KB kiest voor pragmatische opt-out aanpak

Bij het archiveren van Nederlandse websites hanteert de Koninklijke Bibliotheek sinds dit jaar de opt-out aanpak. Websitebeheerders krijgen bericht over de op handen zijnde crawling, archivering en openbaarmaking van de betreffende site. Blijft weigering uit, dan wordt dit beschouwd als een impliciete toestemming. Deze pragmatische aanpak biedt een manier om om te gaan met de ingewikkelde juridische materie van webarchivering.

Marcel Ras, Annemarie Beunen, Elvira Cameron en Tjeerd Schiphof

De omvang van het Nederlandse deel van het World Wide Web groeit fors. Hoewel het .nl-domein in 2006 pas zijn twintigste verjaardag vierde, werd in datzelfde jaar nog de tweemiljoenste .nl-domeinnaam geregistreerd. Inmiddels is het aantal gegroeid tot ruim 2,5 miljoen.[1] Daarmee is Nederland wereldwijd het op drie na grootste country code Top Level Domain.[2]

Nu is er de paradox dat de meeste informatie op het web zeer vluchtig is en een korte levensduur heeft, terwijl we tegelijkertijd het web beschouwen als cultureel erfgoed.[3] Als er niets wordt ondernomen zal dit digitale erfgoed dat voor toekomstig onderzoek naar de ontwikkeling van het web en onze huidige samenleving van belang is, verloren gaan.

De oplossing voor dit probleem is het archiveren van websites. Maar daar kleven onder andere de nodige juridische haken en ogen aan. Zo bestaan sites uit verschillende onderdelen die samen het geheel vormen. Deze onderdelen kunnen elk op zichzelf beschermd zijn door het auteursrecht. Denk aan originele teksten, foto's, zoekprogrammatuur, vormgeving, databanken. Deze auteursrechten kunnen ook nog eens bij verschillende rechthebbenden berusten, zoals een freelancefotograaf of journalist, een uitgever, een softwarefabrikant enzovoort. Vaak rusten er naast auteursrecht ook andere intellectuele eigendomsrechten op websiteonderdelen zoals merkenrecht, naburige rechten, portretrecht en het databankenrecht. Zo kunnen websites een opeenstapeling van rechten vormen.

'Het Internet Archive archiveert en ontsluit websites zonder enige vorm van toestemming'

Nationale bibliotheek archiveert sites

Met deze haken en ogen zag de Koninklijke Bibliotheek zich geconfronteerd toen ze in 2006 startte met het archiveren van een selectie van Nederlandse websites. Als nationale bibliotheek heeft de Koninklijke Bibliotheek niet alleen de taak gedrukte publicaties duurzaam te bewaren, maar ook elektronische. Omdat steeds meer publicaties in laatstgenoemde vorm verschijnen en ook websites als publicaties gezien kunnen worden, is het duurzaam bewaren en toegankelijk houden hiervan een belangrijke taak geworden.

Daar waar de meeste internationale initiatieven zich al in een vroeg stadium richtten op het verzamelen van websites en over het algemeen nog steeds deze aanpak hanteren, richt de KB zich nadrukkelijk op het duurzaam bewaren van gearchiveerde websites. De complexiteit hiervan is de reden waarom de KB pas een jaar geleden met webarchivering is begonnen. Met de ontwikkeling van een e-Depot heeft de KB sinds 2003 een infrastructuur om niet alleen elektronische tijdschriftartikelen op te slaan, maar ook de mogelijkheid om de archivering van websites te kunnen waarborgen.

Gekozen is voor een selectieve benadering. Dit betekent dat er een beperkte selectie van Nederlandse websites gearchiveerd zal worden. Tijdens de eerste fase werden honderd websites gearchiveerd. Dit leverde ruim 360 GB aan data op en ruim 16 miljoen unieke bestanden en tweehonderd verschillende bestandsformaten. Het aantal te archiveren websites zal jaarlijks groeien, waarbij de ge-

selecteerde sites een aantal malen per jaar zullen worden gearchiveerd. De hoeveelheid aan data en unieke bestanden die opgeslagen moeten worden en de veelheid aan bestandsformaten zorgen voor de grootste hoofdbrekens bij het ontwikkelen van een strategie voor duurzame toegang (zie ook het kader hieronder).

Het archiveren van websites is verder een heidens karwei, zelfs wanneer we ons beperken tot een kleine selectie. Sites kunnen dagelijks veranderen en het is dan ook onmogelijk om iedere wijziging in iedere site vast te leggen in een webarchief. Desondanks wordt verwacht dat het archief jaarlijkse ongeveer 15 terabyte zal groeien.

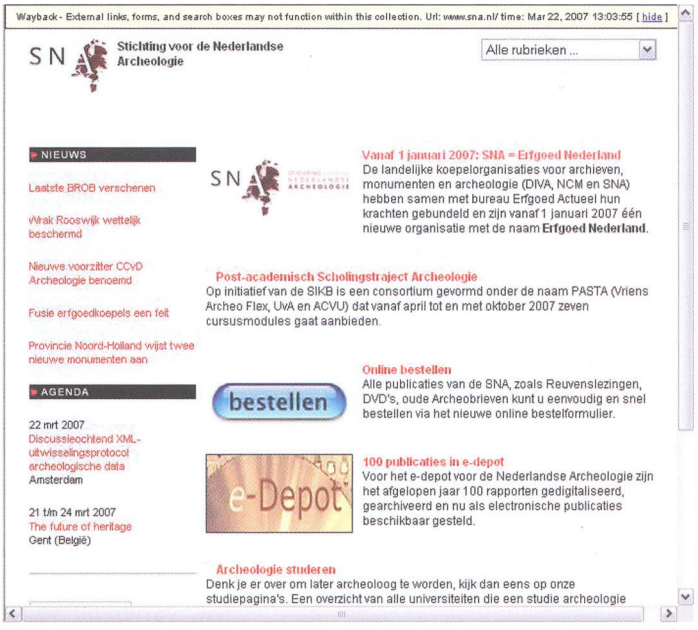
Dankzij de selectieve benadering is het eenvoudiger om de gekozen juridische aanpak te gebruiken. Bovendien is het door het ontbreken van depotwetgeving zo goed als onmogelijk om een domeinharvest uit te voeren. Verder is het Nederlandse domein zeer groot, waardoor het

'Alle siteonderdelen kunnen beschermd zijn door auteursrecht'

Digitale duurzaamheid

Wanneer websites zijn binnengehaald, geïndexeerd en netjes voor de gebruiker toegankelijk zijn gemaakt, begint eigenlijk pas het probleem. Hoe zorgen we ervoor dat deze sites over pakweg vijftig jaar nog steeds voor de gebruiker toegankelijk zijn? We zullen dan geen gebruik meer maken van de huidige browsers en platforms en wellicht dat ook het concept van het web volledig is veranderd. Toch zullen we ervoor moeten zorgen dat wetenschappers in de toekomst over onderzoeksdata kunnen beschikken. Het is reëel om ervan uit te gaan dat een groot deel van die onderzoeksdata afkomstig zal zijn uit webarchieven. Dat websites zijn opgeslagen in het e-Depot is een hele geruststelling, maar niet voldoende. We zullen meer moeten doen. Actief onderzoek naar de wijze waarop we deze sites toegankelijk kunnen houden is noodzakelijk; het bewaren van de juiste metadata om later te kunnen bepalen wat het is en op welke wijze dit dan gepresenteerd moet worden, zijn vereisten. Omdat de presentatie van een website sterk afhankelijk is van de gebruikte browser, maar ook van plug-ins, onmisbaar voor de presentatie van specifieke aspecten van een website (zoals bijvoorbeeld Flash, video en audio), is het noodzakelijk om de meest gangbare browsers en plug-ins te bewaren. De afdeling digitale duurzaamheid van de KB[9] doet intensief onderzoek naar deze aspecten van webarchivering. Daar waar mogelijk samen met andere organisaties wereldwijd, onder andere in het kader van de International Internet Preservation Consortium (IIPC).[10]

kostbaar is om dit in zijn geheel te archiveren. En tot slot is bulkarchivering niet geschikt voor het volledig archiveren van websites. Bulkarchivering richt zich op het maken van een snapshot waarbij strikte grenzen zijn gesteld aan de te crawlen hoeveelheid bestanden en data. Aangezien duurzaam bewaren het uitgangspunt is, lijkt het niet zo heel zinvol om slechts een beperkt deel van websites te bewaren. We bewaren immers ook niet alleen de titelpagina van een boek.



Een van de honderd door de KB gearchiveerde websites

Juridische uitdagingen

Om te bepalen hoe de KB kan omgaan met de ingewikkelde juridische materie van webarchivering, heeft het Centrum voor Recht in de Informatiemaatschappij (eLaw@Leiden) van de Universiteit Leiden een onderzoek uitgevoerd naar de juridische aspecten van webarchivering binnen het Nederlandse recht, met name het auteursrecht en de wet bescherming persoonsgegevens.[4] Om na te gaan wanneer webarchivering in het vaarwater komt van het auteursrecht,[5] maken we hier onderscheid tussen de drie fasen van het proces: het harvesten van websites, het archiveren ervan en het weer beschikbaar stellen aan het publiek.

Harvesten

Het binnenhalen van een website gebeurt door een crawler, die een kopie van de site maakt. Kopiëren is een handeling die onder het auteursrecht valt en in principe is dus vooraf toestemming van de rechthebbende(n) nodig. Een instelling/rechtspersoon kan voor harvesting geen uitzondering in de Auteurswet inroepen (zo geldt de eigen gebruikbeperking alleen voor privé-personen).
Wel zijn er enkele categorieën websites die vrij geharvest mogen worden. Ten eerste zijn dat websites afkomstig van en gevuld door de openbare macht, zoals ministeries en gemeenten. Complicatie is dat vrij kopiëren niet is toegestaan voor werken op deze sites waarop derden het auteursrecht hebben, zoals externe rapporten. Ook het kopiëren van delen van een website die technisch beveiligd zijn, is onrechtmatig. Een tweede groep websites die men vrij mag harvesten, zijn sites met een Creative Commons-licentie die commercieel of niet-commercieel hergebruik toestaat. Echter, sites die in hun geheel onder zo'n CC-licentie openbaar zijn gemaakt, zijn nog schaars. Ten derde mag men ook sites harvesten waarop de rechthebbende(n) expliciet heeft verklaard dat het auteursrecht erop niet zal worden ingeroepen (verklaring van publiek domein).
Kortom, meestal is wel voorafgaande toestemming nodig. Dit zou ondervangen kunnen worden door wetgeving die verplicht tot depot van (elektronische) publicaties bij de nationale bibliotheek, inclusief websites in het .nl-domein en/of gericht op een Nederlands publiek.

Archiveren

Om alle crawlerkopieën van een op diverse momenten geharveste website duurzaam te kunnen bewaren, is het maken van meerdere nieuwe kopieën in diverse formaten soms noodzakelijk. Het auteursrecht is dus weer in het spel, maar de Auteurswet kent uit behoudsoogpunt een uitzondering voor migratiekopieën. Men mag een kopie van een werk maken om deze raadpleegbaar te houden als de technologie waarmee het toegankelijk gemaakt kan worden, in onbruik raakt. Een belangrijk nadeel is dat deze uitzondering niet geldt voor databanken die door het databankrecht worden beschermd. Veel websites kunnen in hun geheel waarschijnlijk als zodanig worden gekwalificeerd, zodat er toch toestemming van de rechthebbenden voor migratiekopieën nodig is.
Bij het kopiëren voor duurzame bewaring moeten de persoonlijkheidsrechten van de auteursrechthebbende(n) worden gerespecteerd. Men mag geen onredelijke wijzigingen in een werk aanbrengen, of een werk verminken of op een andere wijze aantasten als dat de goede naam van de rechthebbende(n) kan schaden.

Beschikbaar stellen aan het publiek

Gezien de missie van de KB om iedereen te laten delen in ons cultureel erfgoed, ligt het voor de hand de websites niet alleen te archiveren maar ook toegankelijk te maken. Dit impliceert openbaarmaking. Maar of dit zonder toestemming van de rechthebbenden op de website(onderdelen) mag, hangt af van de wijze waarop dit gebeurt. Een uitzondering in de Auteurswet staat toe dat werken uit de eigen collectie aan een algemeen publiek beschikbaar worden gesteld in een besloten netwerk dat alleen binnen het gebouw van de bibliotheek te raadplegen is (tenzij met de rechthebbenden iets anders wordt overeengekomen). Voor databanken geldt deze uitzondering weer niet. Voor beschikbaarstelling via een openbaar netwerk als internet, is wél steeds toestemming van de rechthebbende(n) vereist. Daarvoor maakt

'In landen als Denemarken, Frankrijk en Duitsland mag wel zonder toestemming worden gecrawld en bewaard'

het niet uit of men kiest voor openbaarmaking alleen voor geautoriseerde gebruikers via een password of openbare toegang voor een algemeen publiek.

Bescherming persoonlijke levenssfeer

Op websites die verzameld, opgeslagen en beschikbaar worden gesteld, kunnen zich ook zogenaamde persoonsgegevens bevinden. Dit betekent dat men rekening moet houden met de bescherming van de persoonlijke levenssfeer van mensen en de eisen die de wet op dit punt stelt. Dit is de Wet bescherming persoonsgegevens (Wbp). Persoonsgegevens zijn gegevens die een levende persoon betreffen. Dit is een ruim begrip waar veel onder kan vallen, variërend van telefoonnummers, (e-mail)adressen, foto's en nog veel meer. De Wbp legt beperkingen op aan de verwerking van persoonsgegevens. Verwerking is toegestaan wanneer de betrokkene toestemming gegeven heeft of wanneer deze zelf zijn persoonsgegevens duidelijk openbaar heeft gemaakt.[6] Verwerking is echter ook toegestaan als dit noodzakelijk is voor de behartiging van het gerechtvaardigde belang van degene die verwerkt.

Wel of geen toestemming vragen?

Het auteursrecht vormt dus een obstakel als men grote aantallen websites wil archiveren, ondanks de verschillende beperkingen die er in de Auteurswet zijn opgenomen. Mede door het ontbreken van een wettelijk depotkader is een bepaalde mate van toestemming vooraf nodig van de rechthebbenden van de te archiveren website. Tegenover hun auteursrechtelijk belang staat echter het grote (algemene) belang dat webarchivering dient: het behoud van ons digitale cultureel erfgoed ten behoeve van wetenschappelijk onderzoek en het brede publiek. Dit belang wordt benadrukt in het Unesco Charter on the Preservation of the Digital Heritage.

Kijken we naar andere webarchieven, dan zien we daar verschillende oplossingen. Het Internet Archive, het grootste webarchief ter wereld, archiveert en ontsluit websites zonder enige vorm van toestemming. Dit is ook onbegonnen werk omdat het wereldwijd websites archiveert. Wanneer er iemand bezwaar heeft, kan hij dat kenbaar maken. Het harvesten wordt dan gestopt. Er wordt overigens zeer weinig protest tegen aangetekend.

In tegenstelling tot het Internet Archive vraagt de British Library wel expliciet om toestemming via een schriftelijke overeenkomst. In deze overeenkomst wordt ook gevraagd of er auteurs- of databankrecht van derden op siteonderdelen rust en of ook zij toestemming voor gebruik hebben gegeven. Dit is vermoedelijk de reden dat bijna 75 procent van de contracten niet wordt teruggestuurd. Door deze aanpak ontstaat er een enorme administratieve last en wordt webarchivering bijna onmogelijk.

Een nadeel van een dergelijk *opt-in* contract is dat het vrijwel ondoenlijk is van álle rechthebbenden op siteonderdelen toestemming te krijgen; een contract met bovenstaande vraag biedt daarom slechts schijnzekerheid. Daarnaast kan men, als steeds toestemming moet worden afgewacht, niet snel nspelen op 'events' (zoals '9/11' of 'Katrina') die interessante sites opleveren die elk uur veranderen. In landen als Denemarken, Frankrijk en Duitsland, waar er wel gebruik gemaakt kan worden van een uitgebreide depotwetgeving (of waar deze wetgeving in voorbereiding is), kan er zonder enige toestemming worden gecrawld en bewaard. Maar bij het toegankelijk maken stuit men weer op de Auteurswet, waardoor deze archieven alleen binnen de muren van de nationale bibliotheek toegankelijk zijn. Het Deense archief zelfs pas na

Twee basisstrategieën voor webarchivering

Er zijn twee basisstrategieën voor webarchivering. De eerste strategie is gericht op het automatisch harvesten van een grote hoeveelheid websites (meestal een nationaal domein). De tweede strategie selecteert op basis van een specifiek selectiebeleid. Het automatisch harvesten is relatief goedkoop in vergelijking met de selectieve benadering, waarbij meer handmatig werk verricht moet worden. Daar staat tegenover dat bij het harvesten van een beperkt aantal sites meer aandacht besteed kan worden aan technische details. Ook is het mogelijk om websites tot op het diepste niveau te archiveren.

toestemming van het Danish Data Protection Agency.

Pragmatische benadering

Om te voorkomen dat het archiveren van websites blijft steken in langdurige administratieve handelingen, heeft de KB voor een meer pragmatische benadering gekozen, de *opt-out* aanpak. Aan de beheerders van websites wordt een bericht gestuurd waarin is aangegeven dat de KB de betreffende site uit erfgoedoverwegingen wil gaan harvesten, archiveren en openbaar maken. Daarbij wordt er een termijn gegeven waarbinnen men toestemming kan weigeren. Blijft weigering uit, dan wordt dit beschouwd als een impliciete of stilzwijgende toestemming.

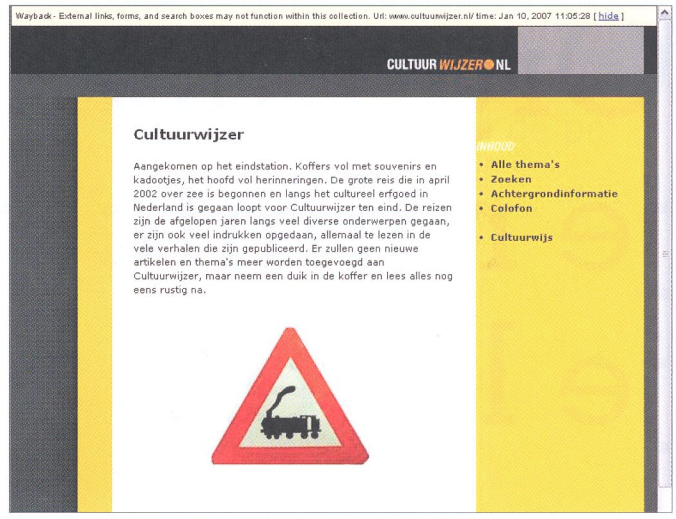
Deze aanpak veronderstelt een impliciete toestemming voor webarchivering wanneer de sitehouder geen anti-harvesting-maatregelen zoals robots.txt heeft toegepast.[7] Omdat het gebruik van *opt-out*, indien de gebruiker geen archivering wenst, zeer gebruikelijk is geworden op internet,[8] valt te verdedigen dat het ontbreken van een zogenaamd robots.txt-file kan worden opgevat als toestemming voor het archiveren. Dat betekent dat de KB zou mogen aannemen dat de websitehouder toestemming geeft, tenzij hij op andere wijze bezwaar maakt.

Op de toegangssite van het webarchief moet duidelijk het doel van de webarchivering worden vermeld, met daarbij de disclaimer dat de inhoud van de websites volledig onder verantwoordelijkheid van de sitehouder zelf valt en dat de archiverende instantie die inhoud ondersteunt noch controleert of wijzigt. Bovendien is een heldere klachtenregeling van belang die rechthebbenden of klagers in verband met de Wbp, portretrecht en dergelijke, de mogelijkheid biedt bezwaar te maken. Na onderling overleg kan men dan bijvoorbeeld besluiten tot een waarschuwende mededeling op de bewuste website of het ontoegankelijk maken van materiaal.

Sinds begin 2007 maakt de KB gebruik van deze benadering. De ervaring met het archiveren van de eerste honderd websites tijdens de eerste fase is overwegend positief. De administratieve last is beperkt tot het versturen van één (mail)bericht naar de beheerder van een website. Dit betekent dat er nog wel gezocht moet worden naar de 'eigenaar' van de site of het aanspreekpunt daarvoor binnen een organisatie. Hoewel deze lang niet altijd gemakkelijk te vinden is, kan er desondanks geconcludeerd worden dat het een besparing oplevert. Van de honderd aangeschrevenen hebben er slechts twee bezwaar gemaakt. Het ene bezwaar had betrekking op het toegankelijk maken van de gearchiveerde website in verband met auteursrechten van derden, het andere bezwaar was niet juridisch van aard maar had te maken met het tijdstip van crawlen. Het is de intentie om zo veel mogelijk 's avonds en 's nachts te crawlen, op momenten dat er minder gebruikelijk dataverkeer van en naar desbetreffende site te verwachten is. De betreffende organisatie gaf aan dat hun website juist in de avond veel bezocht werd.

De KB zal deze benadering blijven hanteren wanneer de selectie van te archiveren websites fors uitgebreid zal worden. Is deze *opt-out* benadering ook in dit vervolg succesvol, dan is het voor het eerst dat voor het archiveren van websites een dergelijke pragmatische aanpak op grote schaal toegepast wordt. <

Marcel Ras is projectleider webarchivering bij de afdeling digitale duurzaamheid van de KB. Annemarie Beunen is universiteit docent aan de juridische faculteit van de Universiteit Leiden bij eLaw@Leiden. Tjeerd Schiphof is universitair docent aan de Erasmus Universiteit Rotterdam en juridisch medewerker bij het KIT. Elvira Cameron is beheerder digitale collectie bij de afdeling e-Depot van de KB.



Voor de presentatie van gearchiveerde websites maakt de KB gebruik van de open source versie van de WayBack Machine

Noten

1] Stand op 12 september 2007: 2,54 miljoen geregistreerde.nl-domeinnamen. www.sidn.nl

2] Na Duitsland (.de), UK (.uk) en Europa (.eu). Verisign, The Domain Name Industry Brief. Volume 4 – issue 3, June 2007. www.verisign.com

3] Zie artikel 1 van de UNESCO charter on the Preservation of the Digital Heritage.

4] Het volledige onderzoeksrapport is te vinden op www.kb.nl/hrd/dd/dd_projecten/projecten_webarchivering.html.

5] Over de basisbeginselen van het auteursrecht is meer te vinden in de *Juridische Wegwijzer Archieven en Musea online* op www.taskforce-archieven.nl/projects/juridischewegwijzer.

6] Naast de 'gewone' persoonsgegevens kan er ook sprake zijn van 'bijzondere persoonsgegevens'. Daarvoor geldt een strenger regime. De achtergrond daarvan is dat bepaalde gegevens beter afgeschermd moeten worden, zoals die met betrekking tot iemands godsdienst of levensovertuiging, ras, politieke gezindheid, gezondheid, seksuele leven, lidmaatschap van een vakvereniging of strafrechtelijke achtergrond.

7] Robots.txt files geven aan dat (onderdelen van) de website niet geïndexeerd en/of gearchiveerd mogen worden. Volgens ongeschreven gedragsregels op internet dient men dergelijke verzoeken te respecteren.

8] Ter vergelijking: in een rechtszaak over indexering en caching door Google (Field/Google) ging een Amerikaanse rechtbank uit van stilzwijgende toestemming, omdat de sitehouder geen gebruik had gemaakt van robots metadata. Verschillen zijn dat Google geen handmatige selectie maar 'alles' automatisch indexeert en de sites tijdelijk in plaats van permanent opslaat.

9] www.kb.nl/hrd/dd/dd.html

10] www.netpreserve.org