De oogst van de digitale Hollandse akker

Webarchivering in Nederland (1)

Kees Teszelszky

Od juli-augustus 2017 (/editie/od-juli-augustus-2017)

Webarchivering komt neer op het oogsten van digitaal gewas op grond van *seeds* binnen een bepaald terrein. Als we meer willen weten over webarchivering in Nederland, dan moeten we de weg weten in het webarchiveringslandschap van Nederland, oftewel het Nederlandse nationale domein waarbinnen webarchivering wordt uitgevoerd.

Een reeks van vijf artikelen schetst een beeld van webarchivering in Nederland in zijn verschillende facetten. In dit eerste artikel zet de auteur uiteen hoe het webarchiveringslandschap eruitziet, wat er aanwezig is en wat daarvan wordt geoogst, wie of welke organisatie die oogst binnenhaalt en met welk doel dit allemaal gebeurt.

Om te beginnen: wat is nu het Nederlandse nationale domein op het wereldwijde web? Omdat het *world wide web* in principe geen grenzen heeft, kunnen we het Nederlandse web omschrijven als alles dat op het web wordt gepubliceerd van of over Nederland. Hierbij kunnen we nog op de geografi sche locatie inzoomen (wat wordt gepubliceerd binnen onze landsgrenzen), op taal (wat wordt op het web gepubliceerd in het Nederlands) of op nationaliteit (welke Nederlanders publiceren wat op het web).

Vrijwel iedereen in Nederland is bekend met de .nl-domein uitgang, het tweede nationale domein in de wereld dat in april 1986 werd geregistreerd. De eerste Nederlandse website van het Nationaal Instituut voor Sub atomaire Fysica kwam in februari 1992 online als de derde website in de wereld. Na deze 'oerknal'" explodeerde het Nederlandse web in 23 jaar tot de 5,7 miljoen Nederlandse domeinnamen die het .nl-domein anno 2017 telt en een onbekend aantal Nederlandse sites of sites over Nederland met een andere extensie. Het Nederlandse .nl-webdomein is gezien het aantal geregistreerde websites per inwoner nog steeds het grootste nationale webdomein van de wereld na het .uk-domein. Desondanks (of misschien wel vanwege dit) wordt nog weinig onderzoek gedaan naar wat het Nederlandse nationale webdomein nu precies omvat.

Webcollectie

Het is beter om in de Nederlandse situatie niet te spreken van een 'webarchief', maar van een 'webcollectie' De verzamelingen gearchiveerde websites in Nederland worden in het dagelijkse spraakgebruik webarchieven genoemd naar het Engelse woord *web archive* en het werkwoord *web archiving*. Strikt gezien is een webarchief in bijvoorbeeld een erfgoedinstelling als de Koninklijke Bibliotheek geen echt archief, in de zin van een verzameling van stukken met een bepaald thema, maar het is een collectie: een verzameling van gelijksoortige objecten, die doelbewust bij elkaar gebracht zijn. Webarchivering is in essentie het bewaren van dynamische digitale objecten van het web die zo veel mogelijk in hun oorspronkelijke samenhang worden opgenomen in een collectie om ooit te kunnen dienen als historische bron van de Nederlandse digitale cultuur van een bepaald moment.

Het is belangrijk om verschil te maken tussen een webarchief en een webcollectie, omdat het proces van webarchivering kan plaatsvinden vanuit een archief- en een collectieperspectief, wat gevolgen heeft voor het proces van selectie, opslag en presentatie.[1]

NA, KB en overheidsinstellingen

Binnen de archiefwereld wordt de volgende definitie van een archief gehanteerd: "Een archief is het geheel van archiefbescheiden, ontvangen of opgemaakt door een instelling, persoon of groep personen."[2] Webarchivering vanuit dit perspectief gebeurt in Nederland binnen het kader van de Archiefwet van 1995. Overheidsdiensten zouden alle door hen gepubliceerde websites of andere uitingen op het web moeten archiveren, conform de wet om verantwoording aan de burger af te kunnen leggen over het handelen van de overheid.[3] Burgers en rechtspersonen kunnen rechten ontlenen aan de informatie op een website. Overheidswebsites moeten dan ook zorgvuldig en met een goed doordachte frequentie worden gearchiveerd om deze informatie beschikbaar te houden. De wettelijke basis voor het archiveren van overheidswebsites is in de eerste plaats de Archiefwet 1995, maar daarnaast speelt ook het auteursrecht en de jurisprudentie daarover, omdat gearchiveerde websites een rol kunnen spelen in juridische geschillen.

Webarchivering vanuit archiefperspectief wordt uitgevoerd door het Nationaal Archief of door de overheidsinstellingen zelf. Dit gebeurt in de praktijk nog niet voldoende, zoals blijkt uit een rapport van de Erfgoedinspectie.[4] Het Nationaal Archief heeft in april 2016 een aantal websites opgenomen in het e-depot.[5]

In tegenstelling tot het Nationaal Archief bouwt de KB haar verzameling gearchiveerde websites op vanuit een collectieperspectief.[6] Het belangrijkste criterium is de waarde van een site als Nederlands cultureel erfgoed: het belang van de verzamelde digitale objecten voor de bestudering van de Nederlandse cultuur nu en in de toekomst.[7]

Lagen in het webdomein

Maar hoe kan nu zoiets dynamisch als een site toch worden opgenomen in een vaste collectie van hetzij een archief, hetzij een bibliotheek of andere erfgoedbewaarplaats? De Deense webarchievenonderzoeker Niels Brügger[8] stelt dat het webdomein vijf verschillende lagen heeft die afzonderlijk op verschillende manieren kunnen worden beschreven, bewaard en onderzocht:

1. De eerste laag is die van de webelementen. De verschillende elementen van een website (broncode waar de pagina uit opgebouwd is, tekst, plaatjes, *stylesheet*, etc.).
2. De tweede laag is die van de afzonderlijke webpagina (alle elementen die onder een zeker webadres te vinden zijn).
3. De derde laag is die van de individuele website (alle webpagina's die onder een bepaalde domeinnaam te vinden zijn).
4. De vierde laag is die van de webomgeving (alle sites die met een bepaalde website verbonden zijn door middel van hyperlinks).
5. De vijfde laag is die van het wereldwijde web (alle websites die op een bepaald moment online zijn en bereikbaar zijn).

Daarbij komen nog twee extra lagen die Brügger niet noemt: die van het *deep web*, dat deel dat niet door *webcrawlers* kan worden bezocht, zoals de delen van websites die worden afgeschermd door het bestand robots.txt en databases, en daarbuiten nog het *dark web*, dat deel van het web dat alleen met een speciale browser kan worden bezocht. De laatste laag valt voor zover bekend buiten de webarchivering zoals die in Nederland door erfgoedinstellingen wordt uitgevoerd.

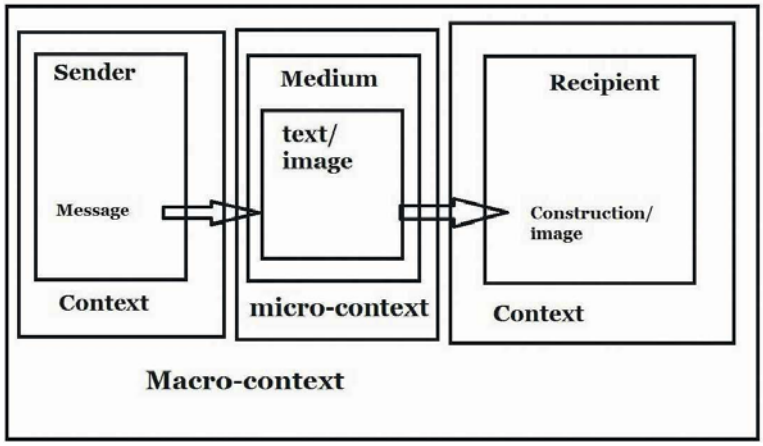
Al deze lagen worden met elkaar verbonden door middel van koppelingen of hyperlinks: zowel op het niveau van de webpagina als die op het wereldwijde web. Deze link structuur kan zichtbaar en aanklikbaar zijn. De linkstructuur kan ook onzichtbaar voor de gebruiker zijn verwerkt in de broncode of juist worden gegenereerd door de website of door de webserver. De linkstructuur is een kenmerkend onderdeel van het wereldwijde web en bovendien een belangrijke bron om de historische context van een website goed te kunnen begrijpen.

Fasen en contexten

Webarchivering is dus veel meer dan het duurzaam opslaan van afzonderlijke websites door erfgoedinstellingen. De interne en externe linkstructuur van een site, de elementen van een pagina en de broncode zijn net zozeer een wezenlijk onderdeel van een erfgoedcollectie, en deze al zijn van belang om een gearchiveerde website ooit te kunnen gebruiken als historische bron of als bewijs. Als Nederlandse webarchiveerders is het cruciaal om te beseffen van welke laag van het web wat wordt bewaard tijdens webarchivering en van welke laag nu juist niet. Het is niet mogelijk om een snapshot te maken van alle lagen van het web samen met alle bijbehorende elementen op een bepaald moment: de techniek laat dat nu nog niet toe.

Op grond van een model over de wijze waarop informatie via het web van de producent naar de gebruiker gaat, dan valt op dat we drie fasen kunnen onderscheiden: zender, medium en ontvanger; en al deze fasen hebben een eigen context waarin een bron tot stand komt.[9]
Deze verschillende contexten corresponderen met de afzonderlijke analytische lagen van het web: de macrocontext stelt het web als geheel voor, de microcontext de inhoud van de webpagina.

De boodschap in de productiefase is de ongepubliceerde website, de website wordt in de consumerfase opgeslagen door de webarchiverende instelling. Webarchivering vindt daarom plaats in het derde hokje van onderstaaande figuur.



Conceptual architecture of communication analysis after Niels Brügger (2010)

Wie in Nederland doen niet aan web archivering?

Op grond van bovenstaande indeling kunnen we een scheiding maken wie zich in Nederland bezighouden met webarchivering en wie niet. Laten we eerst kijken wie hier strikt genomen niet toe behoort. Webarchivering richt zich in de eerste plaats op de website als medium: wij archiveren datgene dat door de zender op het wereldwijde web wordt gepubliceerd. Daarom valt een instelling als het Internationaal Instituut voor Sociale Geschiedenis (IISG), dat bijvoorbeeld digitale bulletin board berichten bewaard, niet in de categorie webarchivering.[10] Wel is het zo dat deze collectie bijzonder waardevol is om contextinformatie te bieden over de geschiedenis van het internet en het begin van het world wide web.
Noch het bewaren van ongepubliceerde websites, databases of Content Management Systems (die horen in het eerste vakje van het communicatiemodel), noch het bewaren van sites die al offline zijn door het redden van een harde schijf is webarchivering. Daarom is het project de Digitale Stad Amsterdam strikt bezien ook geen webarchivering. We spreken hier van webarcheologie, het derde artikel in deze serie zal hier uitvoerig op ingaan. Als een site opnieuw online wordt gezet en dan wordt gearchiveerd, is het wel weer webarchivering. Gebruikersinteracties uit het derde vakje kunnen maar in beperkte mate worden bewaard, vooral omdat dat nog niet technisch mogelijk is. Over de techniek zal ik het later hebben.

Wie in Nederland doen aan webarchiving?

Twee wijzes van archiveren

Al het materiaal van het web kan op twee manieren als bron worden gearchiveerd: door *domain harvesting* (domeinoogst) of *selective harvesting* (selectieve oogst). Het verschil is dat bij de eerste wijze van archivering als het ware een sleepnet met grote mazen door het wereldwijde web wordt getrokken en dat alles gearchiveerd wordt dat op deze manier wordt binnengehaald. Het probleem hierbij is dat een site meestal niet compleet wordt gearchiveerd, omdat de *harvester* een beetje archiveert van heel veel sites. Twee organisaties die websites van het Nederlandse nationale domein op deze manier in een collectie opnemen zijn het Amerikaanse Internet Archive en Common Crawl.[11] Alle Nederlandse webarchiverende instellingen voeren selective harvests uit: dit betekent dat zij bepaalde websites selecteren, maar die wel zo compleet mogelijk proberen te archiveren.

Breed spectrum

Het spectrum van organisaties dat in Nederland aan webarchivering doet is breed. Hieronder volgt een kleine selectie van wie in Nederland aan webarchivering doet, zonder hierbij de pretentie te hebben compleet te willen zijn.

Voor zover bekend is het eerste webarchiveringsproject in Nederland Archipol geweest, dat in januari 2000 werd gestart door het Documentatiecentrum Nederlandse Politieke Partijen in Groningen. De collectie van Archipol omvat ongeveer 1000 websites van politieke partijen, politieke bewegingen en politici die op regelmatige basis worden gearchiveerd en ook na toestemming ter beschikking worden gesteld aan onderzoekers.[12]

Het Nationaal Archief doet aan webarchivering van overheidssites. Beeld en Geluid archiveert websites van omroepen. Daarnaast bestaat Archiefweb, dat websites van overheidsorganisaties archiveert en toont.[13] Ook zijn er een aantal lokale webarchiveringsprojecten, zoals dat van het Regionaal Archief Dordrecht, het Stadsarchief Rotterdam en de Groninger Archieven.[14] De grootste webcollectie in Nederland is die van de Koninklijke Bibliotheek, die sinds 2007 meer dan 12.000 websites heeft bewaard.[15] In totaal zijn ongeveer 15.000 websites van het hele Nederlandse nationale domein opgenomen in een webcollectie.

In het tweede artikel van deze serie zullen we nader ingaan op de verschillende webarchiveringsinitiatieven in Nederland en het overzicht dat daarvan gecreëerd wordt door het Netwerk Digitaal Erfgoed en de Nationale Coalitie Digitale Duurzaamheid.

Constructie van een nieuwe bron

In essentie komt het archiveren van een website dus neer op het bewaren van bepaalde softwarecode die uiteindelijk in de toekomst weer werkend moet worden gemaakt om een resultaat op te leveren dat zo goed mogelijk overeenkomt met de live bron uit het verleden. Webarchivering is daarom niet zozeer het vastleggen van een historische bron, maar de constructie van een nieuwe bron die zoveel mogelijk kenmerken heeft van de oorspronkelijke live website. Tijdens de *harvest* wordt een momentopname gemaakt van de broncode van de site, de elementen als tekst en beeld en een deel van de linkstructuur, en die in de webcollectie weer gereconstrueerd worden tot een nieuwe bron. We herbouwen als het ware een deel van het Nederlandse web voor toekomstig onderzoek.

Het resultaat van de oogst

Wat is nu die bron die wordt bewaard. De meeste Nederlandse webarchiverende organisaties gebruiken de webarchiveringstool Heritrix, net als het Internet Archive. Het resultaat van de harvest is echter niet hetzelfde.

Als we kijken naar een gearchiveerde website in het publieke deel van het Internet Archive en in de collectie van Nederlandse webarchiverende instellingen, dan blijkt dat het Internet Archive een webpagina toont van losse elementen die verzameld zijn uit verschillende momentopnames tijdens aparte *crawls* en die vervolgens samengevoegd zijn tot een nieuwe website. De bron die wordt getoond, is in deze vorm nooit online geweest: het is net alsof we al de verschillende edities van Eline Vere tot op het niveau van de bladzijden en zelfs de zinnen uit elkaar halen. Vervolgens wordt uit al die elementen weer een nieuwe editie gemaakt en die wordt gepresenteerd als de editie in de bekende *Wayback* machine.

Nederlandse webarchiverende instellingen selecteren misschien wel een beperkter aantal websites dan het Internet Archive, maar de oogst van die sites is beter, omdat een momentopname van alleen die sites wordt gemaakt. Nederlandse webcollecties beheren dus authentiekere webbronnen van het Nederlandse nationale web dan het Internet Archive, dat eigenlijk een vergaarbak van losse webelementen is.

In cultuur gebracht

Kortom: als we willen weten hoe het landschap van de Nederlandse webarchivering eruitziet, dan moeten we ook weten hoe dit digitale weblandschap ooit in cultuur is gebracht. Het is essentieel om te weten welke selectiebeleid is gevoerd, welke harvestingstrategie werd gevolgd, welke techniek is gebruikt en tot welk resultaat dat heeft geleid. Het bewaren van informatie hierover is cruciaal om de collectie van waarde te maken voor onderzoekers van de toekomst. Een veelgehoorde opmerking die wij als webarchiveerders horen is dat het Internet Archive 'toch alles al heeft'. Gezien vanuit het collectieperspectief van een collectie van gearchiveerde websites is dat dus maar betrekkelijk.

Tot slot: het is belangrijk dat Nederlandse erfgoedinstellingen op nationaal en lokaal niveau webarchiveren en samenwerken in nationaal verband om het resultaat zo bruikbaar mogelijk te maken voor toekomstige onderzoekers.

Kees.Teszelszky@kb.nl (mailto:Kees.Teszelszky@kb.nl) , Dr. Kees Teszelszky is onderzoeker Webarchivering bij de afdeling Digitale Duurzame Toegang van de Koninklijke Bibliotheek

---

Noten:

[1] Zie: http://ingmarbladertenschrijft.blogspot.nl/2011/04/websitesarchiveren-of-websites.html (http://ingmarbladertenschrijft.blogspot.nl/2011/04/websitesarchiveren-of-websites.html) en de opmerking van Ton de Looijeronder dit blog. Voor de discussie over webarchivering als archiefactiviteit, zie: Aïda Chebbi, Archivage du Web organisationnel dans une perspective archivistique. PhD dissertatie, Université de Montréal. December 2012 (https://papyrus.bib.umontreal.ca/xmlui/bitstream/handle/1866/9203/Chebbi_Aida_2013_these.pdf?sequence=4&isAllowed=y (https://papyrus.bib.umontreal.ca/xmlui/bitstream/handle

/1866/9203/Chebbi_Aida_2013_these.pdf?sequence=4&isAllowed=y)).

[2] Lexicon van Nederlandse Archieftermen. Den Haag, 1983, 1991. 13.

[3] http://wetten.overheid.nl/BWBR0007376/2015-07-18 (http://wetten.overheid.nl/BWBR0007376/2015-07-18) . Zie ook: http://digitaalduurzaam.blogspot.nl/2011/04/webarchivering-innederland-de-status.html (http://digitaalduurzaam.blogspot.nl/2011/04/webarchivering-innederland-de-status.html)

[4] https://www.erfgoedinspectie.nl/actueel/nieuws/2016/12/7/webarchivering-bij-de-centrale-overheid-gebeurt-nauwelijks (https://www.erfgoedinspectie.nl/actueel/nieuws/2016/12/7/webarchivering-bij-de-centrale-overheid-gebeurt-nauwelijks) ; Webarchivering bij de centrale overheid. Het archiveren van websites en uitingen op sociale media. Rapport Erfgoedinspectie.

[5] Jeroen van Luin, Ervaringen met website-archivering in het Nationaal Archief. Rapport, 8 april 2016.

[6] Webarchivering bij de centrale overheid. Het archiveren van websites en uitingen op sociale media. Rapport Erfgoedinspectie. De kracht van het netwerk. KB Beleidsplan 2015-2018. Den Haag, 2014, 21.

[7] De kracht van het netwerk. KB Beleidsplan 2015-2018. Den Haag, 2014. 5

[8] Niels Brügger, Web History, an Emerging Field of Study. In: Niels Brügger ed., Web history. New York, 2010. 3.

[9] Niels Brügger, Website Analysis: Elements of a Conceptual Architecture. Papers from The Centre for Internet Research. Aarhus, 2010, 9.

[10] https://socialhistory.org/sites/default/files/docs/archiving-electronic_messages.pdf (https://socialhistory.org/sites/default/files/docs/archiving-electronic_messages.pdf)

[11] https://archive.org/ (https://archive.org/) ; http://commoncrawl.org/ (http://commoncrawl.org/)

[12] http://www.archipol.nl (http://www.archipol.nl)

[13] www.archiefweb.eu/ (http://www.archiefweb.eu/)

[14] https://www.groningerarchieven.nl/onderzoek/webarchief-groningen (https://www.groningerarchieven.nl/onderzoek/webarchief-groningen)

[15] https://www.kb.nl/organisatie/onderzoek-expertise/e-depot-duurzame-opslag/webarchivering (https://www.kb.nl/organisatie/onderzoek-expertise/e-depot-duurzame-opslag/webarchivering)