



Marcel Ras – projectleider
webarchivering in de KB –
bij een kast met (verouderde)
opslag tapes

Over de duurzaamheid van digitaal

HOE BEWAREN WE HET WEB?

Hoe zag het web er in 2007 uit en hoe communiceerde men toen? Sinds kort wordt ons digitale leven voor de toekomst bewaard. In Den Haag heeft de Koninklijke Bibliotheek een begin gemaakt met het digitaal duurzaam opslaan van het Nederlandse web voor toekomstige generaties. “Je moet eigenlijk meteen maar leren leven met het feit dat er ook dingen kwijtraken.”

Met zijn circa 2,3 miljoen domeinnamen en ruim 1,5 miljoen websites is het Nederlandse webdomein het op drie na grootste ter wereld. Alleen die van de Verenigde Staten, Duitsland en Engeland zijn groter. De Koninklijke Bibliotheek (KB) in Den Haag heeft berekend dat het meer dan een jaar zou duren om het gehele Nederlandse domein te *crawlen*, binnen te halen. “En als je daarmee klaar bent, dan kun je weer opnieuw beginnen. Het heeft op dit moment weinig zin om te doen.” Aan het woord is Marcel Ras, projectleider webarchivering in de KB. Ruim een jaar geleden begon de KB met het ambitieuze project om een selectie van het Nederlandse web op te slaan om het in de toekomst te ontsluiten. Het is niet de eerste instelling die een internetarchief opbouwt, maar wel de eerste waarbij digitale duurzaamheid centraal staat.

Twee manieren

Een krant of boek bewaren is van een geheel andere orde dan een website bewaren. Een krant kan gedeeltelijk zijn vergaan, maar nog wel leesbaar zijn. Als een boek niet meer compleet is, ligt er wel een exemplaar van in een andere bibliotheek. Maar als een website weg is, is hij vaak nergens meer te vinden. “Digitale informatie is er

of is er niet meer. Als er een aantal bitjes is omgevallen, dan heb je vaak niets meer. Het lastige bij websites zijn ook de heel snelle technische ontwikkelingen. We hobbelen er altijd achteraan.”

Er zijn grofweg twee manieren om websites te archiveren. Bij de ene manier wordt een snapshot gemaakt van de homepages en één of twee niveaus daaronder. Bestanden groter dan een x-aantal megabyte worden niet binnengehaald, dit geldt ook voor de meeste afbeeldingen.

Het bekendste voorbeeld van de snapshotmethode is de Wayback Machine van het Amerikaanse Internet Archive. Wil je zien hoe een website er op een bepaald moment ongeveer uitzag, dan is dit een uitstekende bron. Ook landen als Frankrijk, Denemarken, Zweden en Noorwegen passen deze methode toe. De nadelen laten zich raden: de website is incompleet en ongeschikt voor onderzoeksdoelinden omdat de onderliggende inhoud niet is opgeslagen. Bij de andere aanpak wordt een beperkt aantal web-

Als een website weg is, is hij vaak nergens meer te vinden

sites zo gedetailleerd mogelijk gearchiveerd. “Daar willen we dus echt álles van hebben. De bedoeling is dat het archief in de toekomst door wetenschappers gebruikt wordt voor onderzoek”, verklaart Marcel Ras de keuze van de KB voor deze aanpak. “Hoe zag het web er in 2007 uit en hoe communiceerde men toen? Er wordt nu ook al onderzoek gedaan naar bijvoorbeeld taalgebruik op het web, maar deze live webpagina’s verdwijnen. Het internetarchief is voor taalkundigen interessant, maar ook voor historici, sociologen en antropologen.”

Digitale duurzaamheid

In het e-Depot van de KB worden digitale collecties duurzaam opgeslagen. Dat wil zeggen dat de bestanden zodanig in de gaten worden gehouden, dat wanneer een bepaald bestandsformaat in onbruik raakt, er een strategie klaarstaat om deze bestanden ook in de toekomst toegankelijk te houden. Deze strategie wordt per formaat en per collectie vastgesteld. Nederland loopt voorop op het gebied van duurzame opslag. De meeste webarchiveringsprojecten halen websites binnen,

TIPS VAN DE EXPERT

Wat kunnen wij - gewone mensen die hun digitale bestanden willen bewaren - leren van experts die dagelijks te maken hebben met het bewaren van grote hoeveelheden hiervan? We vroegen het Edwin Klijn, specialist elektronisch publiceren bij de European Commission on Preservation and Access (ECPA). De ECPA houdt zich bezig met bewustwording rond het beheer en behoud van cultuur- en wetenschapscollecties.

Bij het bewaren van digitale bestanden is de problematiek waar culturele instellingen en privépersonen tegenaan lopen vaak identiek. Het grote verschil ligt vooral in de hoeveelheid materiaal, dat bij instellingen fors groter is.

Drager en formaat

“Wij onderscheiden de drager en het formaat. Je moet eigenlijk kijken naar de duurzaamheid van beide. Een cd-schijfje is een voorbeeld van een drager en een bestandsformaat is bijvoorbeeld WordPerfect. Er wordt vaak erg gefocust op de houdbaarheid van cd’s, maar meestal doen de bestanden op de cd het sneller niet meer dan de dragers. Dit komt omdat er nieuwe besturingssystemen of nieuwe software

uitkomen, en dat is allemaal weer niet *backward compatible*.

De bestanden die op je cd of dvd staan, zul je op een gegeven moment naar een nieuwer formaat moeten omzetten en dan is het verhaal van de houdbaarheid van die cd of dvd eigenlijk helemaal niet zo interessant meer. Over het algemeen zijn de formaten sneller achterhaald, vooral bij multimedialfiles.”

Digital cliff

“Stel je hebt duizend foto’s op een schijf waar je niks mee doet, dan kan het zo zijn dat je op onverklaarbare wijze één of enkele bestanden niet meer kunt openen. Dat verschijnsel wordt *digital cliff* genoemd. Er zijn dan bitjes

slaan ze op en maken ze al dan niet toegankelijk, zonder te letten op de duurzaamheid.

Emulatie (zie kader) is de meest logische strategie voor een webarchief. “Als je over tien jaar een website uit 2007 wilt bekijken, moet hij draaien op een platform uit 2007 met een browser uit 2007 en plugins uit 2007”, legt Ras uit. “Dat maakt het wel ingewikkeld, want alle gangbare software moeten we bewaren in een zogenaamde *software depository* en het systeem moet weten dat hij een website uit 2007 moet openen met Internet Explorer

versie X, Firefox versie Y of Safari versie Z. Plus dat de bestanden in die site die een plugin nodig hebben, ook getoond kunnen worden. Dat moeten we allemaal netjes gaan bijhouden om ervoor te zorgen dat mensen dat straks echt kunnen.”

In migratie ziet Marcel Ras niet zoveel: “Stel, we hebben straks in ons webarchief 20 miljoen jpeg's die in onbruik raken. Migreren naar een ander formaat zou wel kunnen, hoewel zo'n migratietool een paar maanden bezig is om alles om te zetten. Maar dan kloppen al die links in die

omgevallen, zoals dat wel wordt genoemd. En dat terwijl er niks mee is gebeurd: ze zijn niet gebruikt en ze stonden op een plek waar ze veilig zijn.

Archiefinstellingen die veel met digitale informatie te maken hebben, hebben systemen die regelmatig de integriteit van dat soort bestanden checken. Dat doen ze bijvoorbeeld met een *checksum*: de oorspronkelijke waarde van een bestand wordt met een bepaalde berekening vergeleken met de huidige waarde. Als je dit vertaalt naar advies, dan zou ik zeggen: probeer ze eens in de zoveel tijd te openen.”

Origineel

“Cultuurinstellingen gaan uit van een *archival master*. Dat is een bestand dat niet gecompriemd en niet gecorrigeerd is, bijvoorbeeld een tif-bestand. Je moet heel erg oppassen met compressie, omdat dat ervoor kan zorgen dat je een bestand niet meer uit kunt pakken. Bewaar dus het originele bestand. Dat is meestal het grootste bestand, bij foto's bijvoorbeeld tif of raw. Stel dat je van een foto een tif en een jpeg hebt, bewaar dan altijd de liefst onbewerkte tif en niet de jpeg. Sowieso als je iets voor langere tijden wilt bewaren, moet je kiezen voor een standaardformaat.”

LOCKSS

“Er is een systeem met een grappige naam: LOCKSS, dat staat voor *lots of copies keep stuff safe*. Dat wil zeggen dat je écht belangrijke bestanden op meerdere en verschillende dragers zet. Het is misschien een beetje panisch in sommige gevallen, maar je spreidt zo wel de risico's.

Een aardig advies voor mensen die dingen gaan inscannen: gooi de originelen vooral niet weg. Later kan misschien veel meer met digitalisering.

Of ik deze richtlijnen ook privé toepas? Haha, wat een gemene vraag. Ik vind zelf dat niet alles moet worden bewaard. Het begint met selectie. Van belangrijke bestanden bewaar ik wel verschillende exemplaren. Ja, ook papieren exemplaren. Over honderd jaar is die papieren versie er in ieder geval nog. Hoe we die gaan bewaren is weer een ander probleem. En al die perkamenten objecten zijn er dan nog steeds, die zijn er al vanaf de Middeleeuwen. Zo duurzaam is digitaal nog niet geweest.”

webpagina's niet meer omdat de bestandsnamen zijn gewijzigd. Die moeten we dan ook allemaal gaan aanpassen. Dat is niet te doen. Emulatie lijkt dus de meest logische strategie."

Emulatie versus migratie

Er zijn twee manieren om digitale bestanden duurzaam te bewaren. Enerzijds kan een object via migratie worden aangepast aan nieuwe hard- en software. Met emulatie geven nieuwe hard- en software het oorspronkelijke digitale object weer in de oorspronkelijke context.

Honderd websites

De eerste fase van het webarchiveringsproject loopt in mei 2007 ten einde. De KB heeft bijna honderd websites volledig gearchiveerd. Dit jaar zal dat aantal uitgroeien tot circa duizend, waarna het aantal gestaag zal blijven groeien. Waar nu nog wordt opgehouden bij de grens van een domein, zullen in de toekomst mogelijk ook stappen naar buiten worden binnengehaald. "De links naar andere websites zijn vaak ook heel interessant. Bij honderd websites zou je dan al een paar duizend sites moeten binnengaan. Je verzamelt dan heel onvermoede dingen die je zelf nooit zou selecteren. Dat zou voor onderzoekers erg interessant zijn."

Marcel Ras legt uit dat in het *harvest*-proces - het proces om websites binnen te halen - aangegeven wordt hoe vaak van een bepaalde website gegevens worden binnengehaald: wekelijks, maandelijks of per kwartaal. Daarna heeft de KB er geen omkijken meer naar. "Stel dat er een bomaanslag plaatsvindt, dan zou je kunnen ingrijpen. Dan wil je bepaalde sites of weblogs ineens elke dag gaan archiveren omdat je bepaalde nieuwsberichten wilt vastleggen. Dat is gedaan na de aanslagen van 11 september en er is ook een heel groot webarchief van na de tsunami in Zuidoost-Azië."

Tijdbalk

En hoe wordt het gearchiveerde web in de toekomst ontsloten? "Je zou over tien jaar het

web van bijvoorbeeld april 2007 kunnen bekijken. Je begint met een zoekactie, waarmee je de websites of weblogs uit die periode vindt. Concreet tref je de website aan zoals ie er toen uitzag, met de content uit die periode. Je hebt ook de mogelijkheid te navigeren naar de website zoals deze in maart of mei was. We denken nu aan een soort tijdbalk waarmee je door de tijd kunt browsen."

De webpagina's in het archief worden bij het binnengaan omgezet tot platte html-pagina's. De dynamische functionaliteit en interactie gaan hiermee verloren. Concreet gezegd: de reageerknop doet het niet meer. Ras: "De actie ben je kwijt. De dynamiek is niet te archiveren en valt helaas niet meer te bestuderen."

Moedeloos

De honderd websites die nu zijn gearchiveerd, zijn goed voor 16 miljoen bestanden met een totale grootte van 360 gigabyte. De KB kwam meer dan tweehonderd verschillende bestandsformaten tegen. "We weten aardig wat van de tien, twintig meest gebruikte bestandsformaten zoals html, pdf, jpg en gif. Die beslaan 98 procent van alle bestanden in het webarchief. Daarvan weten we ook redelijk hoe we dat toegankelijk kunnen houden. We richten ons actief op deze meest voorkomende bestanden. En dan zijn er 180 of meer bestandsformaten waarvan je het bestaan niet eens vermoedde. Dat kunnen gekke dingen zijn als een verkeerd ingetikte extensie. Of een formaat als xmlxocted. Je komt de meest verschrikkelijke dingen tegen. Daar kun je nooit meer wat mee." Wordt Ras nooit moedeloos van deze enorme klus? "Het is wel veel, en het probleem is wel ingewikkeld. Je moet eigenlijk meteen maar leren leven met het feit dat er ook dingen kwijtraken. Dat is nu eenmaal zo. Aan de andere kant is het ook heel spannend, het is erg nieuw. Je zou kunnen zeggen dat het kenmerk van het web is dat het vluchtig is, dus laat het alsjeblieft ook vluchtig blijven. Daar is geen onderzoeker het mee eens. Elke onderzoeker zal zeggen: alsjeblieft bewaar! En als het kan alles. Maar dat kan niet." En zachtjes voegt hij eraan toe: "En dat hoeft ook niet."



BEWAREN

VIER FAVORIETEN DIE JE NIET MAG MISSEN

500 terabyte webpagina's

The Internet Archive biedt een interessante terugblik op de ontwikkeling van het internet vanaf 1 januari 1996 tot nu. De site heeft een indrukwekkend archief van 85 miljard websites. Verder vind je hier teksten, audio-opnames, live muziek en bewegende beelden, voorzien van beoordeling en commentaar van internetters. Alleen al de websites kosten de Californische organisatie op dit moment ongeveer 500 terabyte (500.000 gigabyte) opslagruimte.

www.archive.org

93 kilometer nationaal archief

Het Nationaal Archief is naar eigen zeggen de grootste openbare archiefinstelling van Nederland met 93 kilometer archieven, waaronder kaarten, tekeningen en foto's. Op de website is informatie en fotomateriaal te vinden over Nederland en zijn koloniën, het graafschap/gewest Holland en van de provinciale bestuursinstellingen in Zuid-Holland, particuliere instellingen en personen van nationale betekenis.

www.nationaalarchief.nl

Europese bieb

European Archive is een openbaar toegankelijke digitale bibliotheek die in 2006 is gestart met het archiveren van de Europese cultuur. European Archive werkt bij de opbouw van het archief samen met Amerikaanse Internet Archive, bibliotheken, musea en andere instellingen. Op de website vind je onder andere een overzicht van alle Europese politieke partijen, kun je muziekstukken van klassieke componisten (Händel, Haydn, Brahms) beluisteren en filmpjes bekijken.

www.europarchive.org

Gratis dataopslag via het web

Als je extra opslagcapaciteit nodig hebt voor je computerbestanden, dan kun je online bij de Engelstalige site Mozy.com gratis 2 gigabyte aan gegevens stallen. Voor de aanmelding bij deze dataopslagdienst moet je een e-mailadres opgeven. Geef niet het adres op van je eigen provider, maar een gratis webmailadres (gmail, hotmail), want Mozy wil nog wel eens een reclamemailtje sturen.

mozy.com