

Hoe houden we oude bestanden toegankelijk?

Deze tekst verscheen in [Kijkmagazine 17 mei 2014](#)

Wie denkt dat bestanden op een harde schijf voor altijd bewaard blijven, komt bedrogen uit. Omvallende bitjes en veranderende formaten zorgen ervoor dat veel foto's, teksten en wetenschappelijke onderzoeken na verloop van tijd niet meer leesbaar zijn. Wat valt ertegen te doen?

De technologie van de Amerikaanse ruimtevaartorganisatie NASA zorgde er in 1969 voor dat astronauten naar de maan vlogen, er rondliepen en ze ook weer veilig terug naar aarde werden gebracht. Een indrukwekkend staaltje techniek, maar niet alles was op dat gebied zo goed uitgedacht. Want de data van de eerste, legendarische maanmissies zijn al decennia lang onleesbaar. Het is altijd netjes opgeslagen, maar opmerkelijk genoeg bestaan er geen computers meer die de bestanden af kunnen lezen.

Wie denkt dat dit probleem alleen voor dit specifieke geval geldt, komt bedrogen uit. Het is een wijdverspreid probleem. Vint Cerf, een van de vaders van het internet, waarschuwde hier tijdens een lezing vorig jaar tegen. Hij gaf aan dat Microsoft Office 2011 op zijn computer een Power Point presentatie uit 1997 niet meer kon openen. Hij benadrukte dat het vaak ontzettend lastig is om documenten met oude formaten te behouden, doordat ze vaak niet meer afgelezen kunnen worden.

Iedereen die bestanden opslaat en ze later weer af wil lezen dient rekening te houden met deze digitale vergankelijkheid. De gangbare formaten waarmee we informatie opslaan en de programma's waarmee we ze openen veranderen door de jaren heen.

“Stel jij bewaart de belangrijke foto's op een harde schijf en al je muziek staat daar ook op”, zegt programmamanager Marcel Ras van het Nederlands Centrum voor Digitale Duurzaamheid. “Dan is dat waarschijnlijk opgeslagen in de tegenwoordig gangbare bestandsformaten mp3 en jpeg. Je bent verstandig en maakt er een back-up van. Je bestanden zijn veilig opgeslagen en er is nog een kopie ook. Maar daarmee zijn ze absoluut niet bewaard voor de eeuwigheid.”

Denk nu eens twintig jaar verder, benadrukt Ras. Kinderen spelen dan op zolder, waar je de oude schijf hebt bewaard. Een hard schijf? Ze hebben waarschijnlijk geen idee wat dat is. “Net zoals kinderen van nu niet weten wat een floppy is en vrijwel niemand meer een apparaat heeft om het af te lezen. Stel dat een van die kinderen toch weet wat een harde schijf is en deze verbindt met een computer? Denk je werkelijk dat een besturingssysteem van de toekomst dan die bestanden kan aflezen?”

In een notendop is dat het probleem waar veel consumenten nu al mee te maken hebben, zonder dat ze het weten. Ze sloegen ooit foto's, muzieknnummers en presentaties op en dachten dat ze daarmee voorgoed bewaard zouden worden. Van papier wisten ze wel dat het zou vergaan. "Maar ze dachten heel naïef dat een harde schijf het eeuwige leven zou hebben. Niet dus", zegt computer wetenschapper David Rosenthal, die bij Stanford University het lang opslaan van data onderzoekt. Ras vult aan: "digitale vergankelijkheid is een groot probleem en we hebben een groot gebrek aan kennis. Niemand weet precies hoe we dit moeten oplossen."

Het speelt overal: bij mensen thuis, de overheid, wetenschappelijke instellingen en bedrijven. Kunnen we binnenkort nog wel de data lezen en controleren van belangrijke onderzoeken uit het verleden? En hoe zit het met de notulen van vergaderingen van de Tweede Kamer?

Het wordt een grote uitdaging om alle oude bestanden beschikbaar te maken, blijkt uit een bezoek aan verschillende archieven duidelijk. Daar wordt met name onderzoek gedaan naar digitale vergankelijkheid. Neem het Nationaal Archief. Na twintig jaar worden de archieven van onder meer de Eerste en Tweede Kamer, de ministeries en de algemene rekenkamer overgedragen aan het Nationaal Archief. "De afgelopen jaren krijgen we ook computerbestanden, daarvoor met name papieren archiefstukken", zegt informatiearchitect Gijsbert Kruithof van het Nationaal Archief.

Daar zitten de meeste vreemde bestandsformaten bij. Kruithof laat een bestand zien dat bestaat uit vreemde tekentjes met af en toe een halve zin. "Onleesbaar. Het is volstrekt onduidelijk waar het precies over gaat. We ontdekten uiteindelijk dat het een dataperfectbestand was en konden het pas daarna lezen."

Kruithof kijkt er inmiddels niet meer van op. Dit bestand wist hij uiteindelijk te openen, maar dat lukt niet altijd. Bijvoorbeeld bij documenten zonder extensie of een die alleen geopend kan worden met een programma die door een hobbyist ooit is gemaakt. "Dit is slechts het begin. We weten niet wat ons nog de komende jaren te wachten staat."

Gelukkig zijn er wel vrij veel mogelijkheden om de bestanden weer leesbaar te maken. Bijvoorbeeld het bestand openen via een virtual machine. Daarmee kan je op de huidige computers oude (besturings)programma's openen. Zoals Windows 95. Vervolgens kunnen archivariissen vaak wel een oud bestand openen. Dit heet emuleren.

Lukt dat toch niet, dan is er nog een optie. Nagaan of er nog ergens fanatiekelingen zijn, die een bepaald programma gebruiken. "Wij kwamen bijvoorbeeld Wordstarbestanden tegen", zegt Kruithof. "Op internet ontdekte ik dat er een usergroup was die bestanden van deze voorloper van Word en Wordperfect 5.1 nog kon openen."

Maar dan moet je wel weten welk programma nodig is. Kruithof pakt een floppy uit de kast. Het heeft een opmerkelijk groot formaat en is zeer stevig. "We hebben geen idee wat we hier mee moeten. We hebben tot dusver geen

drive gevonden om het ding mee te openen”, zegt hij. Op dit moment werkt het Nationaal Archief aan een pilot project met de Tweede Kamer onder meer om dit soort obstakels te voorkomen.

Ook elders hebben ze last van digitale vergankelijkheid. Neem Engeland. Daar wilden ze in de jaren tachtig een digitale tegenhanger maken van het Domesday book uit 1086. Daarin liet Willem de Veroveraar de bezittingen van het Engelse volk optekenen. De BBC wilde het negenhonderdjarig bestaan van het boek viëren en het onderzoek herhalen door de gegevens van een miljoen burgers op te slaan op de allernieuwste technologie: laserdisk. Destijds waren de laserdisk in opkomst, maar die had niet doorgezet. Eind vorige eeuw gebruikte vrijwel niemand nog deze techniek. Het kostte onderzoekers uiteindelijk flink wat tijd om de gegevens weer van laserdisk te halen en in een andere bestandsformaat om te zetten. Zo handig was de nieuwste technologie dus niet geweest, ruim tien jaar later was de info al zeer lastig toegankelijk. Het geschreven Domesday Book leek toekomstbestendiger.

Het laat zien hoe zo’n uitzonderlijke en tijdrovende klus het kan zijn om oude bestanden weer beschikbaar te maken. Gelukkig gaan de archieven die belangrijke bestanden bewaren tegenwoordig zeer zorgvuldig met de informatie om. “We zetten ze om in een gangbare bestandsvorm. Dat noemen we migreren”, zegt Ras. Die methode is niet helemaal waterdicht. “Wanneer we bestanden van het oude tekstverwerkingsprogramma Wordperfect openen in Word zijn er kleine veranderingen. Met name bij de opmaak zoals kopjes of wanneer iets vetgedrukt is, maar soms valt er ook wat weg bij tabellen. Stel dat een cruciaal cijfer verandert bij de migratie, dan klopt een formule opeens niet meer.”

Het heeft wel wat weg van het monnikenwerk uit de middeleeuwen. Zij schreven boeken over en veranderden daarbij vaak wat aan de tekst. Nu gebeurt dat ook bij het kopiëren naar nieuwe formaten, maar dan door de computer. “En bij gigantisch veel meer documenten”, zegt Ras.

Deze problemen gelden vooral voor heel oude bestanden die nu worden gearhiveerd, zoals de twintig jaar oude gegevens die het Nationaal Archief binnen krijgt. Of de foto’s die je thuis nog op een oude schijf hebt staan. “Wat nu in een archief binnenkomt, moet aan strenge voorwaarden voldoen. Het formaat spreken we van te voren af, als iemand wat aanlevert”, zegt unitmanager infrastructuur, beheer en IT Ernst van Velzen van Beeld en Geluid, dat het audiovisuele erfgoed van Nederland beheert en de audiovisuele bestanden van de publieke omroep bewaard. “In detail volgen de omroepen een bepaald schema, bijvoorbeeld hoe ze de audiosporen opslaan. Alle files moeten voldoen aan de standaard.” Daarbij speelt de metadata een cruciale rol. Daarin wordt precies aangegeven welk bestand het precies is en in welk formaat het is opgeslagen. Vrijwel alle archieven houden dit bij.

Naast het niet kunnen aflezen van bestandsformaten is er nog een dreiging voor digitale documenten: bit rot. Dat is de Engelse naam voor bestanden die

onleesbaar worden, omdat een harde schijf ze niet meer goed opslaat. Daarop staan gegevens opgeslagen in code door middel van enen en nullen. De magnetisatie zorgt daarvoor. Maar na verloop van tijd neemt de kans toe dat de bits omvallen. Een 0 wordt dan een 1 of andersom. Met als gevolg dat data verloren gaat. Er zijn al schijfhulpen die dit soort foutjes opsporen, maar er is nog geen waterdichte methode.

“Ook dat is een grote dreiging voor de opslag van gegevens”, zegt Rosenthal. “Je weet nooit welke bitjes gaan flippen. Stel dat je een petabyte opslaat, dan kan je er van uit gaan dat een bit omslaat. Maar welke? Het is zoeken naar een speld in een hooiberg en dat maakt het ongelooflijk lastig. Ik denk zelfs dat het een probleem is dat onoplosbaar is. We kunnen wel ongelooflijk goede harde schijven maken, maar die zijn te duur om te maken. Daardoor hebben we waarschijnlijk altijd met bit rot te maken.” Concrete getallen en wetenschappelijke onderzoeken op hoe groot dit probleem precies is, zijn er nog niet.

Op dit moment worden wel technieken ontwikkeld om het tegen te gaan. Een van de meest interessante werd bedacht door de Nederlander Dirk Roorda, onderzoeker bij DANS (Data Archiving and Networked Services) waar veel data van wetenschappelijk onderzoek wordt opgeslagen. Roorda past een slim trucje toe. Hij maakt gebruik van zogeheten check sums. “Dat zijn een soort vingerafdrukken van bestanden. Die bestaan uit zo’n honderd bits aan info en zodra er wat verandert in een bestand, bijvoorbeeld doordat een bit omvalt, wordt de check sum ook anders”, legt Roorda uit.

Hij maakte een programma dat automatisch veranderingen in check sums detecteert. Daarna kan een eerder gemaakte kopie worden teruggezet zodat het bestand behouden blijft (toch?). Roorda: ‘de komende vijftig jaar worden waarschijnlijk zo ontzettend veel meer bestanden opgeslagen, waardoor je er vanuit kan gaan dat het probleem nog heel veel groter wordt’.

Een andere methode is om gegevens op meerder plaatsen tegelijk op te slaan. Rosenthal maakte een systeem dat daarvoor geschikt is. Het heet Lockss en was oorspronkelijk bedoeld om wetenschappelijke tijdschriften te bewaren. “Vroeger hadden bibliotheken een exemplaar in de kast staan”, weet de onderzoeker. “Nu is alles digitaal en hebben bibliotheken toegang tot een online archief. Als ze een abonnement stopzetten, hebben ze nog wel recht op de oude nummers. En wat nu als een tijdschrift er mee op houdt of in het online archief een storing is?”

Lockss slaat de academische tijdschriften op meerdere plekken op bij bibliotheken. “Maar het wordt nu ook voor onderzoeksdata gebruikt in Canada en de Verenigde Staten. Daardoor krijgen zij een goede back-up van hun data. Die worden op meerdere plekken onafhankelijk van elkaar opgeslagen. Want een back-up maken van een beschadigd bestand wil je natuurlijk niet, dan neem je de fout gewoon over.”

Is digitaal opslaan niet te omslachtig vanwege deze problemen? Misschien is een terugkeer naar papier verstandig? Dat is overdreven en onnodig volgens deskundigen. Ook papier kan vergaan. Stel we printen e-mails uit, dan is het

document in wezen veranderd en kan de authenticiteit ervan niet gegarandeerd worden. Bovendien moet dan voortdurend gekopieerd worden en is er geen betrouwbare metadata van beschikbaar.

Marcel Ras pakt een rode reisgids van Rome vast. "Dit boek komt uit de jaren zestig. Het is voor iedereen toegankelijk die Italiaans kan lezen. We moeten er voor zorgen dat digitale bestanden net zo toegankelijk blijven. Met nieuwe technieken en door slim om te gaan met de opslag lukt dat wel. Maar we moeten wel voortdurend met onze tijd mee gaan zodat er zo min mogelijk verloren gaat."

Vergankelijkheid van het web

Het wereldwijde web verandert voortdurend. Nieuwe pagina's verschijnen of veranderen door een update, anderen verdwijnen. Een ontwikkeling die razendsnel gaat. Daardoor gaat ook veel verloren. En dat is zonde, vindt Internet Archive. Deze non-profitorganisatie uit San Francisco wil oude pagina's van het web toegankelijk houden. Het werkt als een internet bibliotheek. Ze maken kopieën van webpagina's, 160 miljard pagina's verzamelden ze al. Via hun Wayback Machine kan je bijvoorbeeld zien wat in een bepaald jaar de voorpagina van CNN was. Ook het Amerikaanse Library of Congress werkt aan een digitaal conserveringsprogramma en heeft al meer dan tienduizend sites gearchiveerd.