Websites blijvend beschikbaar maken: een terreinverkenning

Het wereldwijde web werd voor velen een primaire informatiebron in het midden van de jaren negentig. Vele websites veranderen voortdurend en vele verdwijnen ook weer snel. Als websites nergens bewaard worden, is deze informatie voorgoed verloren. Zijn daar organisaties mee bezig en hoe verloopt dat? Is het mogelijk het volledige web te archiveren?

HILDE VANOVERBEKE is stafmedewerker informatie en documentatie bij SoCiuS, het Steunpunt voor Sociaal-Cultureel Volwassenenwerk
hilde.vanoverbeke@socius.be

Een dynamisch medium

Het wereldwijde web is het gedeelte van het internet dat met een webbrowser (bijvoorbeeld Internet Explorer) bekeken kan worden, meestal bestaande uit HTML-pagina's. In het begin van de jaren negentig deed het web zijn intrede. Enkele jaren later was het een belangrijk medium voor informatie, communicatie en ontspanning geworden. Het web is erg dynamisch van aard. Vele websites (verzamelingen samenhangende webpagina's) veranderen voortdurend of bestaan slechts voor korte tijd. Iedere internetgebruiker kent de boodschap 'De pagina kan niet worden weergegeven'. Wie een verzameling van geselecteerde, kwaliteitsvolle weblinks voor de eigen doelgroep onderhoudt, weet hoe vaak gecontroleerd moet worden op *broken links*. De referenties in een interessant bevonden webdocument lopen regelmatig op niets uit. De gemiddelde levensduur van een webpagina bedraagt volgens Brewster Kahle, initiatiefnemer van het Internet Archive, 100 dagen, die van een website 19 maanden. Als het web niet gearchiveerd wordt, zal zeer waardevolle informatie voor altijd verloren gaan. Op basis van literatuurstudie wordt hier getracht een zicht te geven op de manier waarop webarchiveringsinitiatieven vandaag te werk gaan en op de problemen waar we nog voor staan. De acties die organisaties kunnen ondernemen om eigen webinformatie te archiveren, vallen buiten het bestek van dit artikel. Over het archiveren van eigen websites is een uitgebreid handboek verschenen van F. Boudrez en S. Van den Eynde[1].

Belang van archiveren

Het belang van het archiveren van websites ligt op vier niveaus[2]. Websites hebben een *documentaire waarde*, als bronnenmateriaal voor onderzoek naar de evolutie van het medium. Vanzelfsprekend hebben ze een grote *informatieve waarde*. Voorts is er de *culturele waarde*. Websites zijn materiële getuigenissen van onze samenleving en behoren tot het culturele erfgoed. Ten slotte geven websites ook *aanleiding tot archiefbescheide*n. Ze worden gebruikt bij transacties of handelingen waarbij een overheid of een bedrijf betrokken is en die verantwoording vereisen. Een administratieve overheid is ook aansprakelijk voor de informatie die ze via het web aanbiedt.

Verschillende organisaties zien het dan ook als hun taak om websites te archiveren: gespecialiseerde documentatiecentra en universiteiten vanwege de documentaire en informatieve waarde, nationale bibliotheken vanuit hun opdracht het culturele erfgoed te bewaren, overheden en bedrijven omdat ze zich mogelijk moeten verantwoorden voor hun handelen.

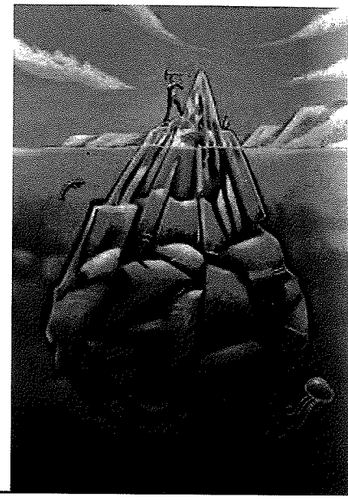
Het diepe web

Om de technische problemen te begrijpen die samenhangen met het archiveren van websites, is het belangrijk te weten dat er heel wat sites en webpagina's zijn waar zoekmachines niet bij kunnen. Men spreekt in dat verband over het diepe web, ook *invisible* of *hidden web* genoemd. C. Brian Smith geeft een lijst van types sites die voor zoekmachines onbereikbaar zijn:

- sites die registratie of paswoorden vragen;
- sites die een betaling vragen;
- sites op een intranet of achter een firewall;
- gearchiveerde pagina's;
- interactieve tools;
- nieuw toegevoegde pagina's;
- sites die *metatags* gebruiken om *crawlers* uit te sluiten;
- informatie die pas bestaat als een zoeker een databank ondervraagt.

Steeds meer websites worden dynamisch opgebouwd en zijn databankgestuurd. Het gaat vaak om heel waardevolle informatie, al dan niet gratis. Volgens een rapport uit 2000 bevatte het 'diepe web' 500 keer zo veel informatie als het *surface web*[4]. Populaire zoekmachines maken dus slechts de spreekwoordelijke top van de ijsberg toegankelijk. Een deel van de informatie in het diepe web betreft wel eerder gegevens dan informatie (bijvoorbeeld reeksen weergegevens) en een deel duplicaten[5].

Realiseren we ons ook goed wat de omvang is van het wereldwijde web. Volgens het Online Computer Library Center (OCLC) in de VS waren er in 2002 8.712.000 unieke websites[6]. Het OCLC houdt echter geen rekening met virtuele webservers, die een belangrijk deel van de webservers vormen. Een onderzoek in 2000 stelde dat er op dat moment meer dan 2 biljoen unieke, publieke web*pagina's* waren[7].



Volgens een rapport uit 2000 bevatte het 'diepe web' 500 keer zoveel informatie als het surface web.
http://www.hmaster.com

Drie verzamelstrategieën

Actuele webarchiveringsinitiatieven gebruiken één van de volgende drie benaderingen[8]: *automatic harvesting*, selectieve benadering of deponering. Bij *automatic harvesting* sporen *crawler*-programma's het *surface web* op via externe links en downloaden ze. Op die manier wordt een momentopname of *snapshot* van het hele web (bijvoorbeeld in een landsdomein) gemaakt. De *harvester* wordt gevoed met een verzameling van duizenden URL's (webadressen) van geschikte webdocumenten, bijvoorbeeld met de domeinnaam van het land[9]. Deze pagina's worden verzameld en geanalyseerd om URL's te vinden die overeenkomen met de selectiecriteria. Vervolgens worden die URL's op hun beurt gebruikt om een tweede reeks documenten te vinden. Dit proces gaat door tot elk document gevonden is en duurt maanden. In het midden van de jaren 1990 werden de eerste *harvesters* gebouwd om webindexen te creëren zoals de zoekmachines Alta Vista. De eerste webarchiveringsinitiatieven gebruikten deze *harvesters*, pas later werden er *harvesters* speciaal voor archivering ontwikkeld, zoals de NEDLIB-*harvester* (zie verder).

Bij de selectieve benadering selecteert de archiverende organisatie websites en wordt er overlegd met de website-eigenaars, waarna de sites met behulp van software binnengehaald worden. Bij deponering worden webgebaseerde documenten op initiatief van de eigenaar overgebracht naar een bewaarplaats, zoals een nationaal archief of bibliotheek.

Voor- en nadelen

Het grote voordeel van *harvesting* is dat het goedkoper is dan de selectieve benadering, doordat er minder personeel ingezet moet worden. Bovendien kan een groot aantal webbronnen verzameld worden[10]. Een groot nadeel is dan weer dat de huidige *crawlers* niet overweg kunnen met databankgestuurde sites, *plug-ins* als Flash en *scripting*technieken als JavaScript. Voor het verzamelen van webpagina's in het diepe web zijn ze dus niet geschikt. De voor- en nadelen van de selectieve benadering zijn zo wat het spiegelbeeld van die van de eerste benadering: een beperkt aantal verzamelde bronnen, duurder, maar meer geschikt voor het diepe web. Om deze websites te kunnen raadplegen moeten immers de oorspronkelijke webserverconfiguratie, serversoftware en het gekoppelde bestandensysteem actief blijven. Een voordeel van de selectieve benadering is dat publieke toegang mogelijk is omdat rechten verkregen zijn van de eigenaars van de websites. Een nadeel is dat in het archief niet genavigeerd kan worden.

Er bestaat een consensus over het feit dat een gecombineerde benadering vaak aangewezen is, om te profiteren van de voordelen van beide strategieën[11]. Het doel van het webarchiveringsinitiatief en de beschikbare middelen spelen eveneens een rol bij de keuze. Deponering, de derde benadering, levert nooit meer dan een fractie van het web op en het webarchief heeft weinig controle op de selectie.

Initiatieven

Het Internet Archive

Het meest ambitieuze initiatief met de *harvesting*-benadering werd ontwikkeld door de Amerikaan Brewster Kahle[12]. Hij richtte het Internet Archive[13] op, dat sinds 1996 websites en nieuwsgroepen wereldwijd archiveert. Deze non-profitorganisatie werkt samen met een commerciële partner, Alexa, die de data verzamelt en om het half jaar oude data naar het Internet Archive doorstuurt. Een volledige *crawl* van het wereldwijde web duurt ongeveer 2 maand. Eind 2003 bevatte het Archive meer dan 200 terabytes data

Eind 2003 bevatte het Internet Archive meer dan 200 terabytes data. http://www.mindjack.com/feature/archive.html.

(1 terabyte = miljoen miljoen bytes), equivalent aan ongeveer 200 miljoen boeken. In 2001 introduceerde het Internet Archive de Wayback Machine, een interface waardoor het publiek de snapshots van websites on line kan raadplegen. De gebruikers moeten een URL invoeren. Er is geen inhoudelijke ontsluiting voorzien. Uit een evaluatie van de Wayback Machine bleek dat in de collectie van het Internet Archive belangrijke inhoud en functionaliteit ontbreken, onder meer links die verwijzen naar verkeerde versies van webpagina's, ontbrekende afbeeldingen en links [14].

Pandora-project (Australië)

De National Library van Australië startte rond dezelfde tijd met het Pandora-project [15] en hanteert daarbij de selectieve benadering. Belangrijk geachte Australische on-linepublicaties (websites, nieuwsgroepen en mailinglijsten) worden gearchiveerd. In juni 2002 waren er dat een 2000-tal. Voor het fysiek verwerven van de webdocumenten worden ofwel afspraken gemaakt met de uitgevers of er wordt software voor automatische verzameling gebruikt. De laatste jaren gebruikt men twee programma's tegelijk: HTTrack en Teleport Pro, omdat dit de kansen op succesvolle replicatie van de sites verhoogt. De frequentie en de diepte van verzamelen is afhankelijk van de soort website. Er bestaat publieke toegang tot het Pandora-archief via het web. Daartoe wordt immers toestemming van de eigenaar van de informatie gevraagd. De sites kunnen opgezocht worden via 15 hoofdonderwerpsdomeinen en een alfabetische titellijst.

Kulturarw-project (Zweden)

In 1997 startte de Koninklijke Bibliotheek van Zweden het Kulturarw-project [16] met als doel het Zweedse web te verzamelen. Het project maakt daarbij gebruik van *automatic harvesting*. Voornamelijk sites in het .se domein worden verzameld. Als *harvester* is gekozen voor een aangepaste versie van de open-bron*harvester* Combine. Die loopt nu twee tot drie keer per jaar. Er is beperkte publieke toegang.

NEDLIB-harvester

Het NEDLIB-project was een samenwerkingsproject van Europese Nationale Bibliotheken, gesteund door de Europese Unie, dat liep van 1998 tot en met 2000. Doel was een infrastructuur te construeren als basis voor het bouwen van een Europese depotbibliotheek. De standaarden die in het kader van dit project zijn gepubliceerd, zijn nu wereldwijd geaccepteerd [17]. De NEDLIB-*harvester* was één van de instrumenten ontwikkeld in dit project. Hij was de eerste die speciaal ontworpen werd voor het archiveren van webdata [18]. Het bouwen was moeilijker dan verwacht: pas eind 2002 was er een versie beschikbaar die voldeed. De software comprimeert de gegevens en kan *incremental harvesten*. Dit betekent dat bij een volgende ronde alleen documenten gedownload worden die gewijzigd of nieuw zijn. De NEDLIB-*harvester* kan de grootte van een webserver schatten en bepalen hoe vaak de server bezocht wordt. Dit maakt de tijd nodig voor een volledige ronde gevoelig korter en kleine servers worden niet overbelast.

Een archiefmodule van de *harvester* genereert archiveringsmetadata (onder meer *archive identifiers*, lokatie- en tijdsinformatie) en verwerkt de verzamelde documenten zodat ze opgeslagen en geïndexeerd kunnen worden. Als 'unieke *identifier*' van een webdocument berekent de NEDLIB-*harvester* een *MD5 checksum*. Die maakt het mogelijk om dubbels te verwijderen en de opslagruimte te beperken. Een test op de IJslandse webruimte leerde dat tot 2/3de van de inhoud van het archief dubbels betrof. De broncode van de NEDLIB-*harvester* is vrij beschikbaar. Eind 2003 werd het programma gebruikt in Finland, in de Tsjechische republiek en in Noorwegen voor het archiveren van de nationale webruimte. De *harvester* werkt echter nog niet feilloos. Er blijven problemen, vooral door slechte data en slecht ontwikkelde http-servertoepassingen, zoals html-bestanden die een reeks binaire data bevatten of URL's langer dan 256 bytes.

Andere initiatieven

Sinds 2000 is het aantal Europese nationale bibliotheken dat experimenteert met archivering van het nationale web snel gegroeid. Onder andere Frankrijk (zie verder), Oostenrijk, het Verenigd Koninkrijk, Noorwegen, Denemarken en Portugal zijn bezig met webarchivering. Naast de initiatieven van nationale bibliotheken zijn er vooral die van nationale archieven en van onderzoeksinstellingen. In Nederland bijvoorbeeld begon in 1995 het Internationaal Instituut voor Sociale Geschiedenis (IISG) onder de naam 'Occasio' met het archiveren van een hele reeks nieuwsgroepen. In 2000 is het Documentatiecentrum Nederlandse Politieke Partijen (DNPP) begonnen met de archivering van websites van politieke partijen. Het archief kreeg de naam 'Archipol' [19] en het is publiek via het web raadpleegbaar [20].

Gecombineerde aanpak van BnF

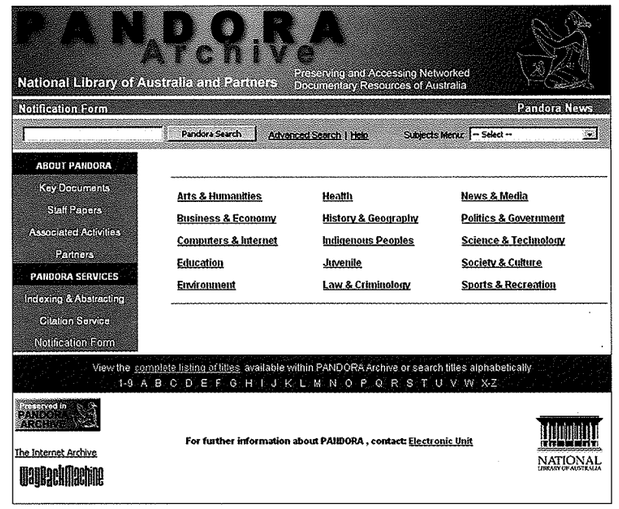
De Bibliothèque Nationale de France (BnF) integreert het beste van de verschillende benaderingen van webarchivering en doet sinds 1999 belangrijk onderzoek terzake [21]. Enerzijds wordt het Franse web verzameld via een *harvester*, anderzijds is een *deposit track* uitgewerkt voor het diepe web. Bij het *harvesten* wordt de *Xyleme crawler* gebruikt, waarbij bepaalde pagina's regelmatig ververst worden *(incremental crawlen)*. De frequentie kan geleid worden door de gebruiker of door een automatische schatting van de veranderfrequentie en het belang van een pagina. Dat belang wordt bepaald door het aantal links naar deze pagina én het aantal links ernaar in belangrijke pagina's. Dit mechanisme wordt eveneens gebruikt om zoekresultaten in zoekmachines als Google te rangschikken. De berekening van belang gebaseerd op linkstructuur is een manier om de aandacht te focussen op dat deel van het web dat het meest gebruikt wordt. Een test waarbij een automatisch gegeneerde rangschikking van sites vergeleken werd met een evaluatie van belangrijkheid door vakbibliothecarissen toonde een behoorlijke mate van correlatie.

De BnF erkent het belang van menselijke input in het collectieproces, onder meer bij het identificeren, selecteren en verzamelen van sites uit het diepe web. In 2002 liep een pilootproject over archivering van het diepe web. Elke stap van het proces van deponeren van webinhoud werd uitgetest. De schatting is dat er voor de archivering van 1000 sites uit het diepe web 16 medewerkers nodig zijn, waaronder 5 vakbibliothecarissen. *Webharvesters* kunnen hulp bieden bij het detecteren van deze websites door analyse van technische kenmerken van verzameld materiaal, bijvoorbeeld aanwezigheid van paswoordbeveiliging.

En België en Nederland?

De lezer zal opgemerkt hebben dat België noch Nederland voorkwamen in het rijtje van nationale bibliotheken met gevorderde webarchiveringsactiviteiten. In Nederland verzamelt het Depot van Nederlandse Publicaties ook elektronische publicaties sinds 1996 [22]. Het zijn dus de uitgevers die hun elektronische publicaties deponeren bij de Koninklijke Bibliotheek (KB). Er werd gestart met off line elektronische publicaties in de vorm van cd-roms, cd-i's en diskettes. Vanaf begin 2003 worden ook on-linepublicaties verzameld. On-linepublicaties die frequent aangepast worden zijn niet in de regeling opgenomen. Van 1998 tot en met 2000 coördineerde de KB het NEDLIB-project (zie hoger). In april 2003 startte de KB een pilootproject over archivering van het Nederlandse web om de mogelijkheden te verkennen en de stand van zaken in het buitenland vast te stellen.

De Koninklijke Bibliotheek Brussel staat nog helemaal aan het begin van een depot voor elektronische publicaties [23]. In het voorjaar van 2003 besliste zij om stappen te ondernemen om een e-depot te installeren. Van 1999 tot 2003 liep in samenwerking tussen het stadsarchief Antwerpen en het Interdisciplinair Centrum voor Recht en Informatica van de KULeuven het DAVID-project [24], het eerste onderzoeksproject over digitale duurzaamheid bij overheden in Vlaanderen. De focus ligt bij archivering van digitale bronnen gecreëerd binnen een (overheids)organisatie. Er werd heel wat expertise ontwikkeld. Een belangrijk resultaat van dit project is een handboek over digitaal archiveren.



Het Pandora-project van de National Library van Australië. Http://pandora.nla.gov.aut.

Gebruikerstoegang

Tot nu toe concentreerden webarchiveringsprojecten zich vooral op het verzamelen van websites. Er is nog niet veel aandacht besteed aan de preservering en de toegang tot webcollecties. Webarchieven zijn extreem grote databanken. Om een webarchief toegankelijk te maken voor eindgebruikers moeten de documenten geïndexeerd worden met een full text zoekmachine. In Scandinavië liep van september 2000 tot juni 2002 een samenwerkingsproject hierover: het Nordic Web Archive (NWA)[25]. De groep besliste begin 2001 om de zoekmachine van het bedrijf FAST aan te kopen. Dat heeft een vrij beschikbare globale webindex ontwikkeld. Om te kunnen indexeren worden alle documenten geconverteerd naar XML. Aan het eind van het project waren de instrumenten nog niet robuust genoeg om ze publiek beschikbaar te maken. Begin 2003 startte NWA II om de ontwikkeling af te werken en in december 2003 werd de NWA *Toolset* vrijgegeven.

Bij het verlenen van publieke toegang tot collecties webdocumenten spelen auteursrechten een belangrijke rol. Zelfs wanneer webdocumenten in alle landen onder het wettelijk depot vallen, is er geen sprake van onbeperkte publieke toegang tot via *harvesting* verzamelde en gearchiveerde websites. Het opsporen van al wie auteursrechten heeft, is een onmogelijke zaak. Denkbare oplossingen zijn dat de gearchiveerde websites slechts een bepaalde tijd na publicatie on line beschikbaar komen, dat sites on line raadpleegbaar zijn maar dat de auteurs het recht hebben om de sites uit het archief te laten verwijderen, of dat de toegang beperkt is tot onderzoek en onderwijs[26]. Een bijkomend probleem voor de toekomst is hoe gebruikers toegang kunnen krijgen tot webcollecties die verdeeld zijn volgens nationaliteit, onderwerp en brontype.

Preservering

De bewaring van webbronnen op langere termijn roept nog vele vragen op. De materialen moeten zo beheerd worden dat ze toegankelijk blijven als de technologie verandert. Verschillende preserveringsstrategieën (migratie, emulatie[27]...) moeten worden overwogen. Over de bruikbaarheid van migratie en emulatie is men het voorlopig oneens. Verder moeten er bewaarplaatsen gerealiseerd worden[28]. Die kunnen gebaseerd worden op het standaard Reference Model for an Open Archival Information System (OAIS). In het NEDLIB-project[29] is het OAIS-model aangepast zodat het bruikbaar is als model voor een digitaal magazijn met een archieffunctie. Een centraal probleem voor bewaarplaatsen zal de verzekering zijn van de authenticiteit van digitale objecten: zijn deze laatste precies wat ze beweren te zijn? Mogelijk komen er op termijn gecertificeerde digitale bewaarplaatsen. In ieder geval zullen de huidige projectactiviteiten van onder meer nationale bibliotheken tot de kernactiviteiten van de gastinstellingen moeten gaan behoren.

Overleg en samenwerking

De opdracht van webarchivering is zo immens dat samenwerking noodzakelijk is. Door de globale aard van het web kunnen de grenzen van het nationale web moeilijk duidelijk getrokken worden. Naast nationale bibliotheken hebben wetenschappelijke bibliotheken, archieven, uitgevers, departementen computerwetenschappen... een rol. De 'ECDL-workshops on webarchiving' en het recent opgerichte 'International Internet Preservation Consortium' zijn in Europa belangrijke overleg- en samenwerkingsorganen.

ECDL-workshops on webarchiving

Sinds 2001 vindt er jaarlijks een workshop plaats over onderzoek en praktijk van webarchieven, verbonden aan de European Conference on Research and Advanced Technologies for Digital Libraries (ECDL). De Nationale Bibliotheek van Frankrijk is de belangrijkste organisator[30]. Opdat de deelnemers in contact zouden kunnen blijven is aan de workshop een discussielijst[31] verbonden. De eerste twee bijeenkomsten focusten op de activiteiten van nationale bibliotheken en de collectiestrategieën en technologieën die in webarchiveringsinitiatieven gebruikt worden[32]. De thema's van de workshop in 2003 waren het ontwerp en gebruik van *persistent identifiers*[33], instrumenten voor webarchivering en verslagen van ervaringen. Voorbeelden van denkwerk rond *identifiers* zijn de *ARK identifier* (California Digital Library, VS), URN:NBN (Uppsala University Library, Zweden) en het werk in het Noorse Paradigma-project. In dat project wordt voorgesteld om *identifiers* toe te kennen op het niveau van '*expression*' en '*work*' zoals gedefinieerd in de *Functional Requirements for Bibliographic Records (FRBR)*. Het Paradigma-project heeft ook een *workflow* ontwikkeld voor het *harvesten* en beheren van een webarchief.

International Internet Preservation Consortium (IIPC)

In 2002 begonnen verschillende Europese nationale bibliotheken en het Internet Archive besprekingen over samenwerking bij de ontwikkeling van nieuwe instrumenten voor webarchivering[34]. In de zomer van 2003 werd het International Internet Preservation Consortium gevormd. Het consortium wil webarchivering bevorderen door standaarden te ontwikkelen, evenals goede praktijken en instrumenten. Zes werkgroepen werden opgericht, waaronder één rond instrumenten voor toegang, het diepe web en vereisten voor onderzoek. Het consortium wordt geleid door Nationale Bibliotheek van Frankrijk. De meeste IIPC-bibliotheken nemen deel aan de ontwikkeling van een *webcrawler* van de tweede generatie, Heritrix genaamd. Dit wordt eveneens een open-brontoepassing.

Besluit

Het archiveren van het web is 15 jaar na het verschijnen van het medium in een fase van reële toepassingen gekomen. Op relatief korte termijn komen verbeterde collectietechnieken ter beschikking. De eerste onderzoeken en experimenten over toegang, standaardprocedures en preservering hebben hun vruchten afgeworpen. In verschillende Europese landen heeft de overheid al het belang ingezien van bewaring van webgebaseerde informatie. Op 17 oktober 2003 heeft de Unesco het *Charter on the preservation of the Digital Heritage* goedgekeurd[35]. Het Charter bepaalt onder meer dat de toegang tot het digitale materiaal, voornamelijk het materiaal dat zich in het publieke domein bevindt, niet belemmerd mag worden door onredelijke beperkingen.

Op het gebied van het archiveren van websites met een dynamische inhoud bestaat nog weinig expertise. Dit leidt Michael Day, verslaggever van de derde 'ECDL workshop on webarchiving', tot de bedenking dat het diepe web misschien buiten de scope van webarchiveringsinitiatieven valt[36]. Webarchivering kan beschouwd worden als één onderdeel van meer omvattende digitale preserveringsstrategieën.

Noten

1. *Boudrez, Filip en Van den Eynde, Sofie, Archiveren van websites. - Antwerpen: Stad Antwerpen, 2002. - (DAVID-rapport; 7). - 104 p. - http://www.antwerpen.be/david/teksten/Rapporten/Rapport7.pdf*
2. *Boudrez, Filip, ibidem.*
3. *Smith, C. Brian, Getting to know the invisible web, in: Library Journal, suppl. to summer 2001, p. 16-19.*
4. *Bergman, Michael, The deep web: surfacing hidden value, in: The Journal of Electronic Publishing, 7 (2001) 1. http://www.press.umich.edu/jep/07-01/bergman.html*
5. *Sieverts, Eric, Het diepe web, in: Informatie Professional, 4 (2000); p. 33-35.*
6. *http://wcp.oclc.org*
7. *Day, Michael, Collecting and preserving the world wide web: a feasibility study undertaken for the JISC and Wellcome Trust. - Bath: UKOLN, 2003. - 85 p. - http://library.wellcome.ac.uk/projects/archiving_feasibility.pdf*
8. *Day, Michael, ibidem.*
9. *Hakala, Juha, Archiving the web: European experiences. Presentation in CONSAL XII, 20-30 October 2003, Brunei. – http://www.lib.helsinki.fi/tietolinja/0203/Webarchive.html.*
10. *Day, Michael, ibidem.*
11. *Day, Michael, ibidem.*
12. *Voerman, Gerrit, Keyzer, André en den Hollander, Frank, Het archiveren van websites, in: Informatie Professional, 5 (2001); p. 16-19.*
13. *http://www.archive.org*
14. *Day, Michael, ibidem.*
15. *http://pandora.nla.gov.au*
16. *http://www.kulturarw3.kb.se*
17. *http://www.kb.nl > kenniscentrum > e-Depot en digitale duurzaamheid*
18. *Hakala, Juha, ibidem.*
19. *den Hollander, Frank en Voerman, Gerrit, Het web gevangen: het archiveren van de websites van de Nederlandse politieke partijen - symposiumbundel. - Groningen: Universiteitsbibliotheek Groningen, 2003. - 55 p. - ISBN 90-367-1760-4*
20. *http://www.archipol.nl*
21. *Masanès, Julien, Towards continuous webarchiving: first results and an agenda for the future, in: D-Lib Magazine, 8 (2002) 12. - http://www.dlib.org/dlib/december02/masanes/12masanes.html*
22. *http://www.kb.nl > kenniscentrum > e-Depot en digitale duurzaamheid*
23. *Boudrez, Filip, ibidem.*
24. *http://www.antwerpen.be/david/website*
25. *Hakala, Juha, ibidem.*
26. *Lyman, Peter, Archiving the world wide web, in: Building a national strategy for digital preservation. - Washington D.C.: Council on Library and Information Resources and Library of Congress, 2002; p. 38-51. -http://www.clir.org/pubs/reports/pub106/Web.html*
27. *'Migratie' betekent dat de bestanden naar een andere omgeving (hardware, besturingssysteem en applicatiesoftware) omgezet worden zodat ze met een nieuwe computerconfiguratie compatibel zijn. 'Emulatie' houdt in dat de vereiste hard- en softwareomgeving wordt nagebootst op een host-systeem.*
28. *Hakala, Juha, ibidem.*
29. *http://www.kb.nl/coop/nedlib*
30. *http://bibnum.bnf.fr/ecdl*
31. *http://listes.cru.fr/wws/info/web-archive*
32. *Day, Michael, 3rd ECDL workshop on webarchiving, in: Ariadne, (2003) 37. - http://www.ariadne.ac.uk/issue37/ecdl-Web-archiving-rpt.*
33. *Een 'persistent identifier' is een permanente, lokatie-onafhankelijke referentie naar een netwerkgebaseerde bron.*
34. *Hakala, Juha, ibidem.*
35. *Nieuwsbrief van het DAVID-project, 2003 (21)*
36. *Day, Michael, ibidem.*

SAMENVATTING

Op basis van literatuurstudie wordt aangetoond hoe webarchiveringsinitiatieven te werk gaan en welke problemen daarbij rijzen. Het belang van het archiveren van websites wordt geschetst, en belangrijke kenmerken van het web worden aangehaald: het bestaan van het 'diepe web' en de grote omvang. Om webbronnen te verzamelen zijn er drie strategieën, elk met hun voor- en nadelen: automatic harvesting, selectieve benadering en deponering. Vervolgens worden de eerste webarchiveringsinitiatieven voorgesteld. De laatste twee jaar krijgen ook de aspecten gebruikerstoegang, preservering en het ontwikkelen van standaarden de nodige aandacht.

ABSTRACT

Based on the study of literature, the article illustrates how web archiving initiatives are functioning and which problems remain to be solved. It dwells on the importance of archiving websites and goes into some major characteristics of the web: the existence of the deep web and its huge size. Three strategies exist for collecting web sources, each with pros and cons: automatic harvesting, selective approach and deposit. The first web archiving initiatives are presented, besides items such as user access, preservation and the development of standards that receive attention in the last two years.