Webarchivering bij het Ministerie van Verkeer en Waterstaat
Verslag van een Onderzoek

17-06-2005 / Definitief

R.J.J.Voorburg
J.L.E.Goutier

Capsis BV
Ministerie van Verkeer en Waterstaat

Ministerie van Verkeer en Waterstaat

Inhoudsopgave

Managementsamenvatting ---- 4

1. Leeswijzer ---- 5

2. Websites archiveren, waarom en hoe ---- 6
2.1 Waarom websites archiveren? ---- 6
Belang voor bedrijfsvoering, verantwoording en rechtszekerheid van de burger ---- 6
Verplichting op grond van de Archiefwet ---- 6
2.2 Hoe websites te archiveren 7
Microarchivering of macroarchivering ---- 7
Bronnen of snapshots ---- 7
Duurzame standaarden ---- 8
Metadata ---- 8

3. Doelstellingen van het onderzoek ---- 9
3.1 Aanleiding ---- 9
3.2 Doelstellingen ---- 9
3.3 Onderzoeksvragen ---- 9

4. Conclusies en aanbevelingen ---- 11
3.4 Conclusies ---- 11
3.5 Aanbevelingen ---- 11

5. Gehanteerde literatuur ---- 13

Bijlagen ---- 14

A. Onderzoeksverslag ---- 15
1 De organisatie ---- 15
2. Het vaststellen van de te archiveren websites ---- 15
Wat is een website ---- 15
Wat zijn de websites van het Ministerie van Verkeer en Waterstaat? ---- 16
Welke sites moeten gearchiveerd worden? ---- 16
3. Het maken van een snapshot ---- 16
Voorbereiding ---- 16
Configuratie van de archiveringsapplicatie ---- 16
Capture en opslag ---- 17
Post-processing ---- 18
4. De toegankelijkheid van de snapshots in Capsis Presurf ---- 18
De gehanteerde ordening ---- 18
Het bekijken van snapshots ---- 20
De zoekfunctionaliteit ---- 20
5. Beoordeling geschiktheid snapshotmethode ---- 22
Analyse technische duurzaamheid gearchiveerde websites ---- 23
Analyse technische belemmeringen gebruik snapshot-methode ---- 24
Herhaalde snapshots, omvang en archiveringsfrequentie ---- 24
Overdracht naar het Nationaal Archief ---- 24
Eisen aan overdracht naar het digitale depot ---- 24
Afhankelijkheid van het gehanteerde bestandssysteem ---- 24
Metadata ---- 25
Terzijde: archivering van intranet ---- 25

Bijlage B. De kwaliteit van de gegenereerde snapshots ---- 27
1. Inleiding ---- 27
2. Algemene beoordeling van de snapshots: problemen en oorzaken ---- 27
Afwijkingen ten gevolge van fouten van de bronserver ---- 27

Afwijkingen ten gevolge van onvolkomenheden in de crawler of viewer van Presurf ---------28
Problemen gerelateerd aan de wijze van gebruik van URL-parameters ----------28
Javascript-gerelateerde problemen----------29
Problemen ten gevolge van de implementatie van 'browser checks'----------30
Problemen gerelateerd aan het gebruik van Macromedia Flash. ----------30
Problemen gerelateerd aan de implementatie van sessie-management----------30
3. Beoordeling door inhoudelijk betrokken medewerkers ----------30
De interpretatie van de vragen ----------30
Over verschillen in de beoordeling----------31
4. Technische analyse ----------31
Kwaliteit gebruikte HTML-opmaak----------31
Gebruikte MIME types of bestandsformaten ----------32
Omvang van sites en snapshots ----------33

Bijlage C. Results----------34
1. Overall assessment----------34
Overview of issues and scores ----------34
Distribution of scores from overall assessment----------38
Frequency of issues----------38
2. Questionnaire for assessment of selected snapshots----------39
3. Extended assessment of selected snapshots----------41
Snapshot capsis.arc/100030/20041112/haringvlietsluizen ----------41
Snapshot capsis.arc/100030/20041123/projectijzerenrijn ----------42
Snapshot capsis.arc/100030/20041124/projectvera ----------43
Snapshot capsis.arc/100030/20041212/a2denboscheindhoven ----------44
Snapshot capsis.arc/100030/20041112/a2denbosch----------45
Snapshot capsis.arc/100030/20041112/aanlega50 ----------46
Snapshot capsis.arc/100030/20050125/aanlega50 ----------47
Snapshot capsis.arc/100030/20041124/randwegeindhoven----------48
Snapshot capsis.arc/100030/20041203/zuid-willemsvaart ----------49
Snapshot capsis.arc/100030/20041112/bouwdienst----------50
Snapshot capsis.arc/100030/20041203/rws-avv----------51
Snapshot capsis.arc/100030/20041022/wetpersonenvervoer ----------53
Snapshot capsis.arc/100030/20041123/nederlandleeftmetwater----------54
Snapshot capsis.arc/100030/20041015/ivw----------55
Snapshot capsis.arc/100030/20041124/rikz ----------57
Snapshot capsis.arc/100030/20041119/kaderrichtlijnwater ----------60
Snapshot capsis.arc/100030/20040916/nieuws----------61
Snapshot capsis.arc/100030/20041112/gordeldier----------63
5. Analyses of a reference website ----------64

Bijlage D. Voorstel metadata voor snapshots van websites 29-04-2005 ----------65
Identificatie----------65
Beschrijving----------65
Beheersgegevens----------65
Organisatiegegevens----------66
Gebruiksvoorwaarden----------66
Gerelateerde bronnen----------66
Metadata----------67

Bijlage E: Suggesties voor verder onderzoek ----------68
1. Openstaande vragen----------68
2. Webontwerp en client side scripts ----------68
3. Ondersteuning van de actie van het genereren van snapshots ----------68
4. Archiveringsfrequentie, omvang en kosten ----------69
5. Integratie met het bestaande regime voor records management ----------69
6. Presurf----------69
Aanpassingen omgeving snapshots----------69
Inzet archivering Intranet----------69
Gebruikte crawler-engine ----------70
7. Overdracht Nationaal Archief ----------70

Managementsamenvatting

Het ministerie van Verkeer en Waterstaat ontwikkelt en beheert al sinds 1995 internetsites. Aanvankelijk werden deze sites vooral gebruikt als voorlichtingsinstrument en waren ze vrij statisch van karakter. Met het naderbijkomen van de elektronische overheid gaan de internetsites van het Ministerie een steeds belangrijkere rol spelen in de communicatie met de burger. Ze worden niet meer alleen gebruikt om informatie te presenteren, maar ook om te communiceren met burgers over het beleid dat wordt gevoerd of wordt ontwikkeld en om transacties uit te voeren. De websites hebben daarmee niet alleen een functie gekregen binnen de voorlichtingsprocessen van het ministerie, maar ook binnen de uitvoerende- en beleidsprocessen.
De inzet van websites bij de uitvoering van deze processen leidt er toe dat websites niet alleen aan te merken zijn als mooie (en ook vluchtige) “publicaties”, maar ook als “archief” dat informatie bevat die voortkomt uit deze processen en die doorgaans niet in een andere, papieren, vorm beschikbaar is .
In een eerder stadium heeft het ministerie van Verkeer en Waterstaat een onderzoek laten uitvoeren naar de vraag of websites onder de archiefwetgeving vallen en, zo ja, hoe ze dan gearchiveerd zouden kunnen worden. Uit dit onderzoek bleek dat de archiefwetgeving op websites van toepassing is en dat er in beginsel verschillende mogelijkheden zijn om ze te archiveren. Eén methode heeft ook de praktijk zijn waarde bewezen: de snapshotmethode

Het ministerie van Verkeer en Waterstaat heeft samen met het bedrijf Capsis een pilotproject uitgevoerd om te onderzoeken of deze methode ook voor het ministerie geschikt is. De snapshotmethode wordt namelijk tot dusver met name gebruikt door nationale bibliotheken die het web van hun eigen land willen vastleggen. De toepassing van de snapshotmethode binnen een organisatie met als oogmerk het voldoen aan belangen van bedrijfsvoering, verantwoording of kennisdeling is nog niet aangetroffen.

De doelen van dit pilotproject waren om
- met behulp van de applicatie Presurf van Capsis ervaring op te doen met het archiveren van de internetsites van het ministerie;
- een beproefde lijst met richtlijnen voor het bouwen van websites te leveren, waardoor archivering van deze websites mogelijk wordt;
- een eerste webarchief op te bouwen.

In het pilotproject is met Presurf een archief opgebouwd van 114 websites van het Ministerie van Verkeer en Waterstaat, met een totale omvang van 13220 MB.

Het voorliggende document geeft een verslag van dit project.

Het rapport concludeert dat het toepassen van de snapshotmethode met behulp van de applicatie Presurf zeer wel mogelijk is. De kwaliteit van de snapshots en de efficiency van het proces van archiveren neemt toe, als het ministerie bij het ontwerpen en opzetten van websites een aantal aanbevelingen ter harte neemt. Deze aanbevelingen stemmen overigens in belangrijke mate overeen met de Webrichtlijnen die door advies.overheid.nl zijn geformuleerd. Zij zijn niet alleen belang voor de archivering van websites, maar ook voor de goede toegankelijkheid ervan.

1. Leeswijzer

Hoofdstuk 2 geeft een beeld van de achtergronden van het onderzoek en beschrijft de mogelijke benaderingen en methoden voor webarchivering.

Hoofdstuk 3 beschrijft de doelstellingen van het onderzoek en de gehanteerde onderzoeksvragen.

De conclusies ten aanzien van de toepasbaarheid van de snapshot-methode zijn te vinden in hoofdstuk 4, evenals een aantal aanbevelingen die een succesvolle toepassing ervan vergemakkelijken.

Bijlage A beschrijft de in het onderzoek gevolgde werkwijze en biedt antwoorden op onderzoeksvragen.

In bijlage B worden de resultaten gepresenteerd van de analyses van de kwaliteit van de snapshots. Ook wordt een analyse gegeven van de oorzaken van eventuele onvolkomenheden in de gegenereerde snapshots.

De Engelstalige bijlage C. laat op een meer gedetailleerde wijze de resultaten van de uitgevoerde analyses zien.

Bijlage D bevat een voorbeeld van de metadata die aan een website gekoppeld zouden moeten zijn.

Bijlage E geeft aan welke vragen binnen het project onbeantwoord zijn gebleven en welke nieuwe vragen zijn opgekomen. Ten slotte wordt hier een aantal aanbevelingen voor vervolgonderzoek gedaan.

2. Websites archiveren, waarom en hoe

1.1 Waarom websites archiveren?

Belang voor bedrijfsvoering, verantwoording en rechtszekerheid van de burger

Het Ministerie van Verkeer en Waterstaat ontwikkelt en beheert al sinds 1995 internetsites. Aanvankelijk werden deze sites vooral gebruikt als voorlichtingsinstrument en waren ze vrij statisch van karakter. Met het naderbijkomen van de elektronische overheid gaan de internetsites van het Ministerie een steeds belangrijkere rol spelen in de communicatie met de burger. Ze worden niet meer alleen gebruikt om informatie te presenteren, maar ook om te communiceren met burgers over het beleid dat wordt gevoerd of wordt ontwikkeld en om transacties uit te voeren. De websites hebben daarmee niet alleen een functie gekregen binnen de voorlichtingsprocessen van het ministerie, maar ook binnen de uitvoerende- en beleidsprocessen.
De inzet van websites bij de uitvoering van deze processen leidt er toe dat websites niet alleen aan te merken zijn als mooie (en ook vluchtige) “publicaties”, maar ook als “archief” dat informatie bevat die voortkomt uit deze processen en die doorgaans niet in een andere, papieren, vorm beschikbaar is . Een website over bijvoorbeeld de ontwikkeling van het Waddenzeebeleid geeft informatie over beleidsplannen en de reacties van betrokken organisaties. Aan deze informatie kunnen rechten worden ontleend. Het blijvend beschikbaar houden van deze informatie in de oorspronkelijke vorm is van belang voor het kunnen afleggen van verantwoording over het gecommuniceerde beleid. Daarnaast vormen de websites een steeds belangrijker onderdeel van het eigen geheugen van de organisatie. Websites tonen bijvoorbeeld de ontwikkelingen in het beleid en in de communicatie over het beleid. Ook zijn de ontwikkelingen in de vormgeving en het gebruik van websites binnen het Ministerie eruit af te lezen.
Kortom: er zijn voldoende redenen voor het ministerie van Verkeer en Waterstaat om zijn websites te archiveren. Bovendien is er een juridische verplichting op grond van de Archiefwet.

Verplichting op grond van de Archiefwet

Overheidsorganen zijn, op grond van de Archiefwet 1995, verplicht hun archieven in goede, geordende en toegankelijke staat te brengen en te bewaren (art. 3) . Dit geldt voor archieven “ongeacht hun vorm”, dus ook voor digitaal archief. De Archiefschool heeft op verzoek van het Ministerie van Verkeer en Waterstaat onderzocht of websites als archief aangemerkt kunnen worden. De conclusie was bevestigend.
Het enkele feit dat websites archiefbescheiden zijn, wil echter nog niet automatisch zeggen dat zij bewaard moeten worden. Veel archiefbescheiden zijn namelijk (op termijn) vernietigbaar. Het is echter niet goed mogelijk om het huidige selectie-instrumentarium voor archiefbescheiden toe te passen op websites. Websites kunnen bijvoorbeeld onder verschillende selectielijsten met verschillende bewaartermijnen vallen. Ook kan het voorkomen dat er binnen één selectielijst onduidelijkheid bestaat over bewaartermijnen (Hokke, 2003)[1]. In het hier beschreven project is dit selectievraagstuk buiten beschouwing gebleven: websites worden beschouwd als één geheel dat voor bewaring in aanmerking komt.

Indien een website permanent moet worden bewaard, geldt de "Regeling geordende en toegankelijke staat archiefbescheiden" (2002). Relevant is met name artikel 2:

“De zorgdrager zorgt ervoor dat van elk van de archiefbescheiden te allen tijde kan worden vastgesteld:
a. de inhoud, structuur en vorm bij het ontstaan, één en ander voor zover de inhoud, structuur en vorm kenbaar moesten zijn voor de uitvoering van het betreffende werkproces; en
b. op welk tijdstip en uit hoofde van welke taak of handeling het door het overheidsorgaan werd ontvangen of opgemaakt; en
c. de samenhang met de andere door het overheidsorgaan ontvangen en opgemaakte archiefbescheiden.”

[1] Zie de in hoofdstuk 5 opgenomen literatuurlijst. Omwille van de leesbaarheid wordt niet iedere literatuurverwijzing van een voetnoot voorzien.

Voor webarchivering is de regeling praktisch gezien helaas van beperkt nut, aangezien deze niet is toegesneden op websites.
Als een vervolg op het onderzoek van Hokke (2003) heeft Voorburg (2004) in opdracht van het Ministerie van Verkeer en Waterstaat een aanzet gemaakt met praktische, ontwerptechnische richtlijnen voor de opzet van duurzame websites. Een uitgangspunt vormden daarbij de door Horsman (1998) geformuleerde kwaliteitseisen voor digitale archiefbescheiden. Horsman stelt dat archiefbescheiden zo lang als noodzakelijk zodanig beheerd moeten worden dat deze volledig, authentiek, betrouwbaar, toegankelijk, beschikbaar en leesbaar zijn.

2.1 Hoe websites te archiveren

Microarchivering of macroarchivering

Bij het archiveren van websites kunnen verschillende benaderingen worden gevolgd.
Brüggen (2005) maakt hierbij onderscheid tussen microarchivering en macroarchivering. Bij macroarchivering[2] gaat het om het om zeer grootschalige vastlegging van websites, doorgaans uit het oogpunt van het bewaren van nationaal en internationaal cultureel erfgoed. Een bekend voorbeeld van is the vinden bij *The Internet Archive*[3]. Het zijn vooral nationale bibliotheken die actief zijn met deze vorm van conservering.

Het onderhavige rapport gaat niet over deze grootschalige vorm van het conserveren van websites, maar over wat Brüggen microarchivering noemt: het meer kleinschalig conserveren of archiveren van een beperkt aantal websites. Specifiek behandelt het het relatief kleinschalig conserveren van websites door (of mede door) de eigenaar zelf met als doel te kunnen voldoen aan het bedrijfsvoerings- en verantwoordingsbelang en aan wettelijke verplichtingen (zoals de Archiefwet of de Wet Openbaarheid van Bestuur)

Bronnen of snapshots

Als een organisatie er voor kiest haar websites te gaan archiveren dan kan ze verschillende methoden volgen. De oudste methode is gebaseerd op het duurzaam opslaan en documenteren van de digitale bronnen die samen de website vormen. Een nieuwere methode is de snapshotmethode, waarbij niet zozeer de bronnen worden geconserveerd, als wel de pagina's zoals ze op het scherm verschijnen.

De allereerste websites waren volledige statisch van aard. Iedere pagina van zo'n site bestaat uit een statisch tekstbestand met daarin opmaakcodes[4]. In de opmaakcodes van dit (HTML) bestand staan doorgaans verwijzingen naar andere statische bestanden die binnen de pagina moeten worden getoond, zoals met name afbeeldingen. Door deze eenvoud hoeft de conservering van een dergelijke site wat betreft de techniek niet veel meer om het lijf te hebben dan het opslaan van die statische bestanden in hun samenhang. Er is slechts een minimum aan technische metadata nodig om de site weer op de oorspronkelijke wijze aan te kunnen bieden.
Webarchivering door het archiveren van bronbestanden is een goede aanpak voor statische websites waarvan de bronnen beschikbaar zijn.

De techniek die tegenwoordig doorgaans gebruikt wordt voor het genereren van websites is vele malen complexer dan de oorspronkelijke opzet met statische bestanden. De meeste websites zijn zeer dynamisch van karakter. De HTML-bestanden van de pagina's van een moderne site worden feitelijk pas bij het opvragen gegenereerd. Dit gebeurt met deels op maat gemaakte programmatuur (scripts) en database-bevragingen. Een consequentie hiervan is dat de archivering van de bronnen exponentieel complexer is geworden.

Voor het op basis van de bronnen aanbieden van (een archiefversie) van een dynamische website is men afhankelijk geworden van ketens van soms zeer specifieke versies van software. Er worden hierdoor zeer hoge eisen gesteld aan de technische documentatie. In sommige gevallen zal de software bovendien niet meer geschikt zijn voor de hardware die wordt gebruikt. Er zal dan dus

[2] In plaats van archivering zou het beter zijn om hier over 'verduurzaming' of 'conservering'te spreken. Voor archivering zo belangrijke meta-informatie zoals bijvoorbeeld over de context van de documenten ontbreekt immers doorgaans.
[3] http://www.archive.org/
[4] Voor deze opmaakcode wordt de opmaaktaal HTML gehanteerd (Hypertext Markup Language). De webbrowser van de bezoeker van de site kan deze codes interpreteren en zorgt op die wijze dat de pagina op de juiste wijze getoond wordt en functioneert.

geschikte oude hardware gevonden moeten worden. Het herstellen of in stand houden van een complexe website door archivering van de bronnen kan zo zeer kostbaar worden.

De complexiteit van het via de bronnenaanpak archiveren van websites kan worden doorbroken door alleen het eindresultaat te archiveren, dat wil zeggen: de uiteindelijke pagina's en afbeeldingen zoals een bezoeker ze te zien krijgt. Het archiveren kan gebeuren met software die enigszins verwarrend vaak een *offline-browser* wordt genoemd. Feitelijk gaat het hier om software vergelijkbaar met de zogenaamde *crawler*[5] van een zoekmachine. Deze aanpak wordt doorgaans de snapshotbenadering genoemd.

Het conserveren van websites door snapshots te maken is beduidend eenvoudiger en goedkoper dan de archivering van de bronnen. Er is wel een probleem inherent verbonden aan deze methode. Een met snapshots geconserveerde website kan een bevroren website worden genoemd, in de zin dat alle dynamische functionaliteit niet meer zal werken die afhankelijk is van specifieke interactie tussen de bezoeker en de server van waaruit de website wordt opgestuurd. Denk hierbij met name aan functionaliteit die pas werkt nadat de bezoeker specifieke informatie in een webformulier heeft ingevuld, zoals bijvoorbeeld een zoekmachine. Het geheel van doorgaans door specialistische databases gegenereerde informatie achter een formulier wordt ook wel het *deep web* genoemd.

Duurzame standaarden

De digitale archivering van websites wordt bij voorkeur zo ingericht dat de websites raadpleegbaar zijn en blijven zonder dat er technieken als conversie of emulatie nodig zijn. Een webarchief is daarom gediend met de inzet van duurzaam toegankelijke bestandformaten en opmaakstandaarden. De bestandsformaten die in artikel 6 van de "Regeling geordende en toegankelijke staat archiefbescheiden" zijn hiervoor onvoldoende geschikt. Voorburg (2004) heeft daarom in "Webontwerp: richtlijnen voor archivering" een reeks aanbevelingen voor bestandsformaten en opmaakstandaarden gedaan[6].

Metadata

Naast duurzaam toegankelijke bestandsformaten zijn voor een geordende en toegankelijke staat goede metadata onontbeerlijk. Zonder metadata is toegankelijkheid moeilijk te verzekeren en wordt het beheer van het website-archief uiterst moeizaam. Voor sommige metadata bestaan er al min of meer gestructureerde overzichten (de zogenaamde schema's) waar men uit kan putten. In bijlage D staat vermeld welke schema's op dit moment door het Ministerie van Verkeer en Waterstaat voor webarchivering gebruikt kunnen worden.

Geïnspireerd door Australische voorbeelden is in dit project een overzicht opgesteld van de metadata die bij een webarchief voorhanden zouden moeten zijn (bijlage D). Dit overzicht is in lijn met de standaard ISO 23081 voor gebruik en implementatie van metadata.

Binnen de Nederlandse overheid wordt op dit moment ook elders gewerkt aan metadata voor websites. Vermeldenswaard is de overheid.nl webmetadatastandaard. Deze nationale standaard voor webmetadata is gebaseerd op onderzoek van RAND Europe Leiden en de ontwikkeling ervan is begeleid door Advies Overheid.nl. Belangrijk is te constateren dat deze metadatastandaard in de huidige vorm enkel gericht is op toegankelijkheid van de websites en daardoor te beperkt voor webarchivering. Advies.overheid.nl werkt echter aan een uitbreiding van haar webrichtlijnen ten behoeve van archivering.

[5] Een crawler van een zoekmachine doorloopt alle links die het op internet kan vinden. Iedere gevonden pagina wordt ook binnengehaald om in de index van de zoekmachine opgenomen te worden.

[6] Een bestandsformaat op zich ken moeilijk als al dan niet duurzaam gekarakteriseerd worden. Bepalend voor de duurzaamheid zijn de antwoorden op vragen als is het een 'open' bestandsformaat, is het een defacto bestandsformaat. Een volledige behandeling van de vraag 'wat is duurzaam' valt niet binnen de scope van deze rapportage.

3. Doelstellingen van het onderzoek

3.1 Aanleiding

Hoewel het besef dat websites archiefbescheiden kunnen vormen steeds breder verspreid raakt, wordt het archiveren van websites door de Nederlandse overheid nog niet of nauwelijks toegepast. De oorzaak hiervan lijkt zeker ook te liggen in het grote aantal praktische vragen dat bestaat rondom de wijze van inzet en de geschiktheid van de snapshotaanpak.

Het Ministerie van Verkeer en Waterstaat en Capsis zijn gezamenlijk een traject gestart om deze impasse te doorbreken: in een pilotproject doen zij praktische ervaring op met de conservering van websites van het ministerie volgens de snapshotmethode. Na de adviezen van Hokke (2003) en Voorburg (2004) is voor het ministerie bovendien een logische volgende stap om in een pilotproject daadwerkelijk websites te gaan archiveren. Capsis, een Nederlands bedrijf gericht op in de archivering van websites, heeft Presurf, haar applicatie voor webarchivering, hiervoor ter beschikking gesteld. Capsis gebruikt de opgedane ervaringen om deze applicatie verder te ontwikkelen.

3.2 Doelstellingen

Het Ministerie van Verkeer en Waterstaat en Capsis hebben als doelstellingen geformuleerd:

- Praktijkervaring opdoen met het archiveren van websites met de snapshotmethode en de applicatie Presurf van Capsis.
- Op basis van deze ervaringen komen tot een beproefde lijst met richtlijnen voor de opzet van websites die archivering op termijn mogelijk maken.
- Het genereren van een eerste archief van websites voor het Ministerie van Verkeer en Waterstaat om daarmee relevante sites te kunnen behouden.

Een bijkomend doel is:

- Het stimuleren van een bredere aandacht voor de archivering van websites.

3.3 Onderzoeksvragen

Er zijn drie groepen onderzoeksvragen onderscheiden.

De eerste groep onderzoeksvragen heeft betrekking op de archiveerbaarheid van de websites van het Ministerie van Verkeer en Waterstaat met behulp van de snapshotaanpak

Hieronder vallen de volgende vragen:

- Welke sites en welke pagina's kunnen probleemloos worden binnengehaald en wat levert problemen op?
- Welke stappen kunnen in het proces van het archiveren volgens de snapshotmethode worden onderscheiden?
- Welke kwaliteitscontrole vindt bij de onderscheiden stappen plaats?
- Wat is een in de praktijk hanteerbare definitie van een website? Wat zijn praktische criteria om websites te kunnen onderscheiden?
- Hoe kan in het archief het beste de overgang van 'gewone web' naar het diepe web (*deep web*) worden ondersteund of gepresenteerd?
- Hoe kunnen problemen met eventuele incorrecte of onvolledig gearchiveerde pagina's of onderdelen worden voorkomen? Welke aanpassingen in wijze van archiveren zijn zinvol? Welke aanpassingen aan de kant van de te archiveren websites zijn zinvol?
- Welke aanvullingen of aanpassingen op de bestaande richtlijnen voor websites (Voorburg 2004) zijn zinvol?
- Welk opslagformaat is gewenst?
- Wat is de omvang van één snapshot? Wat betekent het enkel het opslaan van de wijzigingen bij een herhaalde snapshot voor de benodigde opslagcapaciteit?
- Wat is voor de sites/ pagina's een goede frequentie voor het maken van snapshots?

De tweede groep onderzoeksvragen heeft betrekking op de geschiktheid van de voor archivering en beschikbaarstelling gebruikte applicatie Presurf:

- Welke ordening en wijze van presentatie van de snapshots is wenselijk?
- Wat zijn de voor- en nadelen van een 'viewer[7]' voor het raadplegen van de gearchiveerde websites?
- Welke zoekmogelijkheden zijn bruikbaar en zinvol?
- Welke aanvullende wensen zijn er ten aanzien van de functionaliteit?

In de derde groep worden overige vragen geschaard en vragen die gedurende het project relevant werden bevonden:

- Welke (technische en inhoudelijke) metadata zijn gewenst voor het beheer van de gearchiveerde internetsites?
- Welke metadata moeten worden opgeslagen over het archiveringsproces zelf?
- Welke metadata zijn gewenst voor de toegang tot de gearchiveerde internetsites?
- Op welke wijze worden authenticiteit en betrouwbaarheid van het archief van internetsites gegarandeerd?
- Is het mogelijk om *intra*netsites op een vergelijkbare manier te archiveren?
- Op welke wijze wordt vorm gegeven aan de samenhang met de overige archieven van het Ministerie van Verkeer en Waterstaat?

[7] De viewer is hier een hulpmiddel bij het bekijken van pagina's in een webarchief. Zie bijlage A(4) voor een uitleg van de applicatie.

4. Conclusies en aanbevelingen

Ten behoeve van de pilot is een begeleidingsgroep geformeerd met vertegenwoordigers van de ICT-afdeling van de Shared Service Organisatie en de Directie Communicatie van het ministerie. In de begeleidingsgroep was bovendien het Nationaal Archief vertegenwoordigd.
In de pilot werden ruim honderd websites van het Ministerie van Verkeer en Waterstaat volgens de snapshotmethode gearchiveerd. Het merendeel van de onderzoeksvragen kon aan de hand hiervan worden beantwoord. Voor een uitgebreid verslag van het onderzoek, zie bijlage A. De conclusies en aanbevelingen volgen hierna

3.4 Conclusies

De stelling uit het eerdere onderzoek van Hokke, dat de snapshotmethode geschikt zou zijn voor het archiveren van websites werd bevestigd. Voor het overgrote deel van de in dit onderzoek betrokken websites verliep de snapshot-procedure zonder haperingen.
Het benaderen van snapshots van websites met Capsis Presurf werkt eenvoudig en probleemloos. Een aantal verbeteringen in de applicatie blijkt echter wenselijk, zoals het bieden van de mogelijkheid om op specifieke kenmerken (metadata) van websites of snapshots te zoeken en de mogelijkheid direct vanuit de applicatie vastgelegde metadata te bekijken

De kwaliteit van de snapshots wordt onder andere bepaald door de op de site gebruikte standaarden en formaten. Opvallend is dat de pagina's van de onderzochte websites, op een incidenteel bestand na, niet voldoen aan aanbevolen HTML-standaarden.

De problemen die zijn opgetreden hadden vooral te maken met

- het gebruik van javascripts, bijvoorbeeld in dynamische menu's,
- het gebruik van *Macromedia Flash*,
- zogenaamde *browser checks*[8],
- specifiek gebruik van zogenaamde URL-parameters

Deze problemen resulteerden soms in een snapshot van onvoldoende kwaliteit. Het bleek mogelijk de kwaliteit van de snapshots te verbeteren door site-specifieke aanpassingen in de configuratie van de snapshot-actie aan te brengen. Dit kan echter een arbeidsintensief proces zijn.

Zeker voor nieuwe te ontwikkelen (onderdelen van) websites lijkt daarom de beste aanpak om bij opzet en ontwikkeling technieken en standaarden te hanteren die archivering vergemakkelijken. Het volgen van de aanbevelingen zoals te vinden op de site Richtlijnen voor de toegankelijkheid en duurzaamheid van overheidswebsites (Advies.Overheid.nl 2005) zal er toe leiden dat sites toegankelijker worden en vaker zonder grote problemen met de snapshot-methode kunnen worden gearchiveerd. Een bijkomend voordeel is een te verwachten grotere duurzaamheid van het snapshot.

3.5 Aanbevelingen

Geadviseerd wordt te komen tot een integrale benadering voor zowel ontwerp, bouw als archivering. In het onderstaande volgen hiertoe specifieke aanbevelingen.
De voordelen hiervan strekken overigens verder dan archivering. Volstaan wordt hier op te merken dat deze aanbeveling in sterke mate overeenkomen met de Webrichtlijnen van Advies Overheid.nl (2005), die sterk zijn gericht op toegankelijkheid van websites maar ook voor de duurzaamheid van websites van belang blijken te zijn.

1. Zorg voor een goede centrale registratie van de websites die vallen onder verantwoording van het ministerie. Aansluiting op, of integratie in, bestaande systemen voor records management lijkt hier een voor de hand liggende keuze. De registratie noemt ook de voor de inhoud verantwoordelijke medewerker.
2. Registreer van een website alle domeinen waaronder deze benaderd kan worden. Het gaat hier om het hoofddomein (of de hoofd-URL), alle mogelijke aliassen en *redirects*, en eventuele overige

[8] Onder *browser checks* wordt hier functionaliteit verstaan die op basis de identificatie van de webbrowser bepaalt welke informatie op het scherm getoond of van de webserver opgevraagd moet worden.

bestanden die zich buiten de voorgaande domeinen bevinden maar die wel tot de website behoren[9].

3. Stel een profiel (cf Hokke 2003) op van iedere website die raadpleegbaar is
4. Maak in het geval van oneindige URL-domeinen[10] een keuze ten aanzien van de URL's die bij de snapshot-actie achterwege gelaten kunnen worden. Doe dit op basis van een URL-patroon zodat deze begrenzing eenvoudig in de configuratie van de snapshot-actie kan worden ingesteld.
5. Bepaal bij het vaststellen van het profiel van een website of er *deep web* is en welke voorbeeldpagina's uit het *deep web* in het snapshot moeten worden opgenomen.
6. Zet de site zo op dat deze pagina's van het *deep web* via een specifieke URL benaderd kunnen worden. Neem deze URLs op in de configuratie voor de snapshot-actie.
7. Zorg dat unieke URL 's ook naar unieke bronnen verwijzen.
8. Geef unieke pagina's een unieke URL. Pas dus geen sessie-variabelen[11] toe in URLs en voorkom vervuiling van URLs door bijvoorbeeld repeterende parameters.
9. Ontraad het gebruik van *javascript* of andere *client side scripts* zoals *Macromedia Flash*. Kies waar mogelijk voor *server side* oplossingen in plaats van *client side* technieken[12].
10. Vermijd het gebruik van *brower checks*, zowel op de *client* als op de *server*. Biedt bij gebruik van *browser checks* altijd standaard ('*default*') pagina's conform webstandaarden.
11. Ontwikkel nieuwe websites volgens duurzame standaarden zoals gesteld in Webontwerp: richtlijnen voor archivering (Voorburg 2004) of de, nog in ontwikkeling zijnde, Webrichtlijnen Overheid.nl (Advies Overheid.nl 2005)
12. Controleer consequent op geautomatiseerde wijze of nieuwe of gewijzigde pagina's aan de gestelde standaarden voldoen. Vertrouw hierbij alleen op machinematige controle van gehanteerde syntaxis.
13. Zorg voor uitgebreidere zoekmogelijkheden dan Presurf thans biedt, zodat er op specifieke kenmerken van snapshots kan worden gezocht. Denk hierbij b.v. aan de titel, de URLs (inclusief aliassen en *redirects*) en andere te hanteren metadata.
14. Zorg voor voldoende metadata ten behoeve van toegankelijkheid en beheershandelingen
15. Zorg ervoor dat zichtbaar is dat het bij de snapshots om archief gaat en zorg voor een waarschuwing als men zich buiten het archiefdomein begeeft

[9] Een alias is een alternatieve domeinnaam voor een site. Een *redirect* biedt enkel een (al dan niet automatische) doorverwijzing naar de URL of een alias van de website.
[10] Een website kan een praktisch oneindig aantal URL's beslaan, bijvoorbeeld wanneer het een wekelijkse agenda bevat waarbij elke week een link naar de agenda van de volgende week bevat.
[11] Een sessie-variabele is een variabele die gebruikt wordt om voorgaande acties van een bezoeker van de site te onthouden, bijvoorbeeld om te onthouden wie er ingelogd is.
[12] Een *client side* techniek is een programma binnen een pagina dat niet op de webserver draait ('*server side'*) maar dat door de webbrowser van de bezoeker wordt uitgevoerd. Javascript is het meest voorkomende voorbeeld hier van.

5. Gehanteerde literatuur

Advies Overheid.nl, 2005. *Richtlijnen voor de toegankelijkheid en duurzaamheid van overheidswebsites* (versie 1.1). http://webrichtlijnen.overheid.nl

Brügger, N. 2005. *Archiving Websites. General Considerations and Strategies*. Århus, The Centre for Internet Research.

Hokke, H.A. 2003. Blijvend Beschikbaar. Onderzoek naar de archivering van websites. Amsterdam: Archiefschool.

Hokke, H.A. 2003. *Naar archivering van websites. Implementatieadvies bij onderzoeksrapport "Blijvend Beschikbaar".* Amsterdam: Archiefschool.

Hollander, F. den & G. Voerman (red.) 2002. *Het web gevangen. Het archiveren van de websites van de Nederlandse politieke partijen*. Groningen: Universiteitsbibliotheek.

Horsman, P.J. 1998. *Digitaal Archiveren. Het Recordkeeping System als kader voor het beheer van digitale archiefbescheiden.* Den Haag: Rijksarchiefdienst.

National Archives of Australia. 1999. *Recordkeeping Metadata Standard for Commonwealth Agencies. Version 1.0.*

Rothenberg, J., Graafland-Essers, I., Kranenkamp, H. *et al. 2004 Designing a National Standard for Discovery Metadata. Improving Access to Digital Information in the Dutch Government.* Den Haag: Advies Overheid.nl.

Voorburg, R.J.J. 2004. *Webontwerp: richtlijnen voor archivering*. Amsterdam: Uselab BV.

Wolters, D. 2004. *Het geheime web. MIVD websites op Defensie- en NAVO-netwerken en de Acrhiefwet 1995. Onderzoek naar webarchivering bij de Militaire Inlichtingen- Veiligheidsdienst*. Afstudeerscriptie. Deventer: Saxion Hogeschool IJselland.

2002 *Regeling geordende en toegankelijke staat archiefbescheiden*

2004. *Metadata internetsites van de Rijksoverheid.* http://www.regering.nl/meta/thc/metadataset.jsp

Bijlagen

A. Onderzoeksverslag

1. De organisatie

Het onderzoek is gestart in september 2004 met het eerste overleg van de begeleidingsgroep. De begeleidingsgroep bestond uit de volgende personen:

F. Binnendijk (Ministerie van Verkeer en Waterstaat, SSO /FAC)
F. Geelen (Ministerie van Verkeer en Waterstaat, SSO /ICT)
N. 't Hart (Ministerie van Verkeer en Waterstaat, CEND - DCO)
L.J.M. van Luxemburg (Ministerie van Verkeer en Waterstaat, SSO /FB)
H. Hofman (Nationaal Archief)
J. Slats (Nationaal Archief)
R. Verdegem (Nationaal Archief)

In verband met een betrekking bij een andere werkgever heeft F. Binnendijk tot eind 2004 deelgenomen. N. 't Hart is eind 2004 vervroegd uitgetreden. Zijn rol als vertegenwoordiger van DCO is daarna waargenomen door H. Eeuwes.

De uitvoering was vanuit het departement in handen van H. Goutier (SSO /FAC, voorzitter begeleidingsgroep, afstemming met Capsis). Onderzoek en afstemming en rapportage zijn vanuit Capsis uitgevoerd door R. Voorburg.

De begeleidingsgroep kwam ongeveer eens per twee maanden bij elkaar om zaken te bespreken als doelstellingen, onderzoeksvragen, aanpak en resultaten. De praktische werkzaamheden zoals het genereren van snapshots en beoordeling hiervan hebben plaatsgevonden tot maart 2005.

Om te kunnen komen tot een gezamenlijk doordachte opzet voor de te hanteren metadata is er in december 2004 en januari 2005 een discussie gevoerd, onder andere via een speciaal voor dit onderwerp opgezette mailinglijst. Naast leden van de begeleidingsgroep nam aan deze ook E. Hokke van de Archiefschool deel aan deze discussie. Geprobeerd is om metadata ten behoeve van de toegankelijkheid en metadata ten behoeve van het (langdurig) beheer van websites vast te stellen. Een deel van de metadata bleek automatisch te genereren, zoals metadata over gebruikte programmatuur, tijdstip snapshot en dergelijke. Een deel zou echter ook handmatig worden ingevoegd. Hierbij ging het bijvoorbeeld om metadata die de inhoud van een website beschrijven. Het daadwerkelijk handmatig toevoegen van deze metadata is in het kader van het project niet uitgevoerd, onder andere omdat vel sites al waren afgesloten en er geen personen voorhanden waren die nog op de inhoud konden worden aangesproken.
De discussie heeft geleid tot een uitgebreid rapport over de toe te passen metadata, gebaseerd op de Australische AGLS-standaard. De (op ontsluiting gerichte) metadatasets van het The Hague Core project en van Advies.Overheid.nl zijn hierin geïncorporeerd.
Ten behoeve van de presentatie en omwille van de hanteerbaarheid in de praktijk, is dit uitgebreide rapport weer teruggebracht tot een overzicht waarin de essentiële metadata zijn opgenomen (zie bijlage D).

2. Het vaststellen van de te archiveren websites

Wat is een website

Het streven was om met het onderzoek een archief op te bouwen dat ten minste één snapshot bevat van iedere site die onder de zorgplicht van het Ministerie van Verkeer en Waterstaat valt. Om hiertoe te komen was een overzicht gewenst van alle websites van het departement. Met deze lijst met sites kon niet tot een praktisch bruikbare technische definitie van een website gekomen worden.
Capsis hanteert doorgaans als criterium voor een website '*alle pagina's die gemeen hebben naar één en dezelfde homepage te verwijzen*'. Dit uitgangspunt bleek geen hanteerbaar criterium. In sommige gevallen verwees de link "*home*" op een site naar de startpagina van http://www.verkeerenwaterstaat.nl/ en niet naar de eigen startpagina. Ook het criterium '*alle pagina's onder één domeinnaam*' was niet geschikt. Onder het domein minvenw.nl zijn immers diverse sites te vinden. Voor de archivering van websites zal een bruikbare definitie naar verwachting aansluiten bij de definitie van archief als procesgebonden informatie.

Wat zijn de websites van het Ministerie van Verkeer en Waterstaat?

Op het ministerie bleek geen actueel en compleet overzicht van zijn websites voorhanden te zijn. Als eerste uitgangspunt voor het opbouwen van een archief met websites werd een lijst met sites gebruikt die samengesteld was met het oog op een aanstaande migratie van sites naar een ander content management systeem. Deze lijst bevatte met name de volgende informatie:

- De URL van de site
- Is de site extern beschikbaar of alleen op het intranet?
- Betreft de URL een directory met een website of is het enkel een alias voor een site?
- De naam van de dienst die verantwoordelijk is voor de site

Bij bestudering van de lijst bleken niet alle internet-sites daadwerkelijk extern beschikbaar te zijn. De niet-beschikbare sites gaven foutmeldingen als *'file not found'*, *'host not found'* of *'permission denied'*.

Tijdens het maken van de eerste snapshots ontstond de indruk dat de geleverde lijst verre van compleet was. Om tot een completer overzicht te komen is vervolgens een snapshot gemaakt van de gecombineerde hoofddomeinen verkeerenwaterstaat.nl en minvenw.nl. Op geautomatiseerde wijze is daarna op basis van alle pagina's uit dit snapshot een lijst gegenereerd met alle domeinen waar naar gelinkt werd. Deze lijst is vervolgens aan DCO voorgelegd. Dit leidde tot een aanvulling van de lijst met te archiveren websites. Overigens kon DCO ook niet in alle gevallen met zekerheid aangeven of een site inderdaad onder de verantwoordelijkheid van Verkeer en Waterstaat valt.
De uiteindelijk gehanteerde lijst bevatte overigens níet de overkoepelende website http://www.verkeerenwaterstaat.nl/.

Dat de verstrekte lijst niet compleet was, kan waarschijnlijk deels worden toegeschreven aan het doel waartoe de lijst in eerste instantie was samengesteld, namelijk de migratie van websites. Deze migratie omvatte immers niet de sites van Rijkswaterstaat.
Daarnaast lijkt een oorzaak dat projectwebsites dikwijls buiten DCO om worden opgezet.

Gedurende het maken van de snapshots bleek bovendien dat de lijst met aliassen van de opgegeven sites niet compleet was. Voor een goede configuratie van de archiveringsapplicatie zou het bovendien zinvol zijn geweest als er per site expliciet aangeven zou zijn onder welke aliassen en via welke redirects[13] de site beschikbaar was. Dit onderscheid werd nu niet gemaakt.

Welke sites moeten gearchiveerd worden?

Websites kúnnen te archiveren bescheiden vormen. Binnen dit onderzoek is geen aandacht besteed aan de vraag wat er bewaard moet worden en hoe lang dit bewaard moet worden. Aan dit selectievraagstuk is namelijk al eerder aandacht besteed (Hokke 2003)
Het uitgangspunt dat is gehanteerd is dat 'websites' de eenheid van archivering vormen, en dus niet de losse documenten of pagina's op een site.

3. Het maken van een snapshot

Bij het maken van een snapshot is in dit onderzoek volgens deze stappen gewerkt:

Voorbereiding

Ter voorbereiding werd een site eerst bekeken met een gangbare webbrowser. Daarbij werd op de volgende aspecten gelet:

- Kan de site worden benaderd?
- Zijn er problemen te verwachten met oneindige lussen?
- Bevat de site uitzonderlijke bestandsformaten?
- Kunnen er aliassen of redirects worden onderscheiden?

Configuratie van de archiveringsapplicatie

Als de site kan worden benaderd, kan de archiveringapplicatie worden ingesteld om een snapshot van de site te maken. Daarbij is het van belang de volgende configuratie-opties juist in te stellen:

[13] Een alias van een website is een alternatieve domeinnaam die verder exact dezelfde website oplevert. Een veelvoorkomend voorbeeld is dat de meeste website zowel te benaderen zijn met als zonder de 'www.' voor de domeinnaam.
Een *redirect* is een alternatieve domeinnaam die bij benadering al dan niet automatisch doorspringt naar de hoofddomeinnaam of URL.

- *Wat is de hoofddomeinnaam / URL en wat zijn de apart op te nemen aliassen?*
  De hoofddomeinnaam en de aliassen moeten apart worden ingesteld om de pagina's met de verschillende URLs op te nemen. Als enkel de hoofddomeinnaam wordt ingesteld dan zal een alias niet in het archief worden opgenomen. Er kan gesteld worden dat het opnemen van een alias in het archief niet direct nodig is omdat een alias per definitie exact dezelfde pagina's bevat als de site zoals beschikbaar onder de hoofddomeinnaam. Door een alias wel op te nemen is echter nadrukkelijk duidelijk dat de site ook onder een alias beschikbaar was. Bovendien wordt zo voorkomen dat de kans bestaat dat een snapshot niet compleet is als de site onder de hoofddomeinnaam per abuis expliciet verwijst naar een pagina of bestand onder een alias[14].
  Ten aanzien van aliassen is er in dit project is geen onderscheid gemaakt tussen een domeinnaam zonder en een domeinnaam met 'www.' er voor. Standaard is de domeinnaam met prefix 'www.' gehanteerd. Eventuele aliassen zonder 'www.' zijn genegeerd.
  Bij een site met aliassen is de keuze wat als de hoofddomeinnaam beschouwd moest worden door Capsis bepaald. Hierbij werd waar mogelijk gekozen voor de expliciet op de site zelf genoemde domeinnaam.
- *Wat zijn eventuele aanvullende URLs die moeten worden opgenomen?*
  In sommige gevallen bevond een deel van de documenten of pagina's van de site zich op een ander domein. Omdat de archiveringsapplicatie alleen die documenten opneemt die het vindt en die onder een expliciet opgegeven domein vallen, moesten deze domeinen ook expliciet in de configuratie worden opgenomen. Dit gebeurde door een URL-patroon op te geven. Zo kon bijvoorbeeld door het patroon "*.pdf" te specificeren opgegeven worden dat alle bestanden met die eindigen op ".pdf" in het snapshot meegenomen moesten worden.
- *Wat zijn URLs die niet opgenomen moeten worden in verband met oneindige lussen?*
  Daar waar oneindige lussen gevonden werden, werd op vergelijkbare wijze als onder het voorgaande punt beschreven aangegeven welke URL-patronen expliciet niet door de applicatie gevolgd moesten worden.
- *Welke aanvullende opties te hanteren, bijvoorbeeld gedrag ten aanzien van het gebruik van cookies, instructies voor zoekmachines of gebruik van bandbreedte.*
  In sommige gevallen bleek dat delen van een site afgeschermd waren via het *robots exclusion protocol*. Standaard werd de archiveringsapplicatie door Capsis ingesteld deze aanwijzigingen te negeren. Omdat sommige websites niet volledig te bekijken zijn als zogenaamde *cookies* niet worden geaccepteerd, werd ook standaard ingesteld *cookies* te accepteren.
  In het algemeen werd het maximum door de archiveringsapplicatie te gebruiken bandbreedte ingesteld op 100 of 150 Kb/s.

In sommige gevallen bleek direct na het maken van een snapshot dat een deel van de site niet of niet correct in het snapshot opgenomen was. Dit werd waar mogelijk hersteld door na analyse van de oorzaken de configuratie aan te passen. Veelvoorkomende oorzaken waren niet opgenomen aliassen of redirects, ontbrekende aanvullende URL-patronen of het voorkomen van oneindige lussen.

Capture en opslag

Na configuratie kan de snapshot-actie starten. Hoewel het maken van een snapshot een autonoom proces is werd het in het algemeen 'op het oog' in de gaten gehouden. Aan de hand van de output van de applicatie kan direct inzicht verkregen worden in welke pagina's (URLs) op dat moment binnengehaald worden en welke eventuele fouten daarbij optreden. Zo kon bijvoorbeeld met enige zekerheid[15] gesignaleerd worden of de applicatie in een oneindige lus raakte (in dat geval werd de sessie afgebroken en na analyse de configuratie aangepast).

Bij een correcte configuratie zal het voor een herhaalde snapshot niet nodig zijn de voortgang van het proces te observeren. De duur van de snapshot-actie wordt bij correcte configuratie hoofdzakelijk bepaald door de combinatie van de omvang van de binnen te halen data en de beschikbare bandbreedte.

[14] Hyperlinks in een site zullen doorgaans relatief zijn ten opzichte van de domeinnaam. Het kan voorkomen dat in een link de domeinnaam van de site zelf expliciet is opgenomen. Het document waar de link naar verwijst zou niet opgenomen worden als de link naar een alias verwijst en de alias niet opgenomen is in de snapshotopdracht.
[15] Zonder exacte kennis van de techniek achter de site is het niet altijd mogelijk zeker zijn van een oneindige lus op basis van variaties in de URL en de URL-parameters.

Bij het maken van een snapshot worden automatisch relevante metadata opgeslagen. De definitie van de snapshotactie wordt samen met een 'log' van gebeurtenissen zoals foutmeldingen in een bestand in XML-formaat weggeschreven.
Naast deze essentiële log wordt apart in een bestand een verslag weggeschreven van iedere opvraging die de applicatie doet bij de webserver. Daarbij worden onder andere opgevraagde URL, de naam van het weggeschreven bestand en de bij de communicatie uitgewisselde 'headers' zoals het mime-type van het bestand de bestandslengte weggeschreven.

De snapshots worden in de applicatie Presurf volgens een vaste structuur op het filesystem van de server[16] weggeschreven. De volgende (relatieve) directory-stuctuur wordt gehanteerd:

[klantcode]/[datum]/[sitenaam]/[snapshot]

In andere woorden: de snapshots van één klant van Capsis zijn te vinden in een directory met een klantcode die door Capsis aan deze klant toegewezen is. Binnen die directory bevinden zich subdirectories waarvan de naam gebaseerd is op de startdatum van de snapshot die het bevat. De datum is hierbij opgesteld conform ISO 8601 (zonder specificatie van tijd). De snapshots zelf bevinden zich in een directory met als naam een door Capsis gekozen naam voor de betreffende site.

Ter illustratie is een snapshot van de site te vinden onder de URL www.betuweroute.nl in het webarchief te vinden in de directory:
100030/20041112/betuweroute/

Binnen de directory van een snapshot wordt de structuur behouden van de website zoals die door de URLs wordt weerspiegeld. De namen van bestanden met URL-parameters die meerdere malen voorkomen worden hierbij noodzakelijkerwijs wel herschreven[17]. Dit wordt bijgehouden in het uitgebreide logbestand.

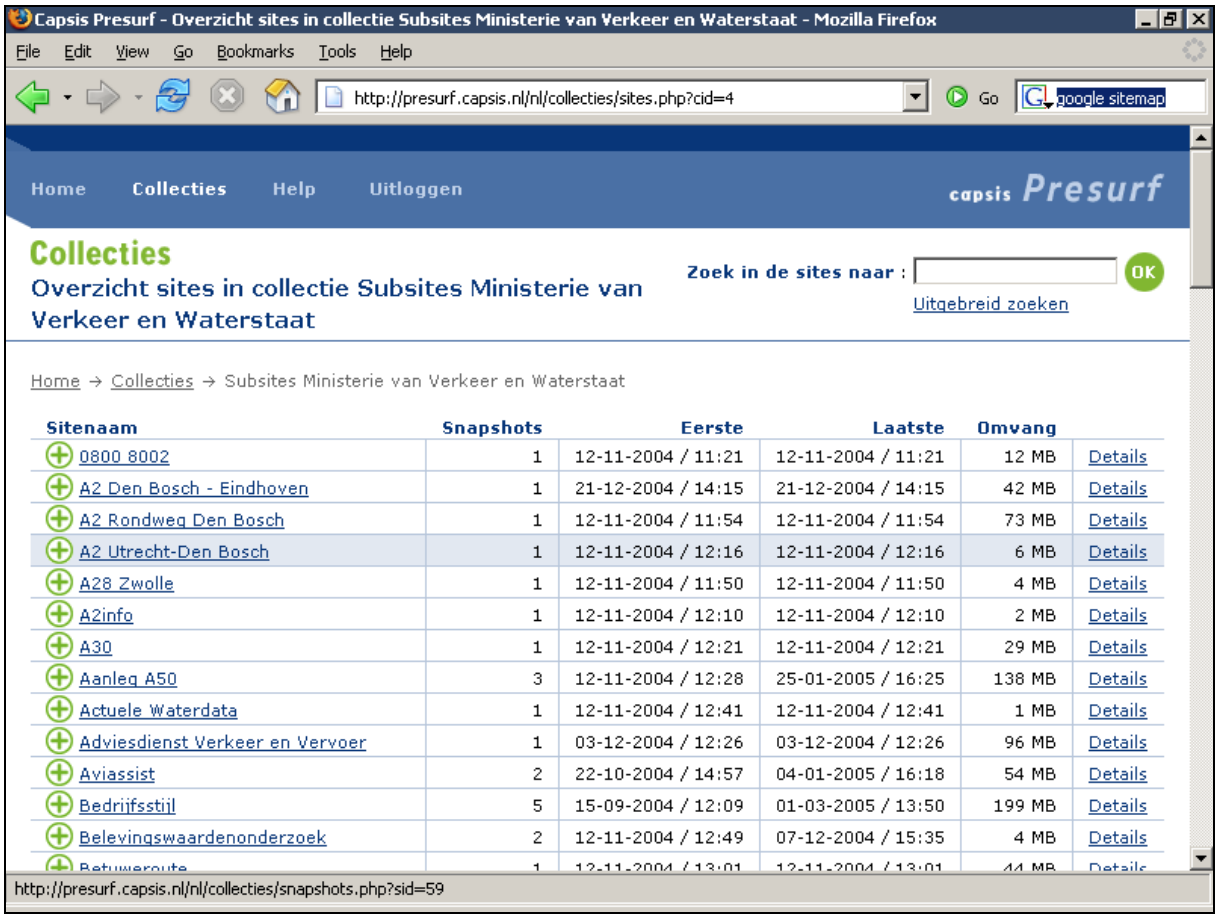
Post-processing
Na opslag van de bestanden van het snapshot vinden er een aantal post-processingen stappen plaats:
- Presurf maakt als engine voor het genereren van snapshots gebruik van de applicatie HTTrack[18]. Deze applicatie voegt aan ieder binnengehaald HTML-bestand een (bij het bekijken onzichtbare) commentaar-regel toe die niet voldoet aan de standaarden van W3C. Na het genereren van een snapshot zorgt Presurf er daarom voor dat dit toegevoegde commentaar weer verwijderd wordt.
- Van ieder opgeslagen bestand wordt als digitale handtekening een zogenaamde MD5-som vastgelegd[19]. Op deze wijze kan later vastgesteld worden of het bestand na de snapshot-actie al dan niet gewijzigd is.
- Van ieder snapshot wordt door de zoekmachine van Presurf een zoekindex gegenereerd. Deze zoekindex wordt bovendien (als kopie) samengevoegd met de overkoepelende zoekindex behorende bij de klant.
- Het snapshot wordt opgenomen in de database van Presurf zodat deze benaderd kan worden via de webinterface van Presurf.

4. De toegankelijkheid van de snapshots in Capsis Presurf
De gehanteerde ordening
Binnen de applicatie Presurf zijn de snapshots op hiërarchische wijze geordend. Deze hiërarchie verschilt uit praktische gronden enigszins van de hiërarchie op het niveau van het bestandssysteem (zoals die gehanteerd wordt bij de opslag van de snapshots).

Per klant worden de snapshots geordend in collecties. Collecties kunnen worden gebruikt om snapshots van sites onder één etiket te groeperen. Binnen een collectie is een overzicht te vinden van sites waar snapshots van gemaakt zijn. Sites worden in het overzicht getoond met hun titel. Deze titel

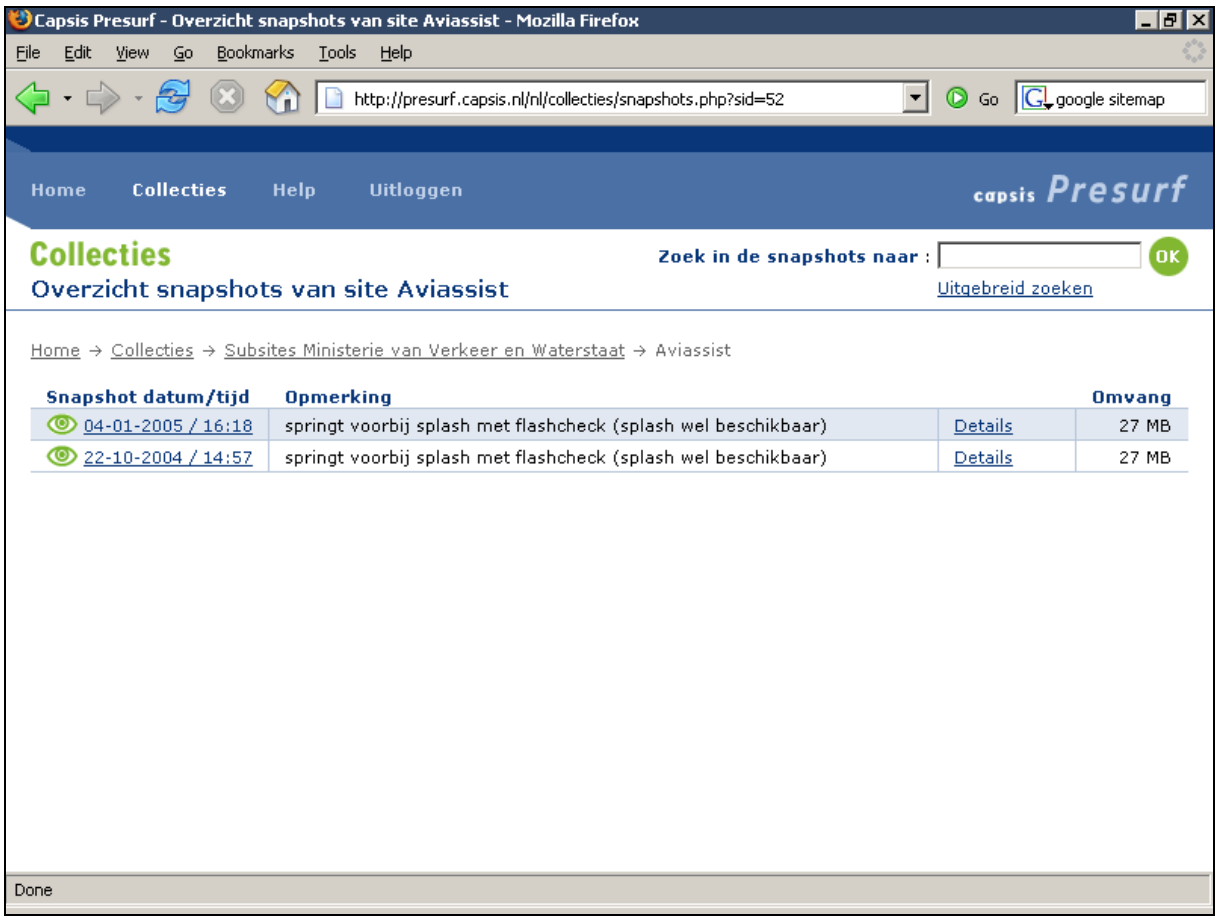
[16] De server is machine die draait onder het besturingssysteem Linux. Er wordt gebruik gemaakt van het Ext3-bestandssysteem.
[17] Dit is nodig omdat bestandsnamen binnen een directory op het bestandssysteem uniek moeten zijn.
[18] Zie http://www.httrack.org/
[19] De MD5-som is een praktisch unieke digitale handtekening van een bestand. Zodra één teken in een bestand verandert, als is het maar een spatie, dan verandert de MD5-som van het bestand .

is door Capsis bepaald op basis van de naam die de site op de beginpagina gebruikt om zichzelf te presenteren en/of de titel zoals die via de HTML-tag '<title>' op de homepage is aangegeven. In die gevallen waar deze titels niet of niet exact overeenkwamen heeft Capsis een keuze gemaakt.



Figuur 1. Overzicht met sites binnen een collectie in Presurf.

Na het aanklikken van de titel van een site wordt een overzicht gepresenteerd met alle snapshots van een die site. In dit overzicht worden de snapshots getoond met als label de datum van de snapshotactie. Door dit label aan te klikken wordt het snapshot in een nieuw venster getoond.



Figuur 2. Overzicht met snapshots binnen een site in Presurf.

Het bekijken van snapshots

De applicatie Presurf is voorzien van een geïntegreerde '*viewer*'-functionaliteit waarmee pagina's van de snapshots als ze opgevraagd worden om te bekijken eerst volgens een van tevoren ingestelde wijze bewerkt worden en pas dan opgestuurd. Er vindt op deze wijze dus een bewerkingsslag plaats (post processing, zie ook 1.4 in deze bijlage). Zo kan 'on the fly' iedere pagina in het archief bijvoorbeeld van de venster-titel "Archiefversie: " worden voorzien zonder dat de opgeslagen bestanden van het snapshot hiervoor aangepast hoeven te worden. Ook is het mogelijk om iedere pagina van een snapshot van een 'balkje' bovenaan te voorzien waardoor het zeer duidelijk is dat het een archiefversie betreft. Relevante metadata kunnen hierin op eenvoudige beschikbaar gesteld worden. Met de viewer is het ook mogelijk om links naar sites buiten het snapshot anders te laten werken, zoals ze buitenwerking te stellen of ze vooraf te laten gaan door een waarschuwing dat het archief wordt verlaten.

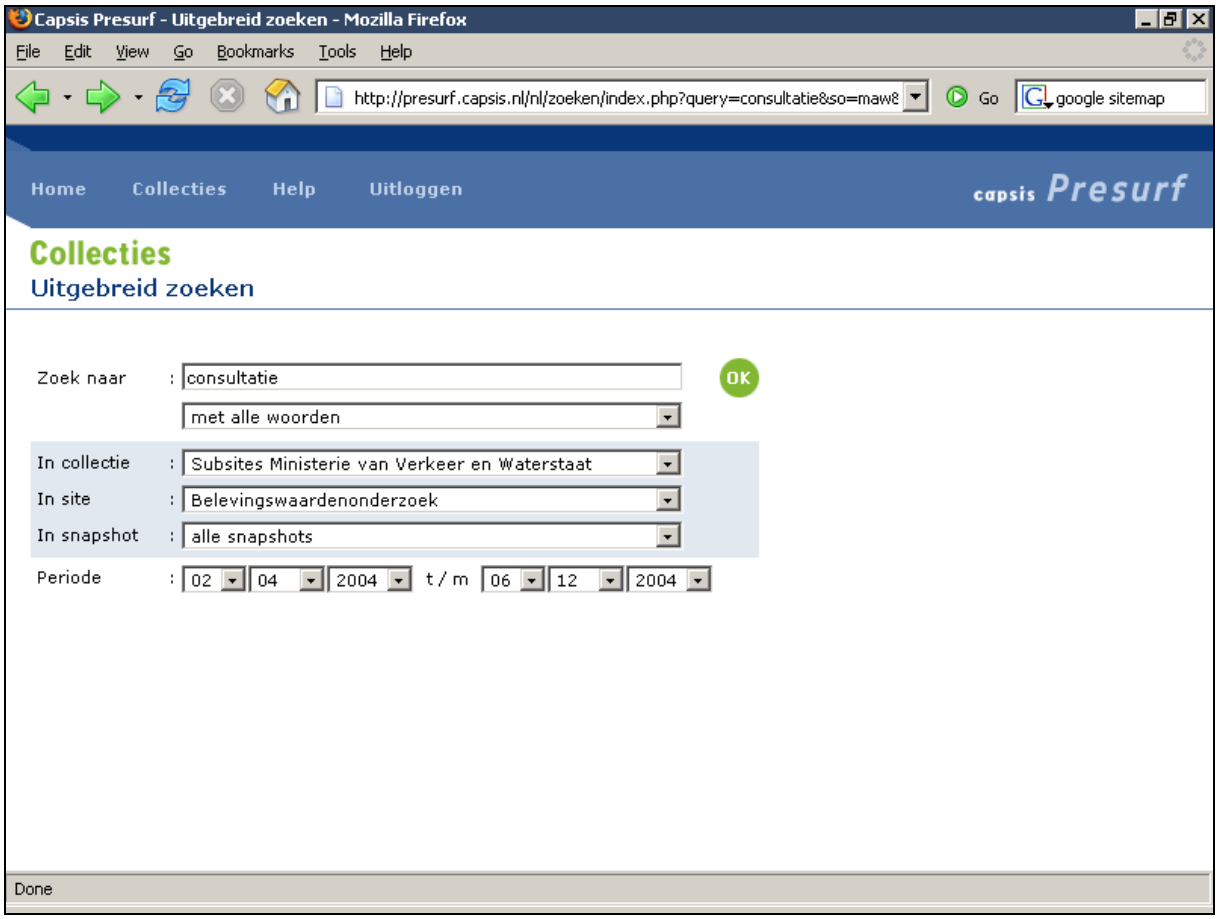
Aan het begin van het onderzoek is de deze functionaliteit van de viewer echter uitgeschakeld omdat de meest gebruikte webbrowser MS Internet Explorer foutmeldingen in de pagina's toonde (op andere veel gebruikte browsers zoals Firefox werkte de functionaliteit wel).
Een consequentie hiervan was dat voortaan enkel de URL-balk van de browser nog liet zien dat niet een originele site maar een snapshot werd bekeken[20].

De zoekfunctionaliteit

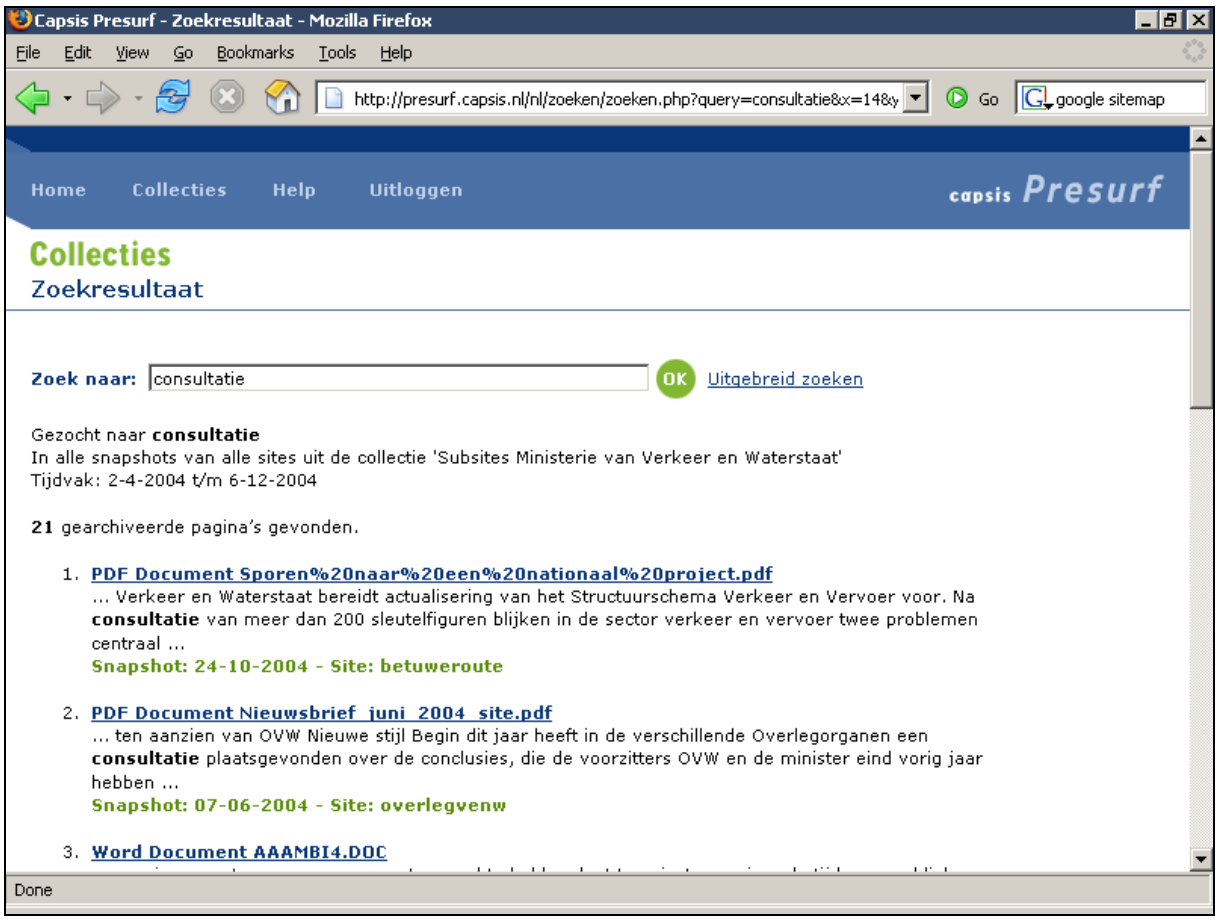
In de voor het onderzoek gebruikte versie van Presurf wordt een snapshot gevonden door door te de hiërarchische ordening heen te klikken. In aanvulling daarop konden specifieke pagina's van een snapshot gevonden worden door gebruik te maken van de *fulltext* zoekfunctie. Het domein waarbinnen gezocht werd kon hierbij begrensd worden tot snapshots binnen bijv. een specifieke

[20] De snapshots draaiden onder het domein presurf.capsis.nl. Zolang dit domein gebruikt wordt bij opvragingen van pagina's wordt het webarchief bekeken. Bij snapshots van sites met frames kan dit problemen opleveren omdat de URL-balk hier geen volledige indicatie biedt van de afkomst van de getoonde frames.

collectie of een specifieke site. Via een uitgebreidere zoekfunctionaliteit konden de zoekresultaten ook worden beperkt tot documenten van snapshots gemaakt in een specifieke periode.



Figuur 3. Interface voor uitgebreid zoeken in Presurf.



Figuur 4. De pagina met zoekresultaten.

5. Beoordeling geschiktheid snapshotmethode

De centrale onderzoeksvraag is 'wat kan er goed gearchiveerd worden met de snapshot-methode / met de applicatie Presurf'. Om deze vraag te kunnen beantwoorden is een model nodig dat gebruikt kan worden ter beoordeling van de kwaliteit van de snapshots.

In het kader van dit onderzoek is de kwaliteit van snapshots op drie aspecten beoordeeld:

- Visuele kwaliteit:
  zien de pagina's van het snapshot er hetzelfde uit als het origineel?
- Functionaliteit:
  werkt alles in het snapshot nog zoals in het origineel?
- Volledigheid:
  bevat het snapshot dezelfde pagina's en documenten als het origineel?

Functionaliteit en volledigheid: het 'deep web'

Het gebruik van de snapshotmethode impliceert dat alle functionaliteit die afhankelijk is van interactie van een gebruiker met processen op de webserver (anders dan het eenvoudig opvragen van een nieuwe pagina) niet meer zal functioneren. Hieronder valt bijvoorbeeld de zoekfunctionaliteit van een site. De informatie op een website die pas benaderd kan worden na specifieke database-bevragingen via een formulier op een webpagina, zoals een zoekfunctionaliteit, wordt doorgaans aangeduid met de term '*deep web*'. Bij *deep web* hangen de aspecten functionaliteit een volledigheid sterk samen: doordat zoekfunctionaliteit in het snapshot niet werkt zal het snapshot (doorgaans[21]) niet volledig zijn.

Deze beperking was voorafgaande aan het onderzoek bekend en is dus niet als zodanig onderzocht. Echter, of de snapshot-methode geschikt is voor de archivering van een website zal in sterke mate

[21] Als zoekresultaten ook expliciet via gewone links op de site te vinden zijn dan zullen deze documenten op normale wijze in het snapshot opgenomen worden.

bepaald worden de waarde van het *deep web* voor de te bewaren website. De keuze voor een specifieke archiverings- of conserveringstechniek hangt daarmee direct samen met het selectievraagstuk.
Hoewel er in dit onderzoek voor gekozen is het selectievraagstuk te laten rusten is er vanwege de bekende beperkingen van de snapshot-methode bij het binnenhalen van *deep web* wel aan de beoordelaars gevraagd hoe zij het belang van het *deep web* achten.

De wijze van beoordeling
Het is helaas niet mogelijk om de visuele kwaliteit, functionaliteit en volledigheid van een snapshot geautomatiseerd te beoordelen. Deze aspecten kunnen het beste worden beoordeeld door iemand die de betreffende site goed kent en precies weet wat er te vinden is, hoe de site functioneert en hoe deze er uit ziet. Er is daarom voor gekozen om een aselecte groep van 20 snapshots te laten beoordelen door personen binnen het departement van wie men gezien hun functie kan verwachten dat ze de betreffende site goed kennen. Hen is niet alleen gevraagd een oordeel te geven over de getrouwheid van het uiterlijk, de functionaliteit en de volledigheid, maar ook om aan te geven hoe belangrijk men het gemis aan *deep web* acht. De gehanteerde vragenlijst is te vinden in bijlage 3.2. De uitkomsten van de ingevulde vragenlijsten zijn te vinden in bijlage 3.3 en in een nader verwerkte vorm in bijlage 2.

Naast een beoordeling van 20 snapshots door inhoudelijk ingewijden is door René Voorburg van Capsis een beoordeling gemaakt van snapshots van alle voor dit project geconserveerde sites. Hierbij is per snapshot één geaggregeerde waardering gegeven, van 1 (snapshot geheel of grotendeels mislukt, onbruikbaar) tot 5 (snapshot lijkt een perfecte afspiegeling van de originele website). Bij de beoordeling is er door het snapshot heen geklikt terwijl de pagina's vergeleken werden met de originele website. Per snapshot is een beperkt aantal pagina's bekeken. Bij het bladeren door het snapshot is gepoogd gevarieerde pagina's te bekijken waarbij nadrukkelijk gelet is op bekende problemen zoals bijvoorbeeld met javascript gegenereerde menu's of gebruik van Macromedia Flash. De resultaten hiervan zijn te vinden in bijlage 3.1.

Analyse technische duurzaamheid gearchiveerde websites

Bij de beoordeling van de snapshots is ook aandacht besteed aan de te verwachten digitale duurzaamheid van de op de sites gehanteerde bestandsformaten. Van een duurzaam opgezette site wordt bijvoorbeeld verwacht dat deze opgemaakt is in correcte HTML, zoals HTML versie 4.01 of hoger, of XHTML versie 1.0 of hoger (Voorburg, 2004).

Voor het testen van de mate waarin HTML-standaarden gevolgd zijn is door Capsis een hulpmiddel ontwikkeld dat gebruik maakt van de vrij beschikbare code van de W3C-validator[22]. Met dit hulpmiddel werd van iedere pagina uit een snapshot op basis van het aangeven documenttype[23] onderzocht hoeveel fouten de opmaakcode bevat. De resultaten van deze tests zijn te vinden in bijlage 3.3. en worden besproken in bijlage 2.

Naast het formaat en de syntaxis van de HTML-bestanden is ook relevant in welke mate er duurzame en open standaarden gevolgd zijn voor bijvoorbeeld afbeeldingen en tekstbestanden (zie Voorburg 2004). In dit onderzoek is daar geen diepgaande analyse van gemaakt. Wel is er voor een de 20 nader onderzochte snapshots geanalyseerd welke MIME-types gebruikt zijn. Dit geeft een aardige indruk van de gehanteerde bestandssoorten maar biedt geen exact beeld van welke formaten er gehanteerd zijn[24]. Zie hiervoor bijlage 3. voor de resultaten en bijlage 2. voor een nadere analyse

De duurzaamheid van een website en daarmee van een snapshot zal ook in sterke mate bepaald worden door de gehanteerde *client side* technieken zoals javascripts en applicaties in bijvoorbeeld Flash en Java. In dit onderzoek is niet gekeken naar de duurzaamheid van dergelijke technieken, anders dan de zijdelinks gerelateerde vraag in welke mate deze technieken de inzet van de snapshot-methode hinderen.

[22] Zie ook http://validator.w3c.org. In het kader van dit onderzoek is specifiek gebruik gemaakt van de versie die beschikbaar is voor de Debian Linux distributie .
[23] Als er niet expliciet een DOCTYPE aanduiding gevonden werd hanteerde de validator de aanname dat er "HTML 4.01 transitional" toegepast was.
[24] Ter illustratie: een digitale video om af te spelen met de Apple Quicktime speler heeft als MIME type ' video/quicktime'. Dit bestand kan echter gecodeerd zijn in een open formaat zoals mpeg-4 dat ook door andere spelers afgespeeld kan worden of in een besloten formaat dat alleen door de Apple Quicktime speler afgespeeld kan worden.

Analyse technische belemmeringen gebruik snapshot-methode

Bij de beoordeling van de snapshots door Capsis is bij eventuele onvolkomenheden ook geanalyseerd waardoor deze onvolkomenheden veroorzaakt zijn. De resultaten van deze analyse zijn te vinden in bijlage 3.2. Zie ook bijlage 2. voor een bespreking van de uitkomsten.

Herhaalde snapshots, omvang en archiveringsfrequentie

Een aantal onderzoeksvragen heeft betrekking op de mate waarin pagina's van een site door de tijd heen veranderen. De mate waarin pagina's veranderen is een belangrijke factor voor het bepalen van de archiveringsfrequentie van de sites. De invloed op de benodigde opslagcapaciteit is daarmee tweevoudig. Ten eerste zal in het algemeen opgaan dat er vaker een snapshot gemaakt zal moeten worden als een website vaker verandert. Ten tweede zal de benodigde opslagcapaciteit voor een herhaalde snapshot groter zijn als in de tussentijd meer bestanden gewijzigd zijn. Bij een herhaalde snapshot hoeft een pagina die niet gewijzigd is immers niet opnieuw in zijn geheel opgeslagen te worden: een verwijzing[25] is voldoende.

Binnen dit project is om praktische redenen uiteindelijk geen onderzoek meer gedaan naar vragen met betrekking tot de verandering van sites. Nadat de eerste reeks snapshots gemaakt was en deze snapshots beoordeeld waren liep de voor dit project gestelde periode tegen het einde. Bovendien had ondertussen een migratie van de sites naar een geheel nieuw content management systeem plaatsgevonden. Voor deze migratie zijn de sites geheel opnieuw opgezet. Een consequentie hiervan is dat bij een geautomatiseerde analyse van de gewijzigde pagina's op basis van de MD5-som van een pagina naar verwachting ieder pagina als gewijzigd aangemerkt zou gaan worden. Dit zou uiteraard geen waardevolle informatie opleveren.

Overdracht naar het Nationaal Archief

Eisen aan overdracht naar het digitale depot

De in dit onderzoek gegenereerde snapshots zullen, indien ze daar op basis van selectiecriteria voor in aanmerking komen, uiteindelijk naar het Nationaal Archief overgedragen moeten kunnen worden. Het Nationaal Archief kent hier momenteel geen regelgeving voor anders dan gesteld in *Regeling geordende en toegankelijke staat archiefbescheiden* (2002). Websites kunnen om diverse redenen onmogelijke voldoen aan deze regeling (zie ook Voorburg 2004). Om websites op te kunnen nemen zal het Nationaal Archief nader te bepalen eisen moeten vaststellen. In dit kader is het relevant dat het Nationaal Archief werkt aan het vaststellen van *Functionial Requirements* (Functionele Eisen) voor het archiefsysteem waarin digitale bescheiden zoals websites uiteindelijk een plaats zullen moeten vinden (het digitale depot).

Uit overleg in de begeleidingsgroep van dit project kwam naar voren dat het Nationaal Archief er veel waarde aan hecht dat te bewaren websites benaderd kunnen worden zonder afhankelijk te zijn van bijzondere software zoals 'viewers'. Dit maakt dat juist de snapshotmethode zeer geschikt is omdat de hiermee gegenereerde snapshots zelfs *'offline'* te gebruiken zijn, dus direct met een webbrowser vanuit ieder gangbaar besturingssysteem en bestandssysteem.

In het kader van dit onderzoek is gesproken met Bill Roberts van het Engelse bedrijf Tessella. Hij werkt voor het Nationaal Archief aan het in kaart brengen van de Functionele Eisen voor het te realiseren digitale depot. De wijze van opslag van snapshots zoals door Presurf gehanteerd wordt (zie onder 4.3) (met behoud sitestructuur, zonder conversies van bestanden, gecombineerd met metadata in een XML-formaat) zou volgens hem naar verwachting naar het digitale depot overgedragen moeten kunnen worden. De precieze wijze waarop dit plaats zou moeten vinden is vooralsnog niet duidelijk.

Afhankelijkheid van het gehanteerde bestandssysteem

De opslag van snapshots en bestanden in die snapshots gebeurt nu in Capsis Presurf door gebruik te maken de het bestandssysteem van de server waarop Presurf draait. Deze aanpak kent risico's gerelateerd aan beperkingen in gehanteerde bestandssystemen en ten aanzien van de wijze waarop bestanden worden opgeslagen. Dit probleem kan met name spelen als snapshots verhuisd worden naar een ander bestandssysteem met meer, of andere beperkingen. De volgende problemen kunnen spelen:

[25] Onder Unix-bestandssystemen kan dit bijvoorbeeld met een zogenaamde symbolische link, te vergelijken met een 'snelkoppeling' onder MS-Windows of een 'alias' onder Mac OS.

- Onder het door Presurf gebruikte Ext3 bestandssysteem kunnen tekens gebruikt worden in de bestandsnamen of directories die onder het besturingssysteem MS Windows niet gebruikt mogen worden.
- De lengte van het pad (de combinatie van directories en de bestandsnaam) van een bestand is onder MS Windows beperkt tot 255 tekens. Unix / Linux kent hier doorgaans geen beperking.
- MS Windows maakt geen onderscheid tussen hoofdletters en kleine letters. Unix/ Linux-bestandssystemen doen dit wel.

Bij migratie van een snapshot van Unix/Linux naar MS Windows kunnen dus problemen verwacht worden. Specifiek bestaat de mogelijkheid dat links niet meer werken doordat de namen van sommige bestanden gewijzigd (ingekort of aangepast) zijn of dat sommige bestanden geheel ontbreken.

Er is niet onderzocht hoe groot de problemen zullen zijn bij het overplaatsen van de gegenereerde snapshots naar een MS Windows-omgeving. Een snelle analyse van migratie van snapshots[26] van het intranet van het Ministerie van Verkeer en Waterstaat (parallel aan dit onderzoek gemaakt) van Ext3 (Linux) naar NTFS (MS Windows 2000) liet hooguit op zeer beperkte schaal problemen zien.

Hoewel het probleem dus niet groot lijkt te zijn, zeker niet binnen een Unix / Linux-omgeving, zou een meer duurzame oplossing kunnen zijn om de opslag van snapshots onafhankelijk te maken van het gebruikte besturingssysteem. Dit kan door de bestanden gecombineerd in één bestand volgens specifiek formaat op te slaan. Deze aanpak wordt ook gehanteerd door het initiatief *the Internet Archive*[27] . Deze organisatie slaat snapshots op in het door hen gedefinieerde open 'ARC'-formaat[28]. Dit heeft wel tot gevolg dat de snapshots niet meer 'offline' benaderd kunnen worden maar enkel nog via een server-applicatie die het gehanteerde formaat 'verstaat'. De voor het toegankelijk maken van de snapshots gebruikte server zal hierdoor naar verwachting aanmerkelijk zwaarder belast worden.

Deze problematiek verdient verder onderzoek.

De bestaande snapshots in Presurf kunnen met de vastgelegde metadata alsnog omgezet worden naar bijvoorbeeld het ARC-formaat zonder dat er relevante informatie verloren gaat.

Metadata

Een goed geordend archief kan zoals gesteld niet zonder metadata. Ook voor het Nationaal Archief zou het wenselijk zijn dat er een uniforme standaard gevolgd wordt met betrekking tot de te hanteren metadata. Richtinggevend zal hier ongetwijfeld ISO 15489:2001 zijn, de *Records Management Standard*. Afgeleid daarvan vormt ISO 23081-1:2004 een hulpmiddel voor het gebruik en de implementatie van metadata. Deze standaard is momenteel nog in conceptvorm.

In dit onderzoek is gekeken of de Australische *Record Keeping Metadata Standard for Commonwealth Agencies* (National Archives of Australia, 1999) als uitgangspunt gebruikt kan worden voor metadata voor websites of snapshots van websites van het Ministerie van Verkeer en Waterstaat.

De velden van de Australische set en de geschiktheid hiervan voor de archivering van websites is uitgebreid besproken (onder andere via een mailinglijst).. Dit heeft geleid tot een aanzet voor een metadataset zoals te vinden is in bijlage D.

Terzijde: archivering van intranet

Parallel aan dit project zijn er met Capsis Presurf sites op het intranet van het Ministerie van Verkeer en Waterstaat geconserveerd. In het algemeen bleek dit niet zeer verschillend van het archiveren van websites.
Ook bij het intranet bleek de lijst met aangeleverde URLs van sites niet volledig. Een belangrijk verschil met het internet was dat er veel vaker redirects voorkwamen. Doordat de redirects en aliassen niet op de aangeleverde lijst vermeld werden resulteerde dit vaak in snapshots die enkel een pagina met een redirect-melding bevatten. Om dit probleem te omzeilen is er voor gekozen om daarna alle top het intranet te vinden websites in één snapshot op te nemen.

[26] Het ging hierbij om een grote set met snapshots, in totaal voor 8 Gb aan data.
[27] http://www.archive.org/
[28] Zie http://www.archive.org/web/researcher/ArcFileFormat.php

De resulterende snapshots zijn niet uitgebreid geanalyseerd.

De snapshots zijn op twee DVD's aan het ministerie aangeleverd. Als formaat is er hier gekozen voor een gecomprimeerd "tar" bestand. Op deze wijze konden de snapshots via een DVD overgedragen worden zonder dat er problemen kunnen ontstaan ten gevolgde van de migratie naar een ander bestandssysteem.

Bijlage B. De kwaliteit van de gegenereerde snapshots

1. Inleiding

Dit hoofdstuk bespreekt de kwaliteit van de verkregen snapshots zoals naar voren kwam bij de beoordeling van de snapshots door Capsis, door inhoudelijk betrokken medewerkers van het departement en de technische analyse.

Verwijzing naar specifieke snapshots of sites gebeurt hier met de door Capsis gehanteerde codering. Zowel bij sites als bij snapshots verwijst het eerste deel van de codering naar zowel Capsis als naar de partij waarvoor webarchieven gegenereerd zijn. Dit gebeurt met de code `capsis.arc/100030` . Het getal verwijst hierbij naar de klantcode die Capsis aan het ministerie toegekend heeft. Voor een site wordt deze code gevolgd door de door Capsis gekozen verkorte titel van de site. Voor een snapshot wordt deze code gevolgd door eerst de datum van het snapshot (ISO 8601 zonder specificatie van tijd) en daarna de verkorte titel van de site. Tussen de te onderscheiden delen van deze identificatie-codes wordt een schuine streep ("/") geplaatst.

Ter illustratie:

capsis.arc/100030/20041123/noordzee	Een verwijzing naar de snapshot van 23 november 2004 van de site met verkorte titel 'noordzee'.
capsis.arc/100030/noordzee	Een verwijzing naar de website (of groep snapshots) met de verkorte titel 'noordzee'.

2. Algemene beoordeling van de snapshots: problemen en oorzaken

De door Capsis uitgevoerde algemene beoordeling van de snapshots (zie bijlage C.1.) geeft het volgende beeld van de kwaliteit van de snapshots (zie bijlage A.5 voor de methode).

Tabel 1. Verdeling van score van de snapshots.

Score:	**Aantal:**	**%:**
1(geheel of grotendeels mislukt)	5	5.3
2 (ernstige afwijkingen)	11	11.6
3 (grotere afwijkingen)	8	8.5
4 (kleine afwijkingen	5	5.3
5 (geen afwijkingen)	66	69.5

Met het oog op het verbeteren van kwaliteit van de snapshots is het zinvol na te gaan waardoor niet van alle websites een perfecte snapshot gemaakt is. Hier is door Capsis een analyse van gemaakt. Het resultaat hiervan kan samengevat worden in de volgende tabel:

Tabel 2: Oorzaken sub-optimale kwaliteit snapshots

Categorie oorzaken:	**Aantal:**	**%:**
afwijkingen ten gevolge fouten bronserver	1	1,1
afwijkingen ten gevolge onvolkomenheden in crawler of viewer van Presurf	3	3,2
problemen gerelateerd aan wijze van gebruik van URL parameters	3	3,2
javascript-gerelateerde problemen	14	14,7
problemen ten gevolge van implementatie 'browser checks'	3	3,2
problemen gerelateerd aan het gebruik van Macromedia Flash.	5	5,3
problemen gerelateerd aan de implementatie van sessie-management	2	2,1

In het volgende worden de oorzaken nader besproken

Afwijkingen ten gevolge van fouten van de bronserver

In slechts één geval werden er grotere afwijkingen in het snapshot geconstateerd die waarschijnlijk toegeschreven zullen moeten worden aan problemen op de bronserver (de webserver) die zich voordoen in combinatie met de gebruikte archiveringsmethode.

Snapshot capsis.arc/100030/20041112/0800-rws (van de site zoals beschikbaar onder http://www.08008002-rijkswaterstaat.nl/) bevat afbeeldingen die fouten bevatten. De aard van de fouten suggereert dat de bronserver op onjuiste wijze *'caching'* geïmplementeerd heeft. De bestanden van de afbeeldingen met fouten bestaan namelijk uit drie delen:

- data, op het oog overeenkomend met de HTML-pagina waar de afbeeldingen in opgenomen zijn,
- 'header'-informatie zoals de webserver met de client uitwisselt voorafgaande aan de overdracht van een eigenlijke bestand,
- binaire informatie, waarschijnlijk de data van de afbeelding zelf.

De bronserver maakt gebruik van Apache webserver en de Tomcat module van het Jakarta-project voor gebruik van zogenaamde Java Servlets voor server-side scripts[29].

Afwijkingen ten gevolge van onvolkomenheden in de crawler of viewer van Presurf

In drie gevallen traden er afwijkingen op die na analyse toegeschreven zijn aan onvolkomenheden in de crawler of viewer van Presurf.

Snapshot capsis.arc/100030/20041123/noordzee (van de site beschikbaar onder http://www.noordzee.org) geeft voor een bestand een *'file not found'* melding. Het bestand is echter wel opgenomen in het archief. Het probleem ontstaat doordat het door Capsis Presurf gebruikte bestandssysteem onderscheid maakt tussen hoofdletters en kleine letters terwijl het bestandssysteem van de bronserver dat niet doet (een systeem dat werkt onder MS Windows 2000).
De snapshots capsis.arc/100030/20041112/actuelewaterdata (van de site beschikbaar onder http://www.actuelewaterdata.nl/) en capsis.arc/100030/20041112/fileplanregiorotterdam (van de site beschikbaar onder http://www.fileplanregiorotterdam.nl/) bevatten hyperlinks die in de snapshot niet meer werken (respectievelijk *'file not found'*-meldingen in een 'cgi-bin'-directory[30] en *'access denied'*-meldingen). Een exacte oorzaak werd in dit geval niet achterhaald.

Overigens deden zich bij de uitgebreide analyse van de sites door contactpersonen van het ministerie nog twee andere problemen voor die gerelateerd waren aan een onvolkomenheid in Presurf (zie bijlage B.3.). Deze problemen waren bij de door Capsis uitgevoerde algemene analyse ondertussen verholpen.

Problemen gerelateerd aan de wijze van gebruik van URL-parameters

In ieder geval drie keer ontstonden er problemen bij het binnenhalen van een URL door de wijze waarop de site gebruik maakt van parameters in de URL[31]. Met URL-parameters is het mogelijk om met een beperkt aantal bestanden een praktisch onbeperkt aantal pagina's te genereren, in de zin van een onbeperkt aantal verschillende URLs. Voor de archiveringsapplicatie is een pagina met een andere URL (inclusief de parameters) altijd een andere pagina (ongeacht de inhoud) en dus een pagina die óók gearchiveerd moet worden.
Het probleem kan zich op verschillende manieren uiten:

De oneindige agenda

Een relatief bekend voorbeeld is de website met een dynamisch gegenereerde agenda. Iedere pagina van de agenda heeft links "volgende" en "vorige" waarmee de site een praktisch oneindige agenda bevat met praktisch oneindig veel pagina's.

Vervuiling van URL-parameters

Een ander voorbeeld is het gebruikt van een URL -parameter als "print=1" om iedere pagina op de site een variant te geven met een vormgeving die specifiek geschikt is om te printen. Dit gaat mis als de printbare variant conform het origineel ook een link naar een printbare variant biedt die met als URL-parameters "print=1&print=1". Deze pagina zal dan vervolgens weer een link bieden naar een pagina met als URL-parameters "print=1&print=1&print=1" en dit gaat zo door *ad infinitum*. Er kan gesteld worden dat hier sprake is van vervuiling van de URL-parameters.

Verrijking met links naar zoekresultaten

[29] Volgens de analyse van Netcraft, zie http://www.netcraft.com/
[30] In directories op een webserver met de naam 'cgi-bin' zijn doorgaans specifiek server-side scripts te vinden.
[31] Een URL-parameter is een variabele die aan een webpagina wordt meegegeven door deze in de URL op te nemen. Een fictief voorbeeld is http://nieuws.nl/bericht.php?id=1 waarbij 'id' de naam van de variabele is. In dit geval zou zo een nieuwsbericht getoond worden dat in de database geregistreerd is als bericht nummer 1.

Een andere variant heeft te maken met de integratie van de resultaten van de zoekmachine van de site. Op de site Interwad (capsis.arc/100030/20041112/interwad) zijn documenten te vinden die een links bieden naar de zoekresultaten van verschillende trefwoorden voor de betreffende pagina. Iedere nieuwe pagina op de site leidt zo met deze pagina's met zoekresultaten in een zeer groot aantal nieuwe unieke pagina's.

Verveelvoudiging door aanbieden van meerdere 'views'
De pagina's van sommige website kunnen op meerdere manieren bekeken worden, bijvoorbeeld met data op verschillende wijzen gesorteerd, gebruikmakend van verschillende vormgevingssjablonen, of andere soms zeer specifieke modificaties. Doorgaans worden deze varianten ingesteld met behulp van URL-parameters. Een gevolg hiervan is dat ieder pagina meerdere varianten kent. Zeker als de ingestelde opties ook kunnen overerven naar dieper gelegen pagina's dan kan een inhoudelijk beperkt onderdeel van een website snel tot een praktisch onbruikbaar groot aantal pagina's leiden.

Javascript-gerelateerde problemen
De meeste problemen zijn gerelateerd aan het gebruik van javascripts. Javascript wordt vooral veel gebruikt om 'client-side' dynamische menu's te maken. Bij de gebruikte archiveringsmethode levert dit met name twee soorten problemen op. Snapshots zijn soms onvolledig ten gevolge van problemen met het *parsen* van de scripts door de applicatie. Ook komt het vaak voor dat het snapshot niet goed getoond kan worden doordat de scripts alleen werken binnen een gespecificeerd 'absoluut' pad[32].

Onvolledige snapshots ten gevolge van problemen met parsen van scripts
De crawler analyseert alle URLs die het op de site vindt en haalt deze vervolgens binnen, voor zover de gevonden URL binnen het gespecificeerde domein valt. De crawler probeert de URLs in javascripts ook te herkennen maar kan dit alleen in relatief eenvoudige situaties. In gevallen waar een URL opgebouwd wordt door bijvoorbeeld javascript-variabelen te combineren, bezit de crawler onvoldoende intelligentie de URL te herkennen. Als gevolg hiervan zal het betreffende bestand niet binnengehaald worden (tenzij de URL elders op de site op een wel te herkennen wijze opgenomen is).

Niet goed functionerende menu's door afhankelijkheid van een specifiek pad.
Javascript-menu's gaven regelmatig problemen doordat deze zo geprogrammeerd waren om met absolute URLs te werken. Doordat de snapshots een andere locatie hebben dan de oorspronkelijke scripts werken de links in script niet meer.
Bij HTML-bestanden herschrijft de crawler eventuele absolute URLs naar relatieve URLs zodat de links in de snapshot ongeachte de locatie van de snapshot werken. De huidige versie van de applicatie is niet in staat de links in javascripts te herschrijven naar relatieve URLs. Bovenal zal het vaak niet goed mogelijk zijn om in een javascript met relatieve URLs te laten werken, dit aangezien de scripts doorgaans vanaf verschillende locaties opgevraagd zullen worden. Hierdoor zal er in een script afhankelijk van de locatie van waaruit het opgevraagd wordt telkens een ander relatief pad nodig zijn. Om dit probleem te voorkomen wordt bij javascripts door ontwikkelaars in het algemeen gebruik gemaakt van absolute URLs of absolute paden.

Overige javascript-gerelateerde problemen
Het bovenstaande biedt geen uitputtend overzicht van geconstateerde problemen met javascripts. Noemenswaardige problemen traden bijvoorbeeld ook op met zogenaamde 'browserchecks' op basis van javascript. Dit probleem wordt apart vermeld en besproken.
Een probleem dat ook vaak voorkomt heeft vooral cosmetische consequenties. Veel sites hebben aan plaatjes met een javascript een zogenaamd *'mouse over'*-effect gekoppeld. Dit houdt in dat er op de site een ander plaatje getoond wordt zodra en zolang de muiswijzer zich boven het plaatje bevindt. Doorgaans werken deze *'mouse overs'* niet meer naar behoren in de snapshots.
Nog een voorbeeld van problemen die met javascript kunnen optreden: een site (snapshot capsis.arc/100030/20041126/riza) bevat een door javascript gegeneerde link naar een bestand waarvan de naam deels bestaat uit de naam van de actuele maand. Dit deel van het snapshot leek prima te functioneren, totdat er een nieuwe maand was aangebroken. Gesteld kan worden dat de functionaliteit van het snapshot hier onvoldoende 'bevroren' is.

[32] Onder een absolute URL word hier zowel een URL als "http://website.org/javascript.js" verstaan als ook een URL waar de servernaam niet in opgenomen is maar wel de webroot, dus "/javascript.js". In het tweede voorbeeld kan ook gesproken worden over een absoluut 'pad' . Een relatieve URL of een relatief pad is bijvoorbeeld "javascript.js" of "../javascript/js" en is dus relatief ten opzichte van de pagina die de verwijzing bevat.

Problemen ten gevolge van de implementatie van 'browser checks'

Sommige pagina's tonen hun inhoud afhankelijk van de opgestuurde identificatie van de bezoekende browser (de '*user agent*'-*string*). Op deze wijze hoopt de webbouwer een voor de bezoekende browser geoptimaliseerde website te kunnen tonen. Bij een onjuiste implementatie[33] levert dit problemen als de site bezocht wordt met een minder gangbare browser. De *crawler* van het archiveringsprogramma is daar een voorbeeld van. Deze situatie kan zich zowel bij *server side* als bij *client side browser checks* voordoen.
Bij client side browser checks in javascript kan zich bovendien het probleem voordoen dat het javascript niet of niet naar behoren werkt[34]. In die gevallen zal er geen pagina ingelezen kunnen worden. Een interessant probleem met javascript en *browser checks* deed zich voor bij snapshot capsis.arc/100030/20041112/aanlega50. Bij de capture van deze site werd via een browser check op basis van de '*user agent*'-*string* van de crawler een speciale 'lage resolutie'-versie van deze site binnengehaald. Bij het bekijken van dit snapshot werd het javascript actief en dat concludeerde op basis van de *'user agent'-string* van de browser dat de 'hoge resolutie'-versie ingeladen zou moeten worden. De bestanden specifiek voor die versie waren echter niet in het snapshot opgenomen. Dit probleem is in een later snapshot overigens verholpen door expliciet ook de speciale 'hoge resolutie'-bestanden op te nemen in de capture-opdracht.

Problemen gerelateerd aan het gebruik van Macromedia Flash.

Als zogenaamde '*client side*'-techniek kunnen de problemen veroorzaakt door het gebruik van Macromedia Flash in hoofdlijnen vergeleken worden met de problemen veroorzaakt door het gebruik van javascripts. De *crawler* probeert Flash te *parsen* maar slaagt hier niet altijd in. De *crawler* kan de URLs in Flash ook niet herschrijven waardoor absolute URLs in Flash niet meer zullen functioneren.

Problemen gerelateerd aan de implementatie van sessie-management

De *crawler* van de archiveringsapplicatie heeft als uitgangspunt dat gelijke URLs ook gelijke pagina's bevatten. In twee gevallen bleek dit niet op te gaan. De betreffende sites bieden hun informatie aan in verschillende talen. Of een pagina informatie in een andere taal toonde dan de standaard taal (Nederlands) werd bepaald door bij te houden[35] of de bezoeker door een specifieke link aan te klikken (een vlaggetje van een land) eerder aangeven heeft de pagina's in een andere taal te willen bekijken. Als gevolg hiervan werden de pagina's van deze sites in het Nederlands binnengehaald totdat de crawler de link koos die leidde tot de keuze van een andere taal. De rest van de pagina's van de site werd vervolgens in deze taal gearchiveerd, totdat weer een andere taal werd gekozen. Een snapshot werd zo in verschillende talen gearchiveerd, maar iedere pagina maar in één taal.

3. Beoordeling door inhoudelijk betrokken medewerkers

De interpretatie van de vragen

Naar aanleiding van de gevarieerde wijze waarop de formulieren ingevuld zijn is er voor gekozen om de beoordeling van de snapshots niet op een geaggregeerde wijze in tabel of grafiek te presenteren. Voor de uitkomsten wordt verwezen naar de individuele reacties zoals deze te vinden zijn in bijlage C.3. De wijze waarop de formulieren ingevuld zijn suggereert namelijk dat de vragen op zeer verschillende wijze geïnterpreteerd zijn. Geregeld werden ook niet alle vragen beantwoord. Met name het onderscheid tussen vormgeving, functionaliteit en volledigheid lijkt door veel beoordelaars als kunstmatig en weinig praktisch te worden ervaren.

De vraag over het belang van het *'deep web*' lijkt door sommigen vooral geïnterpreteerd te zijn met betrekking tot de actuele online website en door anderen met betrekking tot een webarchief. De reactie van één respondent bevestigt dit beeld. Na aangegeven te hebben het *deep web* zeer belangrijk te vinden wordt daaraan als reactie toegevoegd "*Maar voor een archiefexemplaar niet interessant*". Een andere respondent geeft aan " *Ik denk dat het* [deep web - RV] *minder relevant is voor een archief. Hier bekijk je de gepubliceerde informatie en ben je niet geïnteresseerd in interactie. Die zou wel beschreven moeten zijn maar hoeft niet uitvoerbaar te zijn*."

[33] Bij een goede implementatie wordt er door de server een versie opgestuurd die voldoet aan gangbare webstandaarden, tenzij er een specifieke afwijkende browser gedetecteerd wordt.
[34] Ongeveer 5% van de browsers op internet heeft geen javascript of heeft javascript uitgeschakeld staan. In aanbevelingen voor een veilig internetgebruik wordt dikwijls aangeraden javascript in de browser uit te schakelen.
[35] Er is niet uitgezocht op welke wijze dit gebeurde. Mogelijk via gebruik van een zogenaamd *'cookie'* of anders via *server side* sessiebeheer.

Een conclusie is dat het lastig is om de waarde van snapshots te beoordelen zonder daarbij in te gaan op vragen rondom waardering en selectie.

Over verschillen in de beoordeling

Als er voorzichtig geprobeerd wordt de beoordeling zoals die uit de reacties spreekt te leggen naast de beoordeling van Capsis zelf dan lijken er geen grote verschillen te bestaan. Daar waar er grotere verschillen gevonden worden kunnen die op het objectieve vlak waarschijnlijk meestal toegeschreven aan een twee fouten in Presurf[36]. Deze fouten waren ondertussen hersteld toen Capsis zijn analyse uitvoerde, wat een in geringe mate positiever beeld van de beoordeling door Capsis kan verklaren.

Bij het vergelijken van de beoordelingen zal uiteraard ook rekening moeten worden gehouden met een vertekening die ontstaan kan zijn doordat een medewerker van Capsis in het kader van dit project moeilijk volledig objectief kan oordelen. Bovendien speelt enerzijds dat Capsis door de opgebouwde kennis van mogelijke problemen bij de inzet van de techniek eerder afwijkingen zal ontdekken en herkennen. Anderzijds maakte de grote hoeveelheid van door Capsis beoordeelde sites dat de beoordeling wellicht oppervlakkiger is uitgevoerd dan door de inhoudelijk betrokkenen.

Hoewel de meerwaarde van beoordeling door inhoudelijk betrokkenen niet groot lijkt, is in dit project de wijze waarop een snapshot van een site gemaakt is een paar keer op basis van een reactie van een beoordelaar aangepast[37].

Door direct na het maken van een snapshot dit snapshot pagina voor pagina met de site te vergelijken zal ook een niet inhoudelijk betrokken persoon na kunnen gaan in welke mate de site en het snapshot identiek zijn. Dit kan bij grote websites een zeer arbeidsintensief karwei vormen. Een inhoudelijk betrokken persoon zal hierin waarschijnlijk eerder / op efficiëntere wijze een oordeel kunnen vellen. De beoordeling van Capsis heeft plaatsgevonden na de beoordeling van inhoudelijk betrokkenen en daarmee tot soms een paar maanden nadat het snapshot gegeneerd was. Een inhoudelijk oordeel ("is dit de juiste en de volledige content van de site?") was door wijzigingen in de sites niet altijd goed mogelijk.. Puur op basis van een technische analyse (zoals foutmeldingen in scripts of *'file not found'*-meldingen) worden echter belangrijke aanwijzingen verkregen dat een snapshot van een site onvolledig is.

4. Technische analyse

Kwaliteit gebruikte HTML-opmaak

De resultaten van de analyse van de kwaliteit van de HTML-opmaak van de nader geanalyseerde sites zijn te vinden in bijlage C.3. Daar waar bij '*aantal fouten*' het teken "-" is genoteerd kon de software geen analyse uitvoeren. In een paar gevallen is bekeken waardoor dit veroorzaakt werd. Dit bleek in die gevallen het gevolg te zijn van een voor de verwachte tekenset onjuiste codering van de karakters.

Samengevat kan gesteld worden dat foutloze websites niet voorkomen. Foutloze HTML-bestanden zijn zeldzaam. De validator geeft aan dat sommige bestanden erg veel fouten bevatten. Dit zou in sommige gevallen verklaard kunnen worden doordat de validator bij gebrek aan een DocumentType[38] een onjuiste aanname doet ten aanzien van het gehanteerde formaat[39].

Ter referentie van de resultaten van de analyse van de kwaliteit van de HTML is een snapshot van de website op http://webrichtlijnen.overheid.nl gemaakt. Deze website vormt een interessant referentie om het het strikt volgens standaarden zoals HTML 4.01 of XHTML 1.0 promoot. De website geeft aan zelf geheel volgens deze standaarden ontwikkeld te zijn. Zoals ook te zien in bijlage D. bleek zelfs bij deze website een niet te verwaarlozen deel van de pagina's fouten te bevatten. Hieruit mag geconcludeerd worden dat het niet eenvoudig is foutloze opmaakcode te produceren en dat een validatie een geautomatiseerd hulpmiddel daartoe onmisbaar is.

[36] Presurf handelde zogenaamde gecodeerde tekens zoals de "%20" voor een spatie op onjuiste wijze af. Bovendien werden bij MS -Word en PDF bestanden onjuiste MIME-types verstuurd waardoor deze via een webbrowser niet leesbaar getoond werden.
[37] De beoordeling van Capsis heeft plaatsgevonden na deze aanpassingen.
[38] Een HTML-bestand volgens nieuwere specificaties wordt geacht bovenaan in het bestand met een DocumentType specificatie aan te geven welke specificatie van de HTML-standaard gevolgd is.
[39] Bij gebrek aan een gespecificeerd DocumentType wordt de aanname gehanteerd dat 'HTML 4.01 transitional' gevolgd is.

Gebruikte MIME types of bestandsformaten

De onderstaande tabel presenteert op geaggregeerde wijze de frequentietabellen met de gehanteerde MIME-types van de bestanden van de nader geanalyseerde websites (zie ook bijlage C.) .

Tabel 3: Geaggregeerde frequentietabel MIME-types.

Aantal:	*MIME-type*
7758	text/html
2433	image/gif
1762	application/pdf
1443	image/jpeg
570	image/pjpeg
365	application/msword
60	text/css
46	application/x-javascript
30	application/zip
17	video/mpeg
14	application/vnd.ms-powerpoint
12	text/plain
8	application/vnd.ms-excel
7	application/octet-stream
7	application/x-shockwave-flash
6	audio/mpeg
4	application/postscript
4	application/x-zip-compressed
4	video/x-ms-asf
3	application/vnd.rn-realmedia
3	audio/x-wav
3	image/bmp
3	text/xml
3	video/avi
2	video/quicktime
1	image/png
1	video/unknown

Deze tabel geeft de MIME types van de verzonden bestanden. Bij de beoordeling zal met een aantal zaken rekening moeten worden gehouden.

- De tabel geeft bijvoorbeeld aan dat er in totaal 46 javascripts gevonden werden in de geanalyseerde sites (MIME type "application/x-javascript"). Javascript wordt echter vaak geïntegreerd in de HTML-bestanden (MIME type "text/html"). Dit betekent dat Javascript feitelijk frequenter toegepast wordt dan hier gesuggereerd. Het zelfde verhaal speelt voor zogenaamde Cascading Style Sheet (MIME type "text/css").
- Het gaat hier bovendien om de bestanden zoals ze door de archiveringsapplicatie gevonden konden worden. In gevallen waar links niet goed gevonden konden worden door de archiveringsapplicatie zoals in sommige gevallen bij gebruik van javascript of Macromedia Flash ("application/x-shockwave flash") zal er een vertekend beeld kunnen optreden. Een grote site die geheel uit Flash is opgebouwd zal met de gehanteerde methode mogelijk maar één Flash bestand opleveren terwijl er in praktijk zeer veel zouden kunnen zijn).
- Een aanduiding van het MIME type geeft enkel een grove indicatie van het gehanteerde bestandsformaat. Vaker is het eerder een aanduiding van een applicatie die het bestand zou moeten (kunnen) openen of van de applicatie waarmee het bestand geschreven is.

Wat betreft de duurzaamheid van bestandsformaten geeft de tabel aan dat het, naast aandacht voor een standaardgetrouwe opmaak van HTML-bestanden, zinvol is nader te kijken naar de

duurzaamheid van de gebruikte Gif- bestanden (MIME type "image/gif") van Jpeg-bestanden (MIME types "image/jpeg" en "image/pjpeg") en van PDF-bestanden (MIME type "application/pdf").

Vanuit het oogpunt dat met name populaire 'open' bestandsformaten de meeste garanties voor duurzaamheid lijken te bieden (Voorburg 2004) valt het op dat er nog veel gebruik gemaakt wordt van bestanden in het MS Word-formaat. Vaak zal het in de ministeriële regels aanbevolen formaat PDF ook goed gebruikt kunnen worden[40]. Dezelfde constatering kan gedaan worden voor bestanden in het formaat van MS-Powerpoint. Ook hier zal PDF vaak kunnen volstaan en een naar verwachting meer duurzaam alternatief bieden.

Naarmate video belangrijker gaat worden zal de keuze voor een duurzaam formaat voor video belangrijker worden. Nu mag gesteld worden dat er voor video een allegaartje aan MIME types (/bestandsformaten) gebruikt wordt.

Omvang van sites en snapshots

Gedurende het onderzoek zijn er in totaal 114 snapshots gemaakt. Zonder gebruik te maken van compressietechnieken was er hiervoor in totaal 13220 Mb opslagruimte nodig. Een gemiddelde snapshot heeft hiermee een omvang van 116 Mb.

Van de nader geanalyseerde snapshots (zie bijlage C.3) is de omvang gemiddeld 123 Mb. Er kunnen echter grote verschillen optreden, de standaardafwijking bedraagt hier namelijk 125 Mb. Gemiddeld telt een snapshot (op basis van de nader geanalyseerde set zoals te vinden in bijlage C) 784 bestanden.

[40] Een terecht uitzondering zou kunnen zijn wanneer het bestand ook door gebruikers van de site gebruikt om kunnen worden om na download te bewerken.WAT STAAT HIER?

Bijlage C. Results

1. Overall assessment

Overview of issues and scores

The following table presents an overview of issues and scores from snapshots of all sites captured. This assessment was performed by René Voorburg of Capsis.

Table 1: Overview of issues and scores.

Snapshot (capsis.arc/100030/)	Remarks	Issues (*)	Score (**)
20041112/0800-rws	Some images were captured incorrectly. Problems with caching at the origination server seem to be a likely cause. These captured image-files appeared to consist of three parts: • data similar to the HTML-files of the pages of the sites • header-information like it is exchanged when a webserver communicates with a browser • binary data, probably the actual image.	a	3
20041221/a2dbehv	See extended description.		5
20041112/a2denbosch	See extended description.		5
20041112/a2utrechtdenbosch			5
20041112/a28zwolle			5
20041112/a2info			5
20041112/a30			5
20041112/actuelewaterdata	Some hyperlinks don't function. The frameset involved tries to load a page from a 'cgi-bin' directory (A 'cgi-bin' directory is normally preserved for scripts or programs that run on the webserver .). It is unclear yet why this doesn't function.	b	3
20041203/rws-avv	See extended description.	c	3
20041022/aviassist	In the viewer the splash-screen is skipped to prevent browser check issues.	d, e	4
20041004/bedrijfsstijl	The setup of this websites makes extensive use of URL parameter. This results in a relative - compared to the original- large snapshot (in bytes).		5
20041207/belevingswaardenonderzoek	Though the site has it own domain name some files originate from the minvenw.nl domain.		5
20041112/betuweroute			5
20041112/bobjijofbobik			5
20041112/bouwdienst			5
20041004/brieven			5
20041022/cawsw			5
20041112/daarkunjemeethuiskomen	This site makes heavy use of Macromedia Flash.	f	1
20040916/nieuws	Overlaps with 'Brieven aan de Tweede Kamer' ('brieven') which has not been included in this snapshot.		5
20041124/rdij			5

20041123/noordzee	One part of the sites gives a file not found error. However, the file has been included in the snapshot but it cannot be accessed in the viewer because is has incorrect capitalization. The originating IIS-server is tolerant to these errors, the Presurf viewer currently is not.	b	4
20041015/dgg			5
20041015/dodehoek			5
20041112/droogtestudie			5
20041015/dvo-international			5
20040916/aanbestedingen			5
20041112/ei-van-columbus			5
20041112/fileplanregiorotterdam	Various pages and images produced an 'access-denied error'. It is unclear what caused this problem. The assumption that is was related user-agent string dependencies could not be confirmed.	b	2
20041112/fluistertrein			3
20041112/gordeldier	See extended description.	f	1
20041112/haringvlietsluizen	See extended description.		5
20041015/havenraad	The javascript based animation on the homepage does not function. Javascript issues prevent some menu's from working properly.	d	2
20041015/hetnieuwerijden	Appears to be an alternative entrance tot hetnieuwerijden.nl		5
20041112/hetnieuwerijden.nl			5
20041112/hslzuid.info			5
20041112/hslzuid.nl	Javascript-issues.	d	2
20041015/ivw			5
20041005/inspraakpunt			5
20041112/interwad	Snapshot action stopped manually because of 'loops'.	c	2
20041119/ivs90			5
20041119/kaderrichtlijnwater			5
20041119/knmi	Has "File not found"-errors, likely caused by javascripts.	d	3
20040916/locov			5
20041123/luchtvaartbeleid		d, e	1
20041112/maaswerken	A restricted domain has been captured due to problems with URL-parameters. Has "File not found"-errors, likely caused by javascripts.	c	1
20041119/keringhuis	A game implemented in javascript does not function.	d	4
20041123/mainport-pmr	Client side techniques prevent using this snapshot.	d,e,f	1
20041123/meetkundige dienst			5
20041015/mitprojectenboek			5
20041123/mobilion			5
20041123/nederlandleeftmetwater	See extended description.		5
20041124/projecten.nlmw			5
20041022/noodoverloop			5
20041015/overlegvenw			5
20040915/oei	Javascript-issues with the menu's.	d	2
20040916/rsst			5
20041124/psd-online	An image is missing on the homepage, analoge to the original.		5
20041123/projectijzerenrijn	See extended description.		5

20041123/projectrobel	Javascript-based menu does not function.	d	2
20041124/projectvera	See extended description.		5
20041123/projectvnk			5
20041015/raadvenw	Has "File not found"-errors, likely caused by javascripts.	d	2
20041015/radionavigatie	Javascript-loaded images disappear after a 'mouse over' event has been triggered.	d	4
20041212/randwegeindhoven	See extended description.		5
20041124/rdw			5
20041124/rijbewijs			5
20041124/rikz	See extended description. This overall score was given after changing two javascripts in the snapshot.		5
20041124/rijksweg1			5
20041015/ritsen			5
20041126/riza	A javascript tries to load a page that has the current month encoded in its filename. The English part has not been archived because the English pages have the same URLs as the Dutch pages.	d, g	3
20041015/rondjerandstad			5
20041126/ruimtevoorderivier	Some smaller Flash-related issues.	f	3
20041126/sloelijn	Note: the javascript based menu still works.		5
20041015/sociaalfonds			5
20041015/ssz			5
20041015/svov			5
20041015/taxiklacht			5
20041022/taximarkt	Stylesheet information appears to be missing. Likely caused by javascript isues.	d	3
20041015/taxiwet			5
20040914/telewerken	Extensive use of javascripts.	d	2
20041022/toegangov			5
20041203/transactie-modalshift			5
20041015/tunnelveiligheid			5
20041203/twentekanalen			5
20041123/vananaarbeter	Extensive use of Flash prevented creating a proper snapshot.	f	2
20041015/verbondennetwerken			5
20041203/verkeerscentrale			5
20041015/wasco			5
20041126/waterland			5
20041124/wegennaardetoekomst	Menu's don't work properly due to javascript related issues.		2
20041022/wegwijzer			5
20041022/wetpersonenvervoer	See extended description.	d	4
20041203/zeeweringen			5
20041203/zuid-willemsvaart			5
20041022/zuiderzeelijn		g	2

***: Legend for the 'Issues'-field:**

a Issues related to originating server
b Issues related to crawler or viewer
c crawler loops due to use of URL parameters
d Javascript related issues
e Issues related to browser checks
f Issues related to the use of Macromedia Flash
g Issues related to session management

See chapter xxx for a discussion of the various issues.

****: Legend for the 'Score'-field:**

5 No issues
4 Snapshot has minor issues.
3 Snapshot has major issues.
2 Snapshot has serious issues.
1 Snapshot action failed / invalid snapshot.

See chapter 4.5 of the report for a discussion of the applied method.

Distribution of scores from overall assessment

The following table presents the distribution of scores as compiled from the overall overview of issues and scores.

Table 2: Distribution of scores from overall assessment.

Score	#	%
1	5	5.3
2	11	11.6
3	8	8.5
4	5	5.3
5	66	69.5

The average score for the snapshots is 4.2 (n=95, σ = 1.3)

Frequency of issues

The following table presents the frequency of the issues from the overall overview of issues and scores. See chapter 5.2. in the report for a discussion on the interpretation of this table.

Table 3: Distribution of scores from overall assessment.

Category	Frequency	%
Issues related to originating server (a)	1	1,1
Issues related to crawler or viewer (b)	3	3,2
Crawler loops due to URL parameters (c)	3	3,2
Javascript related issues (d)	14	14,7
Issues related to browser checks (e)	3	3,2
Macromedia Flash related issues (f)	5	5,3
Issues related to session management (g)	2	2,1

2. Questionnaire for assessment of selected snapshots

Vragenlijst beoordeling archivering websites februari 2005

Algemeen
naam beoordelaar:
beoordeelde website:

Ik verzoek u uw antwoorden te geven op een 5-puntsschaal, waarbij
de 1 staat voor: helemaal oneens
de 5 staat voor: helemaal eens
Desgewenst kunt u uw antwoord nader toelichten.

Vormgeving

De vormgeving en lay-out van het webarchief zien er hetzelfde uit als de originele website:

1 2 3 4 5

O---------O----------0----------O---------X
oneens eens

Toelichting:

Functionaliteit

Afgezien van functionaliteit die gebaseerd is op formulieren (zoals een zoekmachine, databasefunctionaliteit of een reactieformulier) werkt de archiefversie van de site zoals de originele website:

1 2 3 4 5

O---------O----------0----------O---------x
oneens eens

Toelichting:

Volledigheid

Afgezien van eventuele informatie 'achter een formulier' is de gearchiveerde website een volledige kopie:

1 2 3 4 5

O---------O----------0----------O---------x
oneens eens

Toelichting:

Informatie 'achter een formulier' (zoals een zoekmachine, databasefunctionaliteit of een reactieformulier) vormt een essentieel onderdeel van de originele website:

1 2 3 4 5

O---------O----------0----------O---------x
oneens eens

Toelichting:

Bedankt voor uw bijdrage!

3. Extended assessment of selected snapshots

This appendix presents the results from the extended assessment selected snapshots.

Snapshot capsis.arc/100030/20041112/haringvlietsluizen

Start URL:	http://www.haringvlietsluizen.nl/
Title:	Ander Beheer Haringvlietsluizen
Snapshot date:	2004-11-12
Mirror options:	+ *.css, *.gif, *.jpg, *.png, *.js, *.doc, *.pdf ignore instructions for robots
Errors:	13 'file not found (404)'-errors
Number of files:	1186
Size:	116M

MIME types:

Files	*MIME type*
1	video/unknown
2	application/zip
2	text/xml
6	application/msword
7	text/css
47	application/pdf
128	image/jpeg
321	text/html
659	image/gif

HTML formats (doctype): Not specified.

HTML errors:

Errors	*Files*
no data	87
0	0
1-10	40
11-100	134
101-1000	52
>1000	9

Assessment

Agent: Marjolein Burger, Rijkswaterstaat Zuid-Holland.
Date: January 2005

	Score:	Toelichting:
Vormgeving:	5	De vormgeving en lay-out zijn hetzelfde als de originele site qua opbouw, kolombreedte, lettertype, kleurstelling etc.
Functionaliteit:	4	In de boekenkast (archief) werken een tweetal onderliggende documenten niet (startdocument/beslissing op bezwaarschriften).
Volledigheid	5	-
Gemis deepweb:	5	De site heeft met name een informerend karakter met de mogelijkheid tot het downloaden van rapporten, nieuwsbrieven etc.
Opmerkingen	-	

Assessment

Agent: René Voorburg, Capsis BV
Date: January 2005

The problem with the files in 'boekenkast (archief)' is caused by Presurf handling files with encoded characters (like '%20' for a space') in file names incorrectly. A small change in the viewer will solve the problem. The files are properly included in the snapshot.

Snapshot capsis.arc/100030/20041123/projectijzerenrijn

Start URL:	http://www.projectijzerenrijn.nl/
Title:	Project IJzeren Rijn / IJzeren Rijn
Snapshot date:	2004-11-23
Mirror options:	+ *.css, *.gif, *.jpg, *.png, *.js, *.doc, *.pdf ignore instructions for robots
Errors:	3 'file not found (404)'-errors.
Number of files:	84
Size:	70M

MIME types:

Files	*MIME type*
1	application/x-javascript
1	image/jpeg
1	text/css
4	application/msword
17	application/pdf
22	text/html
34	image/gif

HTML formats (doctype): Not specified

HTML errors:

Errors	*Files*
no data	15
0	0
1-10	0
11-100	6
>100	1

Assessment

Agent: Esther van Engelen, Project IJzerenrijn / ProRail
Date: January 2005

	Score:	Toelichting:
Vormgeving:	4	Logo ProRail komt niet door
Functionaliteit:	5	-
Volledigheid:	5	-
Gemis deepweb:	3	-
Opmerkingen	-	

Assessment

Agent: René Voorburg, Capsis BV
Date: January 2005

The ProRail logo is not displayed due to incorrect handling of files with encoded characters in the viewer of Presurf. A small change in the viewer will solve the problem. The files are properly included in the snapshot.

Snapshot capsis.arc/100030/20041124/projectvera

Start URL:	http://www.projectvera.nl/
Title:	Project Verbinding Roosendaal - Antwerpen (VERA).
Snapshot date:	2004-11-24
Mirror options:	+ *.css, *.gif, *.jpg, *.png, *.js, *.doc, *.pdf ignore instructions for robots
Errors:	1 'file not found (404)' error
Number of files:	107
Size:	81M

MIME types:

Files	*MIME type*
1	application/msword
4	text/css
7	image/jpeg
26	image/gif
32	application/pdf
36	text/html

HTML formats (doctype):

Files	*Format*
17	Unspecified / not tested.
19	-/W3C//DTD HTML 4.0 Transitional//EN

HTML errors:

Errors	*Files*
no data	15
0	14
1-10	5
>10	2

Assessment

Agent: Nienke van Geest, ProRail
Date January2005

	Score:	Toelichting:
Vormgeving:	5	
Functionaliteit:	3	Niet alle links binnen de site werken goed
Volledigheid	5	-
Gemis deepweb:	2	Van deze site vormt deze info geen essentieel onderdeel
Opmerkingen	-	

Assessment

Agent: René Voorburg, Capsis BV
Date January2005

Some links don't work due to incorrect handling of files with encoded characters in the viewer of Presurf. A small change in the viewer will solve the problem. The files are properly included in the snapshot.
Some files are displayed as 'garbage' because Presurf doesn't add the proper MIME types when serving the request. This bug will be fixed.

Snapshot capsis.arc/100030/20041212/a2denboscheindhoven

Start URL:	http:// www.minvenw.nl/rws/dnb/projecten/a2dbehv/
Title:	A2 's-Hertogenbosch en Eindhoven
Snapshot date:	2004-12-21
Mirror options:	+ *.css, *.gif, *.jpg, *.png, *.js, *.doc, *.pdf ignore instructions for robots
Errors:	5 'file not found (404)'-errors
Number of files:	332
Size:	43M

MIME types:

Files	*MIME type*
1	text/css
2	application/x-shockwave-flash
2	video/quicktime
3	application/x-javascript
25	application/pdf
43	image/gif
80	image/jpeg
171	text/html

HTML formats (doctype):

Files	*Format*
171	-//W3C//DTD HTML 4.01 Transitional//EN

HTML errors:

Errors	*Files*
no data	0
1-10	30
>10	141

Assessment

Agent: Melanie Persoons, Webbeheer intra/internet Rijkswaterstaat Noord-Brabant
Date: January 2005

	Score:	Toelichting:
Vormgeving:	5	-
Functionaliteit:	5	-
Volledigheid	-	-
Opmerkingen	-	

Snapshot capsis.arc/100030/20041112/a2denbosch

Start URL:	http://www.a2denbosch.nl/
Title:	A2 Rondweg Den Bosch
Snapshot date:	2004-11-12
Mirror options:	+ *.css, *.gif, *.jpg, *.png, *.js, *.doc, *.pdf ignore instructions for robots
Errors:	-
Number of files:	369
Size:	73M

MIME types:

Files	*MIME type*
1	application/zip
1	text/css
2	application/x-javascript
2	video/mpeg
20	application/msword
33	application/pdf
43	image/jpeg
117	text/html
150	image/gif

HTML formats (doctype):

Files	*Format*
117	-//W3C//DTD HTML 4.01 Transitional//EN

HTML errors:

Errors	*Files*
0	0
1-10	0
11-100	116
>100	1

Assessment

Agent:	Melanie Persoons, Webbeheer intra/internet Rijkswaterstaat Noord-Brabant
Date:	January 2005

	Score:	Toelichting:
Vormgeving:	5	-
Functionaliteit:	5	-
Volledigheid	-	-
Opmerkingen	-	

Snapshot capsis.arc/100030/20041112/aanlega50

Start URL:	http://www.aanlega50.nl/
Title:	Aanleg A50 / A50 Homepage
Snapshot date:	2004-11-12
Mirror options:	+ *.css, *.gif, *.jpg, *.png, *.js, *.doc, *.pdf ignore instructions for robots
Errors:	6 “file not found” (404) –errors
Number of files:	405
Size:	46M

MIME types:

Files	*MIME type:*
1	application/x-shockwave-flash
2	text/css
4	video/mpeg
4	video/x-ms-asf
11	application/msword
17	application/pdf
31	image/gif
159	image/jpeg
170	text/html

HTML formats (doctype): Unspecified

HTML errors:

Errors	*Files*
0	0
1-10	155
11-100	14
>100	1

Assessment

Agent:	Melanie Persoons, Webbeheer intra/internet Rijkswaterstaat Noord-Brabant
Date:	January 2005

	Score:	Toelichting:
Vormgeving:	5	-
Functionaliteit:	4	Ik mis de foto op de homepage
Volledigheid	-	-
Opmerkingen	-	

Assessment

Agent:	René Voorburg, Capsis BV
Date:	January 2005

This site uses user-agent string based redirects (client-side, in javascript). These redirects prevent Presurf to follow all links to files that should be included in the snapshot. When the snapshot is accessed with a browser that has a different user-agent string a redirect follows to a file that is not included in the snapshot.
It is possible to circumvent these errors by manually adding these files to the snapshot instruction.

Snapshot capsis.arc/100030/20050125/aanlega50

Start URL: http://www.aanlega50.nl/
Title: Aanleg A50 / A50 Homepage
Snapshot date: 2005-01-25
Mirror options: http://www.aanlega50.nl/high.htm 'http://www.aanlega50.nl/images/\
de%20nieuwe%20tunnel%20bij%20Nijnsel.jpg'
http://www.aanlega50.nl/images/totaal.jpg \
http://www.aanlega50.nl/images/menuframe.jpg \
+ *.css, *.gif, *.jpg, *.png, *.js, *.doc, *.pdf
ignore instructions for robots
Errors: 6 “file not found” (404) errors
Number of files: 419
Size: 48M
MIME types:

Files	*MIME type*
1	application/x-shockwave-flash
2	text/css
4	text/plain
4	video/mpeg
11	application/msword
19	application/pdf
34	image/gif
162	image/jpeg
176	text/html

HTML formats (doctype): Unspecified
HTML errors:

Errors	*Files*
0	0
1-10	158
11-100	17
>100	1

Assessment

Agent: René Voorburg, Capsis BV
Date: January 2005

The files were missing in the previous snapshot (capsis.arc/100030/20041112/aanlega50) have now been included manually. This solves the problem caused by redirects.
An error exists in the current Presurf viewer that it does not properly handle files with character encoding (for example '%20' for a space). Thus bug prevents an image on the homepage being displayed even though the file has been added to the snapshot.

Snapshot capsis.arc/100030/20041124/randwegeindhoven

Start URL:	http://www.randwegeindhoven.nl/
Title:	Randweg Eindhoven / Uitbreiding Randweg Eindhoven
Snapshot date:	2004-11-24
Mirror options:	http://www.minvenw.nl/rws/dnb/projecten/randwegeindhoven/ + *.css, *.gif, *.jpg, *.png, *.js, *.doc, *.pdf ignore instructions for robots
Errors:	3 “file not found” (404) errors
Number of l files:	1107
Size:	146M

MIME types:

Files	*MIME type*
1	application/x-shockwave-flash
1	text/css
2	application/x-javascript
3	video/mpeg
8	application/msword
27	application/pdf
41	image/gif
402	image/jpeg
618	text/html

HTML formats (doctype):

Files	*Format*
1	Unspecified
617	-//W3C//DTD HTML 4.01 Transitional//EN

HTML errors:

Errors	*Files*
0	47
1-10	7
11-100	563
>100	1

Assessment

Agent:	Melanie Persoons, Webbeheer intra/internet Rijkswaterstaat Noord-Brabant
Date:	January 2005

	Score:	Toelichting:
Vormgeving:	5	-
Functionaliteit:	5	-
Volledigheid	-	-
Opmerkingen	-	

Snapshot capsis.arc/100030/20041203/zuid-willemsvaart

Start URL:	http://www.zuid-willemsvaart.nl/
Title:	Zuid-Willemsvaart
Mirror options:	+ *.css, *.gif, *.jpg, *.png, *.js, *.doc, *.pdf ignore instructions for robots
Errors:	-
Number of files:	234
Size:	56M

MIME types:

Files	*MIME type*
1	text/css
2	application/x-javascript
2	application/zip
20	application/msword
26	image/gif
29	application/pdf
31	image/jpeg
122	text/html

HTML formats (doctype): All -//W3C//DTD HTML 4.01 Transitional//EN

HTML errors:

Errors	*Files*
0	0
1-10	0
11-100	121
>100	1

Assessment:

Agent:	Melanie Persoons, Webbeheer intra/internet Rijkswaterstaat Noord-Brabant
Date:	January 2005

	Score:	Toelichting:
Vormgeving:	5	-
Functionaliteit:	5	-
Volledigheid	-	-
Opmerkingen	-	

Assessment

Agent:	René Voorburg, Capsis BV
Date:	January 2005

This snapshot appears to be a perfect mirror of the original.

Snapshot capsis.arc/100030/20041112/bouwdienst

Start URL:	http://www.bouwdienst.nl/
Title:	Bouwdienst Rijkswaterstaat
Snapshot date:	2004-11-12
Mirror options:	+minvenw.nl/rws/bwd/home/www/* + *.css, *.gif, *.jpg, *.png, *.js, *.doc, *.pdf ignore instructions for robots
Errors:	-
Number of files:	1808
Size:	205M

MIME types:

Files	*MIME type*
3	application/vnd.rn-realmedia
3	video/avi
4	application/x-javascript
8	application/msword
8	text/css
69	image/jpeg
142	image/gif
220	application/pdf
359	image/pjpeg
981	text/html

HTML formats (doctype): Not specified.

HTML errors:

Errors	*Files*
no data	7
0	0
1-10	180
11-100	542
>100	252

Assessment

Agent:	Iris Wieland. Bouwdienst 030-285 79 34, i.m.wieland@bwd.rws.minvenw.nl
Date:	December 2004

	Score:	Toelichting:
Vormgeving:	5	-
Functionaliteit:	5	-
Volledigheid:	5	-
Gemis deepweb:	5	-
Opmerkingen	-	

Assessment

Agent:	René Voorburg, Capsis BV
Date:	January 2005

This snapshot appears to be a perfect mirror of the original.

Snapshot capsis.arc/100030/20041203/rws-avv

Start URL	http://www.rws-avv.nl/
Title:	Adviesdienst Verkeer en Vervoer
Snapshot date:	2004-12-03
Mirror options:	http://www.rws-avv.nl/pls/portal30/portal30.home + *.css, *.gif, *.jpg, *.png, *.js, *.doc, *.pdf ignore instructions for robots
Errors:	Manually aborted for looping URL parameters 106 “file not found” (404) errors
Number of files:	1394
Size:	97M

MIME types:

Files	*MIME type*
1	application/x-shockwave-flash
1	application/x-zip-compressed
1	image/bmp
2	application/zip
4	application/octet-stream
4	application/vnd.ms-excel
8	application/x-javascript
19	text/css
24	application/msword
25	image/jpeg
107	application/pdf
177	image/pjpeg
578	image/gif
1273	text/html

HTML formats (doctype): Not specified.

HTML errors:

Errors	*Files*
-	1
0	0
1-10	9
11-100	121
>100	1142

Assessment

Agent:	Brandsma, mw. drs. J., Communicatieadviseur Adviesdienst Verkeer en Vervoer en Wegen naar de Toekomst
Date:	December 2004

	Score:	Toelichting:
Vormgeving:	-	-
Functionaliteit:	-	-
Volledigheid	-	-
Gemis deepweb	-	-
Opmerkingen	Het archiveren van onze site is niet goed gegaan. IK kom niet op de pagina's zoals ik die normaliter krijg op onze webite. ik krijg allerlei foutmeldingen. Ik adviseer nog een keer de site te archiveren.	

Assessment

Agent:	René Voorburg, Capsis BV
Datum:	January 2005

This sites uses javascrip for its main navigation elements. The functioning of these scripts depend on the paths of the pages that is being linked to. These paths change when the snapshot is included in Presurf, causing the scripts to stop functioning properly.

Snapshot capsis.arc/100030/20041022/wetpersonenvervoer

Start URL:	http://www.minvenw.nl/dgp/wetpersonenvervoer/
Title:	Wet Personenvervoer
Snapshot date:	2004-10-22
Mirror options:	+ *.css, *.gif, *.jpg, *.png, *.js, *.doc, *.pdf ignore instructions for robots
Errors:	3 “file not found” (404) errors
Number of files:	232
Size:	4.3M

Mediaformaten:

Files	*MIME type*
1	text/css
2	application/msword
4	application/x-javascript
12	image/jpeg
44	image/gif
166	text/html

HTML formats (doctype): Not specified.

HTML errors:

Errors	*Files*
no data	96
0	0
1-10	18
11-100	69

Assessment

Agent: Jeroen Mol
Date: December 2004

	Score:	Toelichting:
Vormgeving:	5	“Splash-pagina” begint alleen wel met een javascript-error (rollovers). Mouse-out werkt nl. niet (komt door error) Overige pagina’s hebben ook last van javascript-error, maar heeft geen (zichtbare) nadelige invloed op de werking.
Functionaliteit:	5	-
Volledigheid:	5	Maarrr: de bijlagen (worddocumenten op pagina Aanbesteding regionaal openbaar vervoer) zijn verminkt.
Gemis deepweb:	5	Maar voor een archiefexemplaar niet interessant
Opmerkingen	-	

Assessment

Agent: René Voorburg, Capsis BV
Date: Januari 2005

MS-Word documents are not displayed properly because Presurf currently sends them with a header that has the wrong MIME-type (text/html iso. application/msword).

The splash pages uses a javascript for adding a visual 'mouse-over' effects to the buttons. This script depends on the path of the page, so it stops functioning when the path is changed, as happens when it is served form Presurf.

Snapshot capsis.arc/100030/20041123/nederlandleeftmetwater

Start URL:	http://www.nederlandleeftmetwater.nl/
Title:	Nederland leeft met water
Snapshot date:	2004-11-23
Mirror options:	+ *.css, *.gif, *.jpg, *.png, *.js, *.doc, *.pdf ingnore instructions for robots
Errors:	-
Number of files:	198
Size:	33M

MIME types:

Files	*MIME type*
1	application/msword
1	application/octet-stream
1	text/css
4	application/postscript
4	video/mpeg
6	audio/mpeg
21	application/pdf
40	image/jpeg
59	text/html
61	image/gif

HTML formats (doctype): Not specified

HTML errors:

Errors	*Files*
no data	22
0	0
1-10	0
11-100	37

Assessment

Agent:	Walter Snoei
Date:	December 2004

	Score:	Toelichting:
Vormgeving:	5	-
Functionaliteit:	5	-
Volledigheid	5	-
Gemis deepweb	5	De site wordt binnenkort uitgebreid met een functionaliteit die mensen in staat stelt projecten in de directe woonomgeving te zoeken.
Opmerkingen	-	

Assessment

Agent:	René Voorburg, Capsis BV
Date	Januari 2005

This snapshot appears to be a perfect mirror of the original.

Snapshot capsis.arc/100030/20041015/ivw

Start URL:	http://www.ivw.nl/
Title:	Inspectie Verkeer en Waterstaat
Snapshot date:	2004-10-15
Mirror options:	http://www.minvenw.nl/dgg/si/ + *.css, *.gif, *.jpg, *.png, *.js, *.doc, *.pdf ignore instructions for robots
Errors:	41 “file not found” (404) errors
Number of files:	2182
Size:	204M

MIME types:

Files	*MIME type*
1	text/css
2	application/octet-stream
2	application/x-javascript
3	audio/x-wav
5	text/plain
23	application/zip
54	application/msword
135	image/jpeg
417	image/gif
743	application/pdf
756	text/html

HTML formats (doctype):

Files	Format
255	-//W3C//DTD HTML 4.01 Transitional//EN
501	not specified

HTML errors:

Errors	*Files*
no data	14
0	0
1-10	301
11-100	427
>100	14

Assessment

Agent	Josje Majoor, IVW
Date:	December 2004

	Score:	Toelichting:
Vormgeving:	4	-
Functionaliteit:	3	bestanden (pdf’s) zijn niet meer te downloaden
Volledigheid	3	pdf’s ontbreken
Gemis deepweb	4	deze ivw site had echter geen zoekfunctionaliteit. De nieuwe ivw site (vanaf 22/11/04) wel. Downloadables en formulieren zijn en blijven essentieel voor de dienstverlening van IVW.
Opmerkingen	-	

Assessment

Agent:	René Voorburg, Capsis BV
Date:	January 2005

PDF documents are not displayed properly because Presurf currently sends them with a header that has the wrong MIME-type (text/html iso. application/pdf).

Snapshot capsis.arc/100030/20041124/rikz

Start URL:	http://www.rikz.nl
Title:	Rijksinstituut voor Kust en Zee
Snapshot date:	2004-11-24
Mirror options:	+ *.css, *.gif, *.jpg, *.png, *.js, *.doc, *.pdf ignore instructions for robots
Errors:	205 “file not found” (404) errors, (most of them have a javascript as referrer).
Number of files:	812
Size :	492M

MIME types:

Files	*MIME type*
2	text/plain
4	text/css
7	application/vnd.ms-powerpoint
13	application/x-javascript
56	image/gif
82	image/jpeg
199	application/pdf
246	text/html

HTML formats (doctype):

Files	*Format*
76	not specified
170	-//W3C//DTD HTML 4.0 Transitional//EN

HTML errors:

Errors	*Files*
no data	74
0	0
1-10	3
11-100	168
>100	1

Assessment

Agent:	Groeneveld, G.J.J
Date:	December 2004

	Score:	Toelichting:
Vormgeving:	-	-
Functionaliteit:	-	-
Volledigheid	-	-
Gemis deepweb	-	-
Opmerkingen	Ik heb even gekeken en ondanks dat de website 491Mb groot is werkt niets, dus.....	

Assessment

Agent:	René Voorburg, Capsis BV
Date:	January 2005

This snapshot has problems similar to capsis.arc/100030/20041203/rws-avv.
This sites uses javascript for its main navigation elements. The functioning of these scripts depend on the paths of the pages that is being linked to. These paths change when the snapshot is included in Presurf, causing the scripts to stop functioning properly.

To be able to use this snapshot two scripts in the snapshot have been altered manually. These scripts are:
`www.rikz.nl/home/NL/scripts/init_themas.js'
`www.rikz.nl/home/NL/scripts/init_centraal.js'

The original scripts have been copied to:
`www.rikz.nl/home/NL/scripts/init_themas.js_origineel'
`www.rikz.nl/home/NL/scripts/init_centraal.js_origineel'.

When the relative location of the snapshot changes, these scripts have to be adjusted again.

After adjusting these scripts the snapshot appears to be fine, except for Presurf sending the wrong MIME-type for PDF-files.

Assessment

Agent:	Groeneveld, G.J.J
Date:	February 2005

	Score:	Toelichting:
Vormgeving:	1	Oppervlakkig wel maar bv de pagina's onder projectsites (linker navigatiemenu niet) Aantal vb: http://presurf.capsis.nl/viewer/capsis.arc/ 100030/20041124/ rikz/www.rikz.nl/thema/kust_en_veiligheid/ Beheer/dynbeheer.html http://presurf.capsis.nl/viewer/capsis.arc/100030/20041124/rikz/ www.rikz.nl/home/NL/Organisatie/ projectsites_meten.html http://presurf.capsis.nl/viewer/capsis.arc/100030/20041124/rikz/ www.rikz.nl/home/NL/Organisatie/zee.html http://presurf.capsis.nl/viewer/capsis.arc/100030/20041124/rikz/ www.rikz.nl/home/NL/Organisatie/projectsites_eng.html
Functionaliteit:	4	Werkt goed.
Volledigheid	4	-
Gemis deepweb	2	Ik denk dat het minder relevant is voor een archief. Hier bekijk je de gepubliceerde informatie en ben je niet geïnteresseerd in interactie. Die zou wel beschreven moeten zijn maar hoeft niet uitvoerbaar te zijn.
Opmerkingen	-	

Assessment

Agent:	René Voorburg, Capsis BV
Date:	March 2005

The URLs mentioned present a '404' (file not found) error message. Examining the log files generated when creating the snapshot, it is clear that the application has tried to capture these files using a non-existing URL (for example www.rikz.nl/home/home/NL/Organisatie/projectsites_meten.html and www.rikz.nl/home/home/NL/Organisatie/projectsites_meten.html
www.rikz.nl/home/NL/home/NL/Organisatie/projectsites_meten.html
Apparently, the application was unable to parse the correct URL from the javascript-menu that contains the links.

Snapshot capsis.arc/100030/20041119/kaderrichtlijnwater

Start URL:	http://www.kaderrichtlijnwater.nl
Title:	Europese Kaderrichtlijn Water
Snapshot date:	2004-11-129
Mirror options:	+ *.css, *.gif, *.jpg, *.png, *.js, *.doc, *.pdf ignore instructions for robots
Errors:	17 'file not found (404)'-errors
Number of files:	540
Size:	344M

MIME types:

Files	*MIME type*
1	image/png
2	text/css
3	application/x-javascript
3	application/x-zip-compressed
30	application/msword
55	image/jpeg
69	image/gif
177	application/pdf
182	text/html

HTML formats (doctype): Not specified.

HTML errors:

Errors	*Files*
no data	106
0	0
1-10	34
11-100	394
>100	6

Assessment

Agent:	s.l. ras, venw, dgw t. 8206, saskia.ras@minvenw.nl
Date:	January 2005

	Score:	Toelichting:
Vormgeving:	5	-
Functionaliteit:	5	-
Volledigheid	5	-
Gemis deepweb:	5	-
Opmerkingen	-	

Snapshot capsis.arc/100030/20040916/nieuws

Start URL:	http://www.minvenw.nl/cend/dco/nieuws/
Title:	Venw Nieuws
Snapshot date:	2004-09-16
Mirror options:	+ *.css, *.gif, *.jpg, *.png, *.js, *.doc, *.pdf ignore instructions for robots
Errors:	24 'file not found (404)'-errors
Number of files:	2694
Size:	161M

MIME types:

Files	*MIME type*
1	text/plain
1	text/xml
2	application/x-javascript
2	image/bmp
4	application/vnd.ms-excel
4	text/css
7	application/vnd.ms-powerpoint
12	image/jpeg
23	image/gif
34	image/pjpeg
74	application/pdf
164	application/msword
2340	text/html

HTML formats (doctype): 12 -//W3C//DTD Compact HTML 1.0 Draft//EN
others: Not specified.

HTML errors:

Errors	*Files*
no data	2306
0	0
1-10	9
>10	25

Assessment

Agent:	Wilco van Soest	
Date:	February 2005	
	Score:	Toelichting:
Vormgeving:	5	Zie geen noemenswaardige verschillen tussen gearchiveerde en originele site, behalve daar waar links niet werken.
Functionaliteit:	5	Zie hierboven.
Volledigheid	5	Als onder 'informatie achter een formulier' wordt bedoeld Word- en PDF-documenten, dan klopt bovenstaande stelling. Bij het openen van dergelijke documenten op de gearchiveerde website kom je niet op documenten in de originele opmaak, maar in een lange brij tekst (zonder alinea-indeling) met rommel (codes et cetera).
Gemis deepweb:	5	Zie bovenstaande reactie. Zeker als het om een reactieformulier of Word- of PDF-document gaat dan klopt bovenstaande stelling.
Opmerkingen	Ik heb ook nog even naar de overige DCO-sites gekeken. Bij de campagnesite 'Daar kun je mee thuiskomen' viel me op dat alleen de homepage is gearchiveerd en dat de links naar subsites een lege pagina opleveren. Bij de Bedrijfsstijlsite valt me op dat die op verschillende manieren is gearchiveerd (6, 5, en 3 niveaus diep). Tot slot vraag ik me af wie in de toekomst de websites gaat archiveren (de snapshots maakt) en op welke momenten dat gebeurt.	

Snapshot capsis.arc/100030/20041112/gordeldier

Start URL:	http://www.gordeldier.nl/
Title:	Goochem het Gordeldier.
Snapshot date:	2004-11-24
Mirror options:	+ *.css, *.gif, *.jpg, *.png, *.js, *.doc, *.pdf ignore instructions for robots
Errors:	4 'file not found (404)'-errors
Number of files:	7
Size:	216K

MIME types:

Files	*MIME type*
1	application/x-shockwave-flash
1	image/gif
2	text/html

HTML formats (doctype): 199 -//W3C//DTD XHTML 1.0 Strict//EN
others unspecified

HTML errors:

Errors	*Files*
0	0
1-10	0
>10	1

Assessment

Agent: Zomeren, J. van (Jan) - CEND-DCO
Date: January 2005

	Score:	Toelichting:
Vormgeving:	-	-
Functionaliteit:	-	-
Volledigheid	-	-
Gemis deepweb:	-	-

Opmerkingen

IIk heb al meerdere keren gekeken, maar ik krijg niks. Bij het zoeken naar handsfree beelen krijg ik wel zoekresultaten, maar als ik die aanklik krijg een blanco scherm.
Ik heb wel voor de vakantie een pagina gezien maar daar kon ik het filmpje en audiofragmenten niet aanklikken.

Wat gaat er fout?

Assessment

Agent: René Voorburg, Capsis BV
Date: March 2005

The first page of " http://www.gordeldier.nl/" contains a flash-animation with a link that cannot be read or translated by the crawler.
However, by adding the second page to the list of URLs to start capturing a more complete and more functional snapshot could be created.
This is snapshot capsis.arc/100030/20050309/gordeldier .

5. Analyses of a reference website

Analyses of HTML-errors in a snapshot of http://webrichlijnen.overheid.nl as reported by the W3C-based HTML-validator.

Start URL:	http://webrichtlijnen.overheid.nl/
Title:	Richtlijnen voor de toegankelijkheid en duurzaamheid van overheidswebsites
Snapshot date:	2004-11-01
Mirror options:	
Errors:	-
Number of files:	223
Size:	3.6 M

MIME types:

Files	*MIME type*
1	application/x-javascript
1	text/plain
3	text/css
13	image/gif
206	text/html

HTML formats (doctype): 1 -//W3C//DTD HTML 4.0 Transitional//EN

HTML errors:

Errors	*Files*
no data	7
0	185
1	4
2	4
2	4
42	1
47	1

Bijlage D. Voorstel metadata voor snapshots van websites 29-04-2005

Identificatie

	verplicht	*herhaalbaar* [41]	*schema* [42]	*opmerkingen*
naam website	ja	ja	vrije tekst	
alternatieve naam	nee	ja	vrije tekst	
afgekorte naam	nee	ja	vrije tekst	
unieke identificatie	ja	ja		door systeem gegenereerd
url	ja	nee	URL	conform domeinnamenbeleid VW
alias	nee	ja	URL	
redirect	nee	ja	URL	
aanvullende URL	nee	ja	URL of patroon	indien deel site elders ondergebracht

Beschrijving

	verplicht	*herhaalbaar*	*schema*	*opmerkingen*
beschrijving	nee	nee	vrij tekst	bij voorkeur in vastgesteld format
doel/ functie site	ja	ja	taxonomie**, Lokale handelingenbank	
activiteit	ja	ja	taxonomie**	Proces waarvoor de website wordt gebruikt
aggegatieniveau	ja	ja		
trefwoord	nee	ja	taxonomie**	
periode	nee	nee		periode waarover website gaat
datum snapshot	ja	ja		door systeem gegenereerd
taal	nee	ja		standaard Nederlands

Beheersgegevens

	verplicht	*herhaalbaar*	*schema*	*opmerkingen*
omvang snapshot	ja	ja	in aantal bytes	door systeem gegenereerd
generiek formaat	ja	nee		standaard "compound" bij samengestelde bronnen
data format	ja	ja		b.v. ASCII, JPEG, XML, HTML, PDF etc. versie-informatie is hierbij wenselijk tbv. beheershandelingen
locatie	ja	ja		
beheershandeling				zie onderstaande verfijningen
.capture	ja	nee		gegevens over snapshotopdracht en gegenereerde logs
.kwaliteitsoordeel		ja		
.beschrijving		ja		
.migratie		ja		
.conversie		ja		
.verversing		ja		
.verwijdering		ja		
vernietigings-	ja	ja	BSD's	

[41] Dit element kan meerdere malen worden gebruikt als metagegeven van een website.
[42] De standaard of methode die wordt gebruikt om een metagegeven te coderen.

grondslag				
bewaartermijn	ja	ja		standaard "bewaren", BSD's onvoldoende ingesteld op websites
beschrijving beheershandeling	ja	nee		om accountability en audit op beheer mogelijk te maken
datum beheershandeling	ja	ja		door systeem gegenereerd

Organisatiegegevens

	verplicht	*herhaalbaar*	*schema*	*opmerkingen*
rol	ja	ja	*mogelijke waarden:* archiefvormend orgaan, webregisseur, webmaster, capture-agent, functionaris behoud, kwaliteitsbewaker, recordsmanager, auteursrechthouder	- verantwoordelijk voor inhoud, - verantwoordelijk voor I&I beleid, - belast dagelijks beheer site, - maakt snapshots, - belast behoud en toegankelijkh., - bewaakt kwaliteit snapshots, - verantwoordelijk beheer
organisatienaam	ja	ja	code venW****	archiefvormend orgaan
afkorting organisatienaam	ja	ja	code venW****	
naam functionaris	ja	ja		
contactgegevens	ja	ja	e-maildres, tel- en fax-nummer, postadres,bezoekadres	

Gebruiksvoorwaarden

	verplicht	*herhaalbaar*	*schema*	*opmerkingen*
toegankelijkheid	ja	ja		standaard "geen beperkingen" voor internetsites, eventueel beperkingen ten aanzien van intranetsites (mogelijk ook juridische beperkingen i.v.m. auteursrecht, intellectuele eigendom, privacy e.d.)

Gerelateerde bronnen

	verplicht	*herhaalbaar*	*schema*	*opmerkingen*
unieke identificatie	nee	ja		vorige / voorgaande snapshot van dezelfde website, databases indien ook een rol buiten website, RMA die verwijst naar websitearchief
soort relatie	ja	ja		bevat, is onderdeel van, is oudere versie van, is nieuwere versie van, ontleent informatie aan
beschrijving relatie	nee	nee		indien "soort relatie" de relatie onvoldoende expliciteert

Metadata

	verplicht	*herhaalbaar*	*schema*	*opmerkingen*
`invoerdatum`	ja	ja		
`wijzigingsdatum`	ja	ja		

Legenda:

* Activiteitenindex: conform indeling RVD-notitie 3 november 2004

**Taxonomie: burger- en ambtenarentaxonomie VenW, in ontwikkeling

***BSD: basisselectiedocument

****code VenW: Basiscode Documentaire Informatieverzorging ministerie van Verkeer en Waterstaat 2005

Bijlage E: Suggesties voor verder onderzoek

1. Openstaande vragen

Het onderzoek heeft een aantal direct toepasbare oplossingen opgeleverd, maar ook een aantal vragen nog onbeantwoord gelaten.
In dit project is niet meer geprobeerd om tot een definitie van het begrip website te komen die bruikbaar is voor het archiveren van websites. Er is enkel aangesloten bij wat door het departement als (start-URL van een) website beschouwd werd, zonder te proberen de al dan niet impliciet gehanteerde definitie boven water te krijgen. In ieder geval kan wel gesteld worden dat vanuit de praktijk van het archiveren van sites, technische definities weinig zinvol zijn. Voor de archivering van websites zal een bruikbare definitie naar verwachting aansluiten bij de definitie van archief als procesgebonden informatie.
Er is nog geen onderzoek gedaan naar een geschikte aanpak voor het verduurzamen en presenteren van *deep web*. Vragen rondom de archiveringsfrequentie en benodigde opslagcapaciteit zijn nog niet onderzocht.
Het onderzoek heeft bovendien een aantal aandachtpunten aan het licht gebracht die nader onderzoek verdienen.

2. Webontwerp en client side scripts

Aangeraden wordt dat websites zo gemaakt worden dat ze ook zonder javascript of andere vormen van *client side scripting* goed te gebruiken zijn. Het is niet realistisch te verwachten dat er in de toekomst geen websites meer zijn die javascript of Macromedia Flash gebruiken.
Deze constatering, gecombineerd met de resultaten van dit onderzoek, maakt dat het zinvol geacht wordt te onderzoeken op welke wijze javascript-functionaliteit zo opgezet kan worden dat zij blijft functioneren in de omgeving van webarchief. Hierbij speelt de vraag of de scripts zo opgezet kunnen worden, dat ze goed blijven werken als de relatieve locatie op de webserver verandert.

Hierbij dient de opmerking geplaatst te worden dat het probleem wellicht op een goede manier omzeild kan worden door de snapshots in de archiefomgeving met behoud van het oorspronkelijke pad en mogelijk ook de originele domeinnaam te laten draaien. Zie hiervoor ook 5.6.

Scripts in Macromedia Flash kunnen ook gearchiveerd worden. De wijze waarop de scripts geprogrammeerd zijn, bepaalt of de scripts nog werken na opname in een snapshot. Een inventarisatie en analyse van technieken die problemen opleveren en mogelijk alternatieven lijkt nut te kunnen hebben.

3. Ondersteuning van de actie van het genereren van snapshots

Als de kwaliteit van de snapshots te wensen laat, komt dit dikwijls doordat de crawler niet alle relevante URLs heeft gevonden, en dus niet heeft gearchiveerd. Dit wordt in bijna alle gevallen veroorzaakt door problemen met het *parsen*[43] van *client-side scripts* zoals javascript of Flash en door de aanwezigheid van *deep web*.
Een eenvoudige (maar potentieel arbeidsintensieve) oplossing voor dit probleem is door alle in het webarchief op te nemen URLs expliciet aan de *crawler* door te geven (een voorwaarde is wel dat pagina's in het *deep web* ook echt een eigen en uniek URL hebben). Dit kan gedaan worden door deze URLs expliciet op te nemen in de capture-opdracht.

Een andere mogelijkheid is om op een site een (voor gebruikers normaliter onzichtbare) pagina toe te voegen waarop al deze URLs vermeld staan. Het is voor de hand liggend om op een dergelijke pagina ook expliciet op te nemen welke URLs of welke URL-patronen juist niet door de *crawler* van de archiveringsapplicatie gevolgd moeten worden. Dit leidt tot een constructie vergelijkbaar met de *robots exclusion* standaard, een bestand dat aangeeft wat zoekmachine niet mogen indexeren. Het nadenken over een dergelijke sturing van archiveringsapplicaties zou daarom wellicht gezien moeten

[43] Onder *parsen* wordt het doorlopen van code verstaan, in dit geval met het oog op het achterhalen van links naar andere pagina's / URLs.

worden in het licht van een uitbreiding van de *robots exclusion* standaard. Internationale afstemming, bijvoorbeeld via procedures om te komen tot een RFC-document[44], zijn te overwegen.

Opgemerkt moet worden dat een dergelijke uitbreiding van de *robots exclusion* standaard ook zinvol kan zijn voor internetzoekmachines, bijvoorbeeld omdat het hen in staat stelt ook het *deep web* te indexeren.

4. Archiveringsfrequentie, omvang en kosten

In dit onderzoek is uiteindelijk geen aandacht besteed aan vragen rondom de archiveringsfrequentie. Vragen als welk deel van websites regelmatig verandert, wat een zinvolle archiveringsfrequentie is en wat de gevolgen zijn voor de benodigde opslag bij opslaan van enkel de wijzigingen zijn niet onderzocht.

Omdat juist de archiveringsfrequentie en de benodigde opslagcapaciteit sterk bepalend zullen zijn voor de kosten van webarchivering kunnen er over die kosten nu weinig precieze uitspraken gedaan worden.

5. Integratie met het bestaande regime voor records management

Binnen het onderzoek is de archivering van websites volledig los van bestaande regimes voor archiefbeheer binnen het ministerie beschouwd. In praktijk blijkt bovendien dat websites of pagina's van website niet opgenomen zijn in de bestaande systemen voor records management. Er van uitgaande dat websites te bewaren bescheiden kunnen vormen, behoeft het geen betoog dat het beter en efficiënter zou zijn als websites of pagina's van websites in de bestaande systemen voor recordmanagement zouden worden opgenomen.

Een belangrijke vraag die hierbij speelt is welke eenheid gehanteerd moet worden. Worden websites, pagina's op websites of misschien wel delen van websites als uitgangspunt genomen? Bij een onderzoek naar beantwoording van deze vraag zal rekening moeten gehouden met de wijze waarop selectie plaats zal vinden.

6. Presurf

Aanpassingen omgeving snapshots

Snapshots zijn nu in Presurf beschikbaar onder domeinnaam waaronder de applicatie draait. Doordat alle snapshots binnen Presurf deze domeinnaam gebruiken is noodzakelijkerwijs het pad van een pagina binnen een snapshot anders dan op de originele website. Dit is een bron van diverse problemen met javascripts. Een interessante aanpassing van de applicatie is daarom om de snapshots aan te bieden met het originele pad en zo mogelijk ook de originele domeinnaam.Deze aanpak zal praktische gevolgen hebben. Het *viewer*-gedeelte van de applicatie zal bijvoorbeeld van een eigen DNS[45] gebruik moeten maken. Dit kan ingrijpende gevolgen hebben de wijze waarop de applicatie gebruikt kan worden. Er zal hiertoe daarom een weloverwogen ontwerp gemaakt moeten worden dat recht doet aan zowel de technische eisen als vooral ook aan de eisen en wensen ten aanzien van het gebruik en de inzet van de applicatie.

Inzet archivering Intranet

Gedurende het onderzoek is het deel van Presurf rondom de crawler geïnstalleerd op een laptop en gebruikt om het intranet van Verkeer en Waterstaat te archiveren. Dit bleek mogelijk. De gegenereerde snapshots zijn *off line* te gebruiken of kunnen alsnog in de 'voorkant' van de applicatie Presurf worden opgenomen. Voor een frequente archivering van intranetsites lijkt het zinvol om een oplossing te onderzoeken waarbij Presurf als geheel zonder tussenstappen toegang heeft tot het intranet. Dit vereist een oplossing met een *virtual private network* (VPN) of een oplossing waarbij een server met Presurf daadwerkelijk binnen dezelfde omgeving als het intranet draait.

[44] RFC-documenten (voluit Request For Comments) zijn documenten die de protocollen en andere aspecten van het internet beschrijven. Ze worden uiteindelijk als standaard bekrachtigd door de Internet Engineering Task Force.
[45] DNS oftewel Domain Name Server. Dit is de serverapplicatie die in een intra- of internetomgeving domeinnamen aan IP-adressen koppelt.

Gebruikte crawler-engine

Binnen dit onderzoek bleek dat veel van de geconstateerde problemen in de gegenereerde snapshots samenhangen met de manier waarop de *crawler* van de applicatie omgaat met *client side* scripts. Het is daarom te moeite waard te bestuderen of een andere of aangepaste *crawler engine* betere resultaten oplevert.

7. Overdracht Nationaal Archief

Te bewaren snapshots zullen naar verwachting uiteindelijk naar het Nationaal Archief overgebracht moeten worden. De wijze waarop dit moet gebeuren is nog niet duidelijk en zal samen met het Nationaal Archief bepaald moeten worden.
De kern Presurf met snapshots die in principe *off line* te *browsen* zijn (dus zonder hulpmiddelen te benaderen, anders dan een webbrowser) in combinatie met metadata vastgelegd in bestanden in tekst / xml-formaat maken dat het niet waarschijnlijk is dat er grote problemen te verwachten zijn bij conversie naar een formaat zoals het Nationaal Archief het zou wensen. Te onderzoeken is of een bestandsformaat als ".ARC" valt te overwegen. Een voordeel is bijvoorbeeld de onafhankelijkheid van beperkingen zoals opgelegd door de gebruikte bestandssytemen,