Wie bewaart internet?

Bibliotheken koesteren oude boeken, in archieven liggen belangrijke documenten en musea herbergen oude schilderijen. Maar wie bewaart oude internetpagina's?

Dit jaar begint de Koninklijke Bibliotheek (KB) in Den Haag met het verzamelen en opslaan van internetpagina's. Een speciaal voor dit doel ontwikkeld programma maakt een paar keer per jaar een momentopname ('snapshot') van het Nederlandse internetdeel.

Hoe selecteer je echter interessante pagina's uit de enorme chaos die *world wide web* heet? Trudi Noordermeer, wetenschappelijk medewerker van de KB, vertelt: "Een zinnige keuze maken is erg moeilijk. Daarom proberen we alles te bewaren, ook al bestaat de helft uit persoonlijke pagina's die niet interessant zijn. Sommige collega's vinden dat we alles eerst nauwkeurig moeten selecteren en beschrijven. Maar daarvoor ontbreekt de tijd. Als we nu niets bewaren, is de informatie uit de oertijd van internet over een paar jaar voorgoed verdwenen. Dan zien onze kinderen nooit meer de incunabelen, de wiegedrukken, van internet.

"Op dit moment bevindt internet zich in het stadium waarin de eerste boeken uit de Middeleeuwen zich bevonden. Perkament was zo duur, dat men het volledig vol schreef, zonder ruimte te verspillen aan titel of auteur. Pas in het tijdperk van de boekdrukkunst ontstond de titelpagina en het register. Naar analogie hiervan moet men voor de huidige internetpagina's dit soort zaken nog ontwikkelen."

Internetarchief

De eerste archiveringstesten zijn inmiddels succesvol afgesloten. Als proef haalde de KB een paar honderdduizend webpagina's met wetenschappelijke artikelen van Elsevier Science binnen. Lex Sijtsma, coördinator ICT van de KB: "Op dit moment zijn we bezig om de adressen te inventariseren van alle Nederlandse webservers. We kunnen niet alleen selecteren op de domeinaanduiding .nl, omdat Nederlandse sites ook op .com, .org of .net kunnen eindigen. Daarom gebruiken we de lijst met serveradressen van Stichting Internet Domeinregistratie Nederland en van SURFnet."

Voor het binnenhalen van de webpagina's gebruikt de KB het programma NEDLIB-harvester. Het programma is in de programmeertaal C geschreven voor Linux. Voor de zware computersystemen van de KB ontwikkelde men een eigen Digital Unix-versie. Zo'n twintig miljoen Nederlandse pagina's downloaden is nog niet zo makkelijk. Sijtsma: "Als je ervan uitgaat dat de gemiddelde grootte van een pagina twintig kilobyte bedraagt, zou één enkele momentopname uiteindelijk 400 gigabyte groot zijn."

Overigens is de KB niet de eerste die het internet wil bewaren voor het nageslacht. In de Verenigde Staten begon de internetvisionair Brewster Kahle al in 1996 het project *The Internet Archive*, waarbij zijn doel was om het hele Amerikaanse deel van het internet op te slaan. In Europa timmeren vooral Finland en Zweden sinds een paar jaar hard aan hun eigen digitale archief.

Er is echter nog één groot probleem: een manier om eenvoudig door het snapshotbestand te navigeren is er nog niet. Het uiteindelijke doel van de KB is een hulpprogramma te maken voor standaard webbrowsers, zoals Internet Explorer of Navigator, waarmee je net zo gemakkelijk door het gearchiveerde internet kunt surfen als je normaal gewend bent van het 'echte' internet. Sytsma: "Zoals mensen nu de krant van hun geboortedag inkijken, bekijken ze over twintig jaar misschien wel het internet van die dag."

Stefan Verhaegh

Informatie

Koninklijke Bibliotheek
www.kb.nl
The Internet Archive
www.archive.org