VAN BACKUP TOT GEARCHIVEERDE WEBSITE

De archivering van de eerste versies van de Digitale Metropool Antwerpen

Filip Boudrez
Antwerpen, 2002

0. Inhoudstafel

1. Inleiding 1
2. Recuperatie van de bestanden 1
3. Van absolute naar relatieve linken 3
4. Lessen 5
5. Linken 6

1. Inleiding

Antwerpen was de eerste Belgische stad met een eigen website op het World Wide Web. De website kreeg de naam Digitale Metropool Antwerpen (DMA). Websites van Amerikaanse en West-Europese pioniers zoals Digitale Stad Amsterdam dienden bij de ontwikkeling van de DMA als inspiratiebron. Versie 1.0 ging op 11 juni 1995 on line en drie dagen later werd de Antwerpse webstek tijdens de opening van het cybercafé in de openbare bibliotheek officieel aan de pers voorgesteld[1]. Halverwege december 1995 lanceerde men al versie 2.0. Beide versies hebben een historische waarde die hun archivering verantwoordt. In het najaar van 2001 werd de opname in het archiveringssysteem van het stadsarchief van Antwerpen uitgevoerd. De websites waren ondertussen bijna zes jaar oud. Dit is een relatief jonge leeftijd voor een archiefstuk, maar de archivering had toch heel wat voeten in de aarde. Het relaas van de opening van het eerste Belgische website-archief en de reconstructie van haar oudste archiefstukken[2].

2. Recuperatie van de bestanden

De stadsarchivaris van Antwerpen had de voorbije jaren de archivering van de oudste DMA-versies al herhaaldelijk onder de aandacht van de websiteverantwoordelijken gebracht. De stadsarchivaris kreeg te horen dat ze op beide oren mocht slapen: van elke versie stond er minstens één kopie op tape. Het archiveringsproject en de neerlegging bij het stadsarchief werd gestart met het formuleren van de kwaliteitseisen voor gearchiveerde websites[3]. De tapes zelf kwamen zeker niet voor archivering in aanmerking. Het stadsarchief beschikt niet over de nodige apparatuur om tapes in te lezen en verkiest bovendien optische dragers zoals CD-ROM's als opslagmedium voor digitale archiefbestanden. Bovendien bleek het om backuptapes van de webserver te gaan. Deze tapes hebben als bijkomend nadeel dat ze niet direct toegankelijk zijn, maar

---

1 Stadsarchief Antwerpen, websitearchief: DMA_versie1\deze_www\nieuws\speech.htm; Toespraak van schepen Bruno Peeters op woensdag 14 juni 1995.

2 Het tweede deel van deze bijdrage is grotendeels gebaseerd op het logboek dat Telepolismedewerker Peter Claes bijhield van de acties die hij uitvoerde om de inhoud van de tapes naar CD over te zetten (P. CLAES, *Recuperatie van DMA versies*).

3 De algemene archiveringsstrategie voor websites en de bijhorende kwaliteitsvereisten kunt u nalezen in het DAVID-rapport: *Het archiveren van websites*, Antwerpen-Leuven, 2002.

dat hun inhoud eerst teruggeplaatst en gedecomprimeerd moet worden. Hiervoor is de passende backupsoftware vereist.

De overzetting naar een andere drager van de gedecomprimeerde inhoud van de tapes was dus een eerste noodzakelijke stap in het archiveringsproces. Het was echter de vraag of dit wel zou lukken, want de eerste twijfels rezen: Waren de tapes volledig? Stonden er nog wel gegevens op de tapes? De tapes werden jaren lang zonder enige bijzondere zorg of bewaringsmaatregelen in een doos in de lokalen van de IT-verantwoordelijken bewaard. De laatste vraag die zich stelde, was nog belangrijker: Hoe moest men de oude tapes inlezen? Telepolis, het informaticacentrum van de stad, beschikte niet meer over de nodige hard- en softwareconfiguratie waarmee de tapes op een harde schijf konden worden teruggeplaatst. De backups waren immers gemaakt op een SPARCcomputer[4] met Solaris en later Linux als besturingssysteem en Telepolis gebruikt dit computersysteem ondertussen niet meer. De bestanden op de tape waren opgeslagen volgens het bestandensysteem van de computer waarmee de backups ooit waren gemaakt en deze computer was afgestoten. Daarenboven was er geen enkele documentatie voor handen. Het was dus hoogst onzeker dat het nog mogelijk was om de eerste versies van de DMA opnieuw in te lezen, laat staan te voldoen aan de kwaliteitseisen voor archivering.

De archivering van de DMA richtte zich dus eerst op een zoektocht naar een computer met hetzelfde bestandensysteem als de backuptapes. In de kelders van het Telepolisgebouw vond men nog een oude ZooMOO machine[5] gebaseerd op de SPARCarchitectuur en met Linux als besturingssysteem. Het afstoffen van deze oude computer was echter niet voldoende om de backuptapes te kunnen inlezen: op het rootsysteem stond een paswoord, er was geen diskette-, cd-rom- of tapestation aan de computer aangesloten en de computer maakte geen deel uit van het netwerk. Kreeg men de computer nog aan de praat? Hoe moest men de tapes inlezen? Hoe kon een de passende driver voor het tapestation worden geïnstalleerd? Hoe kon de teruggeplaatste inhoud op CD worden geplaatst? Deze computer was nochtans de enige hoop om de tapes te kunnen inlezen. Waar elders beschikte men nog over de nodige computerconfiguratie? Pas na lang zoeken werd het paswoord teruggevonden zodat de computer kon worden opgestart. Eerst werden een aantal daemons uitgeschakeld want het systeem was te zwaar belast. Achtereenvolgens werd de ZooMOO machine in het netwerk opgenomen en werd de tapedriver geïnstalleerd. Om dit laatste mogelijk te maken, werd de kernel van het besturingssysteem opnieuw geconfigureerd en gecompileerd. De tapedriver werd hierbij als onderdeel van de kernel opgenomen. Na het opnieuw opstarten van de computer werd het tapestation herkend.



Over de originele webserver en de computer waarmee backups werden gemaakt, was nagenoeg geen informatie bewaard. Op de DMA versie 1.5 stond wel een foto van de eerste webserver van Telepolis.

Men was er nu al in geslaagd om de nodige configuratie te reconstrueren, maar daarmee was de inhoud van de tapes nog niet teruggeplaatst. Het inlezen (van de inhoudstafel) van de tapes en het kopiëren van de bestanden naar de harde schijf van de ZooMOO computer verliep niet van een leien dakje. Mits enig stunt en vliegwerk kon men uiteindelijk deze problemen omzeilen en was de inhoud van de tapes naar de harde schijf gekopieerd. Via de netwerkverbinding werden de bestanden met FTP gekopieerd naar de harde schijf van een computer die met een CD-writer was uitgerust. De eerste twee versies van de DMA werden op CD gebrand zoals ze ooit op de webserver hadden gestaan.

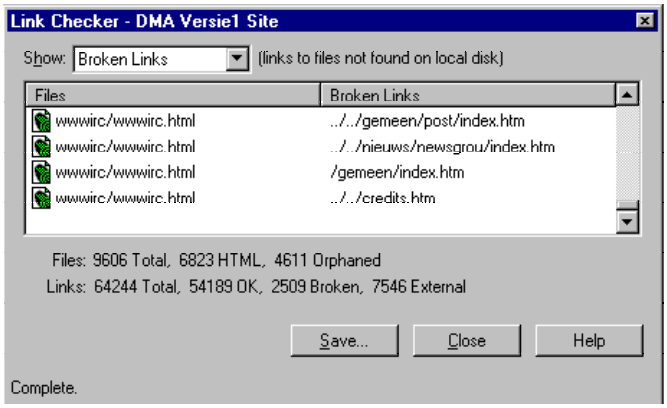
---

[4] SPARC (Scalable Processor ARChitecture) is de multi-processor architectuur van Sun Microsystems die gebaseerd is op RISC-technologie.

[5] ZooMOO staat voor Zoo Mud (multi user dungeon) Object Oriented. De ZooMOO was een telnettoepassing waarmee een virtuele gemeenschap als onderdeel van de DMA werd opgebouwd.

3. Van absolute naar relatieve linken

Beide CD's werden aan het stadsarchief overgedragen. Aangezien het om oude, en dus statische websites gaat, waren ze niet echt afhankelijk van een webserverconfiguratie en zijn ze platformonafhankelijk. De enige uitzondering hierop waren de interne absolute linken en de bezoekerstellers op basis van een CGI-script. Aangezien deze tellers bij archivering toch worden uitgeschakeld, vormde dit geen probleem. De links binnen beide versies zorgden daarentegen wel voor kopzorgen. Voor heel veel interne links was immers een absolute pathaanduiding gebruikt. De links verwezen naar de rootmap (bijv. <a href="/">, <a href="/antabc.htm">), naar de webserver (bijv. <img src="www.dma.be/graphics/redline.gif"> of er werd gewoon verwezen naar een default webpagina (<a href="/p/wijk/">). Voor de correcte omzetting van deze absolute linken is webserversoftware vereist, maar dit beantwoordt niet aan de doelstelling om zo systeemonafhankelijk mogelijk te archiveren. Bij het off line bekijken van de website vanop de CD's kreeg men bijgevolg veel foutmeldingen ("De pagina kan niet weergegeven worden") of vakjes met een rood kruisje op de plaats van de ontbrekende afbeeldingen te zien.



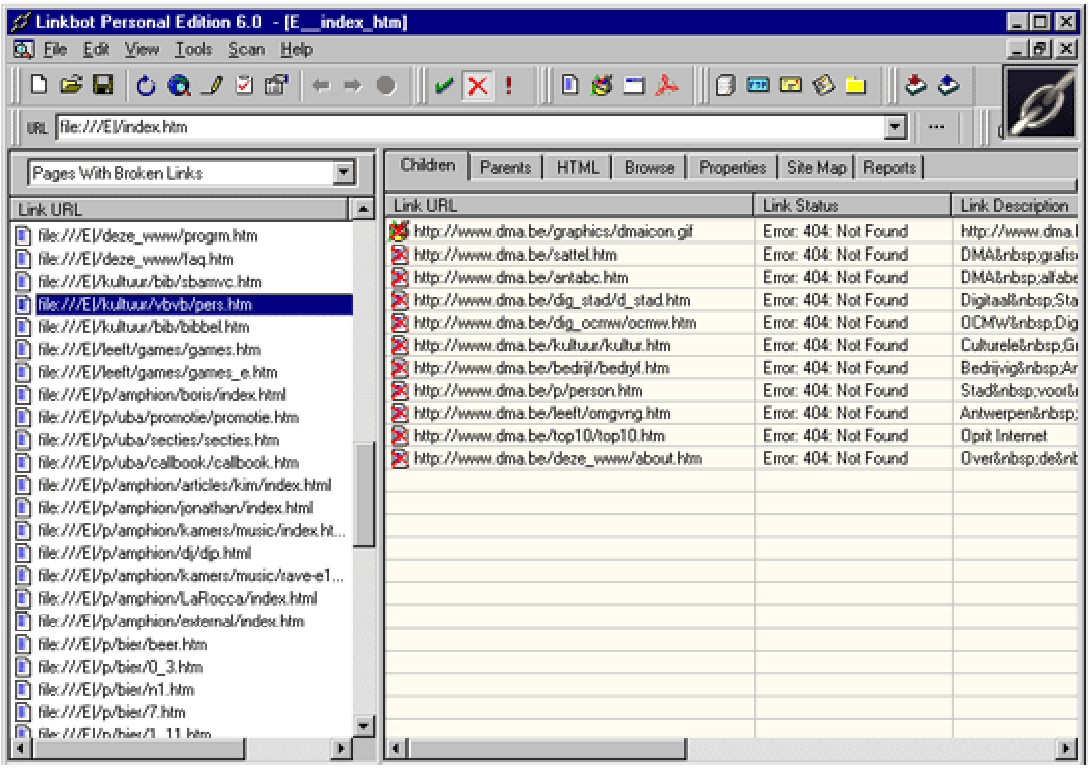
Om dit euvel te verhelpen moesten de interne absolute pathaanduidingen naar relatieve worden omgezet. Hiervoor zijn er twee opties. De eerste mogelijkheid is het terugplaatsen van beide websites op een webserver en ze off line plaatsen met behulp van een off line webbrowser. Het voordeel van deze optie is de automatische omzetting van absolute naar relatieve links. De bestandsomvang van beide versies is echter een grote belemmering. De eerste versie neemt 71,6 megabytes in beslag, de tweede versie 384 megabytes. Het correct downloaden van dergelijke hoeveelheden via het Internet is niet evident. Een tweede mogelijkheid is het manueel aanpassen van de absolute pathaanduidingen. Hiervoor kan op basis van de neergelegde CD's worden verder gewerkt, maar het nadeel van deze oplossing is natuurlijk zijn arbeidsintensiviteit. Met behulp van een linkchecker kon geverifieerd worden om hoeveel broken links het ging[6]. De controle van de eerste versie wees uit dat er 2509 broken links waren. In totaal telde de website meer dan 64000 links (zie afbeelding). Tijdens een oppervlakkige surfsessie werd echter vastgesteld dat de terugkerende 10 à 12 links onderaan elke overzichtspagina absoluut waren en dat deze pathaanduidingen verantwoordelijk waren voor het grootste deel van de broken links.

Uiteindelijk werd er gekozen voor de tweede aanpak en zouden al de broken interne links manueel naar relatieve pathaanduidingen worden omgezet. Twee redenen verantwoorden deze keuze. Ten eerste was er de doelstelling om ook zoveel mogelijk bestanden van de vorige versies te archiveren. Vanwege zijn historische waarde zou het wenselijk zijn om de DMA-versie 1.0 te archiveren, zoals die op 11 juni 1995 on line werd geplaatst. De backup die van de eerste versie was bewaard, heeft eigenlijk betrekking op versie 1.5. Nu is dat in het geval van de eerste versie van de DMA niet zo erg. Naar mate de website werd uitgebreid, werd ook het versienummer aangepast. Versie 1.5 is dus een uitbreiding van versie 1.0 en veel oude bestanden van vroegere versies zijn in de mappen op de webserver blijven staan en dus mee op de CD geplaatst. Deze bestanden zijn niet meer gelinkt aan de versie 1.5 en zouden bij het off line plaatsen met behulp van off line browser niet mee gearchiveerd worden. Door alle bestanden van de backup te archiveren, kan men de bestanden van vorige versies indien nodig nog afzonderlijk raadplegen. Het is alleen spijtig dat de echte allereerste websiteversie van de eerste Belgische stad op het WWW niet meer in surfbare vorm is bewaard.

---

[6] Een link checker kan een onderdeel zijn van een applicatie voor websitedesign (bijv. Dreamweaver, Frontpage) of kan een afzonderlijke applicatie zijn (bijv. SiteXpert, Linkbot, Xenu's Link Sleuth). Sharewareversies van deze laatste programma's zijn gemakkelijk te vinden op het Internet. Er zijn ook een aantal websites die een linkcheckservice aanbieden.

Ten tweede kon met behulp van het programma Linkbot exact opgespoord worden op welke webpagina welke broken interne links stonden. Zonder een dergelijk programma zou het handmatig verbeteren van de links ongebonnen werk zijn. De eerste versie van de DMA telt meer dan 6823 HTML-pagina's en 64244 links. Het handmatig opsporen van de webpagina's met broken links in de HTML-code op basis van trial en error is onmogelijk. Nu duidde het programma de webpagina's met broken links aan en werd er vermeld welke links niet naar een bepaald bestand verwezen. Het verbeteren van de links zelf was handwerk. De read-only CD's werden naar een harde schijf gekopieerd en het verbeterwerk kon starten. Vervelend was wel dat van alle bestanden het attribuut read-only moest worden uitgeschakeld. Eventjes werd nog aan grote "zoek en vervangoperatie" gedacht om steeds terugkerende broken links te verbeteren, maar dit werd afgewezen. De HTML-pagina's met deze links staan immers op verschillende niveau's binnen de mappenstructuur van de website zodat je nooit exact kon opgeven hoeveel niveau's de link moest terugkeren. Een "zoek en vervangoperatie" is hoogstens in één map per keer uitvoerbaar. Dit kon echter maar een beperkt aantal keren toegepast worden. Het verbeteren van alle interne broken links in beide versies van de DMA nam in totaal zo'n 20 uur in beslag. Het resultaat zijn twee websites waarop off line gesurft kan worden.

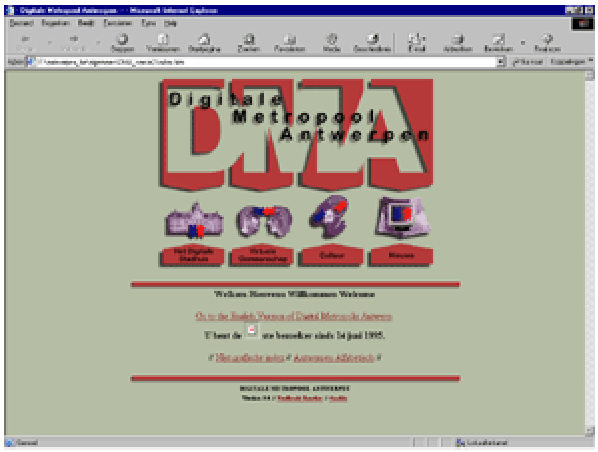


Met de applicatie Linkbot werden de webpagina's met broken links opgespoord. In het rechterscherm verscheen voor elk webpagina een lijst met broken links. Door op de bestandsnaam te klikken werd de broncode van de webpagina geopend. Vervolgens werd één voor één de broken links opgezocht en verbeterd. In dit voorbeeld verwijzen al de broken links naar de webserver en dienden de absolute links naar relatieve omgezet te worden.

Het Linkbotprogramma bracht naast de broken interne links nog een andere mankementen aan het licht. Beide DMA-versies bevatten broken externe links, maar dat is onvermijdelijk en behoeft geen verdere opvolging. Linkbot signaleerde ook webpagina's met broken ankers (links binnen dezelfde webpagina) en met ontbrekende of verkeerde attributen in de HTML-broncode. Beide soorten fouten vormen voor het ogenblik echter geen noemenswaardig probleem bij raadpleging, zodat het verbeteren geen onmiddellijke noodzaak is. Een aantal downloads in een verouderd bestandsformaat werd wel gemigreerd naar een recenter bestandsformaat of een geschikt archiveringsformaat. Zo werden een aantal teksten die in WordPerfect 5.1 beschikbaar waren, naar XML-bestanden omgezet. Na omzetting werden ook de linken naar deze documenten aangepast.



Startpagina van de DMA versie 1
De website kon op twee manieren worden bekeken: een textuele versie (zonder afbeeldingen) en een grafische versie (met afbeeldingen).



Startpagina van de DMA versie 2
Vanaf versie 2.0 werd gebruik gemaakt van frames. Voor de bezoekers wiens browser geen frames ondersteunde, was een frameloze versie beschikbaar.

4. Lessen

Uit de archiveringservaring van de eerste twee DMA-versies kunnen een aantal lessen worden getrokken. Deze conclusies zijn bruikbaar voor het archiveren van digitale bestanden in het algemeen, en niet alleen voor websites.

Vertrouw in geen enkel geval op backuptapes als vorm van digitale archivering. In het geval van de DMA 1.0 en 2.0 liep dit uiteindelijk nog goed af, maar de DMA 3.0 is verloren gegaan[7]. Backuptapes voldoen misschien wel aan de informaticanoden inzake data-archivering op korte termijn, maar beantwoorden zeker niet aan de vereisten voor archivering op lange termijn en dit omwille van meerdere redenen: het gebruik van compressie bij de opslag, de afhankelijkheid aan het backupprogramma voor de correcte decompressie, de afhankelijkheid van het backupprogramma aan een bepaald besturingssysteem, de afhankelijkheid van de tapes aan een bepaald bestandssysteem, de nood aan het juiste type tapestation en tapedriver, de tapes zijn niet rechtstreeks raadpleegbaar. Op grond hiervan moet men eigenlijk elke backuptape voor archivering afwijzen. Zorg er dus voor dat de digitale archiefbestanden zo snel mogelijk in gedecomprimeerde vorm worden overgebracht naar uitwisselbare en duurzame opslagmedia die wel geschikt zijn voor bewaring op lange termijn zoals een CD-ROM. In de meeste gevallen zal men dan nog maatregelen moeten nemen om de gegevens op een platformonafhankelijke wijze te archiveren, want backuptapes bevatten de computerbestanden zoals ze in het actieve informatiesysteem werden gebruikt. In het geval van de DMA was dit niet nodig, maar dienden de interne linken nog aangepast te worden.

De archivering op basis van de backuptapes hield in dat er enerzijds te weinig bestanden en anderzijds te veel bestanden worden gearchiveerd. Bij de websitecontrole met behulp van Linkbot bleek dat er een aantal broken links waren omdat ze verwezen naar bestanden die ontbraken. Deze bestanden stonden om één of andere reden niet op de backuptape. Aangezien de webserver met de originele websites niet meer beschikbaar is, kunnen deze bestanden niet meer gerecupereerd worden en zijn ze bijgevolg onherroepelijk verloren. Dit had vermeden kunnen worden door het archiveringsproces te starten toen de originele configuratie nog beschikbaar was. Ontbrekende bestanden konden dan nog van de webserver worden opgevist.

---

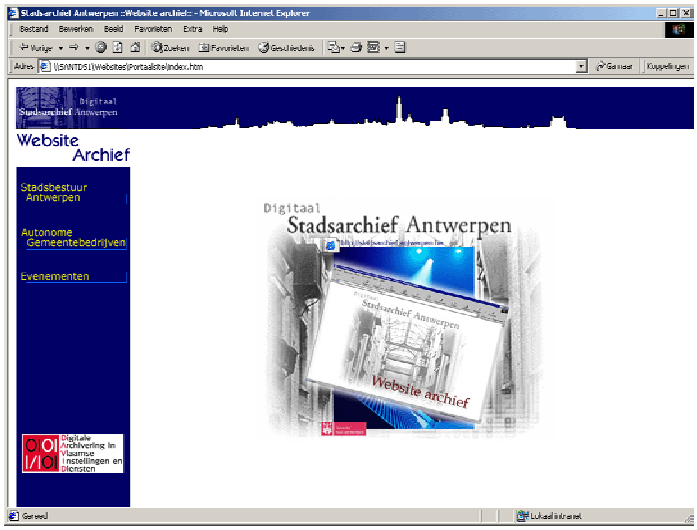
7 Op 18 maart 2002 deelde Wim Verstraeten (Telepolis Antwerpen) aan het stadsarchief Antwerpen mee dat de backuptape waar DMA 3.0 op stond inmiddels overschreven was. DMA 3.0 kan bijgevolg niet meer gerecupereerd en gearchiveerd worden.

De backups waren gemaakt van de harde schijven van de webserver. Dergelijke harde schijven zijn zelden een toonbeeld van rationeel en efficiënt bestandsbeheer. Op de harde schijven staan zowel verouderde als actieve bestanden. In het geval van de DMA betekent dit ongetwijfeld dat er computerbestanden gearchiveerd worden die eigenlijk geen deel meer uitmaken van die welbepaalde websiteversie, maar die op de webserverschijf in hun map waren blijven staan. Er worden dus eigenlijk te veel bestanden gearchiveerd. Het archiveren van de website met behulp van een off line browser kan hiervoor een oplossing bieden. Enkel de gelinkte bestanden worden dan off line geplaatst. Bestanden waar geen enkele andere webpagina naar verwijst, worden niet off line geplaatst want een off line browser volgt systematisch de links die hij aantreft. Uiteindelijk werd deze optie niet gevolgd om nog zoveel mogelijk bestanden van oudere versies te bewaren.

Het starten van het archiveringsproces tijdens de actieve fase van het informatiesysteem moet er niet alleen toe leiden dat men alle bestanden archiveert maar dat men ook onmiddellijk al de nodige contextuele gegevens bijhoudt. Contextuele gegevens over de websites werden nergens expliciet bijgehouden. Het verzamelen van de metadata moest dus volledig achteraf gebeuren en de gearchiveerde websites zelf waren de enige bruikbare bron. Voor de technische metadata is het normaal dat het archiefstuk zelf als bron dient, maar voor de andere metadata is dat meestal een problematische situatie. Voor veel digitale archiefbescheiden betekent dit gewoon dat belangrijke metadata verloren zijn. Dit was ook zo voor de gearchiveerde websites. Op één webpagina werden wel een aantal gegevens met betrekking tot de historiek gepubliceerd. Zo konden we achterhalen wie de ontwerpers van de website waren, wat er nieuw was aan versie 1.5, enz. maar telleraantallen heeft men niet op een statische wijze vastgelegd. Hoeveel mensen dus ooit de eerste twee versies bezochten, blijft dus een raadsel. We hebben er ook het raden naar wat er nieuw was aan de versies 1.1, 1.2, 1.3 en 1.4., welke niet meer gelinkte bestanden tot welke versie behoren en tot wanneer versie 1.5 on line was. Dit toont duidelijk het belang aan om van bij de creatie van digitale documenten actief op te treden en onmiddellijk de nodige documentatie bij te houden. De eerste stap is natuurlijk het definiëren van welke metadata moeten worden vastgelegd. Het feit dat de enige bekende gegevens over de webserver uit de website zelf kon worden afgeleid, illustreert dit.

5. Besluit

In januari 2002 was het website-archief van het stadsarchief Antwerpen operationeel en konden de onderzoekers de gearchiveerde websites en hun metadata in de leeszaal raadplegen. De websites van de stad Antwerpen nemen een prominente plaats in dit website-archief in. De archivering van de eerste versies van de DMA heeft nochtans aan een zijden draadje gehangen. Het opnieuw samenstellen van een computerconfiguratie die de backuptapes kon inlezen was een kleine krachttoer. Bovendien mag men van geluk spreken dat de tapes niet beschadigd waren en nog nagenoeg alle bestanden bevatten. Het manueel aanpassen van de links was een groot karwei. Uit deze ervaring zijn in ieder geval al de nodige conclusies getrokken, want de archivering van de huidige DMA-versie (versie 5, on line vanaf 1 februari 2001) is momenteel al in uitvoering.



De portaalsite waarlangs de gebruikers toegang krijgen tot de gearchiveerde websites en hun metadata.