, net zoals informatie die via een paswoord beveiligd is, bepaalde interactieve elementen en informatie van externe diensten. Ook websites met animaties die een plug-in vereisen om af te spelen, zoals Flash-toepassingen, zullen niet goed gearchiveerd worden.[21]

Een eenvoudige crawler met grafische gebruikersinterface is HTTrack.[22]

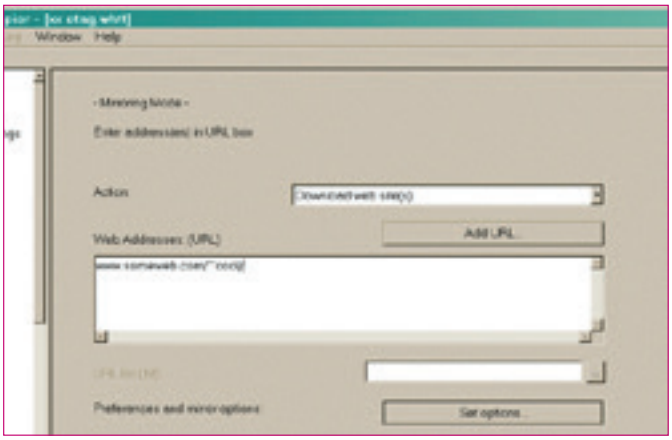
1. Installeer het programma door geef een naam aan het webarchief en kies waar je de gearchiveerde website wil opslaan. Klik vervolgens op 'Next'.



2. Selecteer een actie. Kies voor 'Download web site(s)'.

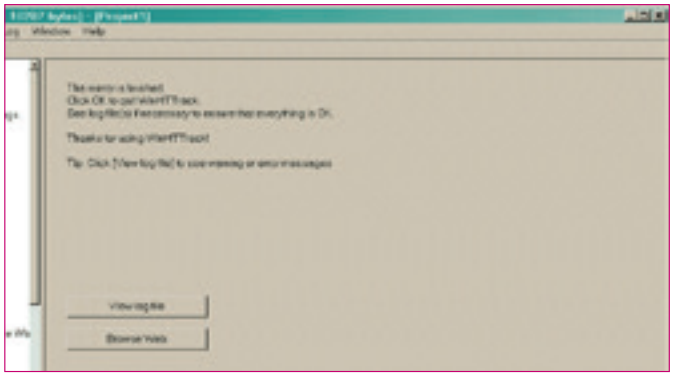


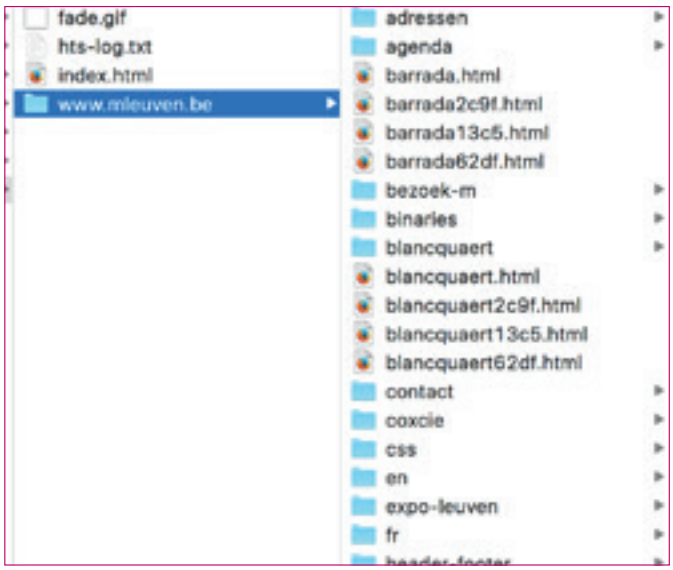
3. Vul de url van je website in. Je kunt meerdere url's downloaden. Kies in dat geval voor 'Add URL' en vul de extra url in. Klik vervolgens op 'Next'.



4. Klik op 'Finish'
5. De crawler is bezig met het downloaden van je website. Laat het venster open zolang deze bezig is.

6. De crawler is klaar.



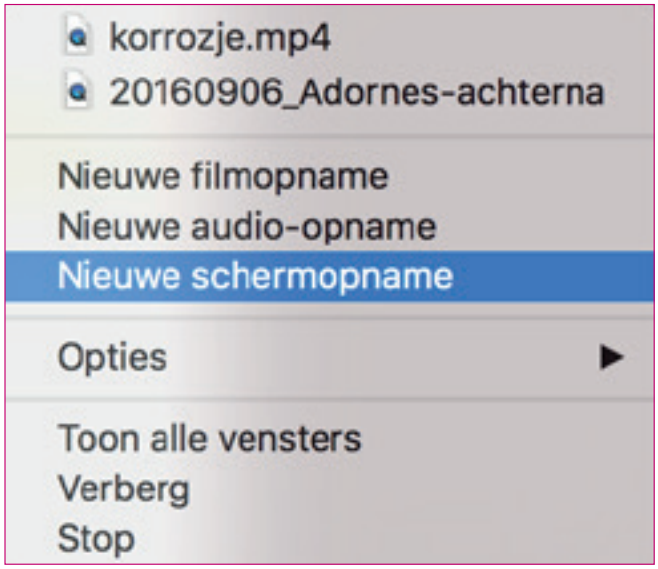


Maak een video van een surfsessie

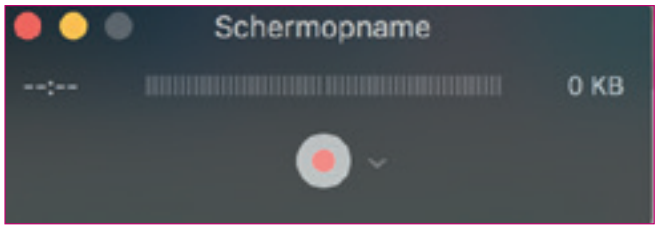
Wanneer je een beeld van een website wil archiveren, maar niet alle pagina's en bestanden wil bewaren, kun je een video maken van een surfsessie op je website. Je kunt dit ook als aanvullende methode gebruiken als de website veel animaties of interactieve elementen bevat of wanneer ze gebruik maakt van externe diensten die moeilijk te capteren zijn.

In dit voorbeeld gebruiken we QuickTime.[23] Als je in een zoekmachine 'screencast' opzoekt, vind je andere software die je kunt gebruiken.

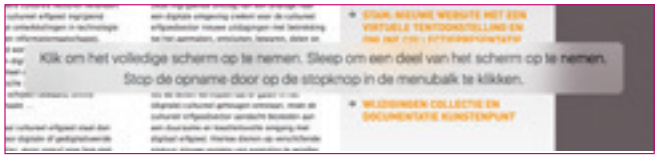
1. Ga naar de website die je wil archiveren.
2. Kies in Quicktime voor 'Nieuwe schermopname'.



3. Een venster verschijnt. Klik op de rode opnamemeknop om de opname te starten.

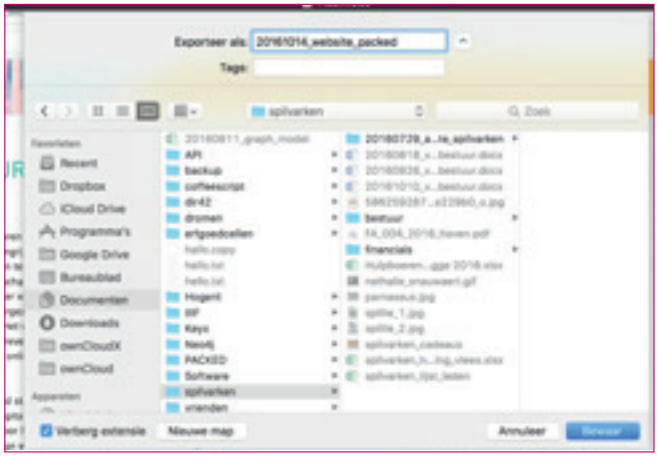


4. Klik om het volledige scherm op te nemen, of sleep het venster over het scherm om slechts een deel van je computerscherm op te nemen.



5. Het venster werd rond de browser gesleept. Klik op 'Start opname'.

6. De opname is bezig. Surf doorheen de website. Als je klaar bent, klik je op de stopknop die in de menubalk staat.
7. Je kunt nu de video bekijken. Klik op 'Bewaar' om de video op te slaan.



Nastasia Vanderperren met de medewerking van Joris Janssens
PACKED vzw

Eindnoten

1. Html is een standaard opmaaktaal voor webpagina's. Zie: https://nl.wikipedia.org/wiki/HyperText_Markup_Language.
2. Een Content Management Systeem is een applicatie die pagina's, afbeeldingen en andere bestanden beheert die samen een website vormen. Voorbeelden van CMS-systemen zijn Wordpress, Drupal en Joomla.
3. F. Boudrez, Archiveren van websites: een kwestie van waardering en 'capture', 5. Zie: http://www.edavid.be/docs/archiveren_websites.pdf.
4. Het deep web is het deel van het web dat niet toegankelijk is voor zoekmachines, zoals databanken die afgeschermd worden door middel van wachtwoorden. De databank achter een CMS-systeem is een onderdeel van het deep web. Zie: https://nl.wikipedia.org/wiki/Deep_web.
5. Boudrez, Archiveren van websites: een kwestie van waardering en 'capture', 7.
6. Flash is software van Adobe en wordt ondermeer gebruikt om animaties, webvideo's en webapplicaties te maken en websites aan te kleden. Je hebt een Flash Player plug-in nodig op je webbrowser om deze bestanden af te spelen. Zie: https://nl.wikipedia.org/wiki/Adobe_Flash.
7. Een plug-in of invoegtoepassing is een uitbreiding van een computerprogramma. In een webbrowser wordt het gebruikt om speciale informatie op een website te kunnen laten zien, zoals flash-animaties.
8. Boudrez, Ibid., 5.
9. Boudrez, Ibid., 7.
10. Boudrez, Archiveren van websites: een kwestie van waardering en 'capture', 7.
11. Beschrijvende metadata zijn metadata die de intellectuele inhoud van een document weergeven. Hieronder vallen aspecten zoals bijvoorbeeld titel, auteur, datum van creatie, keywords en korte inhoud.
12. Een sitemap, soms siteplan, is een pagina of document waarin links naar alle pagina's van een website staan. Dit is een handig hulpmiddel voor bezoekers en zoekmachines om bepaalde pagina's te vinden op een site. Zie: https://nl.wikipedia.org/wiki/Sitemap.
13. Het document kan je hier downloaden: http://www.edavid.be/zelf_aan_de_slag/schema/websiteinfo.xls.
14. Emulatie is een duurzaamheidsstrategie gericht op het weergeven van digitale bronnen zoals ze oorspronkelijk zijn vervaardigd en gebruikt. Door emulatietechnieken toe te passen wordt het mogelijk het gedrag van een verouderde computer na te bootsen op een andere (nieuwere) computer. Op deze manier kunnen bestanden en software die niet meer in gebruik zijn en afspeelbaar zijn op recente computers toch geopend worden.
15. Zie http://www.projecttracks.be/nl/tools/detail/bewaring-van-je-digitaal-archief voor tips over het duurzaam bewaren van je digitale archief.
16. Zie het artikel 'Checksums?! Een instrument voor betrouwbare digitale langetermijnbewaring' in Bladwijzer, nr. 15 (december 2015).
17. M. Pennock, Web-archiving, p. 15-16.
    Zie: http://www.dpconline.org/component/docman/doc_download/865-dpctw13-01pdf.
18. https://archive.org/.
19. https://archive.org/web/
20. Een absoluut pad is een volwaardige verwijzing naar een bestandslocatie en is het volledige adres van de locatie van een bestand, zoals 'http://www.heemkunde-vlaanderen.be/contact/'. Een relatief pad gaat uit van de locatie waar een gebruiker of applicatie zich bevindt. Met een relatief pad kun je verwijzen naar een bestand in een hoger of lager gelegen map zonder het volledige pad te hoeven herhalen. Als je je als gebruiker in de map 'http://www.heemkunde-vlaanderen.be' bevindt, dan volstaat in html een relatieve link naar 'contact' om op het volledige adres 'http://www.heemkunde-vlaanderen.be/contact/' terecht te komen.
21. M. Pennock, Web-archiving, p.11.
22. Zie http://www.httrack.com/, beschikbaar voor Windows, Mac en Linux. Een andere veelgebruikte webcrawler, die ontwikkeld werd door The Internet Archive en een aantal nationale bibliotheken, is Heritrix. Deze kan websites opslaan in het WARC-formaat.
23. https://support.apple.com/kb/DL837?locale=nl_BE.