

Binnen het DAVID-project werd een digitaal archiveringssysteem voor dynamische en interactieve informatiesystemen uitgewerkt. Deze bijdrage¹ is gewijd aan de uitgangspunten en de concepten van dit archiveringssysteem en aan de wijze waarop dit archiveringssysteem werd geïmplementeerd door het Stadsarchief Antwerpen, de archiefpartner in het DAVID-project.

1. Het DAVID-project

Het DAVID-project is het eerste Vlaamse onderzoeksproject over digitale archivering. DAVID is het Nederlandstalige acroniem voor 'Digitale Archivering in Vlaamse Instellingen en Diensten'. Het project heeft tot doel te onderzoeken hoe digitale archiefdocumenten gevormd door Vlaamse instellingen en diensten binnen hun context op een duurzame en authentieke wijze digitaal worden gearcheeerd. Het DAVID-project loopt vier jaar en bereidt tegen het einde van 2003 een handboek over digitale archivering voor. Het project richt zich zowel op digitale kantoordocumenten (tekstverwerking, spreadsheets, e-mails, etc.) als de archivering van archiefbescheiden die binnen dynamische en interactieve informatiesystemen worden gevormd. Voor beide types digitale archiefdocumenten werd ondertussen een archiveringssysteem uitgewerkt.

2. De informatiesystemen als uitgangspunt

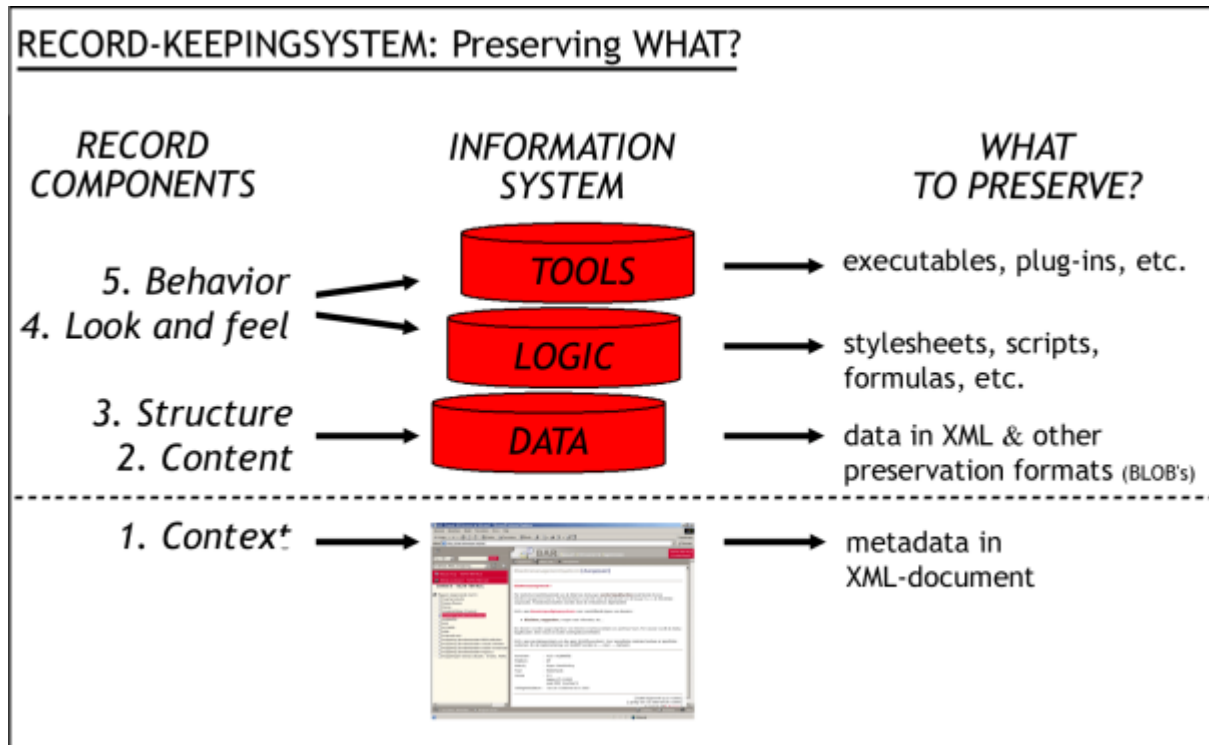
Eén van de initiële doelstellingen van het DAVID-project was het ontwerpen van een typologie van digitale archiefbescheiden die als basis voor de archiveringsstrategieën kan dienen. Het in kaart brengen van alle informatiesystemen die bij de administratie van de stad Antwerpen lopen, leidde tot het besluit dat een perfect sluitende typologie niet kon worden opgesteld, en ook geen zin heeft. Belangrijke vragen zoals WAT en HOE wordt gearcheeerd, kunnen enkel maar beantwoord worden wanneer men het informatiesysteem kent waarbinnen digitale archiefbescheiden worden gecreëerd en beheerd. Functionaliteiten, architectuur, koppelingen met andere informatiesystemen, organisatie van de data, enz. zijn belangrijke parameters die elk een invloed op de archiveringsstrategie kunnen hebben. Het uitgangspunt voor het archiveringssysteem is bijgevolg het informatiesysteem zelf².

Dit houdt onder meer in dat de archivaris op het ogenblik van archivering moet beschikken over gegevens over het informatiesysteem. Metadata over de informatiesystemen worden echter in zeer weinig administraties of IT-departementen systematisch of op gestructureerde wijze bijgehouden. Op het ogenblik van archivering beschikken archivariissen bijgevolg enkel over het informatiesysteem zelf, in het beste geval aangevuld met mondeling verstrekte informatie. Het spreekt voor zich dat dit een onvoldoende basis is voor belangrijke beslissingen zoals de identificatie van archiefbescheiden, de archiefwaardering en het uitstippelen van de archiveringsstrategie.

¹ Deze bijdrage werd geschreven voor de Digicult.Info nieuwsbrief en is gebaseerd op de lezing *Preservation of records from database-driven informationsystems* die werd gegeven op de ErpaWorkshop 'Long-term preservation of databases', Swiss Federal Archives, Bern, 9 april 2003.

² F. BOUDREZ, *Het digitaal archiveringssysteem: beheersinventaris, informatielagen en beslissingsmodel als uitgangspunt*, Antwerpen, 2001. Het concept van de beheersinventaris werd verder uitgewerkt in de DAVID-bijdrage: F. BOUDREZ, *Archief onder controle? De beheersinventaris als sleutelinstrument bij de archivering van digitale archiefbescheiden*, Antwerpen, 2002.

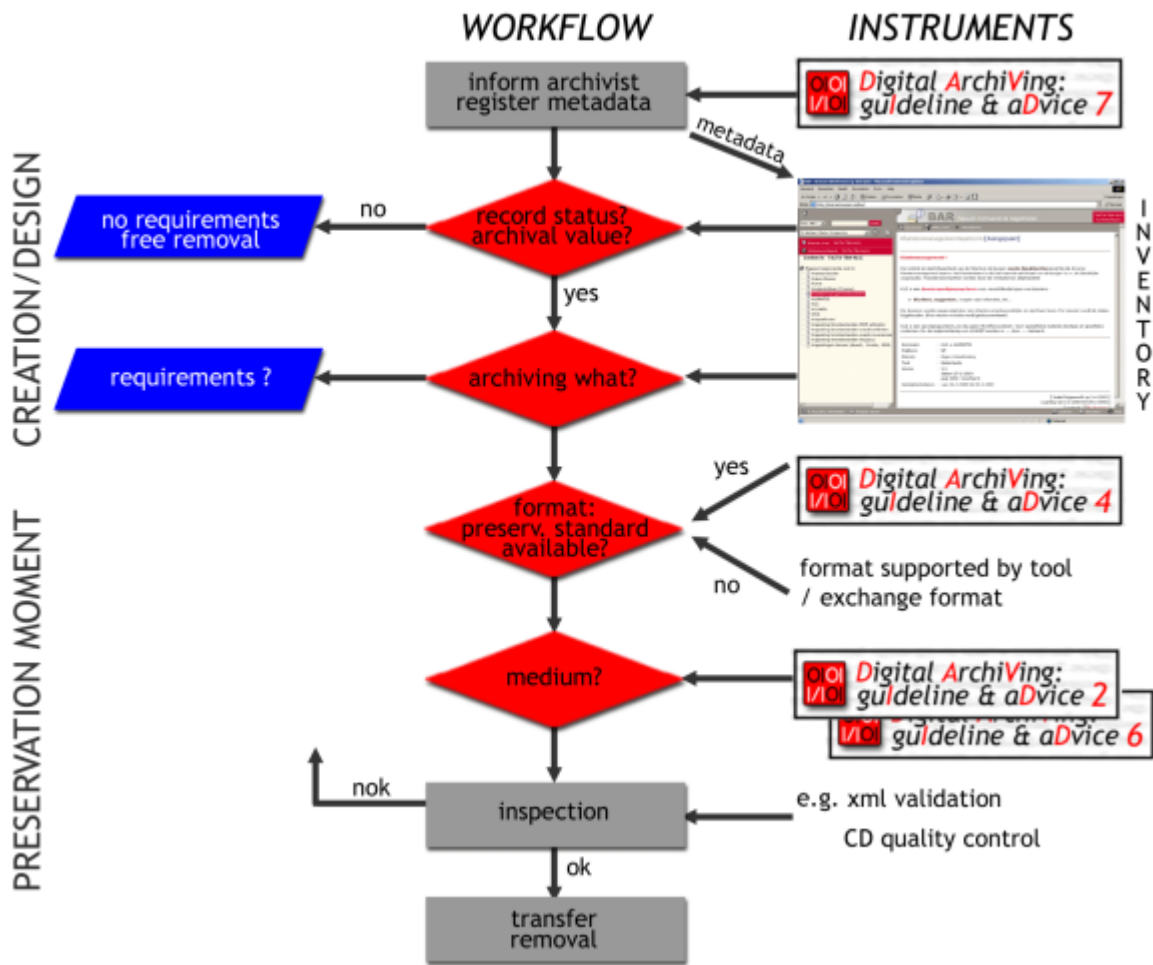
Om dit probleem op te vangen werd een nieuw archiefinstrument gecreëerd: de beheersinventaris van digitale informatiesystemen. In deze beheersinventaris houden administratieve medewerkers, systeemverantwoordelijken en de archivaris vanaf de creatie metadata bij over het digitaal informatiesysteem. De beheersinventaris van de stad Antwerpen is een relationele databank met webinterface en dynamisch datamodel. Het basisdatamodel voor deze beheersinventaris zijn de gegevensvelden die vanuit archiefstandpunt noodzakelijk zijn. Een dergelijke beheersinventaris kan echter ook andere doeleinden dienen zoals de helpdeskfunctie of het beheer van de IT-infrastructuur. Op die manier biedt de beheersinventaris een meerwaarde voor de hele organisatie en is de archivaris niet de enige belanghebbende partij voor het up-to-date houden van de beheersinventaris.



3. Procedure: van creatie tot archivering

Van bij de ontwikkeling van nieuwe informatiesystemen of de aanpassing van bestaande informatiesystemen wordt de beheersinventaris geactualiseerd en wordt de archiefdienst op de hoogte gebracht van de wijzigingen. Dit is een verplicht onderdeel van de IT-procedure bij de stad. De archiefdienst onderzoekt of er binnen het informatiesysteem archiefdocumenten met archiefwaarde worden gecreëerd of beheerd. Op dit ogenblik wordt eigenlijk al vastgelegd WAT op termijn zal worden gearchiveerd, en of er eventueel bijzondere kwaliteitsvereisten voor het informatiesysteem gelden. Deze kwaliteitsvereisten kunnen betrekking hebben op de bestandsformaten, het bijhouden en registreren van gestructuurde metadata, het waarborgen van de betrouwbaarheid, de encoding, enz. zodat er op termijn 'goed archiveerbare' digitale documenten worden gecreëerd. Het is belangrijk dat deze vereisten bekend zijn vooraleer het informatiesysteem of de wijzigingen in productie gaan. De archiefdienst wordt bij voorkeur betrokken bij het evalueren van nieuwe informatiesystemen of bij het samenstellen van het kwaliteitshandboek van het nieuwe of te wijzigen informatiesysteem.

RECORD-KEEPINGSYSTEM: Procedure



Voor de identificatie van de archiefbescheiden en de archiefwaardering worden informatiesystemen beschouwd als een samenstelling van drie interactieve componenten: de inhoud, de logica en de tools. De waardering moet leiden tot identificatie van de componenten die het document de status van record geven. Belangrijke vragen zijn hierbij of de volledige databank, een gedeelte van de data of enkel de gegenereerde output de records van het informatiesysteem zijn. Toegepast op een GIS-applicatie kan dit betekenen dat de data als GML (Geography Markup Language)-documenten worden gearchiveerd of enkel de kaarten als afbeeldingsbestanden worden bijgehouden (bijv. GeoTIFF, SVG). Van de applicatie voor het bijhouden van het bevolkingsregister worden op lange termijn enkel de bevolkingsgegevens gearchiveerd als XML-documenten. Op dit moment worden ook de grenzen van de databank vastgelegd. Immers, meer en meer informatiesystemen zijn aan elkaar gekoppeld en ontleen informatie aan elkaar. Of deze 'externe' data mee worden gearchiveerd hangt af van de vraag of de gekoppelde databank wordt gearchiveerd, en eventueel van de frequentie waarop dit gebeurt.

De frequentie van de archiveringsacties is in grote mate afhankelijk van de wijze waarop wijzigingen in de databank worden aangebracht. Worden de wijzigingen geregistreerd zonder dat de oude gegevens worden overschreven, dan zal de archiveringsfrequentie afhangen van de omvang van de databank en de performantie van het informatiesysteem. Bij databanken waarbij oude gegevens niet afzonderlijk worden bijgehouden, maar worden overschreven, zijn er verschillende opties. Men archiveert de basisversies en vervolgens alle wijzigingen of men archiveert met een hoge frequentie snapshots. In deze laatste optie heeft men natuurlijk het risico dat tussenliggende versies verloren gaan. Voor de archivering van alle versies zal een combinatie van deze methodes wellicht aangewezen zijn.

Typisch voor digitale archiefbescheiden is dat men hard- en software nodig heeft om de archiefdocumenten in de toekomst te reconstrueren. Bij de archiefwaardering mag men dus niet alleen aandacht besteden aan de inhoud van de databank alleen. Ook logica-elementen en tools kunnen voor archivering in aanmerking komen. Dit zal bijvoorbeeld meestal het geval zijn wanneer de originele 'look and feel' en functionaliteiten of gedragingen van archiefdocumenten mee worden gearchiveerd. De logicalaag van een informatiesysteem bestaat uit alle elementen die de input behandelen en de output genereren. De tools zijn dan de instrumenten voor input en output. De identificatie van de componenten die worden gearchiveerd, is dus niet alleen afhankelijk van de klassieke archiefcriteria, maar ook van de technische vereisten om in de toekomst de archiefdocumenten op een getrouwe wijze te reconstrueren.

Op het ogenblik van archivering wordt bepaald in welk bestandsformaat en op welk medium er wordt gearchiveerd. Het stadsarchief Antwerpen heeft zijn archiveringsstandaarden inzake bestandsformaten, -media en bestandssystemen vastgelegd in richtlijnen. Deze zijn gebaseerd op de algemene richtlijnen en adviezen van het DAVID-project. De archiefvormer bereidt samen met het IT-departement van de stad de neerlegging voor. De neergelegde archiefbescheiden en hun dragers worden geïnspecteerd bij hun aankomst op de archiefdienst. Voor de validatie van grote XML-bestanden werd een validerende parser geprogrammeerd. De kwaliteit van de CD's wordt getest met behulp van een diagnostool. Voldoet de neerlegging niet aan de vooropgestelde kwaliteitsvereisten dan keren de digitale archiefbescheiden voor correctie terug naar de archiefvormer. Van elke drager wordt twee exemplaren neergelegd: één exemplaar wordt bewaard in het archief terwijl de veiligheidskopie naar een andere lokatie wordt gebracht. Pas wanneer de neerlegging wordt goedgekeurd, mogen de archiefbescheiden uit het originele informatiesysteem worden verwijderd.

4. XML-archivering van de archiefbescheiden

Bij de archivering van databanken met tekstuele data wordt zoveel mogelijk gebruik gemaakt van eXtensible Markup Language (XML) als archiveringsformaat. XML biedt een interessante voordelen voor het digitaal archiveren van archiefdocumenten: gemakkelijk uitwisselbaar, geschikt voor gestructureerde tekstuele informatie, toepassing van een expliciet documentmodel, in hoge mate zelfbeschrijvend, enz³. Voor databanken die BLOB's (Binary Large Objects) bevatten wordt XML gebruikt als metadataformaat van de gearchiveerde documenten. Voor de digitale archiefdocumenten zelf die als BLOB in de databank werden bewaard, worden geschikte archiveringsformaten gebruikt.

Het migratieproces van archiefdocumenten binnen databanken naar XML-documenten bestaat uit verschillende stappen. De eerste stap is altijd het opstellen van een documentmodel voor de archiefdocumenten. Aanvankelijk werden hiervoor DTD's ontwikkeld, maar nu wordt geleidelijk aan overgeschakeld naar XML Schemas. Dit documentmodel wordt hoofdzakelijk gebaseerd op de structuur van de archiefdocumenten. Deze structuur kan identiek zijn aan de interne databankstructuur, maar dit is niet noodzakelijk. Voor hiërarchische databanken kunnen de COBOL copybooks in veel gevallen als basis dienen. Bij relationele databanken zit het archiefdocument doorgaans verspreid over verschillende tabellen. Een leidraad bij het samenstellen van een documentmodel is de wijze waarop input en vooral output in het actieve informatiesysteem aan de gebruiker werden gepresenteerd. De mapping van het relationele datamodel naar het hiërarchische documentmodel is niet altijd evident. Beide modellen hebben immers een aantal fundamentele verschillen. Door een goede nesting en het toekennen van semantische tags kan men de interne logica van de documenten in de gearchiveerde bestanden steken. Bij deze databanken wordt de unload van de data dan ook meestal voorafgegaan door query- en joinacties. Met stylesheets kan men indien nodig op een meer expliciete wijze bewaren hoe de archiefdocumenten werden getoond aan de gebruikers van het actieve systeem.

De unload van de databanken door tekstbestanden verdient bijzondere aandacht vanwege de encoding van de karakters. De karakters worden bij voorkeur naar Unicode omgezet. Vooral bij data met diakritische tekens

³ De voordelen van XML voor digitale archivering worden hier niet in extenso behandeld. Zie daarvoor: F. BOUDREZ, <XML> en digitaal archiveren, Antwerpen, 2002 (http://www.antwerpen.be/david/teksten/xml_digitaalarchiveren.pdf).

en waarover geen documentatie meer aanwezig is, durft dit nogal eens voor problemen te zorgen. In een volgende stap worden de voorbehouden XML-karakters vervangen door entities en worden de ongeldige XML-karakters (bijv. controlekarakters) weggefilterd. Dit gebeurt met een zelf geprogrammeerde tool. De tagging van de XML-elementen en het toevoegen van de XML-declaraties is vervolgens de laatste stap in het omzettingproces. Bij de keuze van de tagnamen verdient het natuurlijk aanbeveling om semantische tags te kiezen, al kan dit bij grote bestanden en het veelvuldig voorkomen van dezelfde tags tot een grote mate van redundantie leiden. Dit is zeker het geval wanneer alle velden binnen de databankrecords van tags worden voorzien. De tagging van de individuele velden biedt het voordeel dat de data van de velden individueel adresseerbaar is en dat databankfillers verwijderd kunnen worden. Het gebruiken van afkortingen of codes als tagnamen kan een gedeeltelijke oplossing bieden, maar dit betekent wel dat documentatie over de betekenis van tagnamen moet worden bijgehouden. Hierdoor verliezen de XML-documenten deels hun autonomie en zelfvoorziening.

5. Metadata

Op het ogenblik van archivering worden ook de metadata van de archiefdocumenten en van het informatiesysteem gearchiveerd.

Door zoveel mogelijk semantische XML-tags aan de velden en records toe te voegen, worden al belangrijke metadata in de archiefdocumenten zelf gearchiveerd. Hierdoor heeft men bijvoorbeeld geen externe documentatie meer nodig om de functie van karakters op bepaalde posities binnen de databankrecord te kennen. Bij het gebruik van afkortingen als tagnamen of codetabellen blijft men natuurlijk afhankelijk van externe documentatie om de betekenis van bepaalde velden te kennen.

Essentiële metadata over het informatiesysteem zijn al aanwezig in de beheersinventaris. Deze metadata worden geëxporteerd naar een XML-document en verder aangevuld met metadata over de archiveringsactie. Dit XML-document wordt samen met de archiefdocumenten gearchiveerd.

6. Meer informatie

Meer informatie over dit archiveringssysteem en over het DAVID-project is beschikbaar op de website: <http://www.antwerpen.be/david> . (contact: <mailto:david@stad.antwerpen.be>).