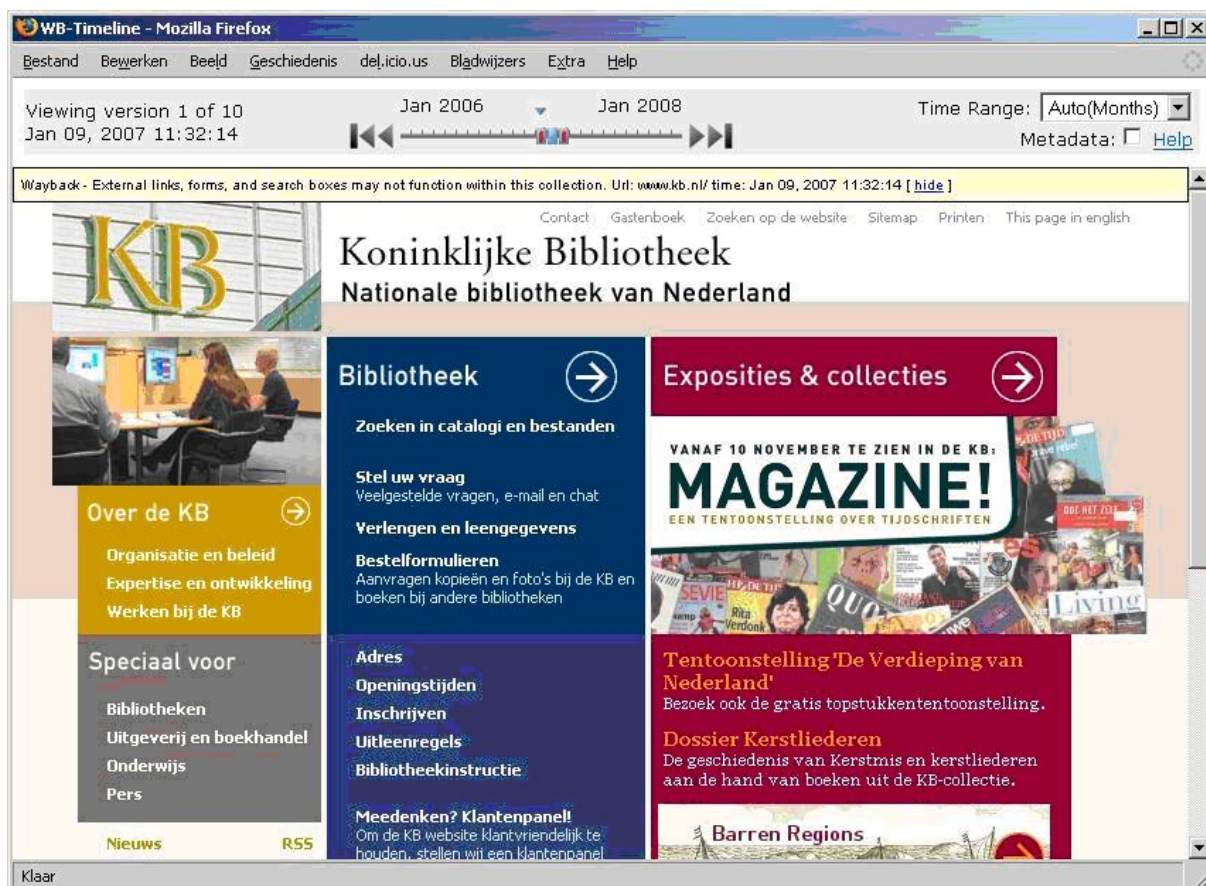


Eerste fase Webarchivering

Koninklijke Bibliotheek



Marcel Ras

September 2007

Evaluatie van de eerste fase van het project webarchivering. Uitgevoerd door de Koninklijke Bibliotheek tussen december 2005 en juni 2007.

Marcel Ras, september 2007.

Inhoudsopgave

1. Inleiding.....	4
2. KB en webarchivering	4
2.1 Webharvesting versus webarchivering.....	5
2.2 Selectie	5
2.3 Eerste fase webarchivering	6
3. Hoe werkt webarchivering?.....	6
3.1 Selecteren en crawlen	8
3.2 Beschrijving van het crawl-proces	8
3.3 Opties voor zoeken en presentatie van gearchiveerde websites	10
3.4 Indexeren, zoeken en presenteren.....	12
3.5 Infrastructuur	13
4. Digitale Duurzaamheid	14
4.1 Het e-Depot.....	14
4.2 Web ARChiving file format.....	15
4.3 Duurzame opslag	17
4.4 Onderzoek bestandsformaten.....	17
5. Juridische aspecten.....	19
5.1 De Auteurswet.....	19
5.2 Bescherming van de persoonlijke levenssfeer	20
5.3 Pragmatische aanpak.....	21
6. Gebruikersonderzoek.....	21
7. Tweede fase webarchivering.....	23

Bijlagen

1. Inleiding

In 2006 koos het Amerikaanse weekblad Time Magazine 'de internetter' tot belangrijkste persoon van het jaar¹. Het blad motiveerde deze keuze door te wijzen op de nieuwe vormen van gemeenschapszin en samenwerking die 'de internetter' aan de dag legt op een wijze en schaal die de wereld nog niet eerder zag. Het ongekende succes van Wikipedia, YouTube en Second Life zijn volgens Time Magazine voorbeelden van 'initiatieven die niet alleen de wereld veranderen, maar ook de wijze waarop de wereld verandert.' De gebruiker is niet langer alleen consument, maar ook een heel belangrijke producent van content. Het internet heeft een enorme impact op het dagelijkse leven, de wijze waarop wij informatie aanbieden en vergaren en de wijze waarop we communiceren en zaken doen.

De invloed van het web is ook in Nederland zichtbaar. Denk maar aan de stemwijzer, die voor vele Nederlanders bepalend was voor hun stemgedrag. Steeds vaker schaffen we artikelen aan via internet of baseren we aankopen op vergelijkingssites. Overbodige bezittingen verruilen van eigenaar via marktplaats en bands worden beroemd via YouTube. Wie de wereld iets te vertellen heeft houdt een weblog bij en ook voor het boeken van een reis en het betalingsverkeer verlaten we ons op internet. Volgens onderzoek van de Stichting Internet Reclame is in de eerste zes maanden het internetgebruik in Nederland gestegen met 8% ten opzichte van 2006. Bijna 11 miljoen mensen van 13 jaar of ouder besteden gemiddeld 8 uur per week op het internet².

Ook binnen de overheid, cultureel erfgoed en wetenschappen is het internet uitgegroeid tot het belangrijkste medium voor het verspreiden, uitwisselen en vergaren van informatie. Meer en meer wordt deze informatie nog uitsluitend op het web gepubliceerd. De omvang van het Nederlandse deel van het World Wide Web groeit fors. Dit .nl-domein bestond in 2006 pas 20 jaar, maar is niet meer weg te denken uit de Nederlandse samenleving. In 2006 werd de 2 miljoenste .nl-domeinnaam geregistreerd³. Daarmee is Nederland wereldwijd het op 3 na grootste "country code Top Level Domain"⁴.

De groeiende afhankelijkheid van het web heeft echter een keerzijde. Het gemak waarmee informatie verwijderd of gewijzigd kan worden maakt het web zeer kwetsbaar. Daar staat tegenover dat we het web steeds meer gaan beschouwen als cultureel erfgoed.⁵ Alle aspecten van onze moderne cultuur zijn vertegenwoordigd op het World Wide Web. Publicaties, debat, kunst, werk en sociale interactie hebben allemaal hun aanwezigheid op het internet en het web. Vaak is informatie op het web zeer vluchtig en heeft het een korte levensduur. 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. Of zoals Malcolm Gillies het verwoordt heeft, "*The daily loss is already huge and we are at risk of losing large parts of our culture*"⁶.

2. KB en webarchivering

In 2006 is de Koninklijke Bibliotheek gestart met het archiveren van een selectie van Nederlandse websites. Als nationale bibliotheek heeft de KB niet alleen de taak gedrukte publicaties duurzaam te

¹ <http://www.time.com/time/magazine/article/0,9171,1570810,00.html>

² <http://stir.web-log.nl/stir/2007/07/internetconsump.html>

³ Stand op 25 juli 2007: 2,48 miljoen .nl domeinnamen geregistreerd. www.sidn.nl

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

⁵ Zie artikel 1 van de UNESCO charter on the Preservation of the Digital Heritage.

http://portal.unesco.org/ci/en/ev.php-URL_ID=13367&URL_DO=DO_TOPIC&URL_SECTION=201.html

⁶ Malcolm Gillies, Born Digital Born Free? The Cultural Impact of the Web. Keynote Address, *Archiving Web Resources Conference, National Library of Australia* (9 November 2004).

http://www.nla.gov.au/webarchiving/about_speakers.html#mgillies

bewaren, maar ook elektronische publicaties. Omdat steeds meer publicaties in elektronische vorm verschijnen, is het duurzaam bewaren en toegankelijk houden hiervan, zoals websites, een belangrijke taak geworden.

Als nationale bibliotheek is de KB verantwoordelijk voor het verzamelen, beschrijven en bewaren van in Nederland verschenen publicaties. Omdat steeds meer publicaties in elektronische vorm verschijnen is het duurzaam bewaren en toegankelijk houden van elektronische publicaties een belangrijke taak geworden. Ook websites kunnen worden beschouwd als publicaties; de KB heeft daarmee een taak in het verzamelen, beschrijven, bewaren en toegankelijk maken hiervan.

Daar waar de meeste internationale initiatieven zich al in een vroeg stadium richtten op het harvesten van websites en over het algemeen nog steeds deze aanpak hanteren, richt de KB zich nadrukkelijk op het duurzaam bewaren en presenteren van gearcheverde websites. Dit betekent dat websites niet alleen geharvest worden, maar dat er tevens een strategie voor de toegang op lange-termijn ontwikkeld wordt.

De complexiteit hiervan is de reden waarom de KB pas in 2006 gestart is met webarchivering. De KB heeft van meet af aan het belang gezien van de digitale uitbreiding van de nationale depotfunctie en heeft ook daadwerkelijk stappen gezet. Vanaf 1995 heeft zij geïnvesteerd in onderzoek naar en de ontwikkeling en inrichting van een elektronisch depot. Met dit 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. Daarnaast heeft de KB nationaal en internationaal een vooraanstaande positie als het gaat om onderzoek naar digitale duurzaamheid van elektronische publicaties. Doordat er eerst geïnvesteerd is in deze technische basisinfrastructuur, kon vervolgens worden begonnen met webarchivering.

2.1 Webharvesting versus webarchivering

De begrippen webarchivering en webharvesting worden vaak door elkaar gebruikt, terwijl er een duidelijk verschil is. De term *webharvesting* wordt gebruikt als overkoepelende term voor het selecteren van relevante informatie en het binnenhalen daarvan.

Het begrip 'webarchivering' (*webarchiving*) duidt binnen de internationale wereld van digitale duurzaamheid doorgaans op het (duurzaam) opslaan van webbronnen (documenten en websites).

Bij webarchivering gaan we dus verder dan alleen het binnenhalen van websites (crawlen of harvesten genoemd), we zorgen er ook voor dat de gecrawelde sites gearcheveerd worden, dat wil zeggen (duurzaam) opgeslagen. Omdat het bewaren voor de lange termijn niet zonder presentatie kan, is de toegang tot gearcheverde websites onlosmakelijk verbonden met het opslaan daarvan.

2.2 Selectie

Er zijn twee basisstrategieën voor webarchivering. De eerste strategie is gericht op het automatisch harvesten van, meestal een grote hoeveelheid, websites (bulkarchivering van bijvoorbeeld een nationaal domein). De tweede strategie selecteert op basis van een specifiek selectiebeleid. Beide strategieën hebben voor- en nadelen. 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 en het mogelijk is om de volledige sites tot op het diepste niveau te archiveren. In een selectieve aanpak kan meer op maat worden gearcheveerd, waarbij ook de frequentie van archiveren per site kan worden bepaald.

De KB heeft er dan ook voor gekozen deze aanpak te gebruiken. Daarbij hebben nog andere redenen een rol gespeeld. Op de eerste plaats is het door het ontbreken van depotwetgeving in Nederland erg lastig om zonder enige vorm van toestemming websites te archiveren. Een selectieve aanpak geeft de

mogelijkheid om voorzichtiger te opereren⁷. Ten tweede richt bulkarchivering zich op het maken van een zg. *snapshot*. Dat betekent dat er strikte grenzen zijn voor de hoeveelheid documenten en de hoeveelheid aan data die gearchiveerd kan worden per website, om te voorkomen dat er een onmogelijke hoeveelheid aan data en bestanden wordt verzameld. Aangezien voor de KB het uitgangspunt voor webarchivering het duurzaam bewaren in het e-Depot 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.

Vooralsnog zal de KB bij de selectie van te archiveren websites zich baseren op haar eigen collectiebeleid⁸. Binnen dit kader zal er een beredeneerde selectie gemaakt worden die bestaat uit een dwarsdoorsnede van het Nederlandse webdomein. Overigens is het Nederlandse webdomein een breed begrip dat zich zeker niet beperkt tot het .nl domein, maar alle in Nederland geregistreerde websites bevat. Primair zullen websites met wetenschappelijke en culturele content geselecteerd worden, maar daarnaast ook websites met een innovatief karakter die exemplarisch zijn voor de huidige trends op het Nederlandse deel van het web. Een volgende stap zal zijn om samenwerking te zoeken met andere kennisinstituten om op die manier de selectie te verbreden en daarbij gebruik te maken van de inhoudelijke expertise van deze organisaties. Verder kan er gedacht worden aan het bieden van een mogelijkheid om zélf websites op te geven ter archivering.

2.3 Eerste fase webarchivering

Van december 2005 tot en met april 2007 heeft de afdeling Digitale Duurzaamheid, Hoofdafdeling Research & Development van de Koninklijke Bibliotheek (KB) het project *eerste fase webarchivering* uitgevoerd. Doel van dit project was het verkrijgen van inzicht in de technische, organisatorische, juridische en functionele mogelijkheden, de kosten en haalbaarheid van een webarchief bij de KB ten behoeve van het duurzaam behoud van een selectie van Nederlandse websites.

Vooraf waren er twee hoofddoelstellingen geformuleerd:

1. Beschrijven van het gehele proces, inclusief het laden en terugleveren van gearchiveerde websites vanuit het e-Depot. Resultaat is een *proof-of-concept*.
2. Opdoen van kennis, ervaring en inzichten die dienen als aanbeveling ten behoeve van besluitvorming en een vervolgfase die gericht moet zijn op het operationaliseren van de activiteit webarchivering (2^{de} fase webarchivering)

De eerste fase is uitgevoerd en geëvalueerd. Op basis van de opgedane kennis en ervaring is gestart met de tweede fase van het project. De tweede fase van het project is gericht op het inbedden van een infrastructuur voor webarchivering en het opschalen van de selectie. Het aantal te archiveren websites zal jaarlijks groeien, waarbij de geselecteerde sites een aantal malen per jaar gearchiveerd zullen worden.

3. Hoe werkt webarchivering?

Nadat is bepaald welke websites er gearchiveerd dienen te worden, is de volgende stap het binnenhalen van deze websites (harvesten of crawlen genoemd) met behulp van speciaal daarvoor ontwikkelde software⁹ (zie voor een schematische weergave van de workflow bijlage 2). Dit is sterk vergelijkbaar met wat de crawlers van zoekmachines als Google doen, met dit verschil dat de crawler

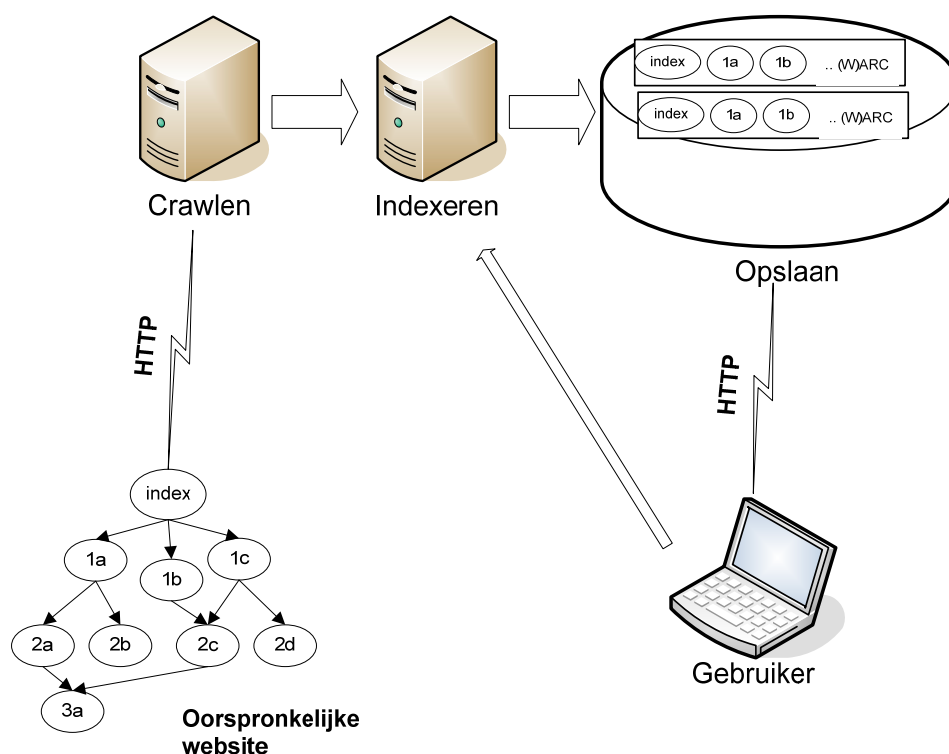
⁷ Online publicaties zijn inbegrepen in de depotwetgeving van ondermeer Denemarken (<http://www.bs.dk/content.aspx?itemguid=%7B332484E6-A5B1-4CEE-B953-059843182050%7D>.) en Duitsland (http://www.ddb.de/aktuell/presse/pressemitte_dnb_neu.htm). In Frankrijk bereidt men momenteel uitbreiding van de bestaande depotwetgeving voor zodat er ook websites onder zullen vallen.

⁸ Zie het collectieplan 2006-2009: <http://www.kb.nl/hkc/collectieplan/collectieplan.html>

⁹ De KB maakt, zoals de meeste webarchiveringsprojecten, gebruik van de Open Source crawler Heritrix. Deze is ontwikkeld door de Internet Archive. Zie: <http://crawler.archive.org/>

van een webarchief daadwerkelijk alle bestanden van een website probeert binnen te halen. Op basis van het te archiveren domein (bijvoorbeeld www.kb.nl) volgt de crawler alle links vanaf de startpagina. In het voorbeeld hieronder volgt de crawler dus vanaf de indexpagina de verwijzingen naar 1a, 2a, 3a, 1b en verder. Totdat de volledige oorspronkelijke website is binnengehaald. De KB streeft ernaar om alle bestanden waaruit een geselecteerde website is opgebouwd te archiveren (voor zover de techniek en eventuele beveiliging dat toelaat). Het is echter ook mogelijk om grenzen te stellen aan de hoeveelheid te harvesten data en het maximale aantal bestanden met betrekking tot één website. Een dergelijke aanpak wordt toegepast bij het maken van een snapshot. De crawler kan zodanig worden geconfigureerd dat deze daar rekening mee houdt. In het hieronder gegeven voorbeeld beperkt de crawler zich tot de documenten binnen het aangegeven domein (kb.nl), maar het is mogelijk om daar buiten te treden en de crawler te vertellen dat hij een aantal stappen naar “buiten” mag volgen. Op die manier wordt de bewust gemaakte selectie verbreed met de links die een crawler tegenkomt.

Wanneer een website (of een aantal websites) is binnengehaald, wordt deze full-text geïndexeerd. Websites zijn over het algemeen opgebouwd uit een grote hoeveelheid losse bestanden, zo bestaat de website van de Eerste Kamer uit ruim 130.000 bestanden. De door de KB gebruikte crawler Heritrix “verpakt” al deze losse bestanden in een soort “container” waardoor de gearchiveerde versie van de site makkelijker te beheren is. Deze verpakking kan worden beschouwd als een soort ZIP bestand met het verschil dat ieder los bestand voorzien wordt van een metadataomschrijving. Deze metadata bevat informatie over het bestandsformaat, tijd en datum van crawlen en de omvang van het bestand. Voordat de gecrawelde websites opgeslagen kunnen worden in het e-Depot wordt er een kwaliteitscontrole uitgevoerd. Daarbij wordt in eerste instantie gekeken naar de kwaliteit van de binnengehaalde; missen er onderdelen en kloppen de links in een site. Vervolgens worden er gegevens met betrekking tot de verschillende bestandsformaten en versies daarvan verzameld. Het zijn vooral deze gegevens die van belang zijn voor toekomstige presentatie. Deze gegevens worden als technische metadata opgeslagen in het e-Depot. De index wordt buiten het e-Depot opgeslagen. Een gebruiker zoekt met behulp van een specifiek voor het webarchief ontwikkelde zoekmachine in de index. Dit kan full-text, maar ook op basis van een specifieke URL. Het resultaat van de zoekvraag wordt opgehaald uit het e-Depot en gepresenteerd in een interface die behalve de gevraagde versie van de betreffende site ook de mogelijkheid biedt om via een tijdbalk eerdere en latere versies van deze website te raadplegen.



3.1 Selecteren en crawlen

Webarchivering begint met de selectie van de te archiveren websites. Selectiestrategie is mede afhankelijk van het doel waarmee websites verzameld worden. De twee hoofddoelen zijn:

1. vanuit een collectieperspectief: bewaren als cultureel erfgoed
2. vanuit een archiefperspectief: bewaren als bewijs en verantwoording van het handelen van overheden

Selectie van te archiveren websites gaat uit van het perspectief van collectie vorming. Voor de KB is dit gebaseerd op het collectiebeleid. Uitgaande van het collectiebeleid zijn er selectiecriteria opgesteld voor het selecteren van te archiveren websites. Een schematische weergave van het selectieproces is gegeven in bijlage 1. Selectie start met de identificatie van een website of aantal websites. Deze worden vervolgens getoetst aan de selectiecriteria, waarna de kwaliteit van de site wordt bekeken. In het geval dat de site voldoet aan de selectiecriteria en aan kwaliteitseisen, wordt deze opgenomen in de selectielijst. De websitehouder wordt bericht dat de KB de betreffende website zal gaan archiveren (zie de paragraaf over juridische aspecten). De URL wordt vervolgens ingevoerd in de crawler die de site éénmalig binnenhaalt. In de eerste fase is dit proces volledig handmatig uitgevoerd. Dat werkt wanneer er niet meer dan 100 websites geselecteerd worden, maar is onbegonnen werk wanneer dit aantal groter wordt. Er zal dan ook een beheerstool gebruikt moeten worden om dit proces uit te voeren en te bewaken. Hiervoor zal mogelijk de Web Curator Tool gebruikt worden¹⁰.

Tijdens de 1^{ste} fase is er een beperkt aantal van 100 websites geselecteerd en gecrawld. In totaal werd er in deze crawl ruim 360 GB aan data verzameld en ruim 16 miljoen bestanden waaronder 200 verschillende bestandsformaten:

- gecrawled: 100 websites
- 364 GB aan ongecomprimeerde data
- 180 GB aan gecomprimeerde data
- deze 180 GB is “ingepakt” in 2.314 ARC-bestanden
- de crawl bestaat uit ruim 16 miljoen unieke bestanden
- waarvan meer dan 200 verschillende bestandsformaten (MIME-types)
- de crawler heeft 92 dagen nodig gehad om dit te crawlen¹¹

Voor het crawlen van websites wordt gebruik gemaakt van de open source Heritrix crawler, versie 1.10¹². Deze wordt door de meeste webarchiveringsprojecten wereldwijd gebruikt en is onderdeel van de zg. IIPC-toolset¹³. Heritrix draait op een Linuxserver¹⁴ en maakt gebruik van het http-protocol om websites binnen te halen. Het resultaat van een harvest is een verzameling *container*-bestanden (ARC-bestanden) waarin de losse files waaruit een website is opgebouwd zijn samengevoegd.

3.2 Beschrijving van het crawl-proces

Een typische crawl van een website begint met het invoeren van het webadres van een bedrijf of organisatie in de webcrawler, <http://www.kb.nl> bijvoorbeeld. Dit start-adres wordt ook wel de *seed*

¹⁰ <http://webcurator.sourceforge.net/>

¹¹ Aangezien er meerdere Heritrix-instanties op de crawlserver actief waren, overstijgt de looptijd van Heritrix de werkelijke testtijd van 45 dagen (vanaf begin januari tot halverwege februari).

¹² Heritrix is ontwikkeld in Java binnen een Open Source licentie en draait bij voorkeur in een Linuxomgeving. Meer over Heritrix, zie: <http://crawler.archive.org/>

¹³ De IIPC is het International Internet Preservation Consortium. De KB is lid van de IIPC. Voor meer informatie over dit consortium, zie: www.netpreserve.org

¹⁴ De technische gegevens van de gebruikte server: 64 bit Dual Core Intel Xeon, 3.0 GHz; 32 bit OS: RedHat Linux 4 ES; 4 GB RAM; 73 GB SCSI harde schijf; Gigabit ethernet NIC.

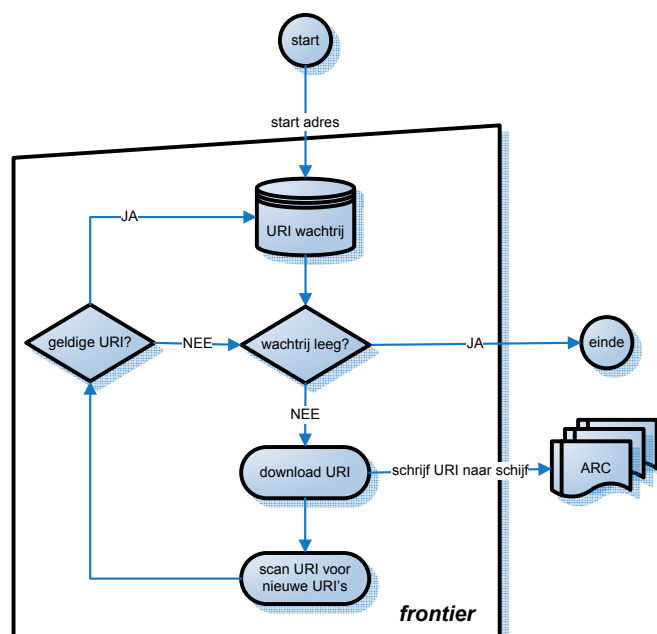
genoemd. Het eerste wat de webcrawler controleert is of de beheerder van de website restricties heeft opgelegd t.o.v. webcrawlers in het *robots.txt* bestand. Dit is een plat tekst bestand op de *root* van de te crawlen website, <http://www.kb.nl/robots.txt> in dit geval, waarin vermeld staat waar bepaalde, of alle crawlers wel en niet mogen crawlen. Het kan voorkomen dat in dit bestand alle crawlers niet toegelaten worden tot de website. In dat geval wordt de crawl direct afgebroken. In alle andere gevallen wordt de *seed* toegevoegd aan de, tot dan toe nog lege, wachtrij van de *frontier*.

De *frontier* is het centrale punt van een webcrawler. Het is verantwoordelijk voor het complete traject dat een digitaal object (URI) zoals een HTML pagina of een JPG bestand aflegt binnen de webcrawler. Om de te crawlen webserver te ontzien wordt altijd maar één URI te gelijk gedownload. Zodra de URI is gedownload, wordt het in een ARC container op een lokale harde schijf geschreven en, indien mogelijk, wordt het gescand op de aanwezigheid van nog niet gecrawelde URI's. Deze lijst nog niet gecrawelde URI's wordt vervolgens door een serie *processors* gehaald welke bepalen of een bepaalde URI wel of niet aan, én op welke plaats in de wachtrij wordt geplaatst.

Een *processor* is eigenlijk niets meer dan een vraag welke met JA of NEE beantwoord kan worden. Vooraf wordt bepaald hoe de *frontier* dient te reageren op de uitkomst(en) van een *processor*. Enkele voorbeelden van *processors* zijn:

- bevindt de URI zich in hetzelfde domein als de *seed*?
 - JA: voeg URI toe aan de wachtrij
 - NEE: negeer URI
- zijn er in totaal meer dan X URI's of X MBytes gedownload?
- eindigt de URI op .AVI?

Hieronder staat een schematische tekening van het hierboven beschreven harvest proces.



Om te voorkomen dat de webserver van de te archiveren site overbelast wordt, zijn er een aantal marges ingebouwd in de crawler. Zo zal de crawler tussen ieder te downloaden digitale object (URI) steeds een korte tijd wachten en is de snelheid waarmee bestanden gedownload worden beperkt, om te veel dataverkeer te beperken. De crawler probeert alle bestanden van een website binnen te halen, maar zal per bestand maximaal vijf maal proberen een URI binnen te halen.

Een van de opvallende zaken tijdens deze eerste crawl was dat een aantal grote websites ook daadwerkelijk héél groot bleken te zijn, veel groter dan verwacht. Dit waren voornamelijk universitaire websites die volledig uit een CMS opgebouwd zijn. Van de verschillende universitaire websites is tussen de 5 à 10 miljoen bestanden gecrawled (100 - 200 GB per website) alvorens de crawl werd afgebroken. In alle gevallen was de website tussen de 30- en 40% compleet, wat betekent dat er meer documenten in de *wachtrij* stonden dan dat er gecrawled waren. Naar verwachting zou een crawl van de gehele website tussen de 300 GB en 1 TB aan data opgeleverd hebben.

3.3 Opties voor zoeken en presentatie van gearchiveerde websites

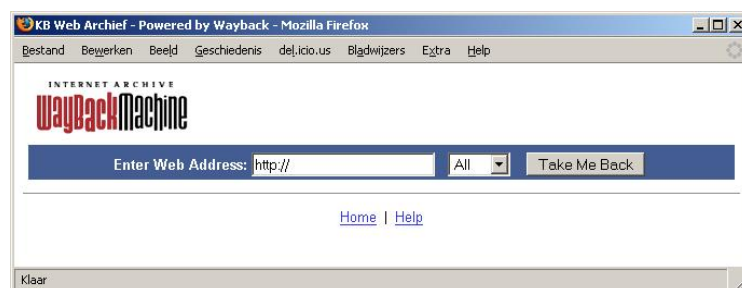
Om een webarchief te doorzoeken zijn er een aantal mogelijkheden:

1. zoeken op specifieke URL (zie Internet Archive);
2. full-text zoeken (de Google-manier);
3. zoeken via de KB-catalogus.

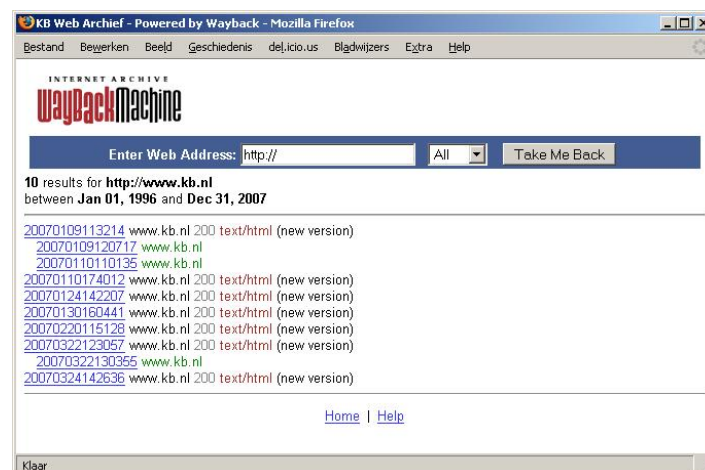
Daarbij zijn er voor de opties 1 en 2 specifieke tools ontwikkeld:

- a. de Open Source Wayback Machine ontwikkeld door de IA;
- b. WERA, ontwikkeld door de Noorse nationale bibliotheek.

Met de Wayback Machine is het alleen mogelijk om te zoeken op specifieke URL's. Dit is een grote beperking omdat de gebruiker deze URL moet kennen om iets te kunnen vinden.



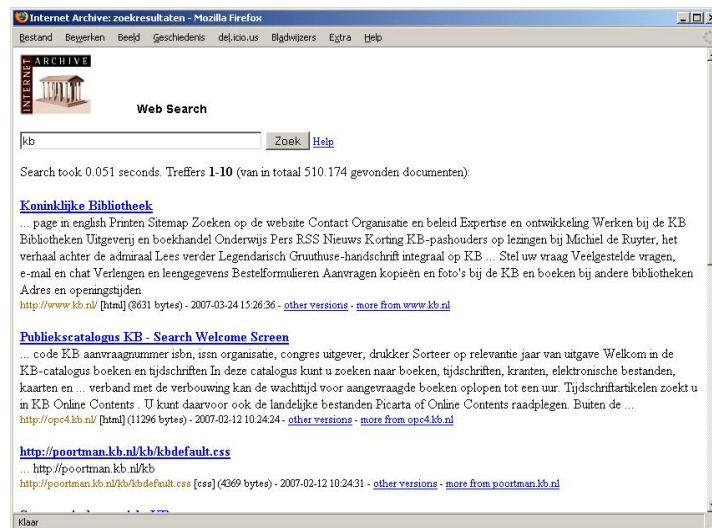
Als zoekresultaat worden de data van de verschillende gearchiveerde versies van de gezochte webpagina gepresenteerd. Datumweergave is een timestamp in jaar, maand, dag, uur, minuut, seconde. Alle documenten die zijn gearchiveerd kunnen worden gevonden. De gebruiker moet echter wel de exacte URL daarvan kennen.



Om te kunnen zoeken op volledige tekst is er in de testfase gebruik gemaakt van een combinatie van NutchWax (voor het indexeren en opbouwen van de zoekfunctionaliteit) en de Wayback Machine (voor het presenteren). Met NutchWax kan de gebruiker gearchiveerde documenten op een Google-achtige wijze vinden op basis van vrije tekst invoer. De zoekresultaten bestaan uit: De titel van de pagina, een

korte tekst, de URL, het bestandsformaat, de grootte (in data), de datum en tijd van archivering en twee extra mogelijkheden: “Other versions”, een lijst van alle data waarop de betreffende pagina is gearchiveerd en “More from this site” waarbij de zoekopdracht binnen de website wordt uitgevoerd.

Het voordeel hiervan is dat de gebruiker kan zoeken in het archief op basis van vrije tekstinput, hij hoeft daarvoor geen URL's te kennen. Het nadeel is echter dat het resultaat vaak een grote hoeveelheid treffers oplevert, vaak treffers die minder relevant zijn voor de oorspronkelijke zoekvraag. Daarnaast kunnen de huidige zoekmachines nog niet overweg met de dimensie tijd die belangrijk is voor een webarchief.

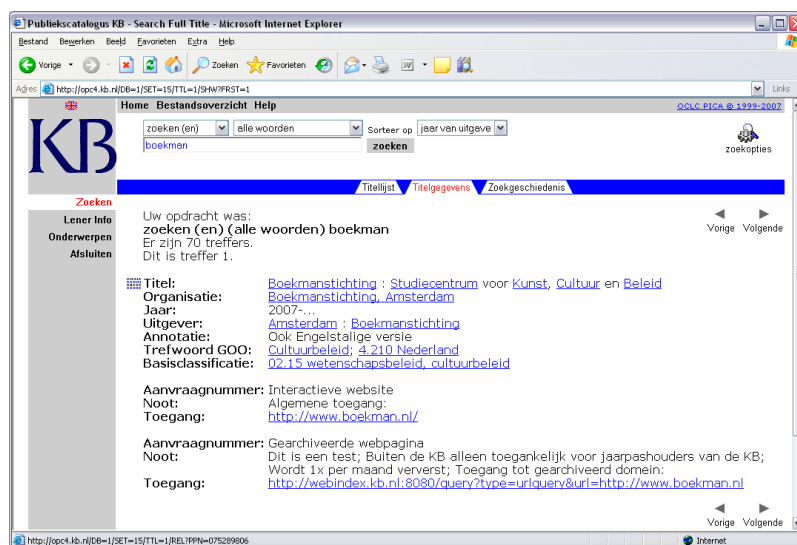


WERA is een portal, gebouwd om de full-text zoekfunctionaliteit van NutchWax heen. Deze zoekinterface stuurt een query door naar NutchWax, ontvangt de RSS-feed van resultaten en verwerkt deze weer in zijn eigen interface. Omdat WERA dit allemaal zelfstandig verwerkt, kan het extra functionaliteit geven ten opzichte van NutchWax. Een interessante functionaliteit van WERA is de tijdbalk die wordt toegevoegd boven in de pagina. Bij meerdere versies van een document kan door de tijdbalk te gebruiken snel andere versies getoond worden, wat erg intuïtief werkt. Het nadeel van WERA is voornamelijk de stabiliteit. Deze laat namelijk helaas nogal te wensen over.



De 3^{de} manier van zoeken in het webarchief gaat uit van de KB-catalogus. Hierin kan gezocht worden op auteur, genre en trefwoorden. Behalve boeken worden hier ook elektronische tijdschriften en digitale bronnen opgenomen. De 100 gearchiveerde websites zijn beschreven op het hoogste niveau

van de website. Er is dus alleen een globale inhoudelijke beschrijving van de website, niet van de specifieke delen waaruit de website bestaat. Het maken van een titelbeschrijving van een gearchiveerde website is gedaan in analogie met die van tijdschriften. De gebruiker kan zoeken in de centrale catalogus van de KB, van waaruit hij doorverwezen wordt naar een webarchief specifieke interface nadat hij een resultaat gevonden heeft.



3.4 Indexeren, zoeken en presenteren

Voor het indexeerproces en het toegankelijk maken van de gearchiveerde website is gebruik gemaakt van de software uit de IIPC toolset. Deze toolset omvat de volgende applicaties:

- Heritrix – crawlen van websites
- NutchWax/Hadoop – *fulltext* indexeren van de crawls
- Wayback Machine / WERA – op URL en *fulltext* doorzoeken van de crawls¹⁵

Deze tools zijn zo ingericht dat ze van elkaars functionaliteit gebruikmaken, het geheel is te beschouwen als een compleet pakket voor webarchivering (met uitzondering van lange termijn opslag).

Ook voor dit aspect van webarchivering geldt dat de KB inzicht wil krijgen in het proces van webarchivering en daarom deze stappen zelf wil doorlopen en zelf de benodigde infrastructuur wil opzetten. NutchWax, Hadoop, WERA en Wayback Machine draaien momenteel op één server. In de testfase werkt dit prima, maar bij het toenemen van de hoeveelheid data wordt het snel duidelijk dat juist deze processen veel vragen van het geheugen van een server. Indexeren en toegankelijk maken zijn dan ook processen waarvoor aparte machines nodig zijn.

Het streven is om het webarchief *fulltext* te indexeren en doorzoekbaar te maken. Dit houdt in dat de gebruiker van het archief, door middel van een aantal woorden op te geven, op een “Google-achtige” manier resultaten ontvangt die aan de zoekvraag voldoen. Een dergelijke indexeertool moet dus in staat zijn een index op te bouwen van woorden, met verwijzingen naar de juiste webpagina. Omdat ons archief met *zg. bitstream containers* werkt (de ARC-bestanden), moet de index ook gegevens bevatten waar en in welke container het bestand te vinden is. Hiervoor wordt gebruik gemaakt van NutchWax¹⁶.

¹⁵ Nutchwax: <http://archive-access.sourceforge.net/projects/nutch/>. WERA: <http://archive-access.sourceforge.net/projects/wera/>

¹⁶ NutchWax is een tak van Nutch, voorheen beter bekend als Lucene, speciaal geschreven met het doel een webarchief te indexeren. NutchWax staat voor Nutch With Archive eXtentions. Tot nu toe hebben we gebruik gemaakt van deze open source toepassing. Standaard indexeerssoftware gebruikt door de KB is Verity. Daarmee zal er de komende tijd geëxperimenteerd worden.

Om de snelheid en de precisie van Google te evenaren, wordt gebruik gemaakt van Hadoop. Deze applicatie optimaliseert de index om snel resultaat te kunnen leveren. Verder biedt Hadoop een omgeving waardoor het indexerende gedistribueerd, over meerdere computers tegelijk, kan verlopen.

Het indexerende van gecrawelde websites kost zeer veel tijd en vraagt veel van een processor. Het gaat daarbij niet om minuten, maar om uren en soms zelfs dagen. Het duurt ongeveer twee uur om een website van 100-200 MB te indexerende. Om de echte grote websites te indexerende, hebben we zeker een dag nodig. De geïndexeerde website wordt vervolgens aan de algemene index toegevoegd. Ook dit is een tijdrovende kwestie. Het is mogelijk om het indexerende proces gedistribueerd uit te voeren, op meerdere machines tegelijk. Het Internet Archive maakt voor haar indexerende projecten gebruik van 300 machines!

Een index zelf heeft ook behoorlijk wat schijfruimte nodig. De eerste crawl bevat ruim 360 GB aan data, de index daarvan is ongeveer 60 GB. Dat betekent dat de index momenteel 1/6 deel van een crawl beslaat. Het is echter geen vast gegeven dat een index altijd 1/6 deel is, dit is sterk afhankelijk van het voorkomen van woorden in een website. Wel moeten we er rekening mee houden dat de index een behoorlijke omvang zal krijgen wanneer we gaan opschalen.

Tot dusver zijn er 100 unieke websites binnengehaald, en volledig doorzoekbaar gemaakt. Maar dit alleen is uiteraard niet genoeg, deze moeten ook getoond kunnen worden. De pagina die de gebruiker in het archief wil zien, moet dan ook opnieuw opgebouwd worden vanuit het archief. Dit levert nog regelmatig problemen op.

Om bepaalde objecten op een website te kunnen zien heeft de gebruiker soms specifieke plug-ins nodig. Access-tools weten niet altijd wat met bepaalde plug-ins aanmoeten. Zo kan het voorkomen dat gedeeltes van een gearchiveerde website niet uit het archief gehaald worden, maar uit het “echte” web. Vaak zijn er *harde* links geprogrammeerd (bijvoorbeeld in een Flash-animatie), waardoor er verwezen wordt naar een locatie buiten het archief. Dit zijn problemen waar momenteel nog hard aan gewerkt wordt.

In de testfase is gebruik gemaakt van twee verschillende versies van de open source versie van de Wayback Machine: 0.6.0 en 0.8.0. Het nadeel van de Wayback Machine is dat er alleen op basis van een URL gezocht kan worden. WERA biedt de mogelijkheid om full-text te zoeken in het archief. Daar staat tegenover dat WERA nog onvoldoende stabiel is en er nauwelijks verder ontwikkeld wordt aan deze tool, terwijl de Wayback Machine voortdurend nieuwe releases kent waarin ook van WERA bekende functionaliteit als presentatie via een tijdbalk geïmplementeerd is.

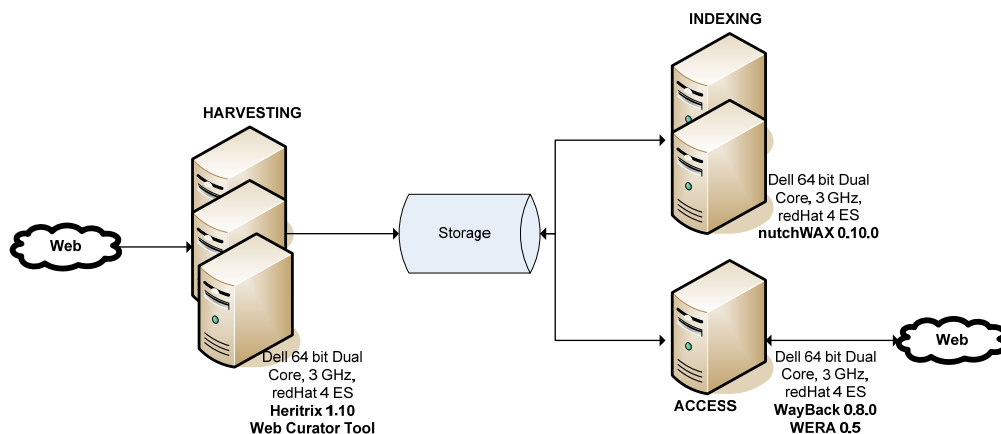
Om te komen tot een goed werkende accesstool, onafhankelijk van de keuze voor WERA of de Wayback Machine, is het van belang om heldere gebruikerseisen vast te stellen. Daarvoor is er een concreet project door de IIPC Working Group on Access geformuleerd waarbij er een lijst van gebruikerswensen opgesteld zal worden die gebaseerd is op verschillende onderzoeken gedaan door de KB, de British Library en de Bibliothèque Nationale de France.

3.5 Infrastructuur

Op dit moment is een groot deel van de hardware nodig voor webarchivering uitbesteedt aan SARA reken en netwerkdiensten in Amsterdam¹⁷. Zij beheren de machines waarop de crawler draait, het indexerende proces en de toegang geregeld is. Opslag vindt plaats in het e-Depot en derhalve binnen de KB zelf. Wel is er een tijdelijke opslagcapaciteit gereserveerd bij SARA om de data te “stallen” direct na het harvesten. Mogelijk dat er voor het indexerende proces gebruik gemaakt gaat worden van Verity. In dat geval zal ook de hardware voor het indexerende proces binnen de KB beheerd worden.

¹⁷ <http://www.sara.nl/>

Tijdens de testfase is er gebruik gemaakt van slechts twee servers waarop de verschillende software draait. Deze infrastructuur zal in de komende fase worden opgeschaald volgens het onderstaande schema.



4. 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 websites over pakweg 50 jaar nog steeds voor die gebruiker toegankelijk zijn? We zullen dan geen gebruik meer maken van de browsers en platforms zoals we die nu gewend zijn te gebruiken en wellicht dat ook het concept van het web volledig is veranderd. Toch zullen we ervoor moeten zorgen dat wetenschappers over 50 jaar hun onderzoek kunnen doen, hun onderzoeksdata kunnen verzamelen en deze kunnen gebruiken. Het is reëel om er vanuit te gaan dan een deel van die onderzoeksdata afkomstig zal zijn uit webarchieven. Dat onze huidige 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 en het bewaren van de juiste technische metadata om later te kunnen bepalen wat het is en op welke wijze dit gepresenteerd moet worden zijn vereisten. Omdat de presentatie van een website sterk afhankelijk is van de gebruikte browser, maar ook van plug-ins noodzakelijk voor de presentatie van specifieke aspecten van een website (zoals Flash, video en audio), is het noodzakelijk om de meest gangbare browsers en plug-ins te bewaren in een software repository. De KB 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)¹⁸.

4.1 Het e-Depot

Een belangrijke taak van de Koninklijke Bibliotheek is het duurzaam bewaren en toegankelijk houden van de nationale productie van elektronische publicaties. Deze productie wordt nu aangeleverd in de vorm van verschillende offline media (cd-roms, tapes, optical disks), en online elektronische publicaties, waaronder elektronische tijdschriften van grote uitgevers. Het gaat hierbij om beschrijving, opslag en beschikbaarstelling van een sterk toenemend aantal elektronische publicaties. Geschat wordt dat het in enkele jaren gaat om honderden terabytes aan data. De elektronische publicaties worden opgeslagen, verwerkt en beschikbaar gesteld door een geautomatiseerd systeem: het e-Depot.

In 1996 sloot de KB een overeenkomst met het Nederlands Uitgeversverbond over het vrijwillig deponeren van offline elektronische publicaties (cd-roms). In hetzelfde jaar werden met de van oorsprong Nederlandse uitgevers Elsevier en Kluwer Academic overeenkomsten gesloten over het

¹⁸ <http://www.netpreserve.org>

duurzaam archiveren van in Nederland uitgegeven wetenschappelijke e-journals. Omdat de plaats van uitgifte in een digitale omgeving minder relevant is, zijn deze overeenkomsten na enkele jaren uitgebreid tot alle elektronische tijdschriften van deze uitgevers. Daarnaast werd afgesproken ook de oudere jaargangen te archiveren.

Voorafgaand aan de realisering van het e-Depot is uitgebreid onderzoek gedaan naar digitale archivering. Zo speelde de KB een belangrijke rol in het NEDLIB-project¹⁹, een samenwerkingsverband van Europese nationale bibliotheken dat van 1998 tot 2000 onderzoek deed naar digitale archivering. De standaarden en richtlijnen die uit dit project zijn voortgekomen worden inmiddels breed gedragen.

Op basis van de NEDLIB-resultaten startte de KB in 1999 een Europese aanbestedingsprocedure voor het ontwikkelen van een elektronisch depotsysteem. IBM heeft vervolgens het systeem ontwikkeld, in nauwe samenwerking met de KB. Het e-Depot is sinds begin 2003 operationeel en ingebed in de organisatie. Het systeem valt onder de verantwoordelijkheid van de afdeling e-Depot, onderdeel van de Hoofdafdeling Verwerking Publicaties, waar ook de papieren publicaties worden aangeleverd en verwerkt.

Het technische hart van het e-Depot is DIAS²⁰, het Digital Information Archiving System. DIAS is gebaseerd op het ISO gecertificeerde OAIS Reference Model²¹ voor digitale archivering. De technische en de organisatorische infrastructuur van DIAS voldoen aan strenge eisen op het gebied van duurzaamheid, flexibiliteit en schaalbaarheid. Aangeleverde publicaties worden met grote hoeveelheden tegelijk gecontroleerd en in het e-Depot geladen.

Inmiddels heeft de KB met diverse (inter)nationale uitgevers contracten afgesloten, waaronder Springer, Blackwell Publishing en Taylor & Francis. Nu al wordt zo'n zeventig procent van het totale aantal wetenschappelijke tijdschriften op het gebied van Science, Technology en Medicine (STM) opgeslagen in het e-Depot. De KB streeft ernaar dit aantal de komende jaren verder uit te breiden. Met elke uitgever worden afspraken gemaakt over de wijze van aanlevering en de toegankelijkheid van bestanden voor het publiek. Het e-Depot is volop in ontwikkeling.

Als beheerder van het e-Depot en dankzij het gespecialiseerde onderzoek dat wordt uitgevoerd, beschikt de KB over de nodige technische en organisatorische expertise op het gebied van digitaal archiveren en permanent access. De KB streeft ernaar met het e-Depot een internationale rol te gaan vervullen in het Safe Place Network, een geografisch gespreid netwerk van digitale archieven die gezamenlijk de verantwoordelijkheid dragen voor het in stand houden van het digitale wetenschappelijke erfgoed.

4.2 Web ARChiving file format

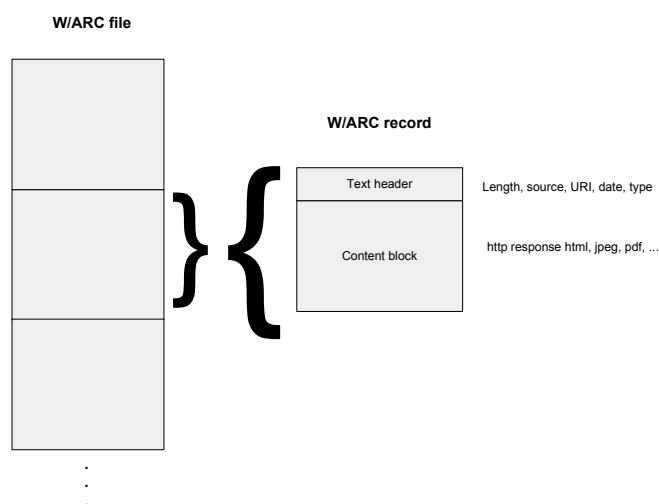
Het resultaat van een harvest uitgevoerd door Heritrix wordt verwerkt in zg. ARC-bestanden. ARC is ontwikkeld door het Internet Archive en voor het eerst gebruikt in 1996 omdat uit ervaring bleek dat het opslaan van miljoenen losse files in een file systeem steeds moeilijker te beheren was.

Een ARC kan beschouwd worden als een *container* waarin de losse bestanden waaruit een website is opgebouwd zijn opgeslagen op een dusdanige wijze dat de onderlinge relaties tussen de bestanden intact blijven. De individuele bestanden worden in *records* opgeslagen, waarbij ieder record bestaat uit een *header* en een *content block*. De *header* bevat metadata met betrekking tot het betreffende bestand dat in het *content block* is opgeslagen. De structuur van een ARC ziet er schematisch weergegeven dan als volgt uit:

¹⁹ <http://nedlib.kb.nl/>

²⁰ <http://www.kb.nl/dnp/e-depot/dm/dias.html>

²¹ http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/oaisbluebook.pdf



Afhankelijk van de omvang van de betreffende websites worden er één of meerdere ARC-bestanden gemaakt. Een ARC bestand kan in theorie een onbeperkte hoeveelheid aan bestanden en data bevatten, maar de ervaring leert dat het beter is om de omvang van een ARC bestand beperkt te houden (100 MB) vanwege het kunnen indexeren en uitpakken van het bestand. Dat betekent dat een website niet altijd zal passen binnen één ARC bestand. In dat geval is het resultaat van een harvest dus een verzameling ARC-bestanden. De 100 websites die in deze 1^{ste} fase zijn geharvest, zijn “verpakt” in 2.314 ARC’s.

De metadata in de headers zorgen ervoor dat de betreffende website kan worden gereconstrueerd. Ieder ARC-bestanden is voorzien van een header op het hoogste niveau, waarin gegevens zijn opgenomen met betrekking tot de betreffende ARC. Deze bevat de volgende metadata:

Bestandsnaam	de naam van het ARC bestand
IP adres	IP adres van de computer die het ARC bestand heeft aangemaakt
Datum	de datum waarop het ARC bestand is aangemaakt
Versie nummer	de versie van het ARC bestand

De *headers* behorende bij de *content blocks* bevatten de volgende metadata velden:

IP adres	het IP adres van het geharveste document
Datum	de datum waarop het document is geharvest
Inhoud type	het type en subtype van het betreffende document (MIME-type)
Offset	de locatie in het ARC bestand waar het document begint (in bytes)
Lengte	de lengte van het document (in bytes)
URL	de URL van het geharveste document

WARC (Web ARCiving file format) is de beoogde opvolger van het ARC formaat, waarbij er een aantal aanpassingen aan de mogelijkheden en structuur van het formaat hebben plaatsgevonden. Motivatie voor deze revisie komt voort uit de discussies en ervaringen binnen de IIPC (International Internet Preservation Consortium). De verwachting is dat WARC een (ISO) standaard zal worden voor structurering, beheer en opslag binnen een webarchief²². WARC zal het standaard output formaat zijn harvesters, zoals Heritrix en tevens input formaat voor indexeer- en acces tools. In WARC zijn een aantal verbeteringen doorgevoerd met betrekking tot harvester support, toegang, uitwisseling, het zélf

²² ISO TC 46/SC 4 N 595 (06/02/2006). Information and documentation – The WARC File Format

kunnen toevoegen van metadata en de opslag van afgeleide bestanden. Hierdoor zou het ook mogelijk zijn om WARC breder te gebruiken dan alleen voor webarchivering²³.

Net als een ARC bestaat een WARC file uit een verzameling van records, waarbij ieder record een digitaal object bevat. Type of aard van een record kan verschillen. Het is mogelijk om additionele informatie op te slaan behorende bij een object, zoals, metadata, informatie met betrekking tot het harvest protocol en resultaten van een conversie.

4.3 Duurzame opslag

Zoals gezegd wordt het resultaat van een crawl opgeslagen in het e-Depot met het doel deze voor de lange termijn te bewaren en te kunnen presenteren. Een belangrijke vraag is de wijze waarop deze websites het beste opgeslagen kunnen worden. Daarvoor zijn er twee mogelijkheden:

1. we slaan de ARC-bestanden op zoals die door de crawler gemaakt worden
2. we slaan de losse bestanden behorende bij een website op in een file-systeem

Beide methoden hebben een aantal voor- en nadelen. Het grote voordeel van opslaan als ARC (en in de toekomst WARC) is dat geen miljoenen losse bestanden in een file-systeem opgeslagen worden, met alle daarbij behorende problemen, maar dat de hoeveelheid bestanden die beheerd moeten worden veel beperkter gehouden kan worden (de 16 miljoen bestanden waaruit de 100 gecrawelde websites zijn opgebouwd kunnen worden opgeslagen in 2.300 ARC-bestanden). Daarbij komt dat ARC gebruikt wordt als een standaard opslagmethode voor webarchieven, het WARC formaat mogelijk een ISO standaard wordt en binnen een ARC bestand de structuur van de website vastgelegd wordt in de bijbehorende metadata.

Het grote nadeel van ARC (en ook WARC) is dat de digitale duurzaamheid daarvan nog onbekend terrein is. Het is een nieuwe laag om de daadwerkelijke files heen. Dit vraagt om een tussenstap wanneer we de bestanden willen presenteren. We zullen dan ook altijd de beschikking moeten hebben over een applicatie waarmee we ARC files kunnen lezen en uitpakken.

Daarnaast moeten we de beschikking blijven hebben over de software die nodig is om een website te presenteren op het scherm. Om een website te kunnen lezen heeft de gebruiker een browser nodig. Dit zorgt nu al voor problemen omdat er verschillende platforms en verschillende browsers gebruikt worden en lang niet alle websites in deze verschillende browsers op dezelfde wijze worden gepresenteerd, ondanks de vele richtlijnen die er zijn. Wanneer we het webarchief online toegankelijk stellen, hebben we dan ook zelf niet of nauwelijks in de hand hoe de gearchiveerde websites getoond worden bij de gebruiker, zeker niet op de langere termijn. Dit is sterk afhankelijk van de door de klant gebruikte browser, de versie en het platform waarop deze draait. We zullen dus moeten leven met het feit dat we het proces van presenteren niet volledig kunnen beïnvloeden.

4.4 Onderzoek bestandsformaten

Zoals gezegd verzamelen we een veelheid aan bestandsformaten, 220 verschillende in de eerste crawl. De meest voorkomende formaten, zowel in hoeveelheid documenten als in hoeveelheid aan data, zijn de voor de handliggende formaten als HTML, JPEG, GIF, PDF en TXT. In de tabel hieronder zijn de tien meest voorkomende bestandsformaten gegeven en de mate waarin ze procentueel gezien voorkomen, in hoeveelheid data en aantallen.

MIME-type	Aantal GB's in %	Aantal URI's in %
text/html	57.85	87.43
application/pdf	11.68	0.74

²³ De Deense nationale bibliotheek heeft onderzocht of WARC een geschikt opslagformaat is. Mads Alhof Kristiansen, *Digital Preservation using the WARC file format*. July 2006.

image/jpeg	10.59	4.19
audio/x-aiff	2.26	>0,01
video/mpeg	1.98	>0,01
application/postscript	1.72	0.12
application/zip	1.29	>0,01
text/plain	1.25	2.13
image/tiff	1.14	>0,01
text/xml	1.10	>0,01

Informatie met betrekking tot de binnengehaalde bestandsformaten wordt in de metadata van de ARC bewaard. We weten dan ook altijd welk formaat het betreffende bestand is en kunnen daar consequenties aan verbinden (met welke software en/of platform te presenteren, wat te doen wanneer het formaat in onbruik dreigt te raken, etc.). Echter, de gegevens met betrekking tot het bestandsformaat is een zg. MIME-type. Deze MIME-type is gebaseerd op de bestandsextensie en informatie uit de header van het http-antwoord van de webserver. Beide gegevens zijn niet erg betrouwbaar en kunnen er gemakkelijk voor zorgen dat er bestandsformaten worden aangemerkt als een formaat wat ze in werkelijkheid niet zijn, wat grote gevolgen voor toekomstige presentatie van deze bestanden kan hebben.

Omdat de informatie van Heritrix niet volledig betrouwbaar is, is het nodig om een controle uit te voeren op de *MIME-type-registratie* van Heritrix. Daarom zijn van 10 websites alle bestanden gecontroleerd met de identificatietool DROID²⁴, welke bijna 800 bestandsformaten identificeert, en met de tool JHove²⁵ welke een 11-tal²⁶ bestandstypen valideert. Let hierbij op het verschil in identificeren en valideren: bij het identificeren van een bestand wordt naar bepaalde punten van het bestand gekeken, waar het valideren van een bestand het gehele bestand controleert. In de onderstaande twee tabellen zijn de resultaten van deze twee controles weergegeven. Beide tools merken voornamelijk (sommige) .XML en .PDF-bestanden aan als *onzeker*, en .CSS, .ICO, .LOG en door PHP gecreëerde HTML pagina's aan als *niet-herkend*.

DROID	tijd (sec.)	gemiddeld URI p/sec.	% van totaal gecontroleerd	herkend		onzeker		niet herkend	
				#	%	#	%	#	%
CBG	8	546	99.5%	4.138	94.3%	14	0.3%	213	4.9%
De Bibliotheken	5	130	27.4%	601	25.4%	29	1.2%	18	0.8%
DEN	14	97	79.4%	1.319	77.1%	19	1.1%	19	1.1%
Edusite	32	538	20.6%	17.012	20.3%	142	0.2%	77	0.1%
De Verdieping van Nederland	31	7	85.9%	204	79.7%	9	3.5%	6	2.3%
Geheugen van Oost	9	443	88.2%	3.979	88.0%	3	0.1%	4	0.1%
Huygens Instituut	10	706	94.1%	6.875	91.7%	143	1.9%	37	0.5%
Museum Boerhaave	6	206	90.2%	1.209	88.1%	18	1.3%	10	0.7%
Tweede Kamer	13	217	94.8%	2.723	91.6%	53	1.8%	42	1.4%
WSF	3	79	88.7%	209	78.3%	14	5.2%	13	4.9%

JHove	tijd (sec.)	gemiddeld URI p/sec.	% van totaal gecontroleerd	herkend		onzeker		niet herkend	
				#	%	#	%	#	%
CBG	19	230	99.5%	768	17.5%	3	0.1%	3594	81.9%
De Bibliotheken	11	59	27.4%	630	26.6%	6	0.3%	12	0.5%
DEN	80	17	79.4%	1.278	74.7%	10	0.6%	69	4.0%
Edusite	81	213	20.6%	9.818	11.7%	94	0.1%	7319	8.7%

²⁴ http://droid.sourceforge.net/wiki/index.php/Development_History

²⁵ <http://hul.harvard.edu/jhove/>

²⁶ ASCII, GIF, HTML, JPEG, JPEG2000, PDF, TIFF, UTF8, WAVE, XML.

De Verdieping van Nederland	6	37	85.8%	184	71.9%	5	2.0%	30	11.7%
Geheugen van Oost	224	18	88.2%	3.706	82.0%	1	0.0%	279	6.2%
Huygens Instituut	185	38	94.1%	4.205	56.1%	130	1.7%	2720	36.3%
Museum Boerhaave	30	41	90.2%	1.153	84.0%	19	1.4%	65	4.7%
Tweede Kamer	34	83	94.8%	2.151	72.4%	21	0.7%	646	21.7%
WSF	13	18	88.7%	174	65.2%	18	6.7%	44	16.5%

De bestandscontrole met DROID en JHove levert veel inzichten op, maar is pas het begin van verder onderzoek naar bestandsformaten binnengehaald via webarchivering en de wijze waarop we daar mee om moeten gaan in de toekomst. Hoewel DROID en JHove relevante resultaten opleveren, zijn beide tools niet ingericht om te gebruiken voor alle mogelijke bestandsformaten die een webharvest oplevert.

De KB blijft verder onderzoek doen naar bestandsformaten, duurzame opslag daarvan en kwaliteitscontrole daarvan. Daarbij zal er bekeken worden welke filechecker geschikt is voor het QA-proces, welke procedures er gevolgd kunnen worden voor de controle van bestanden verzameld via de crawler, op welke wijze we deze gegevens bewaren en welke rol ze zullen spelen bij de vraag hoe we ervoor zorgen dat websites voor de lange termijn bewaard kunnen worden.

5. Juridische aspecten

Een geheel andere uitdaging wanneer we websites willen archiveren is van juridische aard. Aan het archiveren en beschikbaar stellen van websites zijn juridische consequenties verbonden. De handelingen die verricht worden om duurzame toegang van websites te kunnen garanderen, raken o.a. aan het auteursrecht, het databankenrecht, het portretrecht en de bescherming van persoonsgegevens. Om te bepalen hoe de KB kan omgaan met deze ingewikkelde materie, heeft het Centrum voor Recht in de Informatiemaatschappij (eLaw@Leiden) van de Universiteit Leiden in opdracht van de KB een onderzoek uitgevoerd naar de juridische aspecten van webarchivering binnen het Nederlandse recht, met name het auteursrecht en de wet bescherming persoonsgegevens²⁷.

5.1 De Auteurswet

Om na te gaan wanneer webarchivering in het vaarwater komt van het auteursrecht maken we hier onderscheid tussen de drie fasen van het proces:

1. het harvesten van websites
2. het archiveren ervan
3. het weer beschikbaar stellen aan het publiek

Met het crawlen van een website wordt er een kopie gemaakt van de betreffende site. 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 ten behoeve van privépersonen). 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, gemeenten e.d., tenzij het auteursrecht daarop uitdrukkelijk is voorbehouden. 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)

²⁷ A. Beunen, T. Schiphof, Legal aspects of web archiving from a Dutch perspective. Report commissioned by the National Library in The Hague. October 2006.

expliciet heeft verklaard dat het auteursrecht erop niet zal worden ingeroepen (verklaring van publiek domein). Maar meestal is dus wel voorafgaande toestemming nodig.

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 dit raadpleegbaar te houden als de technologie waarmee het toegankelijk gemaakt kan worden, in onbruik raakt. Een belangrijk nadeel is echter dat deze uitzondering niet geldt voor databanken die door het databankrecht worden beschermd. Bij het kopiëren ten behoeve van 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.

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

5.2 Bescherming van de persoonlijke levenssfeer

Op websites die verzamelt, opgeslagen en beschikbaar worden gesteld, kunnen zich ook zogenaamde persoonsgegevens bevinden. Dat 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. Elke nieuwe handeling, oftewel elke 'verwerking' van persoonsgegevens, moet telkens weer aan de eisen van deze wet voldoen. Onder verwerking wordt zo'n beetje alles verstaan wat je met persoonsgegevens kunt doen, dus ook verzamelen, ordenen en weer aan het publiek beschikbaar stellen.

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 gemaakt²⁸. Verwerking is echter ook toegestaan als dit noodzakelijk is voor de *behartiging van het gerechtvaardigde belang van degene die verwerkt*. Lastig is dat er geen harde, eenvoudige regel voorhanden is die zegt of de verwerking aan de eisen van de wet voldoet. Een van de doelstellingen van de Wbp is het verhogen van de transparantie bij de verwerking van persoonsgegevens. Zo eist de wet dat een verwerker zich bekendmaakt bij de personen van wie persoonsgegevens worden verwerkt. Maar dit is niet verplicht als de mededeling van die informatie aan de betrokkenen onmogelijk blijkt of onevenredige inspanning kost, zoals bij webarchivering het geval lijkt te zijn.

²⁸ Naast de "gewone" persoonsgegevens, kan er ook nog sprake zijn van zogenaamde '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. Verwerking van deze gegevens is in principe verboden, tenzij dit gebeurt met toestemming van de betrokkene of als de gegevens door hem/haar zelf duidelijk openbaar zijn gemaakt.

5.3 Pragmatische aanpak

Verscheidende juridische aspecten vormen dus een belemmering wanneer we websites willen archiveren. Een uitgebreide depotwetgeving kan behulpzaam zijn bij de eerste stap, het kunnen crawlen van websites binnen een wettelijk kader. Omdat dit ontbreekt in Nederland moeten we werken binnen de beschikbare juridische kaders als de Auteurswet en de Wet bescherming persoonsgegevens. Tegenover het auteursrechtbelang 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*.

Wanneer we wetgeving als uitgangspunt nemen, dan moeten we een *zg. opt-in* aanpak hanteren. Hierin vragen we vooraf toestemming voor het crawlen, archiveren en toegankelijk maken aan de eigenaar van de website. Bij andere webarchiveringsprojecten is er al enige ervaring opgedaan met deze aanpak, waarbij er schriftelijk om toestemming gevraagd wordt. Deze ervaring leert dat er zeer veel tijd en werk gaat zitten in de administratieve afhandeling hiervan. Op de eerste plaats worden veel verzoeken niet ondertekend teruggestuurd waardoor er herinneringen verstuurd moeten worden. Daarnaast is het onmogelijk om in één overeenkomst van alle rechthebbenden op site-onderdelen toestemming te krijgen, zodat een ondertekening van de eigenaar slechts een schijnzekerheid biedt. Deze aanpak maakt webarchivering in de praktijk dus bijna onmogelijk.

De KB heeft daarom voor een meer pragmatische benadering gekozen, de *opt out* aanpak. Aan de eigenaren van websites wordt een bericht gestuurd dat de KB de site uit erfgoedoverwegingen regelmatig wil gaan harvesten, archiveren en openbaar maken, daarbij wordt een termijn gegeven waarbinnen men toestemming kan weigeren. Blijft weigering uit, dan wordt dit beschouwd als een impliciete of stilzwijgende toestemming²⁹. Het risico op schadeclaims of rechtszaken is via enkele maatregelen te verkleinen. Op de toegangssite 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 geeft 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.

Tijdens de eerste crawl van 100 websites heeft de KB deze opt-out aanpak gehanteerd. Er waren nauwelijks bezwaren tegen de aanpak van de KB. De bezwaren die er gemaakt werden waren hoofdzakelijk van technische aard, zoals de vrees dat de webserver overbelast zou raken vanwege het crawlen. Ondanks de vele juridische haken en ogen is het toch mogelijk gebleken om op een pragmatische manier gebruik te maken van een opt-out aanpak. Het feit dat de KB uitgaat van een vooraf vastgestelde selectie maakt het eenvoudiger te hanteren en verkleint de juridische risico's. De KB zal deze aanpak dan ook hanteren in de volgende fase, waarin de selectie van te archiveren website fors uitgebreid zal worden. Wanneer deze *opt out* benadering ook in dit vervolg succesvol is, dan voorkomt dit een enorme administratieve last en is het voor het eerst dat een dergelijke pragmatische aanpak op grote schaal toegepast wordt voor het archiveren van websites.

6. Gebruikersonderzoek

Omdat de KB nog helemaal aan het begin staat van de ontwikkeling van een webarchief is het mogelijk om op verschillende terreinen te profiteren van de kennis en ervaring opgedaan door andere

²⁹ 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. Niettemin zijn er naast strikt juridisch redenerende rechters dus ook steeds meer pragmatische rechters.

organisaties wereldwijd. Dit kunnen we zeker wanneer het de technische inrichting van een webarchief betreft, maar helaas is er nog weinig bekend over het gebruik en de gebruiker van webarchieven. Een rondgang langs de verschillende webarchieven leverde weliswaar waardevolle informatie op, maar gaf geen coherent beeld van de gebruiker van een webarchief. Wel werd al snel duidelijk dat er een onderscheid gemaakt moet worden tussen archieven, en het gebruik daarvan, die uitgaan van bulkarchivering en de archieven op basis van een beredeneerde selectie. In het eerste geval wordt in principe een heel domein gearchiveerd en kan in potentie iedereen gebruiker zijn (mits het archief toegankelijk is). Een gebruikersonderzoek richt zich in dit geval meer op zogenaamd usability onderzoek, hoe richten we de toegang en zoekfunctionaliteit in. Bij een selectieve aanpak wordt uitgegaan van selectiecriteria, waarbij het potentiële gebruik en de potentiële gebruikers gerelateerd zijn aan de gehanteerde selectie. Gebruikersonderzoek is behalve usability onderzoek ook onderzoek naar doelgroepen en wensen van gebruikers. Inhoud en gebruik van een webarchief zijn dan ook nauw met elkaar verbonden.

De selectieve benadering van de KB vereist dat er kennis van de (potentiële) gebruikers is. Daarom is er een (kwalitatief) gebruikersonderzoek uitgevoerd³⁰. Dit is bedoeld om meer inzicht te krijgen in de wensen van potentiële gebruikers, ten aanzien van de toegang en zoekfunctionaliteit, maar ook ten aanzien van de te selecteren websites. Omdat de inhoud van het webarchief van de KB zich op dit moment nog beperkt tot ongeveer 100 unieke websites, betreft het hier een kwalitatief onderzoek en zijn de resultaten vooral bedoeld om meer inzicht te krijgen in criteria voor selectie en gebruikerswensen. Tijdens het onderzoek is gebruik gemaakt van het webarchief zoals dit in de eerste fase van het project is opgezet. Dit bevat ongeveer 100 unieke websites op het gebied van cultuur, overheid en wetenschap. Binnen het archief kan er gezocht worden op URL (Wayback Machine) en op volledige tekst (NutchWax).

Een vergelijkbaar kwalitatief onderzoek wordt door de Bibliothèque nationale de France (BnF) uitgevoerd. De resultaten van het onderzoek van de BnF en van dit onderzoek zullen worden gecombineerd met een onderzoek naar Access Requirements uitgevoerd in opdracht van de British Library³¹. De resultaten van deze drie onderzoeken zullen op hun beurt worden ingebracht in de Access Working Group van het International Internet Preservation Consortium (IIPC)³². De bedoeling is dat er een definitieve lijst van gebruikerswensen opgesteld wordt aan de hand waarvan bestaande zoekfunctionaliteit en interface voor een webarchief doorontwikkeld kan worden. Daarnaast zal er een methodologie voor gebruikersonderzoek ten behoeve van een webarchief ontwikkeld worden.

De centrale vraag van het onderzoek richt zich op de inhoud en zoekfunctionaliteit van het KB webarchief en de wensen van eindgebruikers en stakeholders ten aanzien hiervan. Deze centrale vraag is opgedeeld in een zestal deelvragen met betrekking tot gebruikers, gebruik en selectie. Om deze vragen te kunnen beantwoorden is er gebruik gemaakt van twee methoden:

1. een doelgroepanalyse
2. observaties van een testpanel

Een van de conclusies van het gebruikersonderzoek was het kunnen zoeken op volledige tekst de absolute voorkeur genoot van de deelnemers aan het onderzoek. Duidelijk is dat voor de deelnemers Google de norm is. Ze verwachten dat een zoekmachine op volledige tekst werkt zoals Google. Een belangrijk punt is de dimensie tijd binnen een webarchief. Dit komt wel tot uitdrukking in de presentatie middels een tijdbalk, maar zal zeker ook in de zoekfunctionaliteit en de presentatie van de resultaten van een zoekvraag verwerkt moeten worden. Daarnaast is een goede hiërarchische presentatie van de resultaten belangrijk.

³⁰ Dit gebruikersonderzoek is uitgevoerd door een student, Sara van Bussel, van de Reinwardt Academie in het kader van een stage. Sara van Bussel, *Gebruikersonderzoek webarchivering*. Juni 2007.

³¹ Simon Wild, *Web Archive Access Requirements*. British Library, June 2007.

³² <http://www.netpreserve.org/about/index.php>

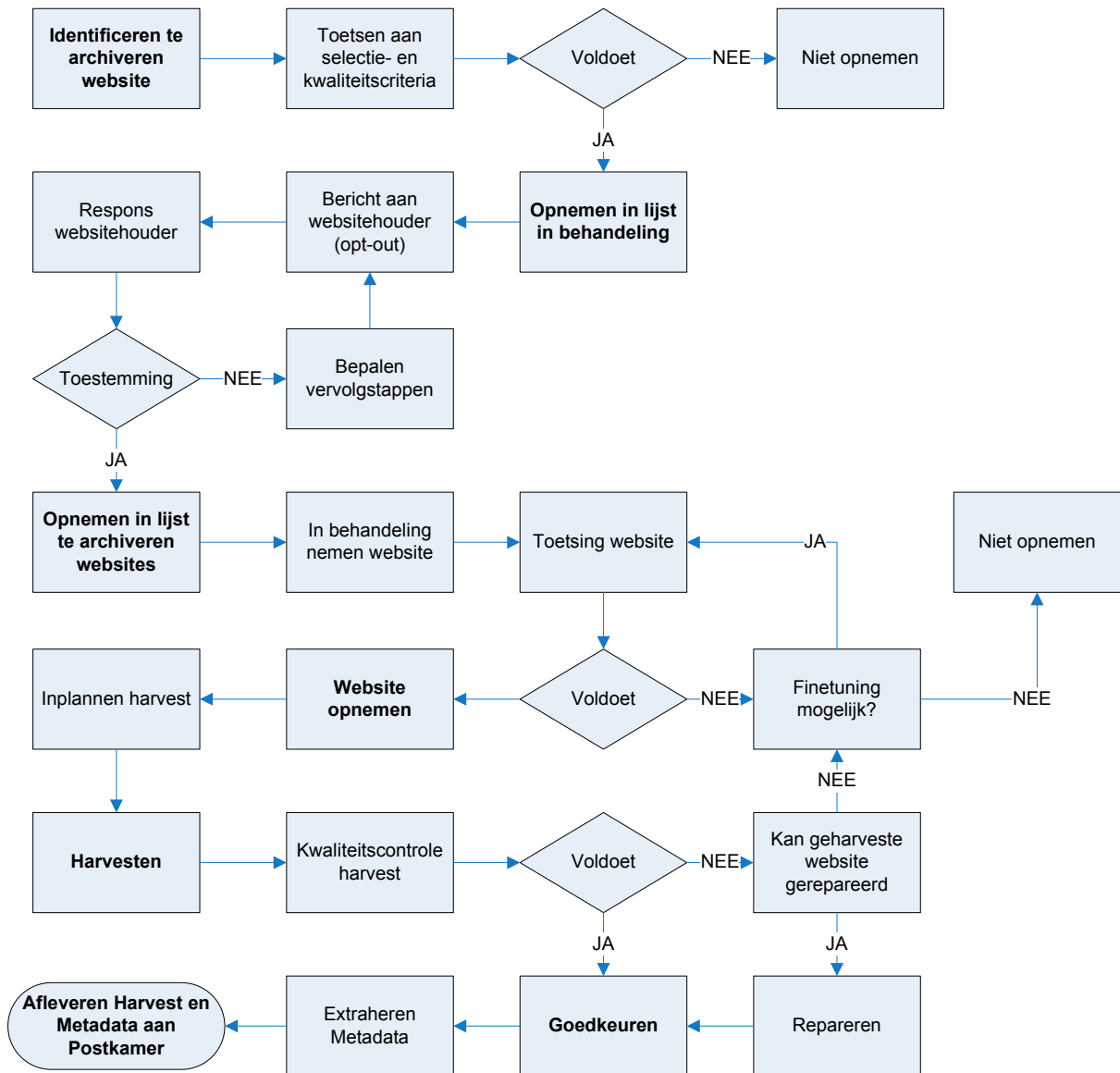
Naar verwachting zal de voornaamste reden voor het gebruik van het webarchief onderzoek zijn. Naarmate het webarchief groeit en ouder wordt zal het steeds meer een bron worden voor (wetenschappelijk) onderzoek. Journalisten en juristen kunnen het webarchief gebruiken voor achtergrondinformatie. Natuurlijk zal het webarchief ook geïnteresseerde 'leken' trekken. Op basis van deze bevindingen zijn er gebruiksscenario's opgesteld. Deze scenario's bevatten concrete voorbeelden voor het gebruik van een webarchief. De hierin genoemde gebruikers zijn: journalist, 'gewone' burger, overheid, jurist, patent aanvrager, student, onderzoeker, genealogisch onderzoeker en interne medewerkers.

7. Tweede fase webarchivering

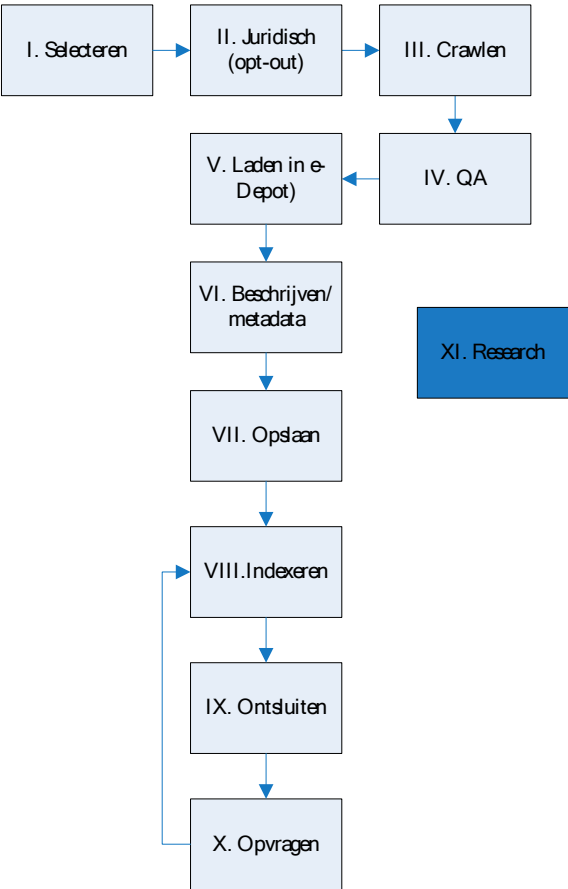
Van december 2005 tot en met mei 2007 is het project *1^{ste} fase webarchivering* uitgevoerd. Doel van dit project was het verkrijgen van inzicht in de technische, organisatorische, juridische en functionele mogelijkheden, de kosten en haalbaarheid van een webarchief bij de KB ten behoeve van het duurzaam behoud van een selectie van Nederlandse websites.

Deze eerste fase voldoende kennis opgeleverd om duidelijke aanbevelingen te kunnen geven voor operationalisering van webarchivering binnen de KB. Deze aanbevelingen vormen de basis voor de uitvoering van de tweede fase van het project. Deze richt zich op de operationalisering van webarchivering in de KB.

BIJLAGE1: SELECTIEPROCES WEBARCHIVERING



BIJLAGE 2: WORKFLOW WEBARCHIVERING



BIJLAGE 3: OVERZICHT VAN GEARCHIVEERDE WEBSITES

Seeds	Aantal x gecrawled	Data gecrawled
80.73.129.154/wijzers/cultuurwijs.nl/ cultuurwijs.nl/i000000.html	1	23-04-2007
80.73.129.154/wijzers/cultuurwijzer.nl/ cultuurwijzer.nl/i000000.html	1	23-04-2007
ahm.nl	1	10-01-2007
ainp.nl	1	20-02-2007
aladin.bibliotheek.nl	1	20-02-2007
annefrank.nl	1	09-01-2007
anno.nl	2	26-01-2007 22-03-2007
archeos.nl	2	10-01-2007 30-07-2007
archeovacature.nl	1	30-07-2007
archiefschool.nl	2	23-01-2007 22-03-2007
archievendag.nl	1	30-07-2007
belvedere.nu	2	10-01-2007 22-03-2007
bibliotheek.nl	2	30-01-2007 20-02-2007
boekbalie.nl	1	20-02-2007
boekenjeugdguids.nl	1	20-02-2007
boekenpret.nl	1	20-02-2007
boekman.nl	2	10-01-2007 22-03-2007
cbg.nl	2	20-02-2007 22-03-2007
cbs.nl	1	12-02-2007
centraalmuseum.nl	2	05-01-2007 22-03-2007
clingendael.nl	2	10-01-2007 22-03-2007
collectbritain.co.uk	1	26-03-2007
cornelisvanderwal.web-log.nl	1	16-01-2007
cpb.nl	1	30-01-2007
cultuurwijs.nl	1	12-01-2007
cultuurwijzer.nl	1	10-01-2007
cwi.nl	1	31-01-2007
dans.knaw.nl	1	30-01-2007
datbewarenl.nl	2	09-01-2007 30-07-2007
dbna.nl	1	30-07-2007
debibliotheek.nl	2	30-01-2007 20-02-2007
dedubbeldepalmboom.nl	2	04-01-2007 22-03-2007
demoanne.nl	1	16-01-2007
den.nl	1	23-01-2007
denhaag.nl	1	30-01-2007
deverdiepingvannederland.nl	2	10-01-2007 22-03-2007
digidiva.nl	1	30-07-2007
digitaalverleden.nl	1	30-07-2007
digitaleatlasgeschiedenis.nl	1	10-01-2007
divakoepel.nl	3	29-01-2007 22-03-2007
divaprofielen.nl	1	30-07-2007
doar.nl	1	16-01-2007
edusite.nl	2	10-01-2007 22-03-2007
eerstekamer.nl	1	24-01-2007
erfgoedactueel.nl	2	12-02-2007 30-07-2007
erfgoedacarte.nl	1	30-07-2007
erfgoedinspectie.nl	1	29-01-2007
erfgoednederland.nl	1	30-07-2007
farsk.nl	1	16-01-2007
geheugenvannederland.nl	1	04-01-2007

geheugenvanoost.nl	1	10-01-2007	
gemeentearchief.amsterdam.nl	1	10-01-2007	
gemeentearchief.rotterdam.nl	2	10-01-2007	22-03-2007
gezondheidsloket.nl	1	20-02-2007	
gezondheidsplein.nl	1	30-01-2007	
henkvanderveer.nl	1	16-01-2007	
hetkenniscentrum.nl	2	04-01-2007	23-03-2007
hetutrechtsarchief.nl	2	29-01-2007	22-03-2007
hmr.rotterdam.nl/sh	1	05-01-2007	
huygensinstituut.knaw.nl	1	30-01-2007	
iiav.nl	1	26-01-2007	
iisg.nl	2	26-01-2007	22-03-2007
ikwilietsleren.nl	2	29-01-2007	20-02-2007
illc.uva.nl	1	04-01-2007	
inghist.nl	1	26-01-2007	
inl.nl	1	22-01-2007	
inpoldering-noordwalcheren.com	1	22-01-2007	
johan-veenstra.nl	1	16-01-2007	
joopboomsma.nl	1	16-01-2007	
kb.nl	1	12-02-2007	
kennisnet.nl	1	12-01-2007	
kich.nl	1	10-01-2007	
koffervolscherven.nl	1	30-07-2007	
koninklijkhuis.nl	1	23-01-2007	
kunsthier.nl	1	16-01-2007	
leesplein.nl	1	20-02-2007	
letterkundigmuseum.nl	1	04-01-2007	
listserv.surfnet.nl/archives/neder-l.html	1	24-01-2007	
literatuurplein.nl	1	20-02-2007	
louiscouperus.nl	1	23-01-2007	
meertalen.nl	1	20-02-2007	
meinderttalma.nl	1	16-01-2007	
mijnstempel.bibliotheek.nl	1	20-02-2007	
minocw.nl	1	29-01-2007	
monumentengevaar.nl	1	30-07-2007	
museumboerhaave.nl	1	04-01-2007	
myneigeneker.nl	1	16-01-2007	
neder-l.nl	1	04-01-2007	
nederlandsfotomuseum.nl	1	22-01-2007	
netkids.bibliotheek.nl	1	20-02-2007	
niwi.knaw.nl	1	12-01-2007	
nlkompas.nl	1	20-02-2007	
nwo.nl	1	10-01-2007	
nyenrode.nl	1	12-02-2007	
onderzoekinformatie.nl/nl/oi/nod	1	29-01-2007	
onzetaal.nl	2	24-01-2007	22-03-2007
openmonumentendag.nl	1	30-07-2007	
overheid.nl	1	04-01-2007	
planjeeigenruimte.nl	1	30-07-2007	
postbus51.nl	1	02-02-2007	
projecten.leesplein.nl	1	20-02-2007	
racm.nl	2	29-01-2007	22-03-2007
regering.nl	2	29-01-2007	22-03-2007
rembrandthuis.nl	1	04-01-2007	
rhc-eindhoven.nl	2	22-01-2007	23-03-2007
rhcl.nl	1	22-01-2007	

rkd.nl	1	22-01-2007	
rmo.nl	1	04-01-2007	
romanadvies.bibliotheek.nl	1	20-02-2007	
romeinsschip.nl	1	30-07-2007	
ru.nl	1	01-02-2007	
schoolbieb.nl	1	20-02-2007	
scp.nl	2	04-01-2007	22-03-2007
siteclx.nl	1	04-01-2007	
sitegenerator.bibliotheek.nl/wsf/home	1	30-01-2007	
sna.nl	1	09-01-2007	
stichtingncm.nl	1	30-07-2007	
telin.nl	1	29-01-2007	
tinbergen.nl	1	22-01-2007	
tno.nl	2	30-01-2007	26-03-2007
tresoar.nl	1	29-01-2007	
tweedekamer.nl	2	04-01-2007	22-03-2007
vng.nl	2	29-01-2007	23-03-2007
waterstaatsgeschiedenis.nl	1	04-01-2007	
weekvandegeschiedenis.nl	2	15-02-2007	22-03-2007
wetwegwijzer.nl	1	20-02-2007	
zoeken.bibliotheek.nl	1	20-02-2007	