ARCHIVEREN VAN WEBSITES
Een kwestie van waardering en "capture"

Filip BOUDREZ
Medewerker, Stadsarchief Antwerpen

- De auteur gaat in op de groeiende behoefte aan het archiveren van websites. Het artikel vertrekt van het belang van websitearchivering onder meer omwille van de informatieve waarde die ze in zich dragen of van het verspreiden van overheidsinformatie door publieke overheden. De impact op het digitaal archiveren van het auteursrecht mag niet vergeten worden. Het meeste aandacht gaat naar het belang van het uitwerken van een archiveringsprocedure voor websites waarbij logischerwijze eerst de archiveringsdoelstelling wordt omschreven om daarna de uiteindelijke archiveringsprocedure te ontwikkelen. Als leidraad kan het beslissingsmodel dat binnen het DAVID-project (Digitaal Archiveren in Vlaamse Instellingen en Diensten) tot stand kwam, genomen worden: wat en hoe archiveren, wie archiveert en wanneer archiveren. Alvorens een website uiteindelijk aan het digitaal archief toe te voegen, dienen de nodige kwaliteitstests te gebeuren.

- L'auteur aborde les besoins croissants nécessaires à l'archivage des sites Web. L'article traite d'abord de l'importance de l'archivage des sites Web, notamment du fait de la valeur informative qu'ils comportent ou de la diffusion d'information officielle par les autorités publiques. L'impact du droit d'auteur sur l'archivage numérique ne doit pas être oublié. La plus grande attention va à l'importance de l'élaboration d'une procédure d'archivage pour les sites Web, par laquelle, logiquement, on décrit en premier lieu l'objectif de l'archivage, pour ensuite élaborer la procédure d'archivage proprement dite. Comme fil conducteur, on peut utiliser le modèle décisionnel développé par le projet DAVID (Digitaal Archiveren in Vlaamse Instellingen en Diensten - Archivage numérique dans les institutions et services flamands) : quoi et comment archiver, qui archive et quand ? Enfin, préalablement à l'ajout d'un site Web à une archive numérique, les tests de qualité requis doivent être effectués.

Het efemeer en uniek karakter van websites deed al vlug na het ontstaan van het World Wide Web de behoefte aan archiveringsoplossingen groeien. The Internet Archive startte al in 1996 met het verzamelen van websites. Dit initiatief inspireerde de bibliotheekwereld die ook naar methoden voor het vastleggen van webpagina's en allerhande webmateriaal zocht. Binnen de bibliotheek-wereld ontstonden diverse internationale en nationale initiatieven voor de archivering van het volledige web, een nationaal domein of geselecteerde websites[1].

Met het uitbreiden van de functies van websites groeide geleidelijk aan ook binnen de archieven het inzicht dat websites en webgebaseerde documenten de status van archiefdocument kunnen hebben en dus voor lange termijnarchivering in aanmerking kunnen komen[2]. Terwijl de bibliotheekwereld streeft naar het vastleggen van zoveel mogelijk webmateriaal en volledige websites die binnen het acquisitieprofiel passen, ligt voor de archiefwereld de klemtoon op het archiveren van de (delen van) websites en webgebaseerde documenten met archiefstatus. Dit veronderstelt dat de archiefdocumenten eerst worden geïdentificeerd en gewaardeerd. Archieven richten zich hierbij niet alleen op de Internet websites, maar ook op de intranetsites, het zogenaamde "deep web" en de neerslag van transacties en handelingen die via websites tot stand komen.

De archivering van websites is voor archivarissen een uitdaging in meerdere opzichten. Een goede archivering van websites is ten eerste slechts mogelijk wanneer met archivering al wordt rekening gehouden nog voor de eigenlijke creatie van de website en wanneer een archiveringsprocedure op het websitesbeheer van toepassing is. De archivaris dient hiervoor niet alleen betrokken te worden bij de ontwikkeling en het beheer van websites, maar veel organisaties zullen hun websitesbeheer hiervoor moeten herstructureren en controleren. Websites zijn ten tweede een heel vluchtig medium die zo snel mogelijk na de creatie dienen vastgelegd te worden ("capture"), zo niet dreigen ze verloren te gaan. Aangezien veel websites een dynamische, interactieve en gepersonaliseerde inhoud aanbieden, is die "capture" bovendien geen evidentie. Websites zijn ten slotte een mooi voorbeeld van digitale documenten waarvan we niet alleen context, inhoud en structuur in tijd willen overbrengen, maar ook de originele "look & feel" en een aantal basisfunctionaliteiten zoals hyperlinking.

In deze bijdrage wordt beschreven hoe archieven websites en webgerelateerd materiaal kunnen archiveren[3]. Na het schetsen van het belang van websitesarchivering worden kort de krachtlijnen van het auteursrecht en vooral de gevolgen voor archivering beschreven. Vervolgens wordt de archiveringsprocedure uitgetekend. Op basis

van de archiveringsdoelstellingen en het DAVID- beslissingsmodel wordt beschreven hoe organisaties hun archiveringsprocedure voor websites en webgerelateerde documenten kunnen gestalte geven. Hierbij wordt op verschillende plaatsen verwezen naar ervaringen en initiatieven van het stadsarchief Antwerpen en het DAVID-project op het vlak van websitesarchivering. Deze bijdrage sluit af met tips en richtlijnen voor het bouwen en onderhouden van duurzame en archiveerbare websites.

Belang van websitesarchivering

De archivering van websites kan met verschillende redenen worden gemotiveerd. Websites:

- zijn een belangrijke (exclusieve) informatiebron
- van publieke instellingen zijn bestuursdocumenten waaraan rechten kunnen worden ontleend en waarover de overheid verantwoording moet kunnen afleggen
- bevatten, publiceren of genereren archiefdocumenten
- behoren tot ons digitaal erfgoed

Informatieve waarde

Websites zijn een belangrijke bron van informatie. Het verspreiden van informatie en het voorlichten van geïnteresseerden was de eerste functie van de sites op het WWW. Aanvankelijk werd dezelfde informatie ook nog in andere vorm en via andere kanalen verspreid, maar ondertussen evolueerden websites tot de exclusieve publicatieplaats en bevatten ze alsmaar meer unieke informatie. Webpagina's en webpublicaties namen geleidelijk aan de rol over van traditionele papieren publicaties, terwijl papieren documenten een andere functie krijgen. Websites en papieren publicaties worden dus meer complementair en overlappen elkaar minder.

Organisaties communiceren via websites op een gestructureerde wijze informatie over hun samenstelling, beleid, bevoegdheden, taken, werkwijze, richtlijnen, dienstverlening, enz. naar de buitenwereld. Websites zijn hierdoor uitermate geschikt als primaire bron voor onderzoek van gelijk welke aard (instellingengeschiedenis, reconstructie van het beleid, geschiedenis van het WWW, sociale geschiedenis, enz.).

Binnen organisaties wordt ook gebruik gemaakt van het intranet voor de communicatie en het beheer van informatie. Het intranet vormt mee het organisatiegeheugen en is onmisbaar voor de bedrijfsvoering. Interne memo's, richtlijnen of werkdocumenten circuleren niet meer op papier maar worden centraal ter beschikking gesteld op het intranet.

Typisch voor websites is de snelheid waarmee informatie kan worden aangepast of waarmee informatie van het web verdwijnt. Dit leidt tot een tegenstelling. Enerzijds worden websites het referentiemateriaal, maar anderzijds zijn websites heel vluchtig. Ze kunnen immers diverse keren per uur worden aangepast. Webpagina's die gisteren nog werden geraadpleegd, zijn vandaag niet meer beschikbaar of bevatten andere informatie. De nood aan aan het vastleggen van deze informatie is groot. Anders dreigen deze mogelijke informatiebronnen voor de toekomst verloren te gaan.

Overheidswebsites zijn bestuursdocumenten!

Publieke overheden maken volop gebruik van websites om overheidsinformatie te verspreiden. Een website is één van de kanalen waarlangs de overheid de actieve openbaarheid van bestuur in de praktijk brengt. De Commissie voor Toegang tot Bestuursdocumenten meent dat een overheid moet beschikken over alle bestuursdocumenten waarvan zij de auteur is. De openbaarheidsreglementering houdt onrechtstreeks de plicht in dat overheden alle bestuursdocumenten ter beschikking stellen aan de burger die erom verzoekt. Administratieve overheden moeten bijgevolg de verschillende versies van hun website tijdens hun administratieve bewaartermijn bijhouden en toegankelijk houden. Het verwijderen of wijzigen of van on line content mag niet betekenen dat deze documenten onmiddellijk permanent worden vernietigd.

Overheidswebsites worden beschouwd als een officiële publicatie waarvoor de overheid verantwoordelijk is en waaraan de burger rechten kan ontlenen. Informatie op overheidswebsites kan immers handelingen en beslissingen van burgers of bedrijven reguleren of beïnvloeden. De overheid moet zich kunnen verantwoorden over de informatie die ze via het web verspreid. Zonder zelf de verschillende versies van zijn website bij te houden, is het voor de overheid moeilijk om zich in te dekken tegen toekomstige aansprakelijkheidseisen.

Websites zijn archiefdocumenten?

De archiefstatus van websites is een tijdlang een punt van discussie geweest waarbij werd afgevraagd of een website een publicatie dan wel een archiefdocument is. Aan deze discussie was ook de vraag inzake de archiveringsverantwoordelijkheid verbonden. Immers, afhankelijk van de status van een website zijn respectievelijk de bibliothecaris of de archivaris verantwoordelijk.

Het antwoord op deze vragen hangt in grote mate af van het profiel van de website van de organisatie. Een overwegend statische website met het on line verstrekken van informatie als hoofddoel kan als publicatie of eventueel als documentatiemateriaal worden beschouwd. Deze redenering geldt hoofdzakelijk voor de oudere generaties websites.

Inmiddels zijn websites danig geëvolueerd en zijn ze meer dan louter en alleen digitale publicaties. Internet- en intranetwebsites worden alsmaar meer in de werkprocessen van de organisatie ingeschakeld. Via websites worden transacties uitgevoerd of websites bevatten informatie die voortvloeit uit de werkprocessen van de organisatie. De documenten die op basis van deze webtransacties ontstaan, worden doorgaans in databanken bijgehouden. Deze websites genereren of bevatten archief en dienen gearchiveerd te worden samen met de andere archiefdocumenten die binnen dat werkproces worden gecreëerd of ontvangen.

In de meeste gevallen hebben websites een gemengde status en worden ze zowel als een publicatie of als een archiefdocument beschouwd. De archivering van websites is in die gevallen een gezamenlijke verantwoordelijkheid van bibliothecaris en archivaris.

Websites als digitaal erfgoed

Websites behoren tenslotte ook tot ons digitaal erfgoed. Bepaalde websites bevatten digitaal erfgoed of hebben zelf een culturele waarde. In het Unesco *Charter on the Preservation of the Digital Heritage* (15 oktober 2003) worden websites expliciet tot ons digitaal erfgoed gerekend (art. 1). Deze vorm van cultureel en digitaal erfgoed dient toegankelijk te blijven voor toekomstige generaties en moet gevrijwaard blijven van verlies ten gevolge van technologische veroudering en andere risico's (art. 2 en 3).

Besluit

Het archiveren van websites en webgebaseerde documenten is vanwege meerdere redenen belangrijk. Deze webgerelateerde archiefdocumenten dienen net zoals alle andere archiefdocumenten in een goede, geordende en toegankelijke staat worden bewaard. Dit is alleen maar mogelijk wanneer organisaties een archiveringsprodure voor websites uitwerken en in de praktijk brengen.

Auteursrecht

De archivering van websites en allerhande webmateriaal dient in overeenstemming met het auteursrecht te gebeuren[4]. Het auteursrecht beschermt de maker van originele werken. De auteur heeft morele rechten en vermogensrechten op zijn werk. De vermogensrechten van de auteur omvatten het exclusieve recht op reproductie, distributie, verhuring, uitlening, produceren van afgeleide werken en mededeling aan het publiek. Het auteursrecht is van toepassing op alle onderdelen van een website (de vormgeving, de inhoud, de functionaliteiten, enz.) en kan dus bij verschillende mensen berusten.

Het digitaal archiveren van een website druist op diverse punten in op het auteursrecht. Digitaal archiveren van een website houdt immers in dat een website wordt:

- gekopiëerd: maken of samenstellen van een exemplaar voor opname in het archief
- aangepast: aanbrengen van een aantal veranderen zodat de website beantwoordt aan de kwaliteitscriteria van een gearchiveerde website
- verspreid: ter beschikking stellen van een gearchiveerde website in de leeszaal, op een medium of via het Internet wat gelijk staat aan her-publiceren.

Het kopiëren, aanpassen of verspreiden is slechts toegelaten wanneer men over de toestemming van de auteur(s) beschikt. De bibliothecaris of de archivaris kan hiervoor een overeenkomst met de auteur afsluiten. Een dergelijke overeenkomst kan de vorm aannemen van een archiveringslicentie en bepaalt ondermeer onder welke voorwaarden het opnemen in het digitaal archief, aanpassen en verspreiden van een website toegelaten is[5]. Deze toestemming is niet vereist wanneer een bibliothecaris of archivaris de website van de eigen organisatie archiveert en wanneer de organisatie drager is van de vermogensrechten op zijn website.

Archiveringsprocedure voor websites

De archiveringsprocedure van websites start met het formuleren van een archiveringsdoelstelling en loopt vanaf de creatie van een website tot en met het verzekeren van de lange termijnbewaring. Vanwege de veroudering van opslagmedia, de informatiesysteemmigraties, het overschrijven van gegevens en de technologische veroudering is duurzame digitale archivering enkel mogelijk wanneer van bij de creatie al met archivering rekening wordt gehouden en wanneer de websites na opname in het digitaal archief raadpleegbaar blijven.

Bij het uitwerken van een archiveringsprocedure voor websites wordt uitgegaan van de digitale archivering van websites. De archiefwetenschap heeft als uitgangspunt dat archiefdocumenten in hun primaire vorm worden gearchiveerd. Een website is immers een digitaal archiefdocument dat digitaal geboren wordt en bijgevolg digitaal wordt gearchiveerd. Bovendien is archivering op papier in het geval van websites geen optie: informatie, opmaak, "look & feel" of gebruikerservaring, essentiële functionaliteit (bijv. hyperlinking), enz. gaan verloren.

De archiveringsdoelstelling

Alvorens een archiveringsprocedure uit te werken, is het belangrijk om eerst duidelijke doelstellingen te formuleren. Het uitwerken van een archiveringsprocedure houdt immers een aantal keuzes in, zoals wat wordt van een website voor archivering vastgelegd en met welke frequentie worden die onderdelen gearchiveerd. Deze vragen kunnen maar op een goede wijze worden beantwoord als duidelijk vastligt waarom een website wordt gearchiveerd.

De doelstelling van de archiveringsprocedure is in grote mate afhankelijk van de aard van de website, meer bepaald het doel waarvoor de website wordt gebruikt en welke informatie via dit kanaal wordt verspreid. Voor het bepalen van de archiveringsdoelstelling baseert men zich best op een analyse van de website:

- het profiel van de website:
  - bevat de website unieke informatie?
  - is het een statische, dynamische of interactieve website?
  - hoe frequent wordt de website gewijzigd?
- het doel van de website
  - waarvoor en binnen welk werkproces wordt de website gebruikt?
  - welke informatie wordt via de website verspreid?
  - welke rechten kunnen aan de website worden ontleend?
- de visibiliteit van de website
  - wat is het doelpubliek van de website?
  - hoeveel bezoekers heeft de website?
- de risico's:
  - welke risico's loopt men wanneer de website of een bepaald onderdeel niet wordt gearchiveerd?

Mogelijke doelstellingen voor archiveringsprocedure voor een website zijn:

- het archiveren van de wijze waarop een organisatie zich presenteert op het web
- het indekken tegen aansprakelijkheidsclaims
- het afleggen van verantwoording over de informatie die de organisatie publiceert
- het archiveren van (alle) informatie die via de website wordt verspreid
- het archiveren van de wijze waarop de on line interactie tussen overheid en burger verloopt
- het archiveren van de transacties die via de website plaatsvinden.

De archiveringsprocedure uittekenen

Als leidraad bij het uitwerken van een archiveringsprocedure kan men zich baseren op het beslissingsmodel dat door het DAVID-project werd uitgewerkt. Dit beslissingsmodel is opgebouwd rond vier vragen[6]:

- WAT archiveren?
- HOE archiveren?
- WIE archiveert?
- WANNEER archiveren?

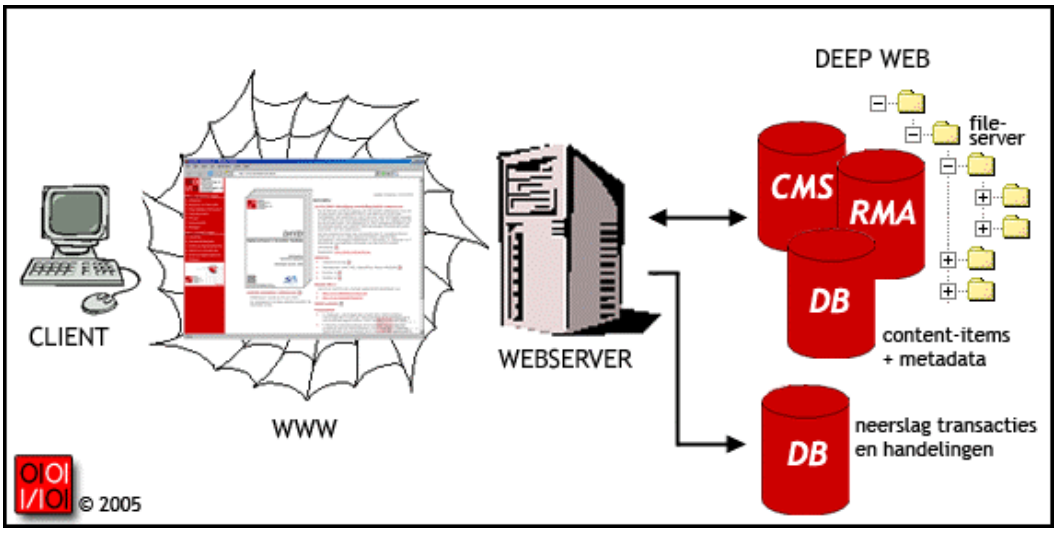
Het uitwerken van deze archiveringsprocedure is de gezamenlijke verantwoordelijkheid van de archiefvormer, de IT-medewerkers en de archivaris. De archiefvormer en de archivaris bepalen ten eerste welke onderdelen van een website of welke webgebaseerde documenten archiefwaardig zijn. Op basis van deze keuze kan in een tweede stap een archiveringsmethode worden gekozen. In functie daarvan legt men vervolgens de verantwoordelijkheden en de archiveringsfrequentie vast. Typisch voor websitesarchivering is dat de WAT- en de WANNEER-vraag sterk met elkaar verweven zijn.

Wat archiveren?

Websites zijn geëvolueerd van statische webpublicaties tot front-ends van dynamische en interactieve toepassingen. Terwijl websites vroeger op zichzelf staande publicaties waren, maken Internet- en intranetwebsites nu meestal

deel uit van een heel informatiesysteem. Een archiveringsprocedure voor websites mag zich bijgevolg niet beperken tot de webpagina's alleen. Voor de archivering van een website komen de volgende onderdelen in aanmerking:

- de webpagina's
- het deep web
- de neerslag van transacties en handelingen
- de metadata



De webpagina's

Een website is een geheel van webpagina's die aan elkaar gelinkt zijn en die via het HTTP-protocol op het Inter- of intranet beschikbaar zijn. Die webpagina's zijn dan weer gelinkt aan allerhande grafische elementen zoals afbeeldingen, animaties en stylesheets. De webpagina en bijhorende bestanden worden door de webserver naar een webclient gestuurd waar ze in een webbrowser worden ingeladen en vervolgens op scherm worden getoond.

De eerste generatie websites hadden een vaste inhoud. Iedereen die een bepaalde website bezocht, kreeg dezelfde inhoud te zien. Ondertussen evolueerden websites grondig en wordt de inhoud van bepaalde webpagina's pas na interactie met de gebruiker "on the fly" of op het moment van het verzoek samengesteld. De inhoud van deze webpagina's is afhankelijk van bepaalde gebruikersvoorkeuren, de geformuleerde zoekopdrachten of de beschikbare informatie in het "deep web". Bepaalde websites zijn zelfs niets meer dan een interface waarlangs "deep web"-informatie beschikbaar wordt gesteld. Dergelijke webpagina's hebben weinig of geen archiefwaarde. De meeste websites zijn echter een tussenvorm van beide soorten. Veel webpagina's bevatten zowel een gedeelte statische inhoud als een gedeelte dynamisch gegenereerde inhoud. De verschillende soorten websites vragen andere archiveringsmethoden.

In theorie hoeven enkel de webpagina's met de status van archiefdocument gearchiveerd te worden. In de praktijk daarentegen gaat men meestal alle webpagina's van een geselecteerde website archiveren, dus ook die onderdelen van een website zonder archiefwaarde of zonder de status van archiefdocument. Immers, dit is in de meeste gevallen nodig voor een getrouwe reconstructie van een gearchiveerde website als één geheel. Het onderscheid tussen de webpagina's met of zonder archiefwaarde is meer van belang wanneer de archiveringsfrequentie van die onderdelen wordt bepaald.

Net als andere digitale archiefdocumenten komen verschillende componenten van een webpagina voor archivering in aanmerking:

- de context: het werkproces waarbinnen een website gebruikt
- de inhoud: zowel statische inhoud als inhoud die dynamisch wordt samengesteld (zie ook *deep web*)
- de structuur: de relatie tussen de webpagina's (via sitemaps)
- de "look & feel": webdesign, opmaak, gebruikerservaring
- de functionaliteiten, het gedrag: hyperlinking, animaties, enz.

Voor webpagina's zullen alle vijf deze componenten na archiefwaardering als essentieel en bijgevolg als te archiveren worden aangeduid.

Het vastleggen van webpagina's houdt een aantal uitdagingen in die voortvloeien uit hun bijzonder karakter. Immers, webpagina's

hebben een aantal typische kenmerken die niet altijd overeenstemmen met het vaste karakter van een archiefdocument. Archiefdocumenten hebben bijvoorbeeld een gefixeerde inhoud en opmaak, terwijl webpagina's:

- soms heel snel aangepast worden: updates kunnen elkaar heel snel opvolgen
- soms geen vaste inhoud hebben: de inhoud is afhankelijk van gebruikersvoorkeuren of –rechten of van de beschikbare informatie in het "deep web"
- geen vaste "look & feel" hebben: de presentatie van een webpagina op scherm is sterk afhankelijk van de interactie met een bepaalde gebruiker (bijv. webbrowser, persoonlijke instellingen, historiek, zoekopdrachten, (taal)voorkeuren, enz.). Wat is de originele website? Welke versie wordt vastgelegd en in het digitaal archief opgenomen?
- sterk met elkaar verweven zijn: websites zijn dikwijls aan elkaar gekoppeld, worden soms op meerdere servers gehost of halen informatie uit externe locaties op. Hoe worden de grenzen van een te archiveren website afgebakend?

Bij het bepalen WAT van een bepaalde website voor archivering wordt vastgelegd, dient rekening gehouden te worden met deze specifieke kenmerken van websites. In veel gevallen zal dit betekenen dat in functie van de doelstellingen van de archiveringsprocedure een aantal keuzes dienen gemaakt te worden. Hiervoor wordt bij voorkeur ook teruggekoppeld naar de functie van de website en de soort informatie die via de website wordt verspreid. Webpagina's met informatie over openbare aanbestedingen hebben vanuit verantwoordingsperspectief een grotere waarde dan webpagina's met nieuwtjes.

In het geval van websites zonder vaste en/of snel wisselende inhoud, is de archivering van de inhoud een belangrijk aandachtspunt voor de gekoppelde "deep web"-applicaties (zie verder). Bij deze websites verloopt het archiveren van de inhoud efficiënter als dit rechtstreeks gebeurt vanuit de applicaties in plaats van dit via de webpagina's te willen doen.

Vereisten voor gearchiveerde websites:

- op basis van de vastgelegde versie is een getrouwe reconstructie van de website mogelijk
- de website wordt binnen zijn context gearchiveerd: informatie over de context waarbinnen de website werd gebruikt, wordt samen met de website vastgelegd.
- de websites worden op een platformonafhankelijke wijze gearchiveerd: voor de reconstructie van een website is enkel een webbrowser met de nodige plug-ins nodig (geen webservertechnologie, geen protocol voor de bestandsoverbrenging, geen DNS)
- alle interne links worden met relatieve pathaanduidingen aangeduid: de archiefgebruiker blijft binnen de gearchiveerde website surfen
- de inhoud van de webpagina's ligt vast: dynamische of interactieve functionaliteiten zoals weergave datum, bezoekersteller, zoekfunctionaliteiten, e-mailadressen, bestelmogelijkheden, enz. worden uitgeschakeld.

Het deep web

Het "deep web" is de verzamelnaam voor alle inhoud die via een website beschikbaar wordt gesteld en die wordt beheerd in content management systemen, databanken, documentbeheerssystemen of op fileservers. Die inhoud kan van een heel uiteenlopende aard zijn: gegevens, tekstdocumenten, spreadsheets, afbeeldingen, geluid, video, applets, multi media, enz. Informatie voor het web wordt in "deep web"-toepassingen bijgehouden om interactie en/of snelle actualiseringen mogelijk te maken. Sommige websites zijn niets meer dan een webinterface voor die "deep web"-toepassingen.

De integratie van websites met "deep web"-toepassingen is technologie gebonden en kan niet op een duurzame manier worden gearchiveerd. De archiefdocumenten binnen de "deep web"-toepassingen worden beter afzonderlijk gearchiveerd. Voorwaarde is wel dat een verwijzing naar de website behouden blijft. Op die manier worden beide aspecten van een website – de webpagina's enerzijds en het "deep web" anderzijds – afzonderlijk vastgelegd.

Voor de "deep web"-archivering is de belangrijkste vraag welke inhoud voor archivering wordt geselecteerd en welke metadata in combinatie met de archiefdocumenten wordt bewaard. Speciale aandacht dient hierbij uit te gaan naar de archivering van de verschillende versies van dezelfde inhoudscomponenten. Welke versies worden gearchiveerd en wat wordt van de versieveranderingen gearchiveerd? Enkel de aangebrachte wijzigingen of de integrale documenten na wijziging? Enkel de wijzigingen archiveren, is economischer inzake opslagverbruik, maar het reconstrueren van een versie op een gegeven tijdstip is moeilijker.

Niet alle content-items of al hun versies dienen gearchiveerd te worden. Voor het beantwoorden

van deze selectievraag worden de archiefdocumenten binnen de "deep web"-toepassing geïdentificeerd en wordt hun archiefwaarde onderzocht. Deze denkoefening vindt bij voorkeur plaats alvorens een "deep web"-toepassing in gebruik wordt genomen. Zo kunnen nog de nodige maatregelen getroffen worden om de inhoud op een bepaalde wijze te beheren zodat de archivering achteraf efficiënt kan uitgevoerd worden. Als bijvoorbeeld ook versiewijzigingen archiefwaarde hebben, dient de "deep web"-toepassing de mogelijkheid te voorzien om verschillende versies en hun metadata bij te houden.

De neerslag van transacties en handelingen via websites

Bij interactieve websites komen ook de transacties en handelingen voor archivering in aanmerking. Bij een aantal websites is de neerslag van transacties en handelingen zelfs de enige archiefwaardige informatie. Via webformulieren stuurt een gebruiker gegevens door naar de webserver of voert hij een bepaalde handeling uit. De schriftelijke neerslag van deze transacties en handelingen wordt doorgaans in één of meerdere databanken bijgehouden. Het archiveren van de neerslag van transacties en handelingen is dus in de meeste gevallen een kwestie van databankarchivering[7].

De metadata

Metadata is de verzamelnaam voor alle informatie die over een (digitaal) archiefdocument wordt bijgehouden. Op basis van metadata wordt een gearchiveerd document geïdentificeerd en toegankelijk gemaakt. In de metadata wordt bijvoorbeeld de administratieve context beschreven waarin het gearchiveerd document wordt gesitueerd. Ter ondersteuning van het beheer en de leesbaarheid op lange termijn, worden doorgaans ook technische gegevens in de metadata opgenomen.

Het bijhouden en verzamelen van metadata is de gezamenlijke verantwoordelijkheid van de webbeheerders, de inhoudsverantwoordelijken (de archiefvormer) en de archivaris. Het is van belang dat van bij het ontwerpen van een website de nodige documentatie al wordt bijgehouden. De metadata worden verder aangevuld tijdens de levensloop van een website. Belangrijke veranderingen zoals nieuwe functionaliteiten, nieuw webdesign, enz. worden in de metadata gedocumenteerd. Op het moment van archivering wordt deze documentatie samen met de website in het digitaal archief opgenomen.

Er dient documentatie bijgehouden te worden over de website, de webserver en de gearchiveerde onderdelen:

- de website:
  - URL
  - naam
  - IP-adres
  - startpagina
  - inhoud
  - functie/doel
  - webdesign
  - webmaster
  - redacteurs
  - on line versies
  - talen
  - grote wijzigingen
  - datum on line
  - datum off line
  - telleraantallen
- de webserver:
  - hardware
  - software
  - koppelingen met "deep web"
- de gearchiveerde onderdelen:
  - de webpagina's (broncode, snapshot, unieke responsen, surfsessie)
    - gearchiveerde versie(s)
    - datum momentopname
    - bestandsformaten
    - aanpassingen bij archivering
    - ontbrekende onderdelen
    - aantal mappen/bestanden
    - vereiste software
  - het deep web
    - toepassingen
    - archiefdocumenten
    - niet-gearchiveerde onderdelen.

In de archiveringsprocedure voor websites bij de stad Antwerpen worden de metadata in een Excel- spreadheet bijgehouden. Hiervoor is een speciaal Excel-sjabloon en een XML Schema ontwikkeld[8]. De archiefvormer en de webbeheerders noteren de metadata in het Excel-document en bezorgen dit aan de archiefdienst die het verder aanvult. Vervolgens worden de metadata naar XML omgezet en samen met de gearchiveerde website in het digitaal archief opgenomen.

Hoe archiveren?

De webpagina's

Het vastleggen van een website voor opname in het archief kan op vier wijzen gebeuren:

- archiveren van de broncode
- archiveren van een snapshot
- archiveren van de unieke webpagina's
- archiveren van een surfsessie

Archiveren van de broncode

Deze eerste archiveringsmethode bestaat uit het kopiëren van de bronbestanden zoals die op de webserver staan. Dit is de gemakkelijkste methode om een website vast te leggen, maar kan echter niet voor elke website worden toegepast. Deze archiveringsmethode is enkel toepasbaar wanneer:

- de webpagina's in (X)HTML beschikbaar zijn vanop de webserver
- de webpagina's reconstrueerbaar zijn zonder webservertechnologie
- de webpagina's en stylesheets aan elkaar gelinkt zijn met relatieve pathaanduidingen.

Mede vanwege deze voorwaarden zal een back-up van een website niet volstaan als archiveringsoplossing[9]. Bovendien zijn back-ups:

- niet platformonafhankelijk:
  - voor het terugplaatsen van de website is de back-upsoftware nodig, en in veel gevallen zelfs de back-upcomputer
  - teruggeplaatste documenten zijn enkel binnen de oorspronkelijk hard- en softwareomgeving operationeel
- helemaal niet ontwikkeld voor lange termijnarchivering: essentiële functionaliteiten zoals duurzame opslag, beheer van metadata en oplossingen voor de lange termijn leesbaarheid ontbreken.

Vooral voor de oudste generaties websites komt archivering van de broncode als archiveringsmethode in aanmerking. Deze websites zijn immers overwegend statisch. Recente websites zijn dynamisch en interactief en kunnen niet op deze wijze worden vastgelegd, want dit zou betekenen dat volledige webserverconfiguraties en gekoppelde "deep web"-toepassingen operationeel moeten blijven.

Voordelen:

- gemakkelijke archiveringsprocedure

Nadelen:

- in de meeste gevallen worden te veel computerbestanden naar het digitaal archief overgebracht. Harde schijven van webservers zijn zelden een toonbeeld van goed beheerde mappenstructuren: veel bestanden die in de on line versie niet meer worden gebruikt, zijn op de harde schijf blijven staan.
- toegang nodig tot de (afgeschermde) harde schijf van de webserver
- grondige kwaliteitscontrole nodig
- arbeidsintensief.

Archiveren van een snapshot

Een snapshot is een momentopname van de webpagina's waaruit een website bestaat. In een snapshot zijn de verschillende webpagina's aan elkaar verbonden. Voor het maken van een snapshot is een bijzonder computerprogramma nodig. Zo'n computerprogramma wordt een off line browser, een webcrawler of een webharvester genoemd. Er zijn zowel commerciële als open source webharvesters beschikbaar. Voorbeelden van open source webharvesters zijn HT Track <http://www.httrack.com>, Heritrix <http://crawler.archive.org> en de NEDLIB-harvester <http://www.csc.fi/sovellus/nedlib>. Zo'n webharvester wordt op een webclient geïnstalleerd en maakt een lokale kopie van de website door alle webpagina's en bijhorende grafische elementen (afbeeldingen, stylesheets, enz.) naar een harde schijf te kopiëren. Op de harde schijf worden niet de oorspronkelijke webserverbestanden gekopieerd, maar wel de (X)HTML-pagina's die op basis van de bronbestanden door de webserver worden samengesteld en naar een webclient worden verstuurd. Voor de raadpleging van zo'n lokale kopie is dan nog een webbrowser en soms ook enkele plug-ins nodig. Op die manier schakelt men de afhankelijkheid van een bepaalde webserverconfiguratie uit om in de toekomst de gearchiveerde website te bekijken. Dit is belangrijk voor een platformonafhankelijke archivering.

Belangrijk bij deze methode is het bepalen van een frequentie waarmee een snapshot wordt gemaakt en wordt gearchiveerd. Meerdere wijzigingen tussen twee snapshotoperaties in worden met deze methode niet mee gearchiveerd (zie *WANNEER archiveren*).

Deze methode is geschikt voor:

- statische websites waarvan de interne linken dmv absolute pathaanduidingen zijn vastgelegd: een webharvester kan automatisch de absolute pathaanduidingen omzetten naar relatieve pathaanduidingen
- dynamische websites: webpagina's die dynamisch worden samengesteld op basis van serverscripts en ingebedde scripting (bijv. asp-, cfm-, jsp-pagina's)
- interactieve websites: websites waarvan de webpagina's na interactie tussen gebruiker en "deep web" worden samengesteld.

Voordelen:

- de website wordt in zijn oorspronkelijke

opmaak en met minimale functionaliteit (hyperlinking)
    - vastgelegd
- de gearchiveerde website kan in zijn geheel geraadpleegd worden: alle webpagina's waaruit een website bestaat, worden vastgelegd
- enkel de computerbestanden die deel uitmaken van de on line versie worden gearchiveerd.

Nadelen:
- de koppeling met back-end toepassingen gaat verloren. Interactie tussen gebruiker en deep web of de serverresponsen wordt niet mee gearchiveerd. Enkel de blanco of vooraf ingevulde (X)HTML-formulieren worden gearchiveerd
- het maken van een goed snapshot vergt technische kennis over websites en webservers
- een snapshot maken van een grote website duurt lang
- kwaliteit is afhankelijk van de snelheid van de webserver en de internetverbinding
- webharvesters worden soms geweerd door webmasters (robot.txt) vanwege overbelasting van de server, time-out foutmeldingen, enz.
- webharvesters kunnen ingebedde hyperlinks (bijv. in oude Adobe LiveMotion of Macromedia Flash-objecten) of links in Javascript-functies maar moeilijk extraheren of omzetten naar relatieve pathaanduidingen[10]
- een "harvest"-operatie is soms moeilijk te definiëren: voorkennis over de websitestructuur is vereist
- dode linken doen het harvestingsproces soms vastlopen
- afgeschermde onderdelen (bijv. met paswoorden) zijn moeilijk vast te leggen
- grondige kwaliteitscontrole nodig
- manueel verbeteren van fouten is heel arbeidsintensief.

Aangezien de kwaliteit van het snapshot sterk afhankelijk is van de gebruikte webharvester, kies je best een webharvester met volgende functionaliteiten:
- afbakenen van de grenzen van het snapshot
- correct omzetten van absolute links in relatieve links
- overnemen van de originele bestandsstructuur van de webserver
- correct volgen van re-directs binnen de website
- extraheren van hyperlinks uit binaire bestanden
- mogelijkheid tot aanpassen externe links (bijv. waarschuwing toevoegen)
- mogelijkheid tot het vermijden van overbelasting van de webserver
- filteren van bestandstypes die al dan niet samen met het snapshot worden vastgelegd
- genereren logbestand van de snapshotoperatie met aantal bestanden en foutmeldingen.

TIPS voor het maken van een snapshot:
- maak een snapshot in een browseronafhankelijke (X)HTML-versie: gebruik een webharvester die zich als een webbrowser kan aanmelden waarnaar de browseronafhankelijke versie wordt verstuurd
- maak een snapshot op een tijdstip met laag serververkeer: spreek een tijdstip af met de webmaster
- bepaal duidelijk op voorhand welke digitale documenten samen met de webpagina's worden vastgelegd. Stem dit af met de archivering van het gekoppelde "deep web". Grote bestanden(tekstdocumenten, afbeeldingen, allerhande multimedia-objecten, enz.) in een snapshot opnemen, vertraagt het maken van een moment-opname aanzienlijk en vergt veel tijd. Beter is om die archiefdocumenten rechtstreeks vanuit de "deep web"-applicaties te archiveren.
- controleer achteraf de waarschuwingen en foutmeldingen die in het logbestand van de webharvester zijn geregistreerd.

Archiveren van de unieke webpagina's

Deze methode bestaat uit het archiveren van elke unieke webpagina. Telkens een webpagina voor de eerste keer wordt aangevraagd, wordt een kopie van deze webpagina in het digitaal archief opgeslagen. Elke wijziging of versieverandering wordt als een nieuwe unieke webpagina beschouwd.

Dit archiveringsproces verloopt volledig automatisch. Hiervoor is een bijzonder computerprogramma nodig dat op de webserver wordt geïnstalleerd. Dit programma controleert alle HTTP-responsen en wanneer de aangevraagde webpagina nog niet werd gearchiveerd, verstuurt het die pagina automatisch naar het digitaal archief of de tijdelijke opslagplaats voor deze documenten. Voorbeelden van dergelijke computerprogramma's zijn pageVault en webcapture[11].

Met deze methode is het mogelijk om alle webpagina's te archiveren die samengesteld worden op basis van de interactie tussen gebruiker,

website en "deep web".

Voordelen:
- elke versie van elke geraadpleegde webpagina wordt gearchiveerd
- zowel de statische als de dynamisch gegeneerde inhoud wordt gearchiveerd
- interactieve en gepersonaliseerde webpagina's worden rechtstreeks als statische (X)HTML- pagina's opgeslagen
- archiveringsproces verloopt volledig automatisch.

Nadelen:
- de website wordt niet als geheel gearchiveerd, enkel individuele webpagina's van de website worden naar het archief verstuurd. Een archiefgebruiker kan geen gearchiveerde website in zijn geheel raadplegen.
- niet-geraadpleegde webpagina's worden niet gearchiveerd
- wie welke webpagina's raadpleegde wordt niet standaard geregistreerd. Dit is technisch wel mogelijk, maar vraagt een interactie met de logbestanden van de webserver (mapping met IP- adressen).

Archiveren van een surfsessie

Een vierde methode om een website vast te leggen, is het filmen van een surfsessie. Terwijl een gebruiker een website bezoekt, registreert een screenrecorder alle webpagina's en instructies tijdens zijn bezoek aan een website. De screenrecorder maakt met een vooraf bepaalde frequentie screenshots van de webpagina's die worden getoond. De sequentie van screenshots wordt vervolgens als een videobestand bewaard.

Deze methode laat toe dat websites in hun originele "look & feel" worden vastgelegd. De website kan bezocht worden met de webbrowser waarvoor hij is ontworpen.

Deze methode is geschikt voor:
- websites die niet op een platformonafhankelijke wijze archiveerbaar zijn (bijv. websites in dhtml (Netscape), DHTML (Internet Explorer) of Flash) en die enkel in een specifieke webbrowser of met bepaalde plug-ins kunnen worden bezocht
- websites waarvan de archivering van de broncode niet mogelijk is, of waarvan geen snapshot kan worden gemaakt.

Voordelen:
- gebruikersinteractie tussen website en "deep web" kan worden vastgelegd
- gemakkelijk te creëren
- enkel lange termijn ondersteuning van het digitale videobestand vereist.

Nadelen:
- enkel geschikt voor het vastleggen van een impressie van hoe een website zich presenteerde in zijn oorspronkelijke omgeving. Deze methode is te omslachtig om grote websites volledig vast te leggen.
- de (ongecomprimeerde) videobestanden zijn (heel) groot.

Besluit

Voor het vastleggen van een website met het oog op archivering zijn meerdere methoden inzetbaar. Men dient in functie van de archiveringsdoelstelling en het type website de passende archiveringsmethode te kiezen. Als de verantwoordingsplicht de drijfveer van de archiveringsprocedure is, dan zal men alle inhoud willen archiveren. De archivering van de unieke responsen is dan één optie. Een andere optie zou het archiveren van momentopnamen in combinatie met de archivering van het "deep web" kunnen zijn. Is daarentegen de archivering van alle inhoud niet het uitgangspunt, dan kan het volstaan om momentopnamen zonder het "deep web" te archiveren. Afhankelijk van het type website kiest men vervolgens tussen de archivering van de broncode of het archiveren van snapshots.

De verschillende methoden sluiten elkaar overigens niet uit, maar kunnen elkaar aanvullen zodat de organisatie over een sluitende archiveringsprocedure voor websites beschikt. Periodieke momentopnamen zoals het archiveren van de broncode of een snapshot kunnen gecombineerd worden met de archivering van de unieke responsen. De vorser beschikt op die manier over een browsebare en volledige momentopname en de organisatie kan over elke versie van de webpagina's verantwoording afleggen. Ook het registreren van een surfsessie kan naast het archiveren van een momentopname worden toegepast. Een mogelijk scenario hierbij is dat de momentopname de platformonafhankelijke versie van een website bevat, terwijl in de surfsessie wordt geregistreerd hoe een website in een bepaalde webbrowser wordt gepresenteerd. Het stadsarchief Antwerpen past deze werkwijze toe voor het archiveren van de websites van de stad Antwerpen.

Het deep web

Voor de archivering van het "deep web" kan niet zomaar één archiveringsmethode worden aangereikt. Veel hangt af van het soort "deep web"-toepassingen die aan de website zijn

gekoppeld. Het uitgangspunt is voor iedere toepassing wel dezelfde: het archiveren van de archiefwaardige informatie die binnen het "deep web" wordt beheerd. Voortbouwend op de WAT-vraag bij "deep web"- toepassingen betekent dit dat de archiefdocumenten in combinatie met hun metadata op een digitaal duurzame wijze worden gearchiveerd.

In de meeste gevallen wordt die archiefwaardige informatie in een databank bijgehouden. Met het oog op lange termijnarchivering is de geijkte archiveringsmethode het exporteren van de archiefdocumenten en hun metadata uit het databanksysteem. Een belangrijk aandachtspunt is het vastleggen van de relatie tussen de archiefdocumenten en hun metadata op een platformonafhankelijke wijze. Mogelijkheden hiervoor zijn:

- inkapseling van metadata in de archiefdocumenten
- archiefdocumenten en hun metadata inkapselen in een XML-wrapper
- onderling verwijzen naar de bestandsnamen van bestanden waarin het archiefdocument en zijn metadata is opgeslagen.

Bij het exporteren van de archiefdocumenten en hun metadata worden beiden ook omgezet naar een geschikt archiveringsformaat. Voor afbeeldingen, geluid en bewegend beeld archiveert men best de digitale moederkopieën. Deze digitale moederkopieën zijn echter maar zelden on line beschikbaar.

Hiervoor zijn de computerbestanden te groot. In de "deep web"-toepassingen zullen slechts afgeleide versies voor on line raadpleging beschikbaar zijn. Deze afgeleiden zijn doorgaans van lagere kwaliteit. Dit geldt zeker wanneer het om streaming-media gaat. Streaming-media objecten zelf zijn overigens niet of heel moeilijk vast te leggen.

Voorbeelden van geschikte archiveringsformaten zijn:

Documenttype	Geschikt archiveringsformaat
Tekst	XML, PDF, TIFF, SXW
Afbeeldingen	TIFF
Geluid	WAV
Video	AAF, MXF

Wie archiveert?

In principe komt het aan de archiefvormer toe om in samenwerking met de IT-verantwoordelijken de archiveringsactie uit te voeren ("push-approach"). De rol van de archivaris richt zich dan voornamelijk op de kwaliteitscontrole en vervolgens op het beschrijven en toegankelijk maken van de gearchiveerde onderdelen van een website. Websites laten anderzijds ook toe dat de archivaris zelf een versie voor opname in het digitaal archief vastlegt ("pull-approach").

Wie welke archiveringsactie uitvoert hangt natuurlijk af van de gekozen archiveringsmethode en de beschikbare competenties. Dit verschilt van organisatie tot organisatie. Het archiveren van de unieke responsen verloopt volledig automatisch. Voor deze methode is enkel een regelmatige kwaliteitscontrole en eventuele bijstelling van de filters nodig. Het archiveren van de broncode of van een surfsessie kan in principe door elke handige pc-gebruiker worden uitgevoerd.

Het maken van een snapshot van een website vraagt in tegenstelling tot de andere methoden wel een gedegen kennis van webtalen en webdesign. De instellingen van de webharvester moeten in functie van de website worden gedefinieerd. Dit vraagt enige vertrouwdheid met de architectuur van de website en eventuele koppelingen met "deep web"-toepassingen. Ook het manueel verbeteren van fouten die eventueel in snapshots voorkomen, vraagt een zekere vertrouwdheid met (X)HTML, stylesheets en scriptingtalen.

Wanneer archiveren?

Het tijdstip waarop (onderdelen van) websites worden gearchiveerd, is voornamelijk afhankelijk van de archiveringsdoelstelling, van het profiel van de website en van WAT van een website wordt gearchiveerd. De wanneer-vraag is grotendeels een kwestie van het vastleggen van de frequentie waarmee wordt gearchiveerd.

Het bepalen van de frequentie is vooral van belang wanneer een snapshot, de broncode of een surfsessie wordt gearchiveerd. Websites met een hoog profiel (belangrijke impact van de verspreide informatie, veel bezoekers, veel aanpassingen) zullen frequenter gearchiveerd worden dan websites met een overwegend statische inhoud of een laag risiconiveau.

Die frequentie is voornamelijk evenementafhankelijk. Voor de hand liggende archiveringsmomenten zijn doorgaans de momenten waarop grote versieveranderingen zoals wijzigingen in de doelstelling, de functionaliteit of het webdesign zich aandienen, maar men kan evengoed ervoor opteren om met een vaste periodiciteit te archiveren.

Wanneer men beslist met een hoge frequentie momentopnamen van de website te archiveren, dan hoeft men niet noodzakelijk iedere keer de volledige website te archiveren. Men kan zich beperken tot de archivering van de webpagina's die sinds een bepaalde datum werden gewijzigd. Dit veronderstelt wel dat men een mechanisme voorziet zodat de webharvester kan controleren of een bepaalde webpagina al werd gearchiveerd of niet. Men kan dit door de wijzigingsdatum in de (X)HTML-header in te bedden, door RSS-feeds[12] te vermelden of door gebruik te maken van MD5-checksums of van een timestamp- of etagmechanisme in de HTTP-headers[13].

In het geval dat ook alle inhoudelijke wijzingen worden gearchiveerd, is de archivering van de unieke responsen in combinatie met de "deep web"-archivering de efficiëntste archiveringsmethode. Het archiveren van de unieke responsen is een continu proces dat volledig geautomatiseerd verloopt. Voor de "deep web"-informatie is dit hoofdzakelijk een kwestie van de archivering van de gekoppelde informatiesystemen (databanken, documentbeheerssystemen of content management systemen). In de meeste gevallen zal hiervoor een archiveringsactie worden uitgevoerd. Die archiveringsactie kan automatisch verlopen (bijv. automatisch exporteren telkens content wordt gewijzigd) of kan met een bepaalde frequentie worden uitgevoerd (bijv. selecteren welke contentitems in bulk worden gearchiveerd).

De kwaliteitscontrole

Alvorens men een website aan het digitaal archief toevoegt, is het belangrijk dat men controleert of de gearchiveerde website voldoet aan de kwaliteitsvereisten voor een gearchiveerde website. Volgende aspecten worden hierbij gecontroleerd:

- is een getrouwe reconstructie van de website mogelijk op basis van de gearchiveerde momentopname, de surfsessie of de webpagina's? Gebruik voor de controle van de momentopname of de webpagina's een computer zonder internetverbinding en verschillende webbrowsers.
- is het snapshot volledig? Controleer vooral de aanwezigheid van:
  - de tweede laag van roll-over afbeeldingen
  - de DTD's of XML Schema's
  - de stylesheets
- worden alle interne linken in de webpagina's, flash-objecten of stylesheets door middel van relatieve pathaanduidingen aangeduid?
- werken alle essentiële functionaliteiten: animaties, applets, Javascripts, enz.
- zijn dynamische of interactieve functionaliteiten zoals datumweergave, bezoekerstellers, on line zoekmogelijkheden, bestelmogelijkheden, enz. statisch gemaakt zodat die niet meer wijzigen wanneer iemand de gearchiveerde versie raadpleegt?
- werden e-mailadressen uitgeschakeld?
- is de nodige documentatie over de website aanwezig?

Met behulp van een tool die de links binnen een website verifieert, kan het grootste deel van die kwaliteitscontrole geautomatiseerd verlopen. Toch loont het de moeite om manueel een visuele controle uit te voeren.

In ieder geval is het heel belangrijk dat de te archiveren website on line beschikbaar blijft tot dat de archiveringsoperatie volledig is afgerond. Tijdens de kwaliteitscontrole kunnen immers fouten of gebreken aan het licht komen waardoor (een deel van) de archiveringsoperatie moet overgedaan worden of bestanden vanop de webserver aan de momentopname worden toegevoegd.

De opname in het archief en raadpleging

Na het voltooien van de kwaliteitscontrole kan de gearchiveerde website aan het digitaal archief worden toegevoegd. Samen met de gearchiveerde website wordt de bijhorende documentatie over de website in het archief opgenomen. Die documentatie wordt in eerste instantie verzameld door de archiefvormer en de IT-verantwoordelijken en wordt verder aangevuld door de archivaris.

In het websitesarchief van de organisatie is het heel belangrijk dat de websites en hun verschillende versies goed van elkaar worden gescheiden. Bij voorkeur bouwt men hiervoor een duidelijke mappenstructuur uit zodanig dat de verschillende websites niet met elkaar vermengd worden.

Idealiter zijn de gearchiveerde websites on line beschikbaar, maar dit is niet zo vanzelfsprekend. Het on line terbeschikkingstellen van een gearchiveerde website:

- mag niet indruisen tegen het auteursrecht
- vraagt een grote opslagcapaciteit van de webserver
- kan moeilijkheden opleveren wanneer de gearchiveerde webpagina's worden geïndexeerd door een zoekmachine.

Vanwege deze redenen kan het meer aangewezen zijn om de gearchiveerde websites enkel in de leeszaal van de archiefdienst te laten raadplegen[14]. De gearchiveerde websites kunnen dan vanop een fileserver of een lokale harde schijf beschikbaar worden gesteld. Een belangrijk aandachtspunt is het maken van een veiligheidskopie die off site wordt bewaard. In principe kan hiervoor elke magnetische of optische drager die geschikt is voor lange termijnarchivering worden gebruikt. Bij bepaalde dragers brengt de toepassing van een uitwisselbaar bestandssysteem echter aanpassingen van de map- en/of bestandsnamen met zich mee. Een voorbeeld hiervan is de toepassing van ISO-9660 bij CD-R's. Men dient hier omzichtig mee om te springen, anders bestaat de kans dat de hyperlinks niet meer werken.

Het raadplegen van de gearchiveerde websites en de bijhorende documentatie kan mogelijk gemaakt worden door een kleine portaalsite te bouwen. Vanuit die portaalsite kan de archiefgebruiker de gearchiveerde websites en hun verschillende versies raadplegen[15].

De lange termijnarchivering

Websites zijn digitale objecten. Voor de raadpleging van digitale objecten is hard- en software vereist. Aangezien hard- en software snel verouderen en in onbruik raken is een digitale bewaarstrategie voor de gearchiveerde onderdelen van een website met middellange of permanente bewaartermijn nodig. Zo niet, dan is het risico groot dat een gearchiveerde website niet meer raadpleegbaar is omdat de nodige hard- en/of software ontbreekt.

Voor de presentatie op scherm van een gearchiveerde website of een webpagina is op zijn minst een webbrowser nodig die niet alleen (X)HTML, maar ook afbeeldingen (GIF, JPEG, PNG) en clientscripts ondersteunt. Voor het bekijken van een bepaalde inhoud van een webpagina zijn soms ook plug-ins vereist (bijv. Flash-player, PDF-reader). De huidige generatie webbrowsers (Microsoft Internet Explorer, Mozilla Firefox, Netscape, Opera, enz.) ondersteunen nog steeds de oudste generaties websites, maar van zodra die ondersteuning dreigt te verdwijnen, is een passende oplossing nodig.

Het voorbije decennium werd al heel wat onderzoek verricht naar digitale bewaarstrategieën voor digitale archiefdocumenten. Algemeen worden migratie en emulatie als geschikte oplossingen voor het digitale duurzaamheidsprobleem naar voor geschoven[16].

Bij migratie worden de digitale documenten omgezet naar geschikte archiveringsformaten. Archiveringsformaten zijn bestandsformaten die aan een aantal kwaliteitsvereisten voldoen. Een geschikt archiveringsformaat is:

- open en gedocumenteerd
- gestandaardiseerd: het beheer is in handen van een standaardiseringsorganisatie
- platformonafhankelijk en uitwisselbaar: onafhankelijk van een producent, een besturingssysteem, een protocol of een computerprogramma
- wijdverspreid.

Het toepassen van de migratiestrategie bij het archiveren van websites zou betekenen dat (bepaalde delen van) de gearchiveerde broncode, snapshots, webpagina's of videobestanden omgezet worden van zodra de nodige softwareondersteuning dreigt te verdwijnen.

De emulatiestrategie bestaat uit het nabootsen van de functionaliteiten van oude computerprogramma's op nieuwe computers. De hardware of de software die men hiervoor nodig heeft, wordt de emulator genoemd. Het ontwerpen of programmeren van een goede emulator is slechts mogelijk wanneer men over voldoende documentatie over het oorspronkelijke computerprogramma en de bestandsformaten beschikt.

Emulatie voor websitesarchivering zou kunnen betekenen dat een emulator voor een webbrowser wordt geprogrammeerd die op een toekomstig platform draait. Met die webbrowseremulator wordt een gearchiveerde website op scherm gereconstrueerd. De gearchiveerde websites zelf hoeven in dit scenario niet omgezet te worden.

Welke bewaarstrategie uiteindelijk het best geschikt is voor de lange termijnarchivering van gearchiveerde websites en hoe die zal worden toegepast, is vooralsnog geen uitgemaakte zaak. Het is wel duidelijk dat standaarden in beide digitale bewaarstrategieën een belangrijke rol innemen. Bij migratieoperaties zijn de doelformaten gestandaardiseerde bestandsformaten. Emulatie heeft de beste kans op slagen wanneer de gearchiveerde documenten in een gestandaardiseerd formaat zijn opgeslagen.

In het geval van websites is het merendeel van alle informatie al in gestandaardiseerde vorm op het web aanwezig. Buitenlands onderzoek toont aan dat de inhoud van websitesarchieven voor 95% of meer HTML-, GIF- en JPEG-bestanden bevatten. Toch is enige waakzaamheid geboden. Veel webontwikkelaars maken gebruik van producenteigen uitbreidingen op de (X)HTML-markuptaal om functionaliteiten of opmaak toe te passen die niet in de (X)HTML-standaarden voorzien zijn. Deze uitbreidingen op de standaard markuptaal zijn doorgaans op maat van één bepaalde webbrowser gemaakt en bijgevolg niet zomaar uitwisselbaar. Dit is bijvoorbeeld het geval met websites die in dhtml (Netscape) of in DHTML (Internet Explorer) werden ontworpen[17]. Voor lange termijnarchivering levert zo'n website bijkomende problemen op. Migratie van zo'n webpagina's of emulatie van een webbrowser die producentspecifieke uitbreidingen op de standaard ondersteunt, is extra moeilijk. Beter is dan om een versie van de website te archiveren die conform de (X)HTML-standaard is.

Het digitale duurzaamheidsprobleem vraagt een continue aandacht waarvoor waarschijnlijk zelfs nooit een permanente oplossing voor handen zal zijn. Telkens een bepaalde technologie in onbruik dreigt te raken, zal hiervoor een alternatief moeten uitgewerkt worden. Mede vanwege deze reden is het van belang dat men goed registreert welke bestandsformaten deel uitmaken van het websitesarchief en dat men de technologische evolutie blijft opvolgen.

Duurzame en archiveerbare websites

De archivering van websites is niet altijd even gemakkelijk en vraagt dikwijls een grote inspanning van personeel en middelen. Door van bij de ontwikkeling van een website met archivering rekening te houden, kunnen de archiveringsacties efficiënter worden uitgevoerd. Dit levert niet alleen tijdwinst bij het archiveren op, maar resulteert ook in een effectief archiveringssysteem voor een website en de informatie die ermee wordt verspreid. Op die manier neemt de kwaliteit van de gearchiveerde websites toe en bereikt men beter de archiveringsdoeleinden.

Onderstaande vereisten voor webpagina's en gekoppelde web content managementsystemen zijn gebaseerd op de resultaten van internationale pilootprojecten voor websitesarchivering en de ervaring die het stadsarchief Antwerpen opdeed bij het archiveren van de websites van de stad Antwerpen[18].

Vereisten voor de webpagina's

Rekening houden met archivering bij het ontwerpen van een website levert niet alleen een archiveerbare website op, maar zorgt er ook voor dat de website heel toegankelijk is. De criteria voor een duurzame en archiveerbare website vallen in grote mate samen met de webtoegankelijkheidsregels[19]. Het naleven van deze regels zorgt ervoor dat zoveel mogelijk mensen toegang hebben tot de informatie op een website.

Belangrijke ontwerpvereisten voor duurzame en archiveerbare webpagina's zijn onder meer:

- werk een duidelijke, uitbreidbare mappenstructuur voor de website uit

- respecteer de officiële standaarden voor mark-uptalen: vermijd het gebruik van de producenteigen of browserafhankelijke uitbreidingen op deze standaarden. Maak geen gebruik van DHTML of van tags die als "deprecated" worden aangeduid. De officiële mark-up standaarden zijn:
  - HTML 4.01: gebruik de strict variant
  - XHTML 1.0: gebruik de strict variant
- mark-up:
  - schrijf structurele mark-up, geen presentatie mark-up
  - schrijf grammaticaal correcte mark-up
  - valideer de webpagina's alvorens ze on line worden gepubliceerd
- zorg voor een duidelijke scheiding tussen enerzijds inhoud en structuur en anderzijds opmaak
  - inhoud en structuur: (X)HTML
  - opmaak: CSS (Level 2.1)
- CSS of XSL-stylesheets worden opgeslagen in afzonderlijke bestanden, en worden niet ingebed in de webpagina's
- respecteer de regels van het gelaagd bouwen: zorg voor voldoende (HTML-) alternatieven wanneer bepaalde elementen (CSS, client-scripts) niet worden ondersteund
- gebruik geen frames in de website: gebruik niet de frameset- of iframesvariant
- hyperlinks:
  - bed geen javascript-functies in hyperlinks in (bijv. a href= "javascript:")
  - bed geen hyperlinks in Flash ActionScript in
  - geef interne links dmv een relatieve pathaanduiding aan
- URL's:
  - maak vriendelijke, menselijk begrijpbare URL's
  - zorg ervoor dat dynamisch samengestelde URL's nog steeds naar dezelfde informatie wijzen, ook al is de informatie gewijzigd
  - vermeld geen sessies of query-strings in de URL's
  - vermijd het gebruik van spaties in map- en bestandsnamen
- gebruik dezelfde karakterset voor de hele website
- bed volgende metadata in elke webpagina in als meta-element:
  - gebruikte karakterset
  - datum on line
  - datum wijziging
- gebruik geen client-side scripts voor essentiële functionaliteit op webpagina's of zorg voor voldoende HTML- of server-side scripting alternatieven
- pas zoveel mogelijk webstandaarden (W3C) toe: vermijd het gebruik van formaten die geen native webtechnologieën zijn en waarvoor bijzondere plug-ins nodig zijn (bijv. PDF en Flash).

Vereisten voor een content management systeem

Websites worden steeds meer aangestuurd vanuit een content management systeem. Met een content management systeem kunnen webpagina's heel snel aangepast worden, zonder dat hiervoor uitgebreide technische kennis nodig is. Idealiter is een content management systeem gekoppeld aan de records managemant applicatie van de organisatie. In de praktijk zijn er echter nog maar weinig organisaties die over een heus records management applicatie beschikken. En als er al een records management applicatie is, dan zal het tot stand brengen van die koppeling in veel gevallen maatwerk met bijkomende hoge ontwikkelingskosten vergen. Voor de meeste organisatie zal de rechtstreekse archivering vanuit het content management systeem een meer realistische en aangewezen oplossing zijn. Hiertoe formuleert men best een aantal vereisten en functionaliteiten om de archivering van de content-items met archiefwaarde en hun metadata mogelijk te maken.

De jongste content management systemen beschikken wel over een aantal documentbeheersfunctionaliteiten, maar deze zijn nog ontoereikend voor de lange termijnbewaring van digitale archief-documenten. Bovendien is het niet de bedoeling dat de content-items met archiefwaarde binnen het content management systeem worden gearchiveerd, maar dat ze samen worden bewaard met de andere (digitale) documenten die binnen hetzelfde werkproces werden gecreëerd of ontvangen. De vereisten en functionaliteiten van een content management systeem zijn bijgevolg gericht op het versie-beheer en het exporteren van content-items en hun metadata zodat ze in een volgende stap in de records management applicatie of rechtstreeks in het digitaal depot kunnen worden opgenomen.

Een content management systeem bestaat uit verschillende componenten. De belangrijkste componenten zijn een beheer- en presentatiemodule. Voor het beheren van de content-items wordt meestal gebruik gemaakt van een databank. Deze content-items worden via een presentatiemodule op het web gepubliceerd. Hiervoor worden onder meer templates gebruikt. Deze templates halen hun inhoud op uit de gekoppelde databank en

voegen die vervolgens in welbepaalde placeholders in. Een aantal vereisten zijn dan ook van toepassing op de databank en de templates zodat archivering efficiënt kan verlopen:

- mbt de databank:
  - de gegevens worden beheerd door een open databank management systeem
  - de gegevens zijn bevraagbaar via SQL-statements en toegankelijk voor andere applicaties via ODBC- of JDBC-koppelingen
  - heeft een gedocumenteerd, overzichtelijk en uitbreidbaar datamodel
- mbt de templates voor webpagina's: op basis van de templates worden webpagina's samengesteld die voldoen aan de vereisten vermeld onder "vereisten voor webpagina's".

Daarnaast zijn er vanuit archiveringsstandpunt ook een aantal gewenste functionaliteiten voor het content management systeem zelf:

- metadatavelden zijn vrij definieerbaar door de organisatie. Volgende metadata voor content- items zijn essentieel:
  - titel / naam
  - uniek webadres
  - versienummer
  - datum redactie van het content-item
  - datum on line beschikbaar (van... tot...)
  - publicatieplaats op de website
- versiebeheer van de content-items en hun metadata:
  - het CMS kan de verschillende versies van content-items en hun metadata bijhouden
  - het CMS voorziet de mogelijkheid dat gepubliceerde content-items pas na versieverandering worden gewijzigd (bijv. check-in enkel mogelijk na versieverandering)
  - het CMS bewaart de verschillende versies van on line content op een statische wijze en als afzonderlijke objecten, in combinatie met hun metadata
- archivering van de content-items en hun metadata:
  - het CMS kan nieuwe en gewijzigde webpagina's automatisch als een statische (X)HTML-pagina bewaren
  - het CMS laat toe dat de sitestructuur en geselecteerde content samen met hun versieveranderingen automatisch wordt gearchiveerd (bijv. via sitemaps voor de sitestructuur of via statische (X)HTML-pagina's voor de webpagina's)
  - het CMS kan geselecteerde content-items en hun metadata in bulk exporteren
  - het CMS biedt ondersteuning om geselecteerde content-items en hun metadata bij export te migreren naar archiveringsformaten
- het CMS kan een gestructureerde audit-trail bijhouden waarin geselecteerde acties worden geregistreerd.

Besluit

In navolging van de bibliotheekwereld is de archivering van websites ook een belangrijke activiteit voor archieven. Voor archieven ligt de klemtoon wel op de archivering van (delen van) websites en webgerelateerde documenten met archiefstatus. Identificatie van de archiefstukken is samen met archiefwaardering een belangrijke taak voor de archivaris.

Voor de archivering van websites en webgerelateerde archiefdocumenten kan geen unieke archiveringsprocedure worden aangereikt. Afhankelijk van het websitesprofiel, de identificatie van de archiefdocumenten en de mogelijke risico's dient elke organisatie zijn eigen archiveringsdoelstellingen te formuleren. In functie van die archiveringsdoelstelling(en) wordt een archiveringsprocedure voor de websites uitgetekend. Het DAVID-beslissingsmodel kan hiervoor als basis dienen.

Het weze duidelijk dat de archivering pas efficiënt en doeltreffend kan verlopen wanneer vanaf de creatie van een website al met archivering wordt rekening gehouden. De archivering moet als het ware ingebed worden in het creatie- en beheersproces van een website. Alleen dan zal de archiveringsprocedure de doelstellingen bereiken en in grote mate geautomatiseerd kunnen verlopen.

Filip BOUDREZ
Stadsarchief Antwerpen
Venusstraat 11
2000 Antwerpen
Filip.Boudrez@stad.antwerpen.be

22 februari 2005

Noten

1 Voorbeelden uit de bibliotheekwereld zijn onder meer: Kulturarw3 (Zweeds domein), Nationale Bibliotheek van Noorwegen (Noors domein), Universiteit van Helsinki en Nationale Bibliotheek van Finland (Fins domein), Koninklijke Bibliotheek van Denemarken met het *Netarchive.dk*-project (Deens domein), *Pandora* (selectie van het Australisch domein), *Minerva* (selectie van het Amerikaans domein). Belangrijke onderzoeksprojecten zijn: the Nordic Web Archive en het International Internet Preservation Consortium. Binnen beide projecten wordt software ontwikkeld voor het archiveren en ontsluiten van websites.

2 Initiatieven uit de archiefwereld voor het archiveren van websites werden onder meer genomen door het Nationaal Archief van Australië en het NARA (US National Audiovisual and Records Administration). The National Archives (UK) werkt voor de archivering van een 50-tal geselecteerde overheidswebsites samen The Internet Archive.

3 Deze bijdrage is de bewerkte tekst van de presentatie: Boudrez, F. Archiveren van websites: een kwestie van waardering en capture, op: Studievoormiddag *Een preserverings- en conserveringsstrategie voor analoge en digitale bibliotheekcollecties en archiefbestanden*, ABD-BVD, Antwerpen, 2 december 2004.
Tenzij anders vermeld is deze bijdrage gebaseerd op: Boudrez, F. ; Van den Eynde, S. *Archiveren van websites*. Antwerpen-Leuven, 2002 (DAVID-rapport nr. 7); Public Record Office (UK). *Managing Web resources. Management of electronic records on websites and intranets: an ERM toolkit.* London, 2001; National Archives of Australia. *Archiving Web Resources: A Policy for Keeping Records of Web-based Activity in the Commonwealth Government*. Canberra. 2001; National Archives of Australia. *Archiving Web Resources: Guidelines for Keeping Records of Web-based Activity in the Commonwealth Government*. Canberra, 2001.

4 Meer informatie over het auteursrecht en de impact op digitaal archiveren is beschikbaar in: Dekeyser, H. *Digitale archivering: een juridische stand van zaken vanuit Belgisch perspectief. Deel 2: Auteursrecht, technische beschermingsmaatregelen en wettelijk depot*. Leuven, 2004 (DAVID-rapport, nr. 8); Dekeyser, H. Auteursrecht, in: Boudrez, F. ; Dekeyser, H. *Digitaal archiveren in de praktijk. Een handboek*. Antwerpen-Leuven, 2004 <http://www.antwerpen.be/david> (login 12/09/05)

5 Een model archiveringslicentie is opgenomen in Dekeyser, H.*, Digitale archivering: een juridische stand van zaken vanuit Belgisch perspectief. Deel 2: Auteursrecht, technische beschermingsmaatregelen en wettelijk depot.* (DAVID-rapport, nr. 8) Leuven, 2004, pp. 56-62.

6 Meer informatie over het DAVID-beslissingsmodel is beschikbaar in: Boudrez, F. Beleid en procedures, in: Boudrez, F. en Dekeyser, H. zie noot 4.

7 Meer informatie over het archiveren van databanken is beschikbaar in: Boudrez, F. Archiveringsprocedures. 3.Informatiesystemen, in: Boudrez, F. ; Dekeyser, H. zie noot 4 ; Boudrez, F. *Een archiveringssysteem voor dynamische en interactieve informatiesystemen, Stadsarchief Antwerpen*, Antwerpen, 2003 ; Testbed Digitale Bewaring. *Van digitale vluchtigheid naar digitaal houvast. Kosten- en beslissingsmodellen. Functionele specificaties Bewaren van databases.* Den Haag, 2002.

8 Het XML Schema is on line beschikbaar op de website van het DAVID-project <http://www.antwerpen.be/david>→ DAVID-website → cases → websites. Het Excel-sjabloon wordt gepubliceerd op de website van het stadsarchief Antwerpen: <http://stadsarchief.antwerpen.be> → 5. Stadsarchief in bedrijf → Digitale archivering → Websites.

9 De oudste versies van de website van de stad Antwerpen waren op back-uptape opgeslagen. Het duurzaam archiveren van deze websites had heel wat voeten in de aarde en hing aan een zijden draadje. Versies 1, 2 en 4 konden mits heel wat inspanningen worden gerecupeerd. Versie 3 ging verloren. Voor meer informatie hierover : Boudrez, F. *Van backup tot gearchiveerde website. De archivering van de eerste versies van de Digitale Metropool Antwerpen.* Antwerpen, 2001.

10 Macromedia lanceerde wel een Flash Search Engine SDK, maar deze dient hoofdzakelijk om zoekrobotten SWF-objecten te laten doorzoeken en indexeren. De ingebedde hyperlinks worden hierbij uit de SWF- animaties geëxtraheerd en naar HTML omgezet (swf2html) Webharvesters kunnen de ingebedde absolute URI's niet omzetten naar relatieve URI's.

11 Fitch, Kent. *Web site archiving – an approach to recording every materially different response produced by a website* <http://ausweb.scu.edu.au/aw03/papers/fitch/paper.html> (login 12/09/05); Fitch, Kent. *An approach to recording every materially different response produced by a Web site*, op: National Library of Australia. Archiving Web Resources, 12 november 2004.

12 RSS-feeds: RDF-samenvattingen die als XML-bestand op de server beschikbaar zijn en waarin de wijzigingen of nieuwe informatie wordt vermeld.

[13] De timestamp in de HTTP-header is de laatste wijzigingsdatum. Etag of Entity Tag is onderdeel van het cache controle mechanisme. (L.R. CLAUSEN*, Concerning etags and datestamps.* 4th International Web Archiving Workshop, Bath, 2004).

[14] De websites gearchiveerd door The Internet Archive en de Nationale Bibliotheek van Australië zijn bijvoorbeeld on line raadpleegbaar. De gearchiveerde websites van het Zweedse Kulturarw3–project zijn enkel in de leeszaal van de Nationale Bibliotheek van Zweden te consulteren.

[15] Een voorbeeld portaalsite van waaruit gearchiveerde websites en hun metadata raadpleegbaar zijn, is beschikbaar op de website van het DAVID-project <http://www.antwerpen.be/david> → DAVID-website → cases → websites

[16] Boudrez, F. Bewaarstrategieën, in: Boudrez, F. en Dekeyser, H. zie noot 4.

[17] Dynamisch HTML is de verzamelnaam voor combinaties van mark-up, stylesheets, DOM en scripting waarmee dynamische en interactieve webpagina's worden gemaakt. DHTML is niet formeel als standaard vastgelegd. DHTML wordt op een verschillende manier in webbrowsers geïmplementeerd, waardoor pagina's met DHTML niet uitwisselbaar zijn.

[18] Boudrez, F. *Van backup tot gearchiveerde website.* zie noot 9 ; Digitaal Archiveren: rlchtlijnen & advies, nr. 5: Websitesbeheer voor archivering.

[19] Web Accessibility Initiative <http://www.w3.org/WAI> (login 12/09/05); Voor een adaptatie van deze richtlijnen voor de Nederlandse overheid: <http://www.webrichtlijnen.overheid.nl> (login 12/09/05)