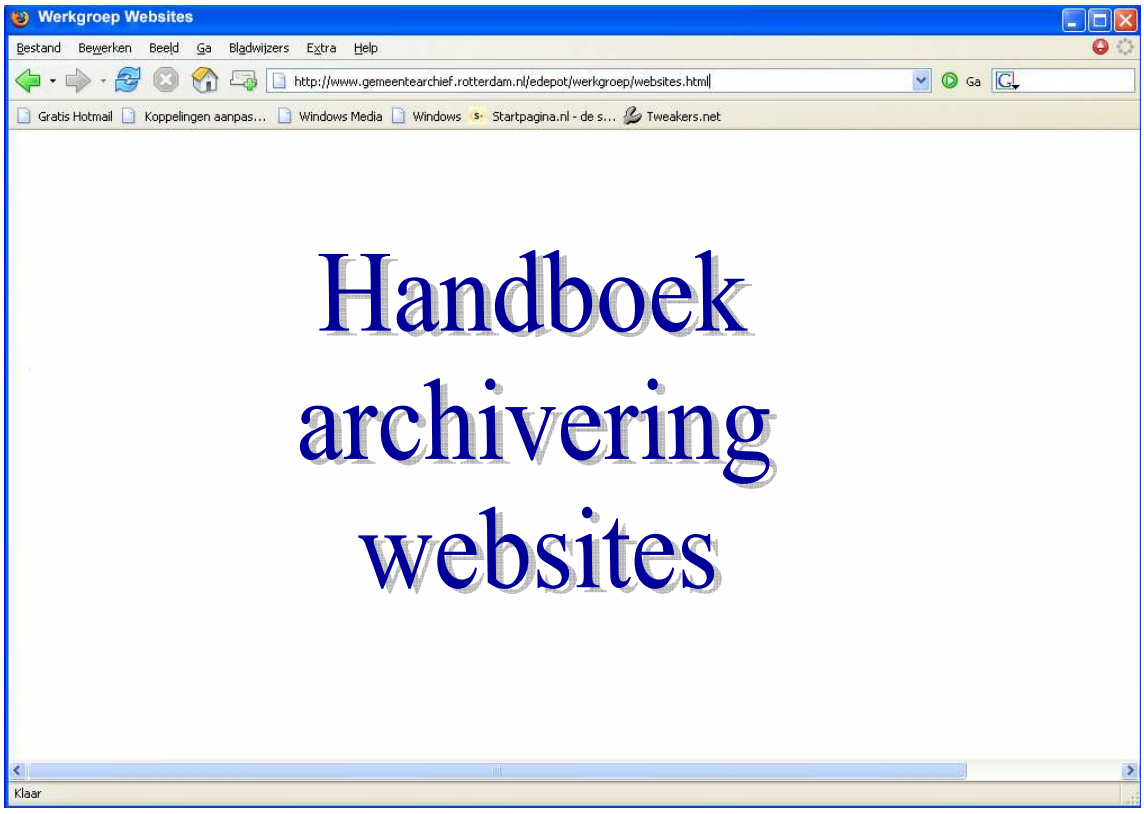
Gemeentearchief Rotterdam
Gemeente Rotterdam

Project E-depot

Werkgroep Websites



Versie: 1.1, concept voor Projectgroep

Datum: 17 november 2005

Auteurs: Jeroen van Oss, Ivar Bermon, Peter van Wijngaarden, Guus van Veldhuizen

Inhoudsopgave

Inleiding

Doel van het handboek

Dit handboek vormt het tastbare product van de werkgroep. Het is enerzijds een handleiding-in-opbouw voor het archiveren van websites en anderzijds een verslag van de activiteiten van het deelproject. In dit verslag worden de bevindingen weergegeven, de keuzes verantwoord, en aanbevelingen gegeven. Het is de bedoeling dat dit handboek na afronding van dit deelproject verder wordt aangevuld met nieuwe of geactualiseerde procedures, instrumenten en methodes.

Doel en resultaten

De werkgroep heeft als opdracht gekregen het ontwikkelen van strategieën, methoden, procedures en technieken voor het selecteren, archiveren en beschikbaarstellen van Rotterdamse[1] websites in het e-depot. Deze opdracht wordt als voltooid beschouwd wanneer de werkgroep:

1. Een keuze heeft gemaakt in de bewaardoelstelling en selectie van websites (Waarom archiveren?);
2. Op grond daarvan kwaliteitseisen heeft geformuleerd aan de authenticiteit en integriteit (context, vorm, structuur, inhoud en gedrag) van gearchiveerde websites (Wat archiveren?);
3. Een archiveringsmethode heeft ontwikkeld waarmee de gestelde authenticiteitseisen kunnen worden gerealiseerd, die geïmplementeerd is in de werkprocessen van de betrokken afdelingen. (Hoe archiveren?)

Ad 3: Dit doel valt uiteen in twee onderdelen:

- De archiveringsmethode zelf, uitgewerkt naar de hoofdprocessen van archivering, zoals aangeduid in het Procesmodel:
    a. Opnemen: er is techniek voor opname geselecteerd, waarmee ervaring is opgedaan; website-specifieke functionele eisen voor e-depot software en procedures zijn benoemd;
    b. Bewaren: er is een strategie voor preservering geformuleerd, toegesneden op het medium websites; website-specifieke metadata voor beheer zijn benoemd; website-specifieke risico's in verband met quarantainebehandeling zijn benoemd;
    c. Fysiek beheer: website-specifieke functionele eisen voor e-depot software en procedures zijn benoemd;
    d. Intellectueel en administratief beheer: website-specifieke metadata zijn benoemd;
    e. Beschikbaar stellen/zoeken en raadplegen: op basis van de authenticiteitseisen is vastgesteld hoe de website gepresenteerd moet worden. Website-specifieke functionele eisen voor E-depot/Digitale Balie software zijn benoemd;
- De implementatie in de organisatie:
    a. Voorstellen voor inpassing van website-archivering in de organisatie van het gemeentearchief, rekening houdend met het Organisatiemodel van het E-depot;

[1] Met 'Rotterdams' wordt bedoeld: 'binnen het werkgebied van het gemeentearchief Rotterdam'.

b. Het bovengenoemde proces van archivering van websites is beschreven en kan als basis dienen voor een op te stellen (deel)procedure website-archivering;
c. Kennis van en ervaring met het archiveren van websites is verworven door de betrokken afdelingen;

Scope

De opdracht dient uitgevoerd te worden binnen de volgende grenzen:

- Naar herkomst: zowel websites van gemeentelijke organen, die deels onder de Archiefwet vallen, als websites van private organisaties binnen het werkgebied van het Gemeentearchief Rotterdam.
- Naar type: alle typen websites, met alle objecten die zich daarin bevinden. Nieuwsgroepen en discussielijsten die niet in een website vervat zijn, vallen niet binnen de scope.
- Naar bewaardoelstelling: zowel vanuit het oogpunt van collectie-opbouw, als vanuit het oogpunt van archiveringsverplichtingen.

Werkwijze en bronnen

De werkwijze van de werkgroep was tweeledig. Omdat in Nederland en internationaal al veel ervaring is opgedaan met websitearchivering, heeft de werkgroep zo veel mogelijk gebruik gemaakt van (web)publicaties. Hierbij ging het zowel om de theoretische basis, als om het leren van elders opgedane ervaringen. Belangrijke bronnen waren de publicaties van het Testbed Digitale Bewaring, het project DAVID van het Stadsarchief Antwerpen, van Erika Hokke van de Archiefschool, van de Deense Koninklijke en Staatsbibliotheken en van het Ministerie van Verkeer en Waterstaat/Capsis. Een complete literatuurlijst is achterin opgenomen.

Daarnaast heeft de werkgroep in beperkte mate zelf geëxperimenteerd met het archiveren van websites. Daartoe hebben de leden een testset van ca. 50 ‘Rotterdamse’ websites samengesteld, evenwichtig verdeeld over de verschillende typen en herkomst. Om haar kennis op zodanig niveau te brengen dat de werkgroep in staat was om te experimenteren, heeft zij een training georganiseerd, met als doelen het verhogen van de algemene kennis over websites in het licht van archivering en het vergroten van vaardigheid in het opnemen (harvesten).

De bevindingen van literatuurstudie, training en experimenteren zijn steeds besproken in de werkgroep en vervolgens vastgelegd in het handboek.

Leeswijzer

Het handboek is ingedeeld naar de deelvragen van het project:

- Waarom archiveren en wat archiveren?: H 1: Bewaardoel, selectie en authenticiteit
- Hoe archiveren?: H 2: Archiveringsmethode en implementatie

In iedere paragraaf komen de volgende aspecten terug:

- Vraagstelling
- Verslag van de activiteiten
- Verworven kennis
- Gemaakte keuzes
- Aanbevelingen
- Vraagstellingen voor verder onderzoek

1. Bewaardoel, selectie, integriteit en kwaliteit

Inleiding

Voordat je kunt beginnen met het ontwikkelen van een archiveringsmethode voor websites, is het belangrijk om te bepalen voor welk doel je ze wilt bewaren. Dit bewaardoel is vervolgens de basis voor de waardering en selectie van websites en van de kwaliteitseisen ten aanzien van authenticiteit en integriteit van het 'document' website. Deze paragraaf behandelt eerst het bewaardoel, daarna het selectiebeleid en tenslotte de kwaliteitseisen.

Wat is een website?

Door zijn dynamiek en open karakter is een website niet makkelijk af te bakenen. De werkgroep hanteert de definitie die Erika Hokke in haar publicatie 'Blijvend beschikbaar' heeft geformuleerd: *Een website is het medium waarmee via het internet informatie gepresenteerd wordt op een statische of dynamische manier, vaak in combinatie met technologieën voor tweezijdige communicatie en transactie, zoals e-mail, nieuwsgroepen of discussielijsten* (H.A. Hokke Blijvend Beschikbaar).
Dat betekent dat niet de website het document/informatie-object is, maar de items die via de website gepresenteerd worden, in hun onderlinge samenhang. Hoe die samenhang wordt afgebakend, moet per website worden bepaald, op grond van een analyse van de inhoud en functie van de geboden informatie. Zie hierover de paragraaf 'Eisen aan authenticiteit'.

Websites als onderdeel van een archief

Vanuit het archiefperspectief kan een website alleen bezien worden als onderdeel van de informatiehuishouding van de archiefvormer. Voor sommige items is de website het enige bewaarmedium, voor andere vervult de website alleen de functie van publicatiemedium en wordt het authentieke document elders bewaard. Een voorbeeld zijn de besluiten van een deelgemeenteraad in het Bestuurlijk Informatie Systeem. De ondertekende exemplaren bevinden zich in dossiers in het papieren archief. De website dient als interface. De neerslag van transacties via de website (denk aan het boeken van een reis of het aanvragen van een vergunning) wordt vastgelegd in een backoffice-applicatie, de website is de gebruikersinterface. Andere items op de website kunnen wel beschouwd worden als archiefbescheiden in zichzelf: een officiële kennisgeving op grond van de Algemene Wet Bestuursrecht, maar ook de hele website als zichtbare neerslag van de wijze waarop een organisatie met haar omgeving communiceert.

Bewaardoel

Om te kunnen bepalen welke websites we willen verzamelen (selectie) en wat we van een geselecteerde website willen bewaren (integriteit en kwaliteit), moeten we eerst het bewaardoel bepalen.
We kunnen twee 'hoofd' bewaardoelen onderscheiden:

1. Vanuit collectieperspectief, primair als cultureel en historisch erfgoed.
2. Vanuit archiefperspectief: primair als ondersteuning van de bedrijfsvoering, als bewijs en verantwoording, secundair als cultureel en historisch erfgoed.

In de opdracht zijn beide invalshoeken opgenomen. Nadere discussie in de werkgroep heeft geleid tot het inzicht dat het verstandig is om voor dit projectjaar een keuze te maken tussen beide benaderingen, omdat deze zeer uiteenlopend zijn. Voor het leerproces leek het de werkgroep verstandig om ons eerst te richten op het archiveren van websites als losse eenheden, als collectie-items, met het oog op het historisch belang, en pas in een vervolgproject het meer gecompliceerde onderwerp van archivering van websites als onderdeel van een organisatiearchief te behandelen. Het voordeel hiervan is, dat de intern opgedane ervaringen met het archiveren van websites, kunnen worden ingezet om de archiefvormers beter te kunnen adviseren over hun archiveringsstrategie.

Selectie

De selectie van websites, of het nu uitgevoerd wordt vanuit archiveringsoogpunt of vanuit collectie-oogpunt, is in principe niet anders dan die van papieren documenten. De huidige collectieprofielen van Atlas, Bibliotheek en Archieven kunnen prima worden gebruikt als grondslag voor de selectie van websites. Tot zo ver hoeft de invoering van acquisitie van websites niet te leiden tot een nieuw selectiebeleid. Wel is het zo, dat websites de mogelijkheid bieden om materiaal te verzamelen van organisaties of personen die anders niet, of alleen op een tijdrovende manier te benaderen zijn. Het verdient daarom aanbeveling om de acquisitieplannen van de collectiebeherende afdelingen aan te passen aan de verzamelmogelijkheden die websites bieden.

Met name vanuit de bibliotheekwereld is een trend ontstaan om een volledig webdomein te archiveren. Het bekendste voorbeeld is The Internet Archive, dat in 1996 begon met het archiveren van het totale WorldWideWeb. In principe zouden deze activiteiten alle lokale initiatieven voor het archiveren van websites overbodig maken. Dit is echter op dit moment niet aan te raden, omdat de integriteit en kwaliteit van de gearchiveerde websites en de continuïteit van dergelijke organisaties niet gegarandeerd is. Bovendien zijn de zoekfuncties zeer beperkt. Bij The Internet Archive kan alleen op de URL-naam worden gezocht. Het is nog niet duidelijk welke rol de Koninklijke Bibliotheek in Nederland ten aanzien van website-archivering op zich gaat nemen: totale bewaring van het Nederlandse webdomein (zoals bij andere Nederlandse publicaties in het kader van het wettelijk depot) of niet. Het verdient aanbeveling om de plannen van de KB nauwlettend te volgen.

Eisen aan authenticiteit

Bij het archiveren van een dynamisch en veelvormig documenttype als websites is het van groot belang af te bakenen wat we van zo'n website willen archiveren. Willen we de chatsessies met de wethouder, de database met stamboomgegevens, de via de website gedane aanvragen voor een kapvergunning bewaren, of gaat het ons alleen maar om de inhoud en de presentatievorm? Willen we het continuüm van mutaties en uitbreidingen 'traploos' kunnen weergeven, of volstaan we met 'snapshots'? Kortom wat is de kwaliteit van de te archiveren website? De keuzes die we daarin maken, vormen het uitgangspunt voor de te hanteren archiveringsmethode.

Vóórdat die keuze gemaakt kan worden, moet eerst worden bepaald wat een website maakt tot wat hij is. Is het een presentatiemedium, een portaal naar andere informatiebronnen of een

transactiemedium? Het antwoord van die vraag hangt af van de functie die de website heeft in één of meerdere werkprocessen. Als dat gedefinieerd is, kun je vervolgens analyseren welke elementen van de informatie-objecten die via de website gepresenteerd worden behouden moeten blijven om er voor te zorgen dat de betekenis daarvan overeind blijft. Welke elementen bepalen de authenticiteit van een website? Het Testbed Digitale Duurzaamheid heeft een methode toegepast om dit voor digitale documenten te definiëren. Daarbij staan twee begrippen centraal: integriteit en verifieerbaarheid.

Met integriteit wordt bedoeld dat het document intact is en niet zodanig veranderd of gecorrumpeerd dat de betekenis ervan niet meer duidelijk is. Wijzigingen zijn tot op zekere hoogte aanvaardbaar, zolang de oorspronkelijke betekenis of functie van het document er niet door wordt aangetast. Verifieerbaarheid betekent dat vast te stellen is dat het document is wat het beweert te zijn. Om dit mogelijk te maken is contextinformatie nodig. Deze informatie wordt vastgelegd in metadata.

De integriteit van een digitaal document hangt af van vijf elementen:
Context: de oorspronkelijke omgeving waarin de website is gemaakt. Zowel de ontstaanscontext (vormer, werkproces) als de relatie tussen de objecten op de website en de relatie met objecten op andere websites. Ook de relatie met de totale informatiehuishouding van een organisatie valt onder het begrip context.
Inhoud: Tekst, plaatjes, filmpjes, geluid, maar ook databases.
Structuur: De samenstelling van een informatieobject op een websites (meestal webpagina) uit allerlei bronnen (backoffice-applicatie, plaatjes, pdf'jes). Dit vindt vaak 'on the fly' plaats, op het moment dat men het oproept, met behulp van een ASP-applicatie.
Vorm (of uiterlijk): de presentatievorm van de website: de vormgeving (lettertype, kleuren, opmaak).
Gedrag (of functionaliteit): De interactieve mogelijkheden t.b.v. de gebruiker. Bijvoorbeeld: zoeken in een database, aanvragen van een vergunning, chatten met de wethouder.

Om dit keuzeproces te structureren heeft de werkgroep het onderstaande formulier ontworpen. Hierin kunnen per element de authenticiteitseisen vastgelegd worden. Dit formulier kan gebruikt worden om per website bepalen welke elementen behouden moeten worden om de betekenis van de websites, bezien vanuit het bewaardoel, overeind te houden en welke elementen als 'accepted loss' beschouwd kunnen worden. Met de aldus geformuleerde authenticiteitseisen kan het archiveringsproces gestuurd worden. De werkgroep heeft, uitgaande van het geformuleerde bewaardoel, namelijk het Rotterdams historisch erfgoed, de in de testset geselecteerde websites beoordeeld en op grond daarvan de volgende generieke kwaliteitseisen geformuleerd. Deze eisen moeten worden gezien als een richtsnoer en niet als ijzeren voorschrift. Analyse van individuele websites kan tot andere keuzes leiden.

2. Archiveringsmethode

Archiveringsproces

Inleiding

Het doel van dit hoofdstuk is het formuleren van een archiveringsstrategie, toegesneden op websites, die kan voldoen aan de in hoofdstuk 1 geformuleerde authenticiteitseisen. Hiertoe heeft de werkgroep de volgende activiteiten ontplooid:

- Literatuur bestudeerd en in de werkgroep besproken, van de Deense Koninklijke en Staatsbibliotheken, de Nederlandse Koninklijke Bibliotheek, DAVID, de Archiefschool, Capsis/Ministerie van Verkeer en Waterstaat en The National Archives UK.
- Deelgenomen aan de klankbordgroep van het project Webrichtlijnen voor overheidsinstellingen versie 2, van het ICTU.
- Geëxperimenteerd met het opnemen en bevriezen van websites door middel van harvesting. Hiertoe zijn de leden getraind door het bureau Capsis.
- Ideeën uitgewisseld met Capsis over preserveringsstrategieën, naar aanleiding van een demonstratie van Capsis over het archiveringspakket Presurf.
- Inbreng geleverd bij het bedenken van de functionaliteit van de Q en de P omgevingen van het E-depot, ten behoeve van het bouwen van een ‘Proof of Concept’ door het bureau CQ2.
- Een concept-metadataset voor websites samengesteld.
- Tijdens het DLM forum oktober 2005 in Boedapest informatie verzameld over de nieuwste stand van zaken.

De meeste webarchiveringsprojecten, zoals Archipol, The Internet Archive en The National Archives van de UK, beperken hun archiveringsstrategie tot het ‘bevriezen’ van een website door middel van een snapshot en het beschikbaar stellen met behulp van de huidige browsers. Het resultaat is een kopie van de website, opgebouwd uit de meest uiteenlopende bestandstypen, zoals JPG-plaatjes, JAVA-scripts, Flash-filmpjes, etc, die met de huidige browsers, en de bijbehorende plug-ins, nog wel te raadplegen is, maar waarvan je nu al kunt zeggen dat toekomstige browsers daar geen raad mee zullen weten. Om websites duurzaam toegankelijk te houden, is ingrijpen in iedere archiveringsfase nodig. In de onderstaande paragrafen worden zij stuk voor stuk behandeld.

Creatie

Bij de bouw van een website moet al rekening gehouden worden met de archiveerbaarheid, bijvoorbeeld door gebruik te maken van gestandaardiseerde bestandsformaten en af te zien van technische hoogstandjes die om zeer specialistische software vragen. Dit aspect is maar in beperkte mate behandeld, omdat het ICTU, in opdracht van het Ministerie van BZK, een nieuwe versie van de Webrichtlijnen voor overheidsinstanties in voorbereiding heeft, waarin de archiveringsaanbevelingen geïntegreerd zijn. Het gemeentearchief was vertegenwoordigd in de klankbordgroep. Zodra deze richtlijnen er zijn, is het zinvol om plannen te bedenken voor de implementatie hiervan bij de gemeente

Rotterdam. De werkgroep adviseert om dit mee te nemen in het E-depot projectjaar 2006. Deze richtlijnen kunnen ook aangepast worden voor gebruik ten behoeve van particuliere leveranciers van archieven en collectiemateriaal.

Opname

Opname van websites betekent vooral het vangen (capture) en bevriezen van een dynamisch complex van informatie-objecten. De toegepaste methode, bijvoorbeeld door middel van snapshots, is van grote invloed op de authenticiteit en integriteit van de te archiveren website. Aan de hand van eerdere bevindingen elders, kunnen een aantal mogelijke methodes of strategieën voor opname van websites worden aangegeven:

- *Rechtstreeks van de broncode*

  Hierbij worden de bestanden waaruit de website bestaat overgenomen en gearchiveerd.

- *Via een snapshot*

  Door middel van een zogenoemde webcrawler wordt een momentopname gemaakt van de betreffende website.

- *Unieke webpagina's*

  Hierbij houdt een programma op de webserver bij welke pagina's worden opgevraagd en opgebouwd. Wanneer deze pagina nog niet eerder is gearchiveerd dan wordt de pagina automatisch alsnog gearchiveerd.

- *Opname van een surfsessie*

  Door middel van een 'screenrecorder' wordt het surfen gefilmd. Wat dus op het scherm verschijnt wordt vastgelegd. De opname kan dan als een videobestand worden bewaard en afgespeeld.

De eerstgenoemde methode is het meest zuiver en biedt de meeste garantie voor authenticiteit. Voor eenvoudige websites, die alleen bestaan uit HTML met wat plaatjes, is deze methode zeer geschikt. Dit type website wordt echter steeds meer verdrongen door websites die 'on the fly' met behulp van een database-applicatie (een ASP of PNP) worden gegenereerd. Bewaren van de bronbestanden, betekent dan ook bewaren van de database-applicatie. Wellicht is het mogelijk dat deze applicaties voorzien worden van een archiveringsfunctie. Zolang deze echter niet voorhanden is, is archiveren aan de bron voor de meeste websites geen haalbare kaart. De werkgroep heeft er voor gekozen om het onderzoek in eerste instantie te beperken tot de archivering van snapshots, omdat deze methode het meest beproefd is. Hoewel de overige methoden zeker niet uit het oog mogen worden verloren. Wanneer wordt aangelopen tegen mogelijke beperkingen van de snapshotmethode dan zullen de alternatieven ook bekeken moeten worden. In de literatuur (DAVID en Capsis) worden de volgende voor- en nadelen van de snapshotmethode genoemd:

Voordelen:

- Vastlegging in de oorspronkelijke opmaak en met minimale functionaliteit opgeslagen
- De website wordt volledig vastgelegd en kan dus in zijn geheel (offline) worden geraadpleegd.
- Alleen die bestanden worden gearchiveerd die deel uitmaken van de online versie van de website.

Nadelen

- De koppeling met applicaties die aan de website hangen gaan verloren.
- Het maken van een goede webshot vereist technische kennis
- Voor grote websites kan het lang duren voor alles binnen is en nogal veel beslag leggen op servers.
- De kwaliteit is ook afhankelijk van de snelheid van de verbinding.
- De mogelijkheid bestaat dat het harvesten door webmasters wordt geweerd.
- Bepaalde soorten links kunnen moeilijk worden omgezet naar bruikbare links.
- Voorkennis over de structuur van de binnen te halen website kan noodzakelijk zijn.
- Niet meer werkende links kunnen het proces van binnenhalen laten vastlopen.
- Afgeschermde onderdelen van websites zijn moeilijk vast te leggen.
- Het is noodzakelijk om na het binnenhalen een grondige controle uit te voeren.
- Het verbeteren van eventuele fouten is arbeidsintensief.

Om te kunnen ervaren in hoeverre deze nadelen van invloed zijn op de gestelde authenticiteitseisen, heeft de werkgroep geëxperimenteerd met het harvesten van websites. Om een aantal redenen is gekozen voor het programma HTtrack. Eén daarvan is dat het programma ook al is gebruikt in 2004 door de toenmalige werkgroep websites. Het is de meest bekende harvestsoftware waar ook al veel ervaring door anderen mee is opgedaan. Tevens is het een Open Source software.

Vanwege de ervaringen van de subgroep websites in 2004, waarbij bleek dat het binnenhalen van websites vaak problematisch was, is besloten om hulp van buiten in te roepen. Het bedrijf Capsis heeft voor het ministerie van Verkeer en Waterstaat al veel ervaring opgedaan met het harvesten van websites met HTtrack en heeft tevens een viewer ontwikkeld voor de gearchiveerde websites. Capsis is ingehuurd om aan de deelnemers van het deelproject een inleiding en cursus te geven over hoe websites te harvesten met HTtrack. De cursus werd verzorgd door René Voorburg en werd positief ervaren onder andere door de korte inleiding over de basisprincipes van internet. Vervolgens zijn de deelnemers aan de slag gegaan met het harvesten van een aantal websites, waarna de resultaten tijdens een terugkomdag zijn besproken.

Bevindingen testen/oefenen HTtrack

- Tijdens het harvesten blijkt dat nog veel fout gaat. Hoewel het grootste gedeelte meestal wel goed gaat blijken er toch bij vrijwel elke website wel iets niet helemaal goed te gaan. Een totaal probleem- en foutloos binnengehaalde website is een zeldzaamheid.
- Het is belangrijk dat voor een website wordt geharvest eerst de site goed bekeken moet worden op eventuele knelpunten, zodat daarop kan worden ingespeeld met HTtrack.
- De voornaamste problemen hebben betrekking op frames, links gekoppeld aan bestanden en bij bepaalde bestanden zoals Java en Flash
- Meestal zal eerst een 'test-run' gedaan moeten worden om eventuele problemen op te sporen waarna de instellingen kunnen worden aangepast.
- Het kan lang duren, zeker bij grotere websites, voordat een site geheel is binnengehaald. Tijdens het binnenhalen is er de mogelijkheid dat een 'loop' optreedt (het binnenhalen blijft dan hangen op een bestand die steeds opnieuw wordt binnengehaald). Het proces moet daarom wel blijvend worden gevolgd.
- Het goed kunnen binnenhalen is vrij specialistisch werk en vraagt onder andere een goede kennis van webpagina's (zoals HTML en Java) en ervaring met het maken van snapshots. Het is de vraag of niet-technisch geschoolde medewerkers van de beherende afdelingen dit goed aan kunnen.
- Er is een concept handleiding voor het werken met HTtrack (zie bijlage 5) welke uitgebouwd kan worden naarmate meer ervaring wordt opgedaan.
- HTtrack slaat de binnengehaalde bestanden wel goed op in een duidelijke structuur. Elke website krijgt een eigen map met daarin twee mappen en een aantal bestanden, waaronder de koppeling naar de index pagina van de gevangen website en het logbestand.
- Wanneer een website goed is binnengehaald blijkt de snapshotmethode een goede manier om websites te archiveren. Aan alle generieke authenticiteitseisen, voorzover zij met de opnamesoftware te maken hebben, kan worden voldaan.

Toch kan het voor bepaalde websites nodig zijn dat ook de bronnen worden gearchiveerd. Dit is vooral van belang voor complexere zaken als bijvoorbeeld databases die als deep-web toepassing bij de website behoren en waarvoor de website het enige bewaarniveau is. Door de technische opbouw van een website is de snapshotmethode niet altijd even bruikbaar. Vooral bij dynamische websites, waarbij de inhoud gegenereerd wordt tijdens de surfsessie, is het maken van een snapshot niet altijd even goed mogelijk. Verder onderzoek zal moeten uitwijzen in hoeverre de beperkingen van de snapshotmethode ook een werkelijke beperking is voor het archiveren van websites. Er is nu nog geen duidelijk beeld van hoe vaak dynamische websites of deep-web toepassingen een rol spelen bij websites die we willen archiveren en of de gegevens in een deep-web toepassing ook te bewaren gegevens zijn.

Doordat veel websites voordurend veranderen is het van belang daarmee rekening te houden bij het archiveren. Er zijn twee benaderingen mogelijk om websites zo volledig mogelijk te archiveren.

- De objectgeoriënteerde benadering waarbij per snapshot de gehele website wordt binnengehaald en bewaard. Het voordeel is dat deze benadering vrij simpel is en door HTtrack makkelijk uit te voeren. Nadeel is het beslag dat wordt gelegd op opslagcapaciteit, omdat alles van een website telkens opnieuw wordt binnengehaald en opgeslagen, zelfs die delen die niet gewijzigd zijn.
- De tweede benadering is de gebeurtenisgestuurde benadering. Hierbij wordt bijvoorbeeld telkens als er een wijziging heeft plaatsgevonden alleen die pagina die gewijzigd is binnengehaald. Ook kan deze benadering gebruikt worden voor zeer dynamische websites, door een snapshot te maken van een pagina zoals die gevormd wordt tijdens het opvragen van een bezoeker. Het is voor deze benadering wel noodzakelijk om gebruik te maken van aanvullende software die HTtrack aanstuurt. Hoewel deze benadering ingewikkelder is dan de objectgeoriënteerde wordt wel voorkomen dat onnodige data wordt opgenomen. Wanneer een site echter met de gebeurtenisgestuurde benadering is gearchiveerd dan zal voor het raadplegen daarvan aanvullende software wenselijk zijn, zodat de afzonderlijk binnengehaalde pagina's binnen de oorspronkelijke context van de website bekeken kunnen worden.

Het harvesten van websites via HTtrack is niet zonder problemen en als de bevindingen van het testen naast de software-eisen in bijlage 4 worden gelegd dan lijkt het programma niet al te best uit de bus te komen, zeker als het gaat om gebruiksvriendelijkheid. Het is daarom zinvol ook te kijken naar mogelijke alternatieven van harvestingsoftware. Als mogelijke alternatieven zou het programma Heritrix onderzocht kunnen worden. Heritrix is ontwikkeld door The Internet Archive en is één van de programma's die gebruikt wordt voor hun wereldwijde World Wide Web database. Het is een Open Source programma ontwikkeld in Java, maar in principe ontwikkeld voor een Linux omgeving. Het is de vraag of het programma niet alleen geschikt is voor grote projecten als The Internet Archive met zijn miljoenen websites maar ook goed functioneert binnen het beperkte kader van Rotterdam. Maar in het vervolg van het deelproject Websites zullen de mogelijkheden van de alternatieve programma's zeker nader bekeken moeten worden.

Capsis heeft een applicatie ontwikkeld, Presurf geheten, die voor de capture gebruikt maakt van de engine van HTtrack met een aantal uitbreidingen/aanpassingen. Deze applicatie neemt een aantal nadelen van HTtrack weg en maakt het tevens gebruikersvriendelijker. De werkgroep beveelt aan om deze applicatie op proef in gebruik te nemen. Gezien de eigen ervaringen en die van Capsis, is ook dan nog veel specialistische inbreng nodig. Bij het experiment dat door Capsis is uitgevoerd bij het Ministerie van Verkeer en Waterstaat werd een score gehaald van 70%. Bij ons eigen experiment, met beperkte specialistische inbreng, kwamen we tot ca. 50%. Deze scores zijn alleen te verbeteren door ervaring op te doen, tools te bouwen en de methodiek te verfijnen. De werkgroep beveelt aan deze expertise in huis op te bouwen en te beleggen bij de e-conservator.

Bewaren

Onder het deelproces ‘Bewaren’ vallen zowel de opslag als de preservering van informatie-objecten.

Opslag

Lopende het E-depot project is de vraagstelling verschoven van het daadwerkelijk realiseren van opslag van websites in het e-depot, naar het opstellen van website-specifieke functionele eisen. De reden hiervoor was dat de geselecteerde e-depotsoftware (Dspace en I-tor) niet geschikt bevonden was en daardoor maatwerk nodig was. Hiervan zijn tot nu toe alleen de quarantaine- en een deel van de preserverings omgeving gerealiseerd.

Om het doel te bereiken heeft de werkgroep deelgenomen aan de sessies met CQ2, waarin de functionele eisen voor de E-depot software werden gedefinieerd. Dit leverde de volgende website-specifieke functionele eisen voor de software op:

- De samenhang tussen de verschillende pagina’s en bestanddelen van pagina’s (plaatjes, geluidsfragmenten, scripts, etc.) blijft intact.
- Bij opvragen wordt het juiste startbestand (meestal de index) aangeroepen.
- Executables in de website, zoals scripts, worden niet door de virusscan geweerd.

Zo lang de e-depot software nog niet beschikbaar is, is het wel mogelijk om een ‘low-profile’ oplossing toe te passen. Het pakket Presurf van Capsis biedt naast functionaliteit voor harvesting, ook faciliteiten voor opslag en beschikbaarstelling. De werkgroep beveelt aan om aan de hand van de authenticiteitseisen en uitvoeringsvereisten na te gaan of dit pakket geschikt is als tussenoplossing.

Preservering

De randvoorwaarden voor langdurig behoud worden bepaald door de deelprocessen creatie en opname, die in het bovenstaande behandeld zijn. In deze paragraaf is de vraag aan de orde: welke maatregelen worden genomen voor duurzame raadpleging als het object eenmaal is opgenomen? Het antwoord op deze vraag kan vervolgens consequenties hebben voor de methode van presentatie en beschikbaarstelling.

Websitepreservering in de bovengenoemde zin wordt, voor zover we konden nagaan, alleen onderzocht door het Deense project en de Nederlandse Koninklijke Bibliotheek. Beide hebben een methode gekozen, die afwijkt van de door het Rotterdamse e-depot gehanteerde migratie-methode, namelijk emulatie.

Uit ons onderzoek blijkt dat migratie in de wereld van website-archivering niet populair is. De reden waarom men hiervoor terugschrikt is dat migratie van de componenten van een website arbeidsintensief en gecompliceerd handwerk betekent. In feite moet de website, die uit veelsoortige, door middel van HTML aan elkaar gelinkte bestanden bestaat, helemaal met de hand worden

gereconstrueerd. Het voordeel van emulatie is, dat het origineel intact blijft en alleen de opslag- en raadpleegomgeving aangepast wordt. Dit kan op verschillende manieren.

-De Deense aanpak
Volgens de Deense aanpak wordt een bibliotheek van en converters en viewers opgebouwd. Bij opvraging wordt een op de benodigde bestandsformaten afgestemde mix van viewers en converters in stelling gebracht om het object op het gewenste platform te representeren. Het nadeel hiervan is dat deze viewers en converters zelf steeds aangepast moeten worden aan de heersende hard- en softwareplatforms van de tijd.

-Aanpak van de KB
De Nederlandse KB onderzoekt samen met het Nationaal Archief de mogelijkheid van emulatie van hardware. Van elke hardware-component (processor, geheugen, scanner, etc.) wordt een emulator aangemaakt, opgeslagen in een bibliotheek en in de juiste combinatie opgeroepen. Men noemt dit ‘modulaire emulatie’. De modules worden aangeroepen door één ‘virtuele machine’. Het voordeel is dat alleen deze ‘machine’ mee hoeft te evolueren met de verandering van platforms.

Keuzes
De conclusie van de werkgroep is dat het preserveren van websites gecompliceerder is dan van de meeste documentsoorten die bijvoorbeeld het Testbed Digitale Bewaring heeft onderzocht (tekst, spreadsheets, databases en e-mails). Het onderzoek staat nog maar aan het begin. De onderzoeksplannen die er liggen, zijn veeleisend en gericht op de middellange termijn. Geen van deze onderzoeken sluit aan op de migratie strategie die het E-depotproject van het Gemeentearchief Rotterdam heeft gekozen. De vraag moet daarom niet zijn: welke preserveringsmethode gaan we op dit moment toepassen, maar welke richting vinden we kansrijk en waar sluiten we ons bij aan en wat past bij de strategie die we voor de overige documenttypen hebben bedacht?

Aanbevelingen
Het is raadzaam om een preserveringsmethode niet alleen te ontwikkelen, maar hiervoor partners te zoeken. Om te verkennen of hier draagvlak voor is, stelt de werkgroep voor om een workshop organiseren over migratie versus emulatie bij websites. Hierbij zouden deskundigen en belanghebbenden van de Koninklijke Bibliotheek, het Gemeentearchief Amsterdam, het Nationaal Archief, Capsis, de Archiefschool, het Ministerie van Verkeer en Waterstaat, het Stadsarchief Antwerpen en Nederlands Documentatiecentrum Politieke Partijen (Archipol) aanwezig moeten zijn. Dit handboek kan mogelijk als startpunt dienen.

Intellectueel en administratief beheer

Onder intellectueel en administratie beheer wordt verstaan het ontsluiten van informatie-objecten door middel van zoekkenmerken en het vastleggen van gegevens ten dienste van de verwerving, het

beheer, de preservering en de beschikbaarstelling van die objecten. Beide soorten gegevens worden metadata genoemd. Metadata dienen tevens voor het garanderen van interpreteerbaarheid door middel van het vastleggen van belangrijke context-gegevens.

De werkgroep heeft zich ten doel gesteld een metadataschema op te stellen, specifiek voor websites. Zij heeft daartoe, op basis van het Handboek webmetadata versie 1.01 van Advies.Overheid.nl, de daarin voorgestelde metadata velden beoordeeld op bruikbaarheid en relevantie voor het GAR. De ISBD (ER) regels zijn hierin goed terug te vinden. Tevens loopt dit model goed in de pas met het Dublin Core model. Een aantal velden die door ISBD (ER) worden ondergebracht in het "annotatieveld" komen wellicht in aanmerking voor "een eigen veld". Het aldus ontstane concept-metadatamodel is opgenomen in bijlage 2. Op het moment dat dit geschreven wordt, wordt er nog gewerkt aan een gemeenschappelijk metadataschema voor Archieven, Bibliotheek en Atlas. Dit metadata voor websites wordt als input voor dit overleg gebruikt.

Beschikbaarstelling

Het beschikbaar stellen van websites zal plaatsvinden middels de Digitale Balie. Websites hebben echter een aantal kenmerken die kunnen leiden tot een aantal specifieke eisen voor het beschikbaar stellen. Zo is het van belang dat het tijdens het raadplegen van een website duidelijk wordt aangegeven wanneer een link naar 'buiten' leidt en niet meer de gearchiveerde website wordt bekeken.

Zoekfunctionaliteit

Voor de zoekfunctionaliteit zal voornamelijk worden gebruik gemaakt van het in ontwikkeling zijnde metadatamodel van het gemeentearchief. Specifiek voor websites is dat ook naar URL's of IP adres gezocht kan worden, waarbij een extra optie zou kunnen zijn dat een URL niet precies hoeft te worden aangegeven. De zoekfunctionaliteit van de Waybackmachine (The Internet Archive) is vrij beperkt omdat bij het zoeken naar een website de URL nauwkeurig moet worden opgegeven. Tevens kan worden onderzocht of het wenselijk is om te zoeken naar woorden in een website, een soort 'googelen' binnen de gearchiveerde websites.

Meestal worden van een website meerdere snapshots gemaakt omdat websites vrij dynamisch zijn en een internetpagina vaak wordt gewijzigd. Het is echter niet zo dat van elke wijziging een snapshot genaakt moet worden, daarvoor kan gebruikt worden gemaakt van de gebeurtenisgestuurde benadering. Het is van belang voor de zoekfunctionaliteit dat van een website de verschillende snapshots worden getoond die in de loop der tijd zijn gemaakt.

Presurf

Door Capsis is het programma Presurf ontwikkeld voor het beschikbaar stellen van websites. Het programma heeft een aantal functionaliteiten welke wellicht ook voor het e-depot interessant zijn. Met Presurf kunnen een aantal 'post-processing' handelingen worden verricht. Zo voegt HTtrack een commentaarregel toe aan het bestand, welke niet voldoet aan de W3C norm, via Presurf wordt deze regel weer verwijderd. Ook wordt door Presurf een index gegenereerd van de gearchiveerde website. Hoewel de manier van opslaan van de websites door Presurf voor het Gemeentearchief waarschijnlijk niet van belang zijn is het tonen van de verschillende snapshots van een websites door Presurf wel een gewenste functionaliteit.

Een sterk punt van Presurf is de viewerfunctionaliteit. Daarbij worden bij het opvragen van een website een aantal bewerkingen uitgevoerd om de website goed te bekijken zonder dat de eigenlijke bestanden worden aangepast. Zo kan bijvoorbeeld in de venstertitel worden vermeld dat het een archiefversie betreft. Ook kan een waarschuwing worden gegeven wanneer bij een link de archiefversie wordt verlaten en men direct het internet opgaat.

Het programma heeft een aan tal mogelijkheden welke het interessant voor het gemeentearchief kunnen maken. Het verdient dus zeker aanbeveling om te onderzoeken of het programma aan de eisen en wensen van het archief kan voldoen en of het in de applicatiestructuur van het e-depot is in te passen.

Auteursrecht

Het (her)publiceren van een website door het gemeentearchief kan auteursrechtelijke consequenties hebben, bijvoorbeeld wanneer deze kunstuitingen van derden bevat. Het project Archipol van het Documentatiecentrum van Nederlandse Politieke Partijen heeft dit probleem opgelost door aan iedere auteur van een website toestemming te vragen voor opname, afspraken te maken over het auteursrecht waar nodig en de toegang tot de websites te beperken tot belangstellenden uit onderwijs en onderzoek door middel van een username/password. De kwestie m.b.t. de auteursrechten betreft niet alleen websites, maar alle digitale objecten die het gemeentearchief beschikbaar wil stellen. Het deelproject Digitale Balie rekent ook dit punt tot haar opdracht. Vanuit het deelproject websites beveelt de werkgroep aan de aanpak van Archipol mee te nemen in de overwegingen. De website van Archipol bevat verwijzingen naar juridische literatuur over nieuwe media en auteursrecht.

Implementatie

Procesbeschrijving in hoofdlijnen

Het proces van website-archivering bestaat uit de volgende hoofd-stappen. Dit overzicht kan dienen als basis voor de werkprocesbeschrijving.

1. Selectie

2. Analyse doel en karakter van de website
3. Bepalen authenticiteitseisen
4. Vertalen van authenticiteitseisen naar snapshotinstellingen en -frequentie
5. Uitvoeren snapshot (zie instructie in bijlage 5)
6. Controle van het resultaat aan de hand van de authenticiteitseisen
7. Opnemen in e-depot quarantaineomgeving, controle op virussen, uitvoeren checksum
8. Toevoegen nieuwe metadata/extraheren ingebedde metadata in e-depot preserveringsomgeving
9. Preserveren
    a. Bij migratiestrategie: signaleren van 'bedreigde' bestandsformaten en migreren naar open standaardformaten.
    b. Bij emulatiestrategie: op peil houden bibliotheek van converters en viewers of hardware-emululatoren.
10. Opnemen in de e-depot bewaaromgeving
11. Beschikbaarstellen in e-depot raadpleegomgeving

Inpassing in organisatie

Bij de uitvoering van dit proces zijn verschillende medewerkers in verschillende rollen betrokken. De werkgroep stelt voor om de rollen als volgt te beleggen:

- Collectiebeheerders (Archieven en Relatiebeheer, Atlas en Bibliotheek)
  Taken: selectie, vaststellen en eventueel bijstellen van eisen t.a.v. integriteit en authenticiteit en controleren van het resultaat op grond van deze eisen, ontsluiting middels metadata
- E-conservator
  Taken: uitvoeren van harvesting, bouwen en onderhouden van tools t.b.v. routinematig harvesten van websites

Kennis

Gedurende de activiteiten van het deelproject hebben de leden grote vooruitgang geboekt in de verwerving van theoretische en praktische kennis ten aanzien van het archiveren van websites. Het is nu duidelijk hoe het archiveringsproces moet verlopen, wat daarbij komt kijken en op welke punten we nog tekort schieten. De tekortkomingen hebben vooral betrekking op de preservering en de specialistische kennis ten aanzien van het harvesten van websites. De training door Capsis heeft de deelnemers inzicht en een basisvaardigheden bijgebracht, maar heeft tevens aan het licht gebracht dat harvesten specialistische kennis vraagt, die in de organisatie belegd moet worden en geleidelijk aan opgebouwd moet worden.

Conclusie

Resultaten

Aan het eind van het deelproject kan worden gesteld dat, na vervulling van enkele randvoorwaarden, de verwerving en archivering van websites in de staande organisatie geïmplementeerd kan worden. Terugkomend op de opdracht van het deelproject zijn de volgende doelstellingen verwezenlijkt:

1. ***Keuze bewaardoelstelling en selectie van websites (Waarom archiveren?)***
   - De werkgroep definieert een website als een platform waarop informatie-objecten in samenhang worden aangeboden. De website zelf is dus niet het informatie-object.
   - De werkgroep heeft er voor gekozen in eerste instantie uit te gaan van het archiveren van websites als collectie-items. Het bewaren van websites als archiefbescheiden wordt doorgeschoven naar een vervolgproject.
   - Ten aanzien van selectie is de werkgroep van mening dat de huidige principes van selectie, zoals die gelden bij Archieven, Atlas en Bibliotheek ook van toepassing zijn op het verzamelen van websites. Plannen van internationale organisaties en de Koninklijke Bibliotheek voor het archiveren van complete webdomeinen maken lokale initiatieven zoals die van het gemeentearchief niet overbodig, omdat deze projecten (nu nog) minder hoge eisen stellen aan authenticiteit en toegankelijkheid. Het verdient wel aanbeveling één en ander af te stemmen met de Koninklijke Bibliotheek.

2. ***Formulering van kwaliteitseisen aan de authenticiteit en integriteit (context, vorm, structuur, inhoud en gedrag) van gearchiveerde websites (Wat archiveren?)***
   De werkgroep heeft een set van generieke eisen aan de authenticiteit van websites geformuleerd (zie bijlage 3). Per website kan hiermee, afhankelijk van de aard en functie van de website, worden bepaald welke elementen wezenlijk zijn en moeten worden gepreserveerd en welke als ‘accepted loss’ achterwege kunnen worden gelaten. De eisen kunnen per website worden aangepast.

3. ***Ontwikkeling van een archiveringsmethode waarmee de gestelde authenticiteitseisen kunnen worden gerealiseerd, die geïmplementeerd is in de werkprocessen van de betrokken afdelingen. (Hoe archiveren?)***
   a. Archiveringsproces en tools
   - Archiveringsrichtlijnen voor creatie van websites worden opgesteld door het ICTU. De implementatie hiervan bij de gemeente Rotterdam en particuliere instelling kan een onderwerp zijn voor een vervolgproject;
   - Voor de opname van websites is de snapshotmethode op dit moment de best haalbare aanpak, maar niet ideaal;
   - Het geteste pakket HTtrack heeft een aantal tekortkomingen, waarvan de belangrijkste kunnen worden ondervangen door het gebruik van de door Capsis ontwikkelde tool Presurf.

- Voor de opslag van websites zijn een aantal specifieke functionele eisen geformuleerd. Zo lang de e-depot software nog niet beschikbaar is, zijn een aantal alternatieven mogelijk. Het is moeite waard om na te gaan of Presurf aan de eisen kan voldoen.
- De problematiek van preservering is groter dan het gemeentearchief zelf kan oplossen. De werkgroep beveelt aan om het initiatief te nemen om hiervoor bondgenoten te zoeken;
- Voor websites hoeft geen geheel nieuw metadatamodel ontworpen te worden. De door de werkgroep als noodzakelijk aangegeven metadata kunnen worden geïntegreerd in het metadatamodel van het gemeentearchief.
- Voor de beschikbaarstelling zijn functionele eisen opgesteld ten aanzien van de software, die door het deelproject digitale balie meegenomen kunnen worden, alsmede voorstellen voor maatregelen ter bescherming van het auteursrecht.

b. Implementatie

- Middels dit handboek is een methodiek ontwikkeld voor de archivering van websites;
- De stappen in het proces van website-archivering zijn beschreven;
- Er is een instructie voor het harvesten opgesteld;
- Het verzamelen van websites zal zowel bij archieven, atlas en bibliotheek kunnen plaatsvinden en kan wat betreft de werkgroep het best worden uitgevoerd in wisselwerking tussen de collectiebeheerders en de e-conservator;
- Er is nog niet voldoende kennis in huis om websites zonder problemen te kunnen harvesten. Deze zal al doende moeten worden opgebouwd, met hulp van deskundigen.

Hoe verder?

Naast de bovengenoemde resultaten, zijn er ook een aantal zaken die de werkgroep nog niet opgepakt heeft, enerzijds omdat zij buiten de scope van het deelproject vielen, anderzijds omdat zij voortgekomen zijn uit nieuw opgedane inzichten. Deze vervolgacties hoeven niet noodzakelijk in projectvorm plaats te vinden. Zij kunnen ook in de lijn opgepakt worden.

1. Een thematische wervingscampagne van websites organiseren. Mogelijke thema's: migranten in Rotterdam, gemeentelijke projecten, wijkorganisaties, lokale politiek. Dit kan ook naar aanleiding van een evenement, zoals The Internet Archive websites heeft verzameld rond de orkaan Katrina, zo zouden in Rotterdam het Sportjaar of de gemeenteraadsverkiezingen van maart 2006 uitgekozen kunnen worden;
2. Opzet van een low-profile archiefomgeving voor websites, zodat de oogst van de wervingscampagne ook kan worden bewaard, ontsloten en beschikbaar gesteld. De eisen voor zo'n softwareomgeving zijn vastgelegd in dit handboek. Presurf van Capsis is een interessante optie;
3. Opbouwen van expertise over websiteharvesting. Deze taak beleggen bij de beoogde e-conservator;

4. Instrueren van collectiebeheerders van het gemeentearchief in basisprincipes van website-archivering. Niet om zelf te gaan harvesten, maar om te kunnen begrijpen wat de mogelijkheden en (on)mogelijkheden zijn. Dit zou door de leden van de werkgroep uitgevoerd kunnen worden, bij wijze van interne kennisoverdracht, met behulp van dit handboek;
5. Voorlichting over de Webrichtlijnen voor archivering van Advies.Overheid.nl, zo nodig in samenwerking met bijvoorbeeld CMS-diensten Rotterdam een Leidraad opstellen voor de gemeentediensten en deelgemeenten. Het externe lid van de werkgroep, Harm Janssen van de gemeente Apeldoorn, is op dit moment bezig met vooronderzoek om te komen tot richtlijnen voor zijn gemeente, die ook toegepast zouden kunnen worden voor Rotterdam;
6. Een discussiebijeenkomst beleggen over websitepreserving met andere belanghebbenden in Nederland, met als doel strategische bondgenoten te vinden en tot een taakverdeling te komen;

Literatuurlijst

Filip Boudrez
Archiveren van websites: een kwestie van waardering en 'capture'. Stadsarchief Antwerpen. Antwerpen, 2005
http://www.antwerpen.be//david

Filip Boudrez en Sofie Van den Eynde
DAVID. Archiveren van websites. Versie 1.0 (mei 2005)
http://www.antwerpen.be//david

Andrew Boyko
Test Bed Taxonomy for Crawler. Version 1.0. Netpreserve.org international internet preservation consortium.
www.netpreserve.org.

Adrian Brown
Archiving Web Resources. The National Archives, 24 nov 2004
http://www.homeoffice.gov.uk

Niels Brügger
Archiving websites. General considerations and stategies. The Centre for Internet Research. Aarhus, 2005.
CFI - Archiving websites.htm
http://cfi.imv.au.dk/eng/pub/webarc/

Capsis
Specificaties Presurf v. 1.0. Dé oplossing voor het duurzaam archiveren en gecontroleerd beschikbaar stellen van websites (29-07-2004).
http://www.capsis.nl

Lars C. Clausen
Handling file formats. Aarhus, Kopenhagen, The State and University Library / The Royal Library, 2004.
http://netarchive.dk/fase2index-en.php

Niels H. Christensen
Towards format repositories for web archives. Dept. of Documentation & Digitalisation, Royal Library of Denmark, Copenhagen. 4e International Web archiving Workshop 2004
http://netarchive.dk/fase2index-en.php

H.A. Hokke
Blijvend beschikbaar: onderzoek naar de archivering van websites. Eindrapport. Amsterdam, 2003.

H.A. Hokke
Naar archivering van websites: implementatieadvies bij het onderzoeksrapport "Blijvend Beschikbaar" Amsterdam, 2003

Jennifer Marill, Andrew Boyko, Michael Ashenfelder
Web Harvesting Survey. Version 1. IIPC
www.netpreserve.org.

Gordon Mohr, Michael Stack, Igor Ranitovic, Dan Avery and Michele Kimpton
An Introduction to Heritrix. An open source archival quality web crawler. (14 juli 2004)
http://crawler.archive.org/An%20Introduction%20to%20Heritrix.pdf

René Voorburg en Hans Goutier

Naar archivering van websites. In: Archievenblad, mei 2004.

Testbed Digitale Bewaring
Van digitale vluchtigheid naar digitaal houvast dl. 1-4 (Den Haag, 2003)
http://www.digitaleduurzaamheid.nl

R.J.J. Voorburg en J.L.E. Goutier
Webarchivering bij het Ministerie van Verkeer & Waterstaat. Verslag van een onderzoek (17-06-2005 / definitief). Capsis BV / Ministerie van Verkeer en Waterstaat
http://www.capsis.nl/files/verslag_onderzoek_webarchivering_minvenw_capsis_publicatie_20050822.pdf

Webmetadata
Webmetadata. In: Het verbeteren van de toegankelijkheid van digitale informatie binnen de Nederlandse overheid. Handboek. Versie 1.01 (16 december 2004). Den Haag, 2004.
http://www.advies.overheid.nl/

Websites

http://www.advies.overheid.nl/

http://www.antwerpen.be//david

http://www.archive.org/web/web.php

http://www.archiefschool.nl/onderzoek/projecten.htm#web

http://www.capsis.nl

http://crawler.archive.org/

http://cfi.imv.au.dk/eng/pub/webarc/

http://www.homeoffice.gov.uk

http://netarchive.dk/fase2index-en.php

http://www.netpreserve.org

http://www.nla.gov.au/padi/topics/92.html

http://www.iwaw.net/05/index.html

Bijlagen

Bijlage 1: Testset websites

URL	Doel(en)	Organisatie(s)	Plaats	Communicatie	Transacties	Audiovisuele media	Databases
	Bedrijven / commerciële instellingen						
www.rotterdamsdagblad.nl	voorlichting, reclame, communicatie t.b.v. product + organisatie	Rottterdams Dagblad	Rotterdam	Email	advertenties opgegeven		krantenartikelen
www.woonnet-rijnmond.nl	voorlichting en mogelijkheid tot transactie t.b.v. product/dienst	woningcorporaties	Rijnmondgebied	Email	reageren op woningenaanbod	nee	beschikbare woningen
www.smit-tak.com	voorlichting, reclame, communicatie t.b.v. product + organisatie	Smit-Tak		Email	aankoop video's		winkelproducten
www.pietjebell.nl	voorlichting, reclame, communicatie t.b.v. product	Shooting Star BV	Amsterdam	Email	nee	ja	
http://www.rotterdam-airport.nl/	luchthaven	Rotterdam Airport	Rotterdam	e-mail in website	ja	nee	nee
http://www.010publishers.nl/	uitgeverij	010 uitgevers	Rotterdam	e-mai in website	ja	nee	ja
http://www.rijnmond.nl/	regionale omroep	RTV Rijnmond	Rotterdam	e-mail		webcam, video, dia's	
	GemeenteRotterdam / deelgemeente, diensten etc.						
www.gemeentebelastingen.rotterdam.nl	gemeentelijke belastingdienst	Gemeentebelastingen Rotterdam	Rotterdam				
www.ggd.rotterdam.nl	voorlichting, communicatie t.b.v. organisatie en volksgezondheid	GGD	Rotterdam	Email	nee		overzicht medicijnen
www.rhrr.nl	Brandweer, rampenbestrijding	Regionale Hulpverleningsdienst Rotterdam Rijnmond	Rijnmondgebied				
http://www.wimby.nl/	promotie voor Hoogvliet	Stichting Wimby	Rotterdam	Email	nee	nee	nee
http://www.woneninrotterdam.nl	promotie voor Rotterdam als woonstad, zoeken van woningen	Ontwikkelingsbedrijf Gemeente Rotterdam	Rotterdam	Email	nee	dia's in intro	woningaanbod
http://www.portofrotterdam.com/NL/	Informatie en promotie voor de haven van Rotterdam	Havenbedrijf Rotterdam	Rotterdam	E-mail/vragenformulier	nee	ja (o.a. interactieve havenkaart)	scheepsbewegingen (koppelingen met dirkzwager)
http://www.deelgemeenten.rotterdam.nl/	portal naar de websites van de verschillende deelgemeenten (hier ljsselmonde)	Gemeente Rotterdam	Rotterdam	e-mail	ja, met DigiD m mogelijk zaken aan te vragen	nee	nee
http://www.aktiegroepoudewesten.nl/	Wijkgebonden aktiegroep	Aktiegroep Het Oude Westen	Rotterdam	e-mail in website		nee	
http://www.norastorm.nl/	daklozen- en verslaafdenzorg	Stichting De Stormvogel	Rotterdam	e-mail in website		nee	nee
http://www.gemeentearchief.rotterdam.nl/	archiefdienst	Gemeentearchief Rotterdam	Rotterdam	e-mail in website	ja	nee	ja
http://www.leefbaarrotterdam.nl/	politieke partij	Leefbaar Rotterdam	Rotterdam	e-amil in website	nee	ja	
	Cultuur / musea						
http://www.dedoelen.nl/	concert- en congresgebouw	Concert- en congresgebouw De Doelen	Rotterdam	e-mail		nee	
http://www.boijmans.nl/	voorlichting, reclame m.b.t. de organisatie	Museum Bojimans van Beuningen	Rotterdam	Email	nee	nee	nee
http://www.rotterdamdooft.nl/	aktiegroep tegen teloorgang cultuur in Rotterdam	Rotterdam dooft	Rotterdam	e-mail lin de website		nee	nee
http://www.rotterdamsemuseumnacht.nl/	promotie voor eenmalig evenement	Musea Rotterdam	Rotterdam	Email	nee	nee	nee
	Onderwijs						
www.eur.nl	voorlichting, reclame, communicatie t.b.v. organisatie(s)	Erasmus Universiteit + andere organisaties	Rotterdam	Email	reserveren/verlengen boeken		bibliotheekcatalogus
http://www.hogeschool-rotterdam.nl/	voorlichting, reclame, communicatie t.b.v. organisatie(s)	Hogeschool Rotterdam	Rotterdam	e-mail in website		nee	
http://www.avbr.nl/	voorlichting, reclame, communicatie t.b.v. organisatie(s)	Academie van bouwkunst	Rotterdam	e-mail in website		nee	
	Diversen			-			
http://welcome.to/Rotjeknor	Privé website voor Rotterdam fans	A. van der Struijs, Rotterdam	Rotterdam	gastenboek	nee	webcam, geluid	nee
http://www.trompenburg.nl/	Bomenpark	Stichting Arboretum Trompenburg	Rotterdam	e-mail	nee		
http://www.stoomschiprotterdam.nl/	Informatie over de s.s. Rotterdam ivm met mogelijke terrugkeer naar r'dam van het schip	Stichting behoud stoomschip Rotterdam	Heemstede	e-mail (in de webpagina)		nee	nee
http://www.rotterdamzoo.nl/	diergaarde	Diergaarde Blijdorp	Rotterdam	e-mail		webcam	
http://www.plaswijckpark.nl/	pretpark	Plaswijckpark	Rotterdam	e-mail in website	nee	nee	
http://www.vvvrotterdam.nl/	toeristische voorlichting	VVV Rotterdam	Rotterdam	e-mail in website	nee	nee	
http://www.feyenoord.nl	voetbalclub	Sportclub Feyenoord	Rotterdam	e-mail in website	ja	ja (video + webcam)	
http://www.erasmusmc.nl/	ziekenhuis	Erasmus Medisch Centrum	Rotterdam	e-mail in website	nee	nee	
	Regio			-			
www.schielandendekrimpenerwaard.nl	voorlichting communicatie t.b.v. organisatie	Hoogheemraadschap van Schieland etc.	Rotterdam	Email	nee	nee	
http://www.ridderkerk.nl	voorlichting, communicatie	Gemeente Ridderkerk	Ridderkerk	e-mail in website	ja Digitaal loket	nee	nee
http://www.berkelenrodenrijs.nl/	voorlichting communicatie t.b.v. organisatie	Gemeente Berkel en Rodenrijs	Berkel en Rodenrijs	e-mail in website	nee	nee	
				-			
http://weblog.r-win.com/iffr	weblog over internationaal filmfestival Rotterdam	Erwin?	?	e-mail, icq, msn	reactie op log	nee	nee

Bijlage 2: Metadataschema

Project E-depot / Gemeentearchief Rotterdam
Deelproject Websites : metadata websites GAR.2
Versie: 1.1 21-6-2006
Status: Concept
Eigenaar: Werkgroep websites
Doel: Samenstelling metadata voor het toegankelijk maken van websites in het Gemeentearchief Rotterdam.

Door de werkgroep websites zijn op basis van het Handboek webmetadata versie 1.01 de daarin voorgestelde metadata velden beoordeeld op bruikbaarheid en relevantie voor het GAR. De ISBD (ER) regels zijn hierin goed terug te vinden. Tevens loopt dit model goed in de pas met het Dublin Core model
De vraag of de interpunctie van de ISBD(ER) regels van toepassing zijn op de metadatasets moet nog worden beantwoord.

Een aantal velden die door ISBD (ER) worden ondergebracht in het “annotatieveld” komen wellicht in aanmerking voor “een eigen veld”.

Veldnaam	Opmerking / Relatie
Metadata in Q / P	
Aanwinstnummer	Wordt automatisch gegenereerd
Datum binnenkomst	Datumveld
Herkomst	Vrij tekstveld
Collectie	Archieven – Bibliotheek – THA (keuzeveld)
Subcollectie	Archieven (nummer en naam archief) THA (deelcollectie, beeld, geluid, portretten etc.)
Rechtstitel	Aankoop – Schenking – Bruikleen - Download
Zorgdrager	Vrij veld
Aantal files en omvang	Wordt gegenereerd door het systeem
Velden volgens Dublin Core	
01.Titel	Vrij tekstveld
02. Auteur / maker	Vrij tekstveld
03. Onderwerp / Trefwoorden	
04. Omschrijving	Vergelijkbare ISBD term is: annotatie
05. Uitgever	
06. Andere medewerkers	In ISBD opgenomen in auteursveld. Zie: 02)
07. Datum / jaar van uitgave	
08. Bestandstype	ISBD term is: Algemene materiaalaanduiding
09. Formaat / format	
10. Bestandsidentificatie	URL
11. Bron	ISBD norm: annotatieveld ??
12. Taal	
13. Relatie	ISBD norm: annotatieveld ??
14. Dekking	Niet van toepassing ??
15. Copyright.	Ook in Q ??
… . Annotatieveld	ISBD norm

BEWARING (Q- en P- server)

Definitie	Betreft vooral de Q en P periode.
Verplichting	Administratief
Doel / motivatie	
Opmerkingen	Bewaring zal voornamelijk worden gebruikt door records managers en andere betrokkenen bij de langdurige opslag van officiële documenten. Dit element zal worden gebruikt om aspecten van de herkomst van de bron te bewaren bij overdracht van het beheer (zorgplicht) van een archiefvormer naar het GAR. Een deel van deze informatie zal later mogelijk worden opgenomen in een archiefbeschrijving of documentatie over het beheer van de documenten.

STATUS (annotatieveld??)

Definitie	De toestand of status van de bron.
Verplichting	Aanbevolen
Doel / motivatie	Maakt het de gebruiker mogelijk te zoeken naar een document aan de hand van de status daarvan. Kan ook worden gebruikt als referentie door een gebruiker die de status van de bron wil weten.
Opmerkingen	De status van een document omvat onder andere: De mate waarin het ontwikkeld of afgerond is, b.v. eerste concept, laatste document, afgerond document. Is het in afwachting van goedkeuring? Als het goedgekeurd is, door wie? Versienummer
Voorbeelden	Voor een serie documenten geschreven voor de ontwikkeling van een beleidsstandpunt. status: Concept versie 0.4 Voor openbare consultatie status: Versie 1.0 Voor publicatie

01. TITEL

Definitie	Naam van de website.
Verplichting	Verplicht
Doel / motivatie	Maakt het de gebruiker mogelijk een document met een bepaalde titel te vinden. De titel wordt veelal gebruikt als belangrijkste referentiepunt in de lijst van zoekresultaten. In voorkomende gevallen kan een ondertitel worden toegevoegd Als de titel ook beschikbaar is in een andere taal kan deze worden toegevoegd als paralleltitel.
Opmerkingen	In het algemeen is de titel de naam waaronder de site (in)formeel bekend staat. Als de formele titel van een document onbegrijpelijk is dan wordt aanbevolen een aanvullende, duidelijke naam te geven aan de bron.
Verfijningen	Elke vorm van de titel die als vervanging of alternatief voor de formele titel van de bron wordt gebruikt. Als alternatieve titel kan elke vorm van de titel die wordt gebruikt als alternatief worden toegevoegd, zoals de naam waaronder de site algemeen bekend is, afkortingen en vertalingen. De alternatieve titel kan ook worden opgenomen in het annotatieveld

02. AUTEUR / MAKER

Definitie	Eenheid met de verantwoordelijkheid voor de creatie van de inhoud van de site
Verplichting	Verplicht indien van toepassing
Doel / motivatie	Maakt het de gebruiker mogelijk documenten te vinden die geschreven of anderszins zijn gecreëerd door een bepaalde persoon of organisatie.
Opmerkingen	Alle personen en organisaties die een belangrijke rol hebben bij het samenstellen van de website, worden opgenomen in het auteursveld. Op al deze personen/zaken wordt een zoekingang gemaakt. Als auteur kunnen worden genoemd: een persoon, organisatie of dienst (corporatie). Alle personen en organisaties (subsequent statements) op die een belangrijke rol speelden bij het samenstellen van de inhoud van het document, maar die niet als auteur/maker (statement of responsebility) worden beschouwd, worden opgenomen in het auteursveld. Op al deze personen/zaken wordt een zoekingang gemaakt
Niet te verwarren met	Uitgever: de auteur/maker is degene verantwoordelijk voor de intellectuele of creatieve inhoud van het document, de uitgever is de persoon of organisatie die het document beschikbaar stelt. In een aantal gevallen zullen de auteur en de uitgever identiek zijn. Een co-auteur speelt wel een belangrijke rol maar heeft geen primaire of algemene verantwoordelijkheid voor de inhoud.

03. ONDERWERP / TREFWOORDEN

Definitie	Onderwerp van de inhoud van de site
Verplichting	Verplicht
Doel / motivatie	Maakt het de gebruiker mogelijk te zoeken op het onderwerp van de bron.
Opmerkingen	De aanbevolen “best practice” is het kiezen van een trefwoord uit een gecontroleerde woordenlijst, i.c. de trefwoordenlijst van het GAR. De waarden voor alle onderwerp-verfijningen moeten worden gekozen uit: de trefwoordenlijst van het GAR, en evt. daar aan toe te voegen gecontroleerde woordenlijsten, thesauri etc.
Niet te verwarren met	Trefwoorden verwijzen naar het onderwerp van de website, niet wat het is. Voorbeeld: gebruik “kaart/map” niet als subject (onderwerp) term indien de bron een landkaart is, maar neem “kaart/map” op in het type element. Gebruik “kaarten” als subject (onderwerp) term indien de bron gaat over kaarten, kaarten maken, cartografie, enz.
Verfijningen	Ttrefwoord: woorden of termen gebruikt voor het zo nauw mogelijk beschrijven van het onderwerp van de bron. Dienen gekozen te worden uit een gecontroleerd vocabulaire of lijst Persoon: dient gebruikt te worden als de bron over een persoon gaat. NB: niet te verwarren met auteur/maker

04. OMSCHRIJVING (zie Annotatie)

Definitie	Nadere omschrijving van de inhoud van de bron.
Verplichting	Optioneel
Doel / motivatie	Helpt de gebruiker te bepalen of de bron aan de behoefte voldoet.
Opmerkingen	Niet-limitatieve voorbeelden van de omschrijving: samenvatting, inhoudsopgave, verwijzing naar een grafische voorstelling van de inhoud, of een vrije tekst.
Verfijningen	abstract / samenvatting van de inhoud van de bron.
Inhoudsopgave	Table of contents (inhoudsopgave): Lijst van onderdelen van de inhoud van de bron.

Voorbeelden	Omschrijving, inhoudsopgave : Documentgeschiedenis, inleiding, achtergrond, lijst van onderdelen, algemene principes, onderdelen

05. UITGEVER

Definitie	Eenheid verantwoordelijk voor het beschikbaar maken van de bron.
Verplichting	Verplicht indien van toepassing
Doel / motivatie	Maakt het mogelijk een bron te vinden die is uitgegeven door een bepaalde organisatie of persoon.
Opmerkingen	Voorbeelden van uitgever zijn: persoon, organisatie, dienst. In het algemeen dient de naam van de uitgever te verwijzen naar de eenheid. Uitgever) wordt hier gebruikt in de breedste zin, zodat een organisatie die een informatiebron op een website plaatst de uitgever is, zelfs als de bron niet in gedrukte vorm beschikbaar is. De uitgever is de persoon of organisatie met welke een gebruiker contact dient op te nemen voor toestemming voor het opnieuw uitgeven van de informatie in de bron, of het verkrijgen van exemplaren in een ander formaat.
Niet te verwarren met	auteur/maker De uitgever is de organisatie of persoon die de bron beschikbaar maakt voor het publiek door het plaatsen van de bron op een website. De auteur/maker is verantwoordelijk voor de inhoud van de bron. In een aantal gevallen zullen de uitgever en de auteur/maker identiek zijn.

06. ANDERE MEDEWERKERS

	Zie veld: 01

07. DATUM / JAAR VAN UITGAVE

Definitie	Datum / jaar van uitgave .
Verplicht	Verplicht
Doel / motivatie	Maakt het de gebruiker mogelijk een document te vinden door de zoekactie te beperken tot een bepaalde datum, b.v. de datum waarop een bron beschikbaar werd gesteld.
Opmerkingen	In de meeste gevallen verwijst datum naar de creatie of beschikbaarstelling van de bron.
Verfijningen	Beschikbaar: datum of periode waarin de site beschikbaar komt of kwam. Created (creatie): creatiedatum van de bron Issued (uitgegeven): Datum waarop de bron formeel werd uitgegeven Modified (gewijzigd): Datum waarop de bron werd gewijzigd Next version due (volgende versie verwacht):Datum waarop de bron zal worden vervangen. Updating frequency (hoe vaak wordt de bron bijgewerkt): Hoe vaak de bron wordt bijgewerkt Valid (geldigheid): Datum/periode waarop/gedurende welke de bron geldig is. NB: ter bespreking in deelprojectgroep: onderverdeling maken in het veld 'jaar van uitgave' of deze gegevens opnemen in veld annotatie
Voorbeelden	datum.creatie) : 2003-03-20 datum.beschikbaar : 2003-03-30 datum.uitgegeven : 2003-04-10 datum.geldig) : 2003-04-10/2003-05-30 Voor een home page die op 6 januari 2000 bereikbaar werd datum.uitgegeven: 2000-01-06 Dezelfde home page in mei, na wijziging datum.uitgegeven: 2000-01-06 datum.gewijzigd: 2000-05-01

08. ALGEMENE MATERIAALAANDUIDING / TYPE

Definitie	Aard of soort van de inhoud van het document
Verplichting	Verplicht indien van toepassing
Doel / motivatie	Maakt het de gebruiker mogelijk te selecteren op soort document: boek, tijdschrift, website etc.
Opmerkingen	De aanbevolen “best practice” is het kiezen van een waarde uit een gecontroleerde woordenlijst.

09. FORMAAT

Definitie	Fysieke of digitale kenmerken van de website
Verplichting	Optioneel
Doel / motivatie	Maakt het de gebruiker mogelijk te zoeken naar bronnen in een bepaald format.
Opmerkingen	In de meeste gevallen zal format de digitale afmetingen van het document aangeven.
Verfijningen	Omvang: grootte of tijdsduur van de bron. Medium: Materiaal of fysieke drager van de bron
Voorbeelden	Voor een webpagina in HTML: Format (format): Text/html

10. VERWIJZING (URL)

Definitie	Eenduidige verwijzing naar de bron binnen een bepaalde context.
Verplichting	Verplicht indien van toepassing
Doel / motivatie	Maakt het een gebruiker mogelijk te zoeken naar een specifieke bron of versie.
Opmerkingen	De aanbevolen “best practice” is identificeren van de bron door middel van een cijfer- of karakterreeks op basis van een formeel identificatiesysteem. Voorbeelden van dergelijke systemen zijn de Uniform Resource Identifier (URI), waaronder de Uniform Resource Locator (URL), Digital Object Identifier (DOI) en het International Standard Book Number (ISBN).
Voorbeelden	Verwijzing (identifier): [URI] http://www.advies.overheid.nl /contactus/contact.asp

11. BRON (ISBD = Annotatieveld?)

Definitie	Verwijzing naar een bron waarvan de huidige bron van is afgeleid.
Verplichting	Optioneel
Doel / motivatie	Maakt het de gebruiker mogelijk bronnen te vinden die ontwikkeld zijn met behulp van de inhoud van een bepaald ander document (b.v. alle documenten gebaseerd op een bepaalde verzameling statistieken).
Opmerkingen	De huidige bron kan geheel of gedeeltelijk zijn afgeleid van het brondocument. De aanbevolen “best practice” is identificeren van het document d.m.v. een cijfer- of karakterreeks op basis van een formeel identificatiesysteem.
Verfijningen	Onderverdeling maken (op basis van ISBD-ER??
Voorbeelden	Voor een rapport gebaseerd op cijfers verzameld gedurende een onderzoek source (Bron): Cijfers afgeleid van de studie van het Nationaal Archief door Het Convent van Archivarissen 1998 http://www.nationaalarchief.nl/onderzoek/2001/jan/03.html

12. TAAL

Definitie	Taal van de intellectuele inhoud van de bron.
Verplichting	Verplicht indien van toepassing
Doel / motivatie	Maakt het gebruikers mogelijk hun zoekactie te beperken tot documenten in een bepaalde taal.
Opmerkingen	Het gebruik van taalcodes vereenvoudigt het invoeren van het language-element. De meeste gebruikers zullen de betreffende codes snel leren. De meeste systemen kunnen ingesteld worden op de volledige naam van de taal, dit vergroot de gebruikersvriendelijkheid. Het gebruik van het language-element is vooral van belang voor documenten die op het Internet beschikbaar zijn.
Voorbeelden	Voor een document geschreven in het Engels language (taal) Voor een Poolse vertaling van een oorspronkelijk in het Engels geschreven bron (gebruik relatie om een koppeling te maken met de oorspronkelijke Engelse versie) language (taal): [ISO 639-1] pl

13. RELATIE (ISBD = Annotatieveld)

Definitie	Verwijzing naar een gerelateerd document.
Verplichting	Optioneel
Doel / motivatie	Maakt het een gebruiker mogelijk andere documenten te vinden die in verband staan met dit document, of om afzonderlijke documenten te combineren tot een verzameling.
Opmerkingen	Kan door een “gewone” verwijzing c.q. kruisverwijzing in het annotatieveld
Verfijningen	Verwijzing naar een bestaande standaard waaraan de bron voldoet. Het beschreven document bestond eerder dan het verwijsdocument, het is in principe dezelfde intellectuele inhoud in een ander formaat Het beschreven document heeft een versie-editie of -aanpassing: het verwijsdocument Het beschreven document omvat het verwijsdocument fysiek of logisch Het beschreven document heeft dezelfde intellectuele inhoud als het verwijsdocument, maar in een ander formaat Het beschreven document wordt aangehaald, geciteerd of anderszins naar toe verwezen door het verwijsdocument Het beschreven document is vervangen door het verwijsdocument Het beschreven document is vereist door het verwijsdocument om daarvan de functie, levering of samenhangendheid van de inhoud te ondersteunen Het beschreven document is een versie-editie of -aanpassing van het verwijsdocument. Een wijziging in versie geeft een significante wijziging van de inhoud aan, niet slechts een verschil in formaat. Het beschreven document haalt het verwijsdocument aan, of citeert het of verwijst er naar Het beschreven document vereist het verwijsdocument ter ondersteuning van de functie, levering of samenhangendheid van de inhoud Het beschreven document vervangt het verwijsdocument.
Voorbeelden	Voor een website welke een eerdere website met vergelijkbare inhoud vervangt Vervangt: www.ministerieVenW.nl Voor een document dat nummer 7 vormt in de serie ‘Informatie Management’ : Is onderdeel van: Informatie management reeks, nr 7 Voor een HTML document dat oorspronkelijk gedrukt was: Is format van: ISBN 0711504237

14. DEKKING

	Niet van toepassing ??

15. COPYRIGHT / RECHTEN

Definitie	Informatie over rechten in en over de bron.
Verplichting	Administratief
Doel / motivatie	Geeft aan wie gerechtigd is de bron geheel of gedeeltelijk in te zien, te kopiëren, opnieuw te distribueren, opnieuw uit te geven of anderszins te gebruiken.
Opmerkingen	In het algemeen bevat een rechten-element een verklaring betreffende het rechtenbeheer van een document, of een verwijzing naar een dienst welke dergelijke informatie levert. Rechten-informatie omvat vaak intellectuele eigendomsrechten, auteursrechten en diverse andere eigendomsrechten. Indien het rechten-element ontbreekt mag men geen aannames maken betreffende de status van deze en andere rechten met betrekking op de bron.
Verfijningen	Copyright: Informatie en identificatie met betrekking tot het juridische eigendom en rechten betreffende (her)gebruik van het gehele of gedeeltelijke document.
Voorbeelden	Voor een document waarvan het auteursrecht berust bij RAND-Europe: “Copyright RAND-Europe”

Comparative outline of the ISBD (ER)

1. Title and statement of responsibility area		1.1 Title proper
	[]	1.2 General material designation **(optional)**
	=	*1.3 Parallel title
	:	*1.4 Other title information
	/ ;	1.5 Statements of responsibility First statement * Subsequent statement
2. Edition area		2.1 Edition statement
	=	*2.2 Parallel edition statement **(optional)**
	/ ;	2.3 Statements of responsibility relating to the edition First statement * Subsequent statement
	,	*2.4 Additional edition statement
2. Edition area	/ ;	2.5 Statements of responsibility following an additional edition statement First statement * Subsequent statement
3. Type and extent of resource		3.1 Designation of resource
	()	3.2 Extent of resource **(optional)**
4. Publication, distribution, etc.	;	4.1 Place of publication, production and/or distribution, etc. First place * Subsequent place
	:	*4.2 Name of publisher, producer and/or distributor, etc.
	[]	*4.3 Statement of function of distributor **(optional)**
	,	4.4 Date of publication, production and/or distribution, etc.
	(	*4.5 Place of manufacture **(optional)**
	:	*4.6 Name of manufacturer **(optional)**
	,)	4.7 Date of manufacture **(optional)**
5. Physical description area		5.1 Specific material designation and extent of item
		5.2 Other physical details
	:	5.3 Dimensions
	; +	*5.4 Accompanying material statement **(optional)**
6. Series area		6.1 Title proper of series or sub-series
Note: A series statement is enclosed	=	*6.2 Parallel title of series or sub-series
by parentheses. When there are two or more series statements, each	:	*6.3 Other title information **(optional)**

is enclosed by parentheses.	/	6.4 Statements of responsibility relating to the series or sub-series First statement * Subsequent statement
	;	6.5 International Standard Serial Number of series or sub-series
	;	6.6 Numbering within series or sub-series
7. Note area		
8. Standard number (or alternative) and terms of availability area		8.1 Standard number **(or alternative)**
		8.2 Key title
	=	8.3 Terms of availability and/or price **(optional)**

General notes on the outline of the ISBD(ER)

A.	Optional elements are indicated as such (see 0.1.3).
B.	Elements preceded by an asterisk can be repeated when necessary.
C.	Areas 6 (Series), 7 (Note) and 8 (Standard number, etc.) can be repeated when necessary. In addition, area 5 (Physical description) can be repeated under certain circumstances (see area 5, Introductory note).
D.	In the above outline, the terms "first statement ...", "subsequent statement ..." and the like denote the order in which these statements are given and have no other connotation.
E.	No provisions are included in the ISBD(ER) for element 8.2 of the ISBD(G) outline (Key title). Provisions regarding qualifications to standard number (or alternative) or to a statement of terms of availability and/or price (element 8.4 of the ISBD(G)) are included in elements 8.1 and 8.3 respectively, rather than as a separate element.
F.	Whenever information normally associated with one area or element appears in the item linked linguistically as an integral part of another area or element, it is transcribed as such.

DUBLIN CORE Simple Format

Dit document beschrijft welke Dublin Core metadata elementen er zijn en geeft een korte definitie van elk element. Voor uitgebreide uitleg en gebruik van de Dublin Core elementen wordt u verwezen naar de gebruikershandleiding.
Deze vertaling geeft ook de keuzes aan voor standaarden (taalcodes, datum codes) die in DONOR verband worden gehanteerd. In zoverre kan deze vertaling afwijken van het oorspronkelijke document "Description of Dublin Core Elements" http://purl.org/metadata/dublin_core_elements"

Inhoud
Intellectueel eigendom
Fysieke weergave

Title
Creator
Date

Subject
Publisher
Type

Description
Contributor
Format

Source
Rights
Identifier

Language

Relation

Coverage

1. Titel Label:Title
De naam van de Internetbron. Meestal gegeven door de auteur, maker of uitgever

2. Auteur of maker Label: Creator
De auteur, maker of organisatie die primair verantwoordelijk is voor de intellectuele inhoud van het werk. Bijvoorbeeld auteurs in geval van geschreven documenten, artiesten, fotografen of illustrators in geval van visuele bronnen.

3. Onderwerp en trefwoorden Label: Subject
Het onderwerp van de Internetbron. Het onderwerp wordt beschreven in korte zinnen of door het gebruik van trefwoorden. Het gebruik van gecontroleerde trefwoorden of classificatie schema's wordt aanbevolen.

4. Omschrijving Label: Description
Een omschrijving van de inhoud van de Internetbron. Abstracts of inhoudsomschrijvingen.

5. Uitgever Label: Publisher
De entiteit die verantwoordelijk is voor het beschikbaarstellen van de Internetbron in de huidig vorm. Bijvoorbeeld een uitgeversbedrijf, een universiteitsafdeling.

6. Andere medewerkers Label: Contributor
Een persoon of organisatie die een belangrijke bijdrage heeft geleverd, maar die secondair is aan de persoon of organisatie die bij "Creator" genoemd wordt. Gedacht kan worden aan editors, vertalers en illustrators.

7. Datum Label: Date
De datum van de totstandkoming of beschikbaarstelling van een Internetbron. Deze datum moet niet verward worden met de datum in "Coverage". De datum in "Coverage" geeft de tijdsperiode van de inhoud van de Internetbron aan. Voorgeschreven wordt de ISO 8601 standaard voor datum, zoals beschreven in W3C Technical Note http://www.w3.org/TR/NOTE-datetime. Hierin wordt het formaat YYYY-MM-DD beschreven. Het formaat YYYY-MM-DD wordt de voorgeschreven standaard. 3 September 1998 wordt in dit geval 1998-09-03.

8. Bestands type Label: Type
Het soort Internetbron, bijvoorbeeld home page, gedicht, working paper, technical report, essay, woordenboek etc. De DC working groups zijn momenteel bezig met twee voorstellen; een minimalist draft: http://sunsite.berkeley.edu/Metadata/minimalist.html en een structuralist draft: http://sunsite.berkely.edu/Metadata/structuralist.html. Het wordt sterk aanbevolen om "Type" van deze lijsten te selecteren.

9. Format Label: Format
Het data formaat van het Internetbestand. Dit gegeven kan gebruikt worden om systeemvereisten (hard- en software) te identificeren die nodig zijn voor raadpleging of bediening van de Internetbron. "Format" is nog in ontwikkeling bij de DC workshop. Uit compatibiliteitsoverwegingen wordt aanbevolen formaatgegevens te selecteren uit een gecontroleerde lijst. Formaten kunnen van de Internet Media Types (MIME types) lijst gekozen worden: http://www.isi.edu/in-notes/iana/assignments/media-types/media-types

10. Bestandsidentificatie Label: Identifier
Een unieke string of nummer ter identificatie van de Internetbron. Voorbeelden zijn URL's of URN's Andere voorbeelden zijn International Standard Book Number (ISBN) of International Standard Serial Number (ISSN)

11. Bron Label: Source
Informatie over een tweede bron waar de huidige Internetbron is van afgeleid. Bij "Source" mag een datum, auteur of maker, formaat, identificatie of andere metadata voor de tweede bron worden ingevuld. Voorbeeld: het is mogelijk om een "Source" datum 1603 in een omschrijving van een 1996 film bewerking van een Shakespeare stuk te gebruiken. "Source" is niet bruikbaar wanneer de huidige bron de orginele is.

12. Taal Label: Language
De taal waarin het document is beschreven. Voorgeschreven wordt de ISO639-1 norm, een twee karakter taalcode van de ISO 639(639-1 en 639-2) norm. Taalcodes kunnen van een lijst worden gekozen: http://www.ics.uci.edu/pub/ietf/http/related/iso639.txt, Eventueel kan ISO3166 (landencodes) gebruikt worden om aan te geven in welk gebied de taal wordt gesproken. Deze kunnen ook van een lijst worden gekozen: http://www.ics.uci.edu/pub/ietf/http/related/iso3166.txt

13. Relatie Label: Relation
Een identificatie van een tweede bron en de relatie met de huide Internetbron. Voorbeelden zijn een editie van een werk (IsVersionOf), een vertaling van een werk (IsBasedOn), een hoofdstuk van een boek (IsPartOf) en een mechanische transformatie van een dataset in een plaatje (IsFormatOf). De DC Working Groups zijn bezig met een lijst met de meest verwachte relaties. Aanbevolen wordt om de relaties van deze lijst te gebruiken: http://purl.oclc.org/metadata/dublin_core/wrelationdraft.html

14. Dekking Label: Coverage
De ruimtelijke of tijdelijke karakteristieken van de intellectuele inhoud van de Internetbron. Ruimtelijke dekking verwijst naar een geografisch gebied. Gebruik coordinaten (longitude (lengtegraad) en latitude (breedtegraad)) of plaatsnamen die van een gecontroleerde lijst komen of volledig uitgeschreven zijn. Dekking in de tijd geeft de tijdsperiode aan waarop de inhoud van de Internetbron betrekking heeft en niet wanneer de Internetbron is gemaakt of beschikbaar is gesteld (dit laatste hoort ingevult te worden bij "Date"). Gebruik dezelfde datum/tijd formaat als bij "Date"; W3C Technical Note http://www.w3.org/TR/NOTE-datetime of tijdsperioden die van een gecontroleerde lijst komen of volledig uitgeschreven zijn. Coverage is in ontwikkeling bij de Dublin Core Working group http://www.alexandria.ucsb.edu/docs/metadata/dc_coverage.html

15. Copyright Label: Rights

Een copyright verklaring of een link naar een copyright verklaring of een link naar een service die informatie geeft over de copyright van de Internetbron.

Laatst bewerkt: 1998-10-30

Bijlage 3: Authenticiteitseisen

NR	EIS	UITVOERING
Context		
1	De gebruiker krijgt een signaal wanneer hij naar een externe link gaat	Functionaliteit in opname-techniek
2	Organisatorische herkomst is reconstrueerbaar	Vastleggen in metadata
3	Rol in werkprocessen is reconstrueerbaar	Vastleggen in metadata
4	Plaats in informatie-architectuur is reconstrueerbaar	Vastleggen in metadata
Inhoud		
5	Tekst wordt integraal weergegeven	Functionaliteit in opname-techniek
6	Afbeeldingen worden integraal weergegeven	Functionaliteit in opname-techniek
7	Bewegend beeld en geluid worden integraal weergegeven	Functionaliteit in opname-techniek
8	Neerslag van transacties worden niet weergegeven. De interface wordt als afbeelding weergegeven	Functionaliteit in opname-techniek
9	Deepweb toepassingen zoals databases worden niet weergegeven. De interface wordt als afbeelding weergegeven	Functionaliteit in opname-techniek
10	Neerslag van chatsessies wordt niet weergegeven. De interface wordt als afbeelding weergegeven	Functionaliteit in opname-techniek
11	Meenemen van mutaties wordt per website bepaald. Informatieverlies wordt daarbij geaccepteerd	Bepalen frequentie van opname
Structuur		
12	Structuur website wordt integraal weergegeven	Functionaliteit in opname-techniek
13	Interne links worden integraal weergegeven	Functionaliteit in opname-techniek
Vorm		
14	Opmaak, vormgeving wordt integraal weergegeven	Functionaliteit in opname-techniek
15	Beeldfunctionaliteit, zie 19	-
Gedrag		
16	Functionaliteit tbv transacties wordt als afbeelding weergegeven	Functionaliteit in opname-techniek
17	Zoekfunctionaliteit wordt als afbeelding weergegeven	Functionaliteit in opname-techniek
18	Chatfunctionaliteit wordt als afbeelding weergegeven	Functionaliteit in opname-techniek
19	Beeldfunctionaliteit, zoals in- en uitzoomen, menugestuurde afbeeldingen, aanklikbare afbeeldingen, veranderende kleuren bij cursorbeweging, wordt werkend weergegeven	Functionaliteit in opname-techniek

Bijlage 4: Programma van eisen voor harvestingsoftware

Programma van eisen voor software bestemd voor het harvesten van websites

De eisen voor de software zijn tweeledig

- eisen waaraan iedere software moet voldoen ongeacht het doel waarvoor het wordt gebruikt
- eisen die specifiek gelden voor het soort software

Algemeen

- De software moet gebouwd zijn in open source, conform de uitgangspunten van het project E-depot
- De software moet stabiel kunnen draaien op de aanwezige apparatuur, het besturingssysteem en het netwerk.
- De software ontwikkelaar c.q. leverancier moet betrouwbaar zijn, waardoor continuïteit zoveel mogelijk is gewaarborgd en problemen ('bugs') snel worden verholpen.
- Ondersteuning van de software, door o.a. goede documentatie en helpfuncties.
- Het programma moet gebruiksvriendelijk zijn. De mogelijkheden kunnen ernstig worden beperkt wanneer voor het gebruik van de software veel technische kennis is vereist.

Specifiek 2

- Het programma moet de te archiveren website zo volledig mogelijk kunnen binnenhalen.
- Het moet mogelijk zijn om aan te kunnen geven wat wel en wat niet van een website binnengehaald moet worden.
- De functionaliteit van een website moet met het binnenhalen zoveel mogelijk intact blijven. Onderdelen die dynamisch of interactief zijn met bijvoorbeeld deep-web toepassingen moeten wel uitgeschakeld kunnen worden met zo min mogelijk verlies van functionaliteit van de website. Wanneer veranderingen zijn aangebracht moet dat aangegeven worden.
- Hyperlinks moeten zo correct mogelijk werken: de interne links moeten relatief gemaakt worden, zodat ze werken ongeacht hoe en waar de website is opgeslagen. Links die zijn ingebed in bestanden moeten kunnen worden geëxtraheerd.
- De metadata die aanwezig zijn in een website moeten ook meegenomen kunnen worden.
- De website moet op zodanige wijze worden opgeslagen dat de verschillende soorten bestanden bijeen blijven en de structuur van de website bewaard blijft.
- De wijze van archiveren moet het mogelijk maken om de website zo functioneel mogelijk te raadplegen. Indien veranderingen hebben plaatsgevonden moet dat aangegeven kunnen worden.
- Het moet mogelijk zijn om te bepalen welke bestandstypen die aanwezig zijn in de website ook te harvesten.
- Er moet een worden logbestand gegenereerd van het binnenhalen van de website waarin vermeld wordt welke bestanden en pagina's zijn binnengehaald en de eventuele foutmeldingen

[2] Deels afgeleid van: Filip Boudrez, *Archiveren van websites: een kwestie van waardering en 'capture'* (Antwerpen 2005)

Bijlage 5: (Concept) handleiding HTtrack

1. Hoe gebruiken we HTtrack

1.1 Analyse vooraf

1. Bepalen domein op basis van URL's, extensies en parameters
2. Zijn er potentiële knelpunten voor HTtrack ? Bekijk hiervoor bijvoorbeeld de bron code van de website
3. Bekijk de website eventueel ook met meerdere browsers (Firefox, Internet Explorer) om te zien of daar eventuele afhankelijkheden in zitten.

Bepalen domein:

- Let op adresbalk
- Pas op voor frames. Kijk naar statusblak, adresbalk, rechtermuisknop
- Let op bestanden met minder gangbare extensies (standaard bepalen!)

Onderzoek mogelijke knelpunten:

- Java, javascript en flash, parameters, (loops!!)
- Robots.txt en robots metatags

1.2. Instellen HTtrack: eerste run

- Kies / beperk het domein
- Kies aanvullende URL's (patronen)
- Beperk uit voorzorg de diepte (zie instellingen / beperkingen)
- Beperk de bandbreedte (zie instellingen / beperkingen)
- Blijf alert tijdens de spiegeling, indien loops lijken op te treden:stoppen!

1.3. Analyseer de offline site en stel HTtrack opnieuw in

- Bekijk de 'logs'
- Is alles gespiegeld? (is het niet onverhoopt afkomstig van de 'live'site ?)
- Let op browserchecks, bekijk de bronnen

1.4. Pas de instellingen aan

- Bijvoorbeeld: aanvullende startpagina's

15. Run HTtrack

- Repeteer…!

1.6. Sla de kopie veilig op

- Eventuele post-processing

Eerste aanzet tot beschrijving werkproces spiegeling websites via HTtrack

Start HTtrack		Indien voor de eerste maal: stel de gewenste taal in
		Ga naar volgende venster
Projectnaam		Naam van de websitie / organisatie
Project categorie		Is pas van toepassing als een aparte opslagserver beschikbaar is. Dan kan een categoriale indeling worden gemaakt. De websites kunnen ook via een numeriek systeem worden opgeslagen en de scheiding in collectie of project wordt dan via metadata aangebracht
Basispad		Staat standaard ingesteld op: C:\Mijn Web paginas. Ieder gewenst pad kan worden ingesteld.
		Ga naar volgende venster

Acties		Stel de gewenste actie in d.m.v. de onderliggende keuzelijst (doorgaans: ‘download websites’ maar er zijn diverse andere mogelijkheden
Toevoegen van een URL		Vul de URL in van de te downloaden website
Webadres (URL)		
URL List		Bevat onderliggende keuzelijst
Instellingen definiëren		Bij aanklikken verschijnt een mappenlijst
	Proxy	instellen op: proxy.rotterdam.nl : 8080 Als je klikikt op instellen verschijnt een schermpje waarop een log-in en een wachtwoord kan worden ingevuld. Indien dit wordt ingevuld ……
	Filters	Kunnen naar behoefte worden ingesteld
	Beperkingen	Maximale diepte: doorgaans instellen op 5 Maximale externe diepte: meestal 1 Maximale transfersnelheid: 25000 bi/sec. Overige instellingen: naar behoefte
	Stroomcontrole	n.v.t.
	Links	Kunnen naar behoefte worden ingesteld
	Structuur	Naar behoefte
	Spider	Met onderliggende keuzelijst. In voorkomende gevallen kiezen voor: geen robot.txt.regels
	Mime types	Aanpassen naar behoefte
	Browser ID	Aanpassen naar behoefte
	Protocol, index, cache	Aanpassen naar behoefte
	Expert	Aanpassen naar behoefte
		Ga naar volgende venster
Instellen remote verbinding		“Geen gebruik van een remote verbinding”
Voltooien		De spiegeling wordt uitgevoerd