# Het archipol-project

Keyzer, Andreas; den Hollander, Frank; Voerman, Gerrit

*Published in:*
Archievenblad



*Publication date:*
2002

# Het Archipol-project

Het archiveren van de websites van Nederlandse politieke partijen

Door A.K. Keijzer, F.J. den Hollander en G. Voerman

**Het in kranten en boeken vastgelegde 'papieren geheugen' mag dan op vele plekken – in bibliotheken, documentatie-instellingen en archieven – worden verzameld en beheerd, het bewaren van het 'digitale geheugen' staat daarentegen nog volledig in de kinderschoenen. Met het 'digitale geheugen' wordt hier niet gedoeld op digitale bestanden - waarvoor wat betreft archivering inmiddels meer aandacht is gekomen - maar op de bouwstenen van het World Wide Web (WWW): de websites.**

Voor zover bekend wordt het WWW nog nergens in de wereld systematisch ontsloten en gearchiveerd, en dat terwijl de websites continu veranderen en er voortdurend nieuwe sites bijkomen en oudere verdwijnen. Wel is er een aantal interessante initiatieven waarbij getracht wordt om periodiek *snapshots* (momentopnames) te maken van (delen van) het Web. Voorbeelden hiervan zijn te vinden in Zweden en Frankrijk, waarbij men het 'nationale' deel van het WWW wil opslaan. Het meest in het oog springende project is het Amerikaanse Internet Archive. De door deze instelling ontwikkelde *Wayback Machine* maakt regelmatig een download van het gehele Internet (voor zover dat althans mogelijk is), waardoor nu al van vele websites meerdere compleet gearchiveerde versies te bekijken zijn. Van het systematisch verzamelen en ontsluiten van websites is echter bij dit project en bij de andere genoemde initiatieven nauwelijks sprake. Dat is echter wel het geval met het Archipol-project, waarmee het Documentatiecentrum Nederlandse Politieke Partijen (DNPP) en de Universiteitsbibliotheek van de Rijksuniversiteit Groningen in september 2000 zijn gestart. In het kader van dit project worden de websites van de Nederlandse politieke partijen en hun nevenorganisaties (en die van de lijsttrekkers) gearchiveerd en ontsloten en in een *online* digitaal archief aangeboden.
Bij de start van het Archipol-project werden twee hoofddoelstellingen geformuleerd: enerzijds de inrichting van een digitaal archief van de websites van de Nederlandse politieke partijen ten behoeve van wetenschap (onderzoek en onderwijs) en media; anderzijds de ontwikkeling van een model van digitale archivering van websites waarvan andere (niet-commeriële) instellingen desgewenst gebruik kunnen maken. Het project bevindt zich momenteel in de afrondende fase; de officiële afsluiting is voorzien in mei van dit jaar. Inmiddels is er in het Archipol-project de nodige kennis en ervaring opgedaan over het archiveren en ontsluiten van websites. Deze kennis heeft tijdens de projectperiode geleid tot een aantal keuzes op organisatorisch, technisch en juridisch gebied.

## Organisatorische aspecten

Om een website volledig te kunnen downloaden voor opslag en archivering kunnen twee wegen worden bewandeld: de beheerder van de betreffende website kan gevraagd worden om eenmalig de gewenste data te leveren en daarna periodiek alle mutaties door te geven, of de 'webarchivaris' haalt de data zelf op vanaf de te archiveren site.
De eerste optie lijkt in theorie de simpelste manier om de gewenste data te verkrijgen, maar in de praktijk ligt dat ingewikkelder. Het eenvoudigweg kopiëren van het gedeelte van een

webserver waarop de website geplaatst is, kan vaak ook een aanzienlijk aantal oudere - niet meer in gebruik zijnde - pagina´s opleveren. Het selecteren van de juiste pagina´s (bij de eerste vulling en daarna periodiek) zal voor een beheerder extra inspanning betekenen. Het project zou daardoor te afhankelijk worden van de goede wil van de webmasters van de partijen. Daar komt nog bij dat een aantal sites wordt opgebouwd via een databasesysteem. Dat wil zeggen dat de feitelijke opbouw van de pagina's van de site *on the fly* tot stand komt, dus op het moment dat de pagina's daadwerkelijk worden opgeroepen. Het in zijn geheel aanleveren van een dergelijke site inclusief de boomstructuur is daardoor zeer moeilijk.
De tweede mogelijkheid – de data als webarchivaris zelf ophalen – is weliswaar arbeidsintensief, maar heeft toch een enkele duidelijke voordelen. Men kan als archivaris zelf beslissen, welke data worden opgehaald en met welke frequentie. Voor het Archipol project is gekozen voor deze werkwijze. Hierbij moet worden opgemerkt dat bij het downloaden van een partijsite wel de aanwijzingen van de webmaster worden opgevolgd.
Een ander organisatorisch aspect van het archiveren van websites dat aandacht verdient, is het gegeven dat men een *kopie* maakt van een bestaande website (de juridische implicaties komen hieronder aan bod). Wanneer deze kopie raadpleegbaar gemaakt wordt, dient het voor de gebruiker duidelijk te zijn dat het hier niet om de actuele website gaat, maar om een gearchiveerde kopie van vroegere datum. Het is dus zaak om dit bij de presentatie van een gearchiveerde website helder aan te geven in de zogeheten 'grafische schil' die om de geraadpleegde site wordt geplaatst.

**Technische aspecten**

Zoals gezegd is er in het Archipol-project voor gekozen om de data van de te archiveren websites zelf 'op te halen' en te verwerken. Voor dit downloadproces was bestaande programmatuur voorhanden, zowel commercieel als gratis (*public domain*). Deze zogeheten *off-line browsers* zijn op bruikbaarheid onderzocht en vergeleken. Daarbij kwamen grote onderlinge verschillen aan het licht. Eén overeenkomst hadden zij wel, namelijk dat zij lastig delen van websites konden downloaden die gebruik maakten van nieuwe webtechnologie als Flash. Daarom is bij het Archipol-project gekozen voor de ontwikkeling van een eigen downloadprogramma. Deze zelf ontwikkelde programmatuur ('Archipol**.**cgi') voor het verkrijgen van de gewenste data bleek in de praktijk een behoorlijk resultaat te geven. Tijdens de projectperiode kwam daarnaast nog een ander goed werkend programma beschikbaar, namelijk HTTrack 3.0. Voor het downloaden van de websites wordt nu zowel HTTrack als Archipol.cgi gebruikt.
Het gebruik van twee programma's creëert ook een controlemogelijkheid. Na het downloaden van een site wordt de volledigheid van de gearchiveerde data gecheckt: is de kopie zoals die wordt opgeslagen in het webarchief een volledige afspiegeling van de actieve site? Helaas bleek dit in lang niet alle gevallen zo te zijn. Daarom wordt elke gedownloade versie vergeleken met de vorige versie. Zo is het mogelijk duidelijke afwijkingen (en dus een mogelijke lacune) direct te signaleren. Daarnaast wordt periodiek de site met een tweede programma gedownload, waarna de uitkomsten worden vergeleken.
Maar zelfs bij een geslaagde, complete download is het niet mogelijk om alle onderdelen van een website in een archief op te slaan. Van interactieve aspecten van een website (zoals een antwoordformulier, een *chatbox* of de zoekmachine) kan niet meer dan een momentopname worden vastgelegd; die interactiviteit werkt in het archief eenvoudigweg

niet meer. Dat wil overigens niet zeggen dat in het archief alleen de tekstbestanden van de websites zijn opgeslagen. Zo wordt bijvoorbeeld ook het *Tomaatwerpspel* van de Socialistische Partij bewaard.
Een ander essentieel technisch probleem is de aanpassing van de interne verwijzingen binnen een site (de interne links) na de download, aangezien deze nu naar het webarchief dienen te verwijzen en niet meer naar de oorspronkelijke site. Ook de externe links (dat wil zeggen de verwijzingen naar webpagina's buiten de site) worden aangepast. Hier wordt een waarschuwing geplaatst dat men het webarchief verlaat en dat deze links naar andere externe pagina's kunnen verwijzen dan op het moment van archiveren, of dat deze zelfs geheel verdwenen kunnen zijn. Om toch een exacte kopie te hebben van de oorspronkelijke website, wordt op gezette tijden een versie opgeslagen waarbij de interne en externe links ongewijzigd zijn gebleven. Overigens wordt een verantwoording van alle noodzakelijke aanpassingen in een metadata-file opgeslagen. Deze file bevat verder o.m. de datum van de download en informatie over de gebruikte apparatuur en programmatuur voor archivering en opslag.
De gearchiveerde sites worden opgeslagen op de eigen documentserver van Archipol. Voldoende beschikbare schijfruimte is hierbij een aanhoudend punt van zorg; de site van het CDA alleen al telt zo'n 3500 pagina's met een totale grootte van bijna 50 Megabyte!

Nu het archief vorm begint te krijgen, moet ten slotte ook nagedacht worden over het zogeheten migratieprobleem. Kan een website die bijvoorbeeld is gemaakt om bekeken te worden met Internet Explorer 5.5, over tien jaar nog steeds bekeken worden met de standaard browsers van 2012? Het is zaak bijtijds naar oplossingen te zoeken voor dit mogelijk toekomstige probleem.

**Juridische aspecten**

Aan het downloaden en archiveren van websites kleven ook juridische problemen. Bij professioneel opgezette websites zoals die van de politieke partijen kan er sprake zijn van een commercieel (grafisch) ontwerp waarop copyright rust. Ook audio- en visuele onderdelen van de site of artikelen, kunnen auteursrechtelijk beschermd zijn. Bij het kopiëren en via een archief beschikbaar stellen van een website dient men met deze rechten rekening te houden.
Archipol werkt dan ook vooralsnog met een gesloten archief. Het bevindt zich weliswaar op het Web maar is niet openbaar. Alleen geautoriseerde gebruikers kunnen toegang krijgen tot het eigenlijke archief. Met de besturen van de politieke partijen vindt overleg plaats over de beschikbaarstelling van het archief en de bijbehorende copyright-kwesties.

**Stand van zaken**

Het Archipol-archief zelf is op verschillende wijzen ontsloten. In de eerste plaats kan men de websites vinden via de online publiekscatalogus van de RUG. In het archief zelf kan men zoeken op partij (of nevenorganisatie) en archiveringsdatum. Ook zijn alle gearchiveerde sites doorzoekbaar via de eigen Archipol-zoekmachine. Verder is er de mogelijkheid een beperkte inhoudelijke analyse te maken van de websites die zijn opgenomen. Zo kunnen de verschillen tussen twee verschillende versies van een website van een partij weergegeven worden en is er de mogelijkheid om woordtellingen te doen. Aan een verdere uitbouw van deze analysetools wordt nog gewerkt.

Op dit moment is het periodiek downloaden, archiveren en ontsluiten van de websites van de Nederlandse politieke partijen in volle gang. Behalve de reguliere partijsites worden ook de sites van nevenorganisaties (bijvoorbeeld de Jonge Socialisten) en van individuele personen (zoals minister Van Boxtel) gearchiveerd. In de aanloop naar de Tweede-Kamersverkiezingen 2002 zullen ook de sites van niet in het parlement vertegenwoordigde partijen en algemene verkiezingssites worden gedownload. Op deze wijze komt er een archief tot stand van de eerste verkiezingscampagne in Nederland die ook gedeeltelijk digitaal gevoerd zal gaan worden.

**Links:**
De projectsite van Archipol:
http://www.archipol.nl
The Internet Archive (the Wayback Machine): http://www.archive.org
Informatie over Archipol: info@archipol.nl
(tevens kunt u hier een username/passwordcombinatie aanvragen om het archief te raadplegen)

G. Voerman is hoofd van het Documentatiecentrum Nederlandse Politieke Partijen van de Rijksuniversiteit Groningen.
A.K. Keijzer en F.J. den Hollander zijn respectievelijk stafmedewerker elektronische dienstverlening en voorlichtingsfunctionaris van de Universiteitsbibliotheek Groningen.
De projectgroep Archipol bestaat behalve uit de auteurs tevens uit H.C.G. Druiven en M.S. van Delden, beiden ook werkzaam bij de Universiteitsbibliotheek Groningen.