stap 3: hoe archiveren?

--> de webpagina's

Er zijn verschillende methoden om websites te archiveren:

1. archiveren van de broncode. Een eenvoudige methode, maar geen duurzaam oplossing en bovendien enkel geschikt voor statische websites.
2. archiveren van de unieke responsen, waarbij elke aangevraagde pagina met een wijziging of versieverandering als een nieuwe unieke webpagina wordt beschouwd en vervolgens volledig automatisch gearchiveerd.
3. archiveren van een surfsessie: terwijl een gebruiker een website bezoekt, registreert een screenrecorder alle webpagina's en instructies tijdens het bezoek aan een website. De sessie wordt vervolgens als een videobestand bewaard. Deze methode laat toe dat de originele "look & feel" van moeilijk archiveerbare (zoals flash-sites) websites in een open formaat worden vastgelegd.
4. archiveren van een snapshot: hiervoor is een *webharvester* of *webcrawler* nodig.
Open Source voorbeelden zijn *HTTrack* (http://www.httrack.com) en *Heritrix* (http://crawler.archive.org). Deze programma's maken een lokale kopie van een website door webpagina's en bijbehorende grafische elementen naar de harde schijf te kopiëren. Van dynamische websites wordt zo een statische momentopname gemaakt die bewaard kan worden. Deze methode is doorgaans de meest geschikte voor het *capturen* van zowel statische als dynamische websites.

--> de metadata

Voor de op te nemen metadata wordt een uitbreidbaar en flexibel XML schema opgemaakt. De metadata worden waar mogelijk automatisch geëxtraheerd.

--> deep-web archivering

Archiefdocumenten binnen de **deep web** applicaties - CMS, databanken, enz. - kunnen beter rechtstreeks vanuit die applicaties gearchiveerd worden, aangezien de bestaande automatische archiveringstools die een snapshot nemen van een website nog gebrekkig zijn op dit vlak. Om wille van de grote verscheidenheid aan applicaties kan er voor het archiveren van het deep web kan geen eenvormige oplossing worden aangereikt.

Meer informatie in: "*Archiveren van websites: een kwestie van waardering en 'capture'*, Antwerpen."

stap 4: toegankelijkheid

De gearchiveerde websites krijgen een plaats in een digitaal depot. De grote waarde van een dergelijk depot wordt pas ten volle gerealiseerd wanneer het ook toegankelijk is voor een publiek. Reeds van bij ontwerp van het depot moet met de toegankelijkheid rekening worden gehouden zodat gebruikers later op een veilige en flexibele manier de archiefdocumenten kunnen raadplegen:
- zonder dat de authenticiteit van de archiefdocumenten in het gedrang komt
- waarbij de gebruikers ten allen tijden de nodige metadata ter beschikking hebben om de documenten correct te kunnen interpreteren
-...

Meer informatie over de kwaliteitsvereisten voor een digitaal depot vindt u in:
"tekst risico-analyse/kwaliteitsvereisten"

Ook roept het toegankelijk maken van een digitaal depot via intra- en/of internet een aantal juridische vragen op. De inhoud van een digitaal depot online toegankelijk maken is geen vanzelfsprekendheid omdat aan de belangen van verschillende betrokkenen geraakt wordt.
Voor archieven, bibliotheken en musea is door een uitzondering in de auteurswet inmiddels mogelijk de inhoud van een digitaal depot ter ter plaatse ter beschikking te stellen.

In het kader van het Digitaal Depot project verschenen recent een aantal publicaties die licht proberen te werpen op deze en andere juridische vragen. Ook zijn een modelarchiveringslicentie voor websites en een aantal modelclausules i.v.m. de verwerking van persoonsgegevens opgesteld.

Alle eDAVID publicaties zijn online terug te vinden:

expertisecentrum eDAVID vzw

websites archiveren in 4 stappen

Digitaal Depot

een bijzondere uitdaging

Websites zijn overal en nergens. Ze zijn dynamisch, multimediaal en interactief. Allerhande bedrijven, organisaties en particulieren hebben een eigen stek op het web, ook culturele instellingen en kunstenaars. In het Unesco 'Charter on the Preservation of the Digital Heritage' (2003) worden websites expliciet tot ons digitaal erfgoed gerekend.

Het aantal sites groeit aan een razend tempo maar oude pagina's verdwijnen net zo snel: ze worden vervangen, geupdate of staan plotseling niet langer online.

Juridisch bevindt websitesarchivering zich nog in een sche-merzone. Privacy en auteurswetgeving hebben de technologische evolutie niet helemaal bijgebeend. Toch is er geen tijd te verliezen, hoe langer we wachten hoe meer websites verloren gaan. Steeds meer bewaarinstellingen trachten dan ook om uiteenlopende redenen en op verschillende manieren een stukje van het internet te vereeuwigen.

In deze folder overlopen we 4 belangrijke stappen die bij het archiveren van websites komen kijken:

stap 1 - wat archiveren?: we starten met het bepalen welke elementen bij archivering zeker moeten bewaard blijven.

stap 2 - kwaliteitsvereisten: door van bij de ontwikkeling van een website met archivering rekening te houden en een aantal kwaliteitsvereisten in acht te nemen, kunnen de archiveringsacties efficiënter worden uitgevoerd.

stap 3 - hoe archiveren?: belicht een aantal methodes om websites te archiveren, zoals het nemen van snapshots en de tools die daartoe gebruikt kunnen worden. Ook gaan we in op het vastleggen van het *deep web*.

stap 4 - toegankelijkheid: enkele aandachtspunten bij het toegankelijk maken van je digitaal depot.

stap 1: wat archiveren?

Vooraleer we van start gaan is het belangrijk te beslissen waarom, en bijgevolg wat, we willen archiveren. Welke aspecten van de website willen we zeker bewaren? Welke elementen zijn bepalend en maken de website tot wat hij is?

--> de webpagina's

Websites hebben een informatieve waarde. Die informatie is ingebed in **inhoud en structuur** van de afzonderlijke pagina's en de relatie tussen de pagina's onderling.

De pagina's bevatten ook visuele en interactieve elementen. Vaak is het net deze specifieke **'look & feel'** in combinatie met bepaalde **functionaliteiten** zoals hyperlinks en animaties, die een bepaalde site uniek maken.

--> de metadata

Voor een latere archiefgebruiker kan ook bepaalde **context** informatie onontbeerlijk zijn - bv. het werkproces waarbinnen een website wordt gebruikt - om de website volledig te kunnen plaatsen en begrijpen.

Naast contextgegevens zijn er tal van andere technische en administratieve **metadata** die samen met de webpagina's kunnen worden bewaard: datum dat de pagina online en offline ging, wie webdesigners en webmaster waren, informatie over hard- en sofware van de webserver, enz.

--> deep-web archivering

Veel websites zijn tegenwoordig dynamisch. De webpagina zelf is niet meer dan een interface. De inhoud en lay-out van de pagina's worden *on the fly* opgebouwd. Inhoud, functionaliteit en structuur worden aangeleverd door **deep web** applicaties: Content Management Systemen (CMS), databanken, documentbeheersystemen en fileservers.

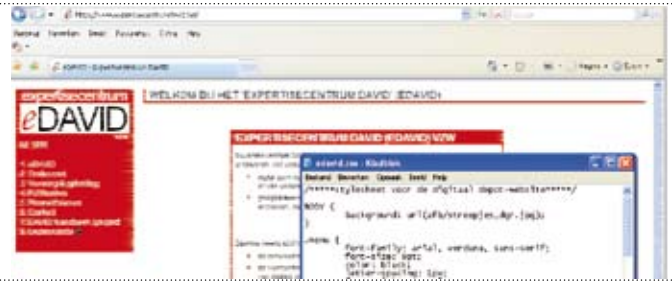
Bij interactieve websites komen ook de **transacties en handelingen** voor archivering in aanmerking, bijvoorbeeld de zoekopdrachten die door gebruikers worden ingegeven. De neerslag van deze transacties en handelingen wordt doorgaans in één of meerdere databanken bijgehouden. Het archiveren van deze elementen is dus in de meeste gevallen een kwestie van databankarchivering.

Het archiveren het deep web dat schuil gaat achter dynamische en interactieve websites brengt extra uitdagingen met zich mee (--> *zie stap 3).*

stap 2: kwaliteitsvereisten

De criteria voor een duurzame en archiveerbare website vallen in grote mate samen met de webtoegankelijkheidsregels. Zo levert het naleven van bepaalde kwaliteitsvereisten niet alleen **tijdwinst** en een archiveerbare website op, maar zorgen ze er ook voor dat de website **goed toegankelijk** is.

Enkele eenvoudig toepasbare ontwerpvereisten voor duurzame en archiveerbare **webpagina's** zijn:
- werk een duidelijke, uitbreidbare mappenstructuur voor de website uit
- zorg voor een duidelijke scheiding tussen enerzijds inhoud en structuur (HTML) en anderzijds opmaak (CSS)
- gebruik geen frames in de website
- maak vriendelijke, menselijk begrijpbare URL's
- gebruik absolute pathaanduidingen voor externe links en document-relatieve pathaanduidingen voor interne links



Websites worden steeds vaker aangestuurd vanuit een **CMS**. Ook voor het CMS zijn er richtlijnen. De beschikbare metadatavelden (titel/naam, uniek webadres, versienummer, enz.) moeten bijvoorbeeld vrij definieerbaar zijn door de organisatie. Andere voorbeelden van kwaliteitsvereisten voor het CMS zijn:
– het is mogelijk de verschillende versies van content-items en hun metadata bijhouden
– het vooziet de mogelijkheid dat gepubliceerde content-items pas na versieverandering worden gewijzigd
– het kan de verschillende versies van on line content op een statische wijze en als afzonderlijke objecten, en in combinatie met hun metadata, bewaren

Andere vereisten zijn van toepassing op de aan het CMS gekoppelde **databank**:
- de gegevens worden beheerd door een open databank management systeem
- de databank heeft een gedocumenteerd, overzichtelijk en uitbreidbaar datamodel

Meer informatie over deze kwaliteitsvereisten vindt u in: *"Digitaal Archiveren: Richtlijn en Advies nr.5: Websites-beheer en content management voor digitale archivering"*