capsis

Naar archivering van websites

René Voorburg (Capsis, r.voorburg@capsis.nl)
Hans Goutier (Ministerie van Verkeer en Waterstaat, Hans.Goutier@sso.minvenw.nl)
Verschenen in Archievenblad, mei 2004

Inleiding

Ook websites kunnen te archiveren bescheiden vormen. Om aan de eisen van de archiefwet te voldoen zal daarom bij de opzet van websites al rekening gehouden moeten worden met eisen ten gevolge van eventuele toekomstige archivering. Voor het Ministerie van Verkeer en Waterstaat vormde dit een reden om een verkenning te laten maken naar kwaliteitseisen voor websites ten behoeve van archivering. De vraag die bij de verkenning centraal stond was 'Wat kan er bij ontwerp en publicatie van een website al gedaan kan worden om de website later op een goede wijze te kunnen archiveren?'

Uitgangspunten voor de verkenning

Het antwoord op de vraag wat er bij het ontwerp van een website gedaan kan worden ten bate van eventuele archivering wordt in sterke mate bepaald door het antwoord op de volgende vragen:
1. Aan welke eisen moeten de bescheiden van een webarchief volgens de wet voldoen?
2. Hoe kan een website het beste gearchiveerd worden?

Bij de verkenning is ten aanzien van de wettelijke eisen de 'Regeling geordende en toegankelijke staat archiefbescheiden 2001' als uitgangspunt gekozen. Als snel kon echter geconcludeerd worden dat deze regeling voor websites weinig bruikbaar is. Als bijvoorbeeld de voorgeschreven bestandsformaten gebruikt zouden worden, dan zou de aard van de website bij het archiveren dermate veranderen dat van vervanging gesproken zou moeten worden. Dat lijkt een slecht uitgangspunt. Dit beknopte artikel gaat echter niet in op de discussie aangaande de toepasbaarheid van de Regeling maar bespreekt de tweede subvraag: “Hoe kan een website het beste gearchiveerd worden?” Welke methoden en technieken lenen zich daar het beste voor?

Benaderingen bij het archiveren van websites

Er kunnen ruwweg drie benaderingen onderscheiden worden bij het archiveren (hier met name gebruikt in de zin van 'capture') van websites:
1. het archiveren van de achterliggende bronnen
2. het archiveren van het eindresultaat (snapshot-methode)
3. integrale recordkeepingfunctionaliteit

Deze drie benaderingen worden beknopt behandeld waarna conclusies volgen.

Het archiveren van achterliggende bronnen

De allereerste websites waren volledige statisch van aard (statisch in de zin van 'gefixeerd', niet-veranderlijk). Iedere pagina van zo'n statische site bestaat uit een tekstbestand met daarin opmaakcodes. In de opmaakcodes (geschreven in de opmaaktaal HTML) van dit bestand staan doorgaans verwijzingen naar andere statische bestanden die door de webbrowser binnen de pagina getoond moeten worden, of die na een klik met de muis getoond moeten worden. Denk hierbij aan respectievelijk afbeeldingen en zogenaamde hyperlinks. Door deze eenvoud van opzet hoeft de archivering van een dergelijke site qua techniek niet veel meer om het lijf te hebben dan het opslaan van al 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. Bij een passende opzet1 is zelfs geen webserver2 nodig om de gearchiveerde bestanden opnieuw als integrale website te kunnen benaderen. Webarchivering door het archiveren van bronbestanden is zo een geschikte aanpak voor statische websites waarbij de bronnen beschikbaar zijn.

Het archiveren van het eindresultaat (de snapshot-methode)

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 nu zeer dynamisch van karakter. De pagina's van een moderne site worden doorgaans feitelijk pas bij het opvragen door de bezoeker gegenereerd. Dit gebeurt met deels op maat gemaakte programmatuur (zogenaamde scripts) en database-bevragingen. Een consequentie van deze techniek voor archivering is dat de bronnenbenadering exponentieel complexer is geworden. Om een gearchiveerde dynamische website weer op basis van de bronnen op oorspronkelijke wijze aan te kunnen bieden krijgt men vaak te maken met ketens van afhankelijkheden van soms zeer specifieke versies van software. In sommige gevallen zal deze software zelfs niet altijd geschikt zijn voor de meest actuele hardware. Er zal dan geschikte oude hardware gezocht moeten worden3. Een zeer onwenselijke situatie.
Deze exponentieel toenemende complexiteit kan doorbroken worden door alleen het eindresultaat van de complexe techniek te archiveren, dat wil zeggen de uiteindelijke pagina's en afbeeldingen zoals een bezoeker ze te zien krijgt (vergelijkbaar met het fotografisch bevriezen van de site, vandaar ook de naam snapshot-methode). Dit archiveren kan gebeuren met behulp van een applicatie die enigszins verwarrend vaak een *offline browser* wordt genoemd.

Ook deze methode van archivering kent beperkingen. In sommige gevallen kunnen webpagina's via deze aanpak niet gearchiveerd worden. Door bij de opzet van websites met de beperkingen van deze aanpak rekening te houden, kunnen veel problemen goed voorkomen worden. Hiertoe zijn in de rapportage aan Verkeer en Waterstaat aanbevelingen gedaan. Waar ook dit geen soelaas biedt kan voor die

---

[1] *Bij gebruik van enkel relatieve verwijzingen in plaats van absolute verwijzingen naar andere bestanden op de site.*
[2] *De applicatie die zorg dat de bezoeker van een website de gewenste informatie (pagina's) via het internet toegestuurd krijgt.*
[3] *Of er moet gebruik gemaakt worden van de techniek van emulatie.*

specifieke onderdelen desgewenst alsnog toevlucht genomen worden tot het archiveren van bronnen als scripts en databases.

Integrale recordkeepingfunctionaliteit

Een website staat nooit op zichzelf. Verschillende processen en systemen binnen een organisatie leiden tot de pagina's van de website van een organisatie. Essentieel onderdeel hierin vormt het zogenaamde Content Management Systeem (CMS), een applicatie voor het beheer van de inhoud en in zekere mate ook de vormgeving van een website. Een logische benadering voor het archiveren van website zou dan ook zijn om dergelijke systemen uit te breiden met recordkeeping functionaliteit. Een CMS met integrale recordkeepingfunctionaliteit zou dan bijvoorbeeld in staat zijn om de website terug te toveren zoals die er bijvoorbeeld 3 jaar geleden uit zag.
Dergelijke uitgebreide CM systemen bestaan echter nog niet. Als ze zouden bestaan, dan is het de vraag of ze voor archivering in de zin van de Archiefwet een goede oplossing zouden bieden. Immers, het digitaal archiveren van een compleet CMS zal vele malen duurder uitvallen dan het archiveren van de uiteindelijke pagina's van de website, zoals bij de andere twee methodes gebeurt. Bedenk bovendien dat een CMS een beperkte levensduur heeft en er dus over bijvoorbeeld 20 jaar meerdere van dergelijke systemen in het digitale archief opgenomen moeten worden. Wellicht kunnen de kosten behapbaar blijven als er gebruik gemaakt kan worden van generieke en gestandaardiseerde recordkeeping functionaliteit. Websites stellen echter weer bijzondere eisen aan systemen voor recordkeeping, wat maakt dat ook van deze oplossing op korte termijn geen heil verwacht mag worden.

Conclusie

Afsluitend lijkt voor complexe websites met de huidige stand van de techniek alleen de snapshot-methode een goede aanpak voor de archivering of de 'capture' van websites. Deze methode heeft als belangrijkste nadeel dat het in beperkte mate aanvullende eisen stelt aan de opzet van de te archiveren website. In een advies aan het Ministerie van Verkeer en Waterstaat is een aanzet gemaakt tot richtlijnen voor de opzet van archiveerbare websites.
Bij eenvoudige, statische websites kan ook goed de bronnenmethode gevolgd worden.
In de wat verdere toekomst wordt misschien een oplossing geboden door in Content Management systemen geïntegreerde recordkeepingfunctionaliteit.

www.capsis.nl - info@capsis.nl