FROM BACKUP TO ARCHIVED WEBSITE

Preserving the legacy websites of the city of Antwerp

Filip Boudrez
Antwerpen, 2002

0. Table of content

1. Introduction

Antwerp was the first Belgian city who had an internetsite on the World Wide Web. The website was called *Digitale Metropool Antwerpen* (DMA). Pioneer websites of American and West-European cities such as *Digitale Stad Amsterdam* served as source of inspiration for the development of DMA. Version 1.0 went online on June 11th 1995 and was introduced officially to the press three days later, on the occassion of the inauguration of the public library 's cybercafé[1]. Version 2.0 had allready been launched halfway through December 1995. Both versions have an historical value which justifies their preservation. These versions were transferred to the Antwerp City Archives in the fall of 2001. Meanwhile, these websites were almost six years old. Allthough this is relatively young for a record, the record-keeping process was nevertheless difficult. This is the story of the start of the first Belgian website archive and the reconstruction of her oldest records[2].

2. Recuperation of the computerfiles

The city archivist of Antwerp had discussed the problem of archiving the oldest DMA-versions several times with the webmasters in the past. The city archivist was told not to worry; for every version at least one copy was safely saved on tape. The real archivingproject and the transfer to the city archives was started with the formulation of the quality demands for archived websites in 2001[3]. The tapes themselves certainly didn't qualify for record-keeping purposes. The city archives don't have the necessary equipment at their disposal to read tapes and furthermore, prefers optical carriers such as cd-rom's as storage medium for electronic records. Moreover, when we had a closer look at the content of the tapes, they turned out to be backuptapes of the webserver. An additional disadvantage of these tapes is that they aren't directly accessible. Firstly, their content needs to be replaced and decompressed, fow which the appropriate backupsoftware is required.

So, transferring the decompressed content from the tapes to another carrier was the first necessary step in the record-keeping process. It was questionable whether this would work because allready the first doubts arose: Were the tapes complete? Was there any data left on the tapes? The tapes were left several years without any special care or preservation measures in a box in the offices of the IT-staff. The final question that remained was even more important: How could the content of the old tapes be restored? Telepolis, the IT center of the city, didn't dispose of the necessary hard- and softewareconfiguration anymore

[1] Antwerp City Archives, website archives: DMA_versie1\deze_www\nieuws\speech.htm; Speech of alderman Bruno Peeters on June 14, 1995.

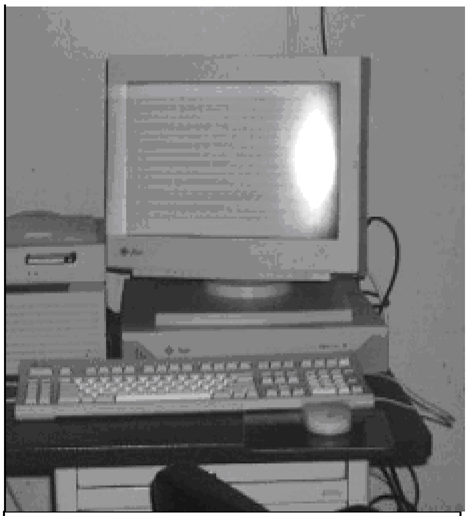
[2] The second part of this article is based mainly on the logbook that Peter Claes (Telepolis) kept during the actions he took to transfer the tapes to cd's. (P. CLAES, *Recuperatie van DMA versies*).

[3] The general archival strategies for websites and the quality demands can be found in the DAVID-report: *Archiving websites,* Antwerpen-Leuven, 2002.

to restore the tapes to a harddisk. The backups were made in a vendor dependant backupformat on a SPARCcomputer[4] with Solaris and later an old Unix version as operatingsystem. Telepolis wasn't using this computersystem and backupsoftware anymore. The files on the tape were also recorded according to the filesystem of the computer with which the backups were once made and a similar computer wasn't running anymore. On top of this, there was no documentation available. So, the possibility to reconstruct the first versions of DMA was highly doubtful; let alone that they could live up to the quality demands for record-keeping.

To be able to archive the DMA, the first thing that needed to be done, was to search for a computer with the same filesysteem as the backuptapes. An old ZooMOO machine, based on the SPARCarchitecture and with a Linux operatingsystem, was found in the cellar of the Telepolis building. Dusting the old computer did not suffice to read the backuptapes: the rootsystem had a pasword, there was no disk-, cd-rom- or tapestation joined to the computer, and was not connected to a network. Would it be possible to make the computer work again? How would it be possible to read the tapes? How could a suitable driver be installed for the tapestation? How would it be possible to transfer the content that had been put back to a cd?

Nonetheless, this computer was the only hope to re-read these tapes. Where else would such computerconfiguration be available? Only after an extensive search, the pasword was found so the computer could be restarted. First, a number of daemons were disactivated because the system was overloaded. Subsequently the ZooMOO machine was integrated in the network and the tapedriver was installed. To enable this, the kernel of the operatingsystem was re-configurated and compiled. Hereby, the tapedriver was installed as a part of the kernel. After restarting the computer, the tapestation was recognized.



Almost no information was retrieved on the original webserver and -computer. A photo of the first webserver of the city of Antwerp was retrieved on DMA 1.5

The necessary configuration was reconstructed, but this did not mean that the content of the tapes was put back in its place. Reading the table of contents on the tapes and copying the files to the harddisk of the ZooMoo computer didn't go without any problems. These problems were solved by some unorthodox solutions, and the content of the tapes was transferred to the harddisk. The files were copied with FTP to the harddisk of a computer with a cd-writer through the network. The first two versions of the DMA were burned onto cd as they had once been available on the webserver.

3. From absolute to relative links

Both cd's were transferred to the city archives. Seeing that these were old, and therefore statical websites, they didn't depend on webserverconfigurations and were therefore for 99 % platform independent. The only exception to this, were the internal, absolute links and the counter which was based on a CGI-script. Seeing that these counters are inactivated when archiving, this wasn't a problem. The links within these versions did cause some concernes. For many internal links, absolute paths were used. These links referred to the rootmap (for instance <a href="/">, <a href="/antabc.htm">), the webserver (for instance <img src="www.dma.be/graphics/redline.gif"> or there was a simple reference to a default webpage (<a href="/p/wijk">) To interpretet these absolute links correctly, webserversoftware is needed but this goes against the goal to archive as system independent as possible. When watching the website off line on cd, a lot of error messages were shown ("the requested page could not be found") or boxes with a red cross on the place of the missing image appeared.

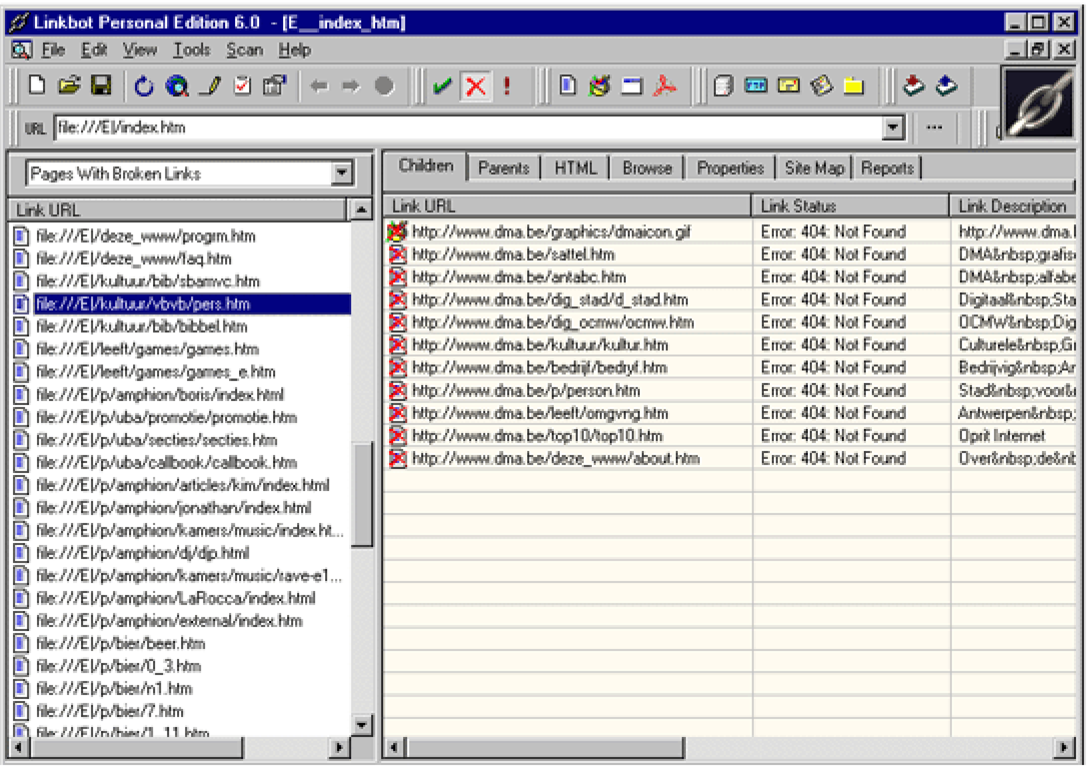
[4] SPARC (Scalable Processor ARChitecture) is the multi-processor architecture of Sun Microsystems which is based on RISC-technology.

To overcome this problem, internal absolute paths needed to be converted to document-relative links. Two options could be used. The first possibility was to place both websites back on a webserver and to place them off-line with the aid of an off-line webbrowser. This offered the advantage that the absolute links were converted automatically to relative links. The fileseize of both versions is a large impediment. The first version takes up 71,6 megabytes, the second version 384 megabytes. Correct downloading such large quantities through the net wasn't obvious. A second possibility is adjusting the absolute paths manually. The deposited cd's could be used for this, but the disadvantage was the labour-intensiveness. The amount of broken links was verified with the aid of a linkchecker.[5] The control of the first version revealed 2509 broken links. The website had a total of more than 64000 links (see image). However, during a superficial surfsession it was determined that the re-occuring links on the bottom of every overview page were absolute, and that these paths were responsable for the major part of the broken links.

Finally, it was deceided to chose the second approach and all the broken, internal links were converted manually. Two reasons accounted for this choice. Firstly, their was the goal to archive as many files of the previous versions as possible. Because of its historical value, it would be desirable to archive the DMA version 1.0 as it was placed online on June 11th, 1995. The backup of the first version that was saved, was actually version 1.5. This is not a bad situation in the case of the first DMA version. As the website expanded, the version number was adjusted. Version 1.5. was more extensive than version 1.0 and a lot of the old files of the previous version were left in folders on the webserver and are therefore burned onto the CD. These files aren't linked to the 1.5 version and, when placed off line with the aid of an off line browser, wouldn't be archived along. When archiving all the files of the backup, it is possible to consult the previous versions seperately. It is regretable that the very first real websiteversion of the first Belgian city on the WWW isn't consultable in a surfable way.

Secondly, it was possible to trace the webpages with internal broken links with the aid of the programme Linkbot. Without such a programme, the manual correction of the links would be an hopeless assignement. The first version of the DMA existed of more than 6823 HTML-pages and 64244 links. Searching the webpages with a broken link in the HTML-code manually on the basis of trial and error is impossible. Now the programme indicated the webpages with broken links and it mentioned which links didn't refer to a certain file. Correcting the links itself, was manual labour. The read-only cd's were copied to a harddisk so the correctionwork could start. Annoying was that with all the files, the read-only attribute needed to be switched off. A large "search and replace operation" was considered to adjust returning broken links, but this idea was rejected. The HTML-pages with these links were situated on several levels within the filestructure of the website so that it was impossible to indicate how many levels the links had to return. A "search and replace operation" can be executed in one folder at the time at the most. This process could only be applied a limited number of times. Correcting all the broken links in both DMA versions took about 20 hours. The result are two websites on which are off line browsable.

[5] A link checker can be a functionality in an application for websitedesign (bijv. SiteXpert, Linkbot, Xenu's Link Sleuth). Sharewareversions of these latest programs are easy to find on the Internet. Several websites offer a link checkservice.


Linkbot Personal Edition 6.0 - [E__index_htm]
File Edit View Tools Scan Help
URL file:///E|/index.htm
Children Parents HTML Browse Properties Site Map Reports

Pages With Broken Links

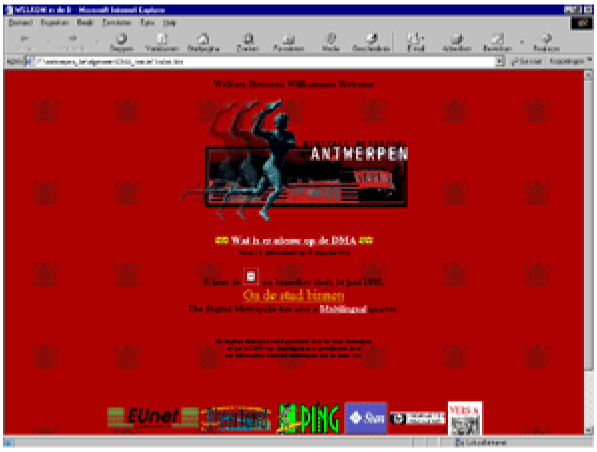
Link URL
file:///E|/deze_www/progm.htm
file:///E|/deze_www/faq.htm
file:///E|/kultuur/bib/sbamvc.htm
file:///E|/kultuur/vbvb/pers.htm
file:///E|/kultuur/bib/bibbel.htm
file:///E|/leeft/games/games.htm
file:///E|/leeft/games/games_e.htm
file:///E|/p/amphion/boris/index.html
file:///E|/p/uba/promotie/promotie.htm
file:///E|/p/uba/secties/secties.htm
file:///E|/p/uba/callbook/callbook.htm
file:///E|/p/amphion/articles/kim/index.html
file:///E|/p/amphion/jonathan/index.html
file:///E|/p/amphion/kamers/music/index.ht...
file:///E|/p/amphion/dj/dp.html
file:///E|/p/amphion/kamers/music/rave-e1...
file:///E|/p/amphion/LaRocca/index.html
file:///E|/p/amphion/external/index.htm
file:///E|/p/bier/beer.htm
file:///E|/p/bier/0_3.htm
file:///E|/p/bier/n1.htm
file:///E|/p/bier/7.htm
file:///E|/p/bier/1_11.htm

Link URL	Link Status	Link Description
http://www.dma.be/graphics/dmaicon.gif	Error: 404: Not Found	http://www.dma.l
http://www.dma.be/sattel.htm	Error: 404: Not Found	DMA grafis
http://www.dma.be/antabc.htm	Error: 404: Not Found	DMA alfabe
http://www.dma.be/dig_stad/d_stad.htm	Error: 404: Not Found	Digitaal Sta
http://www.dma.be/dig_ocmw/ocmw.htm	Error: 404: Not Found	OCMW Dig
http://www.dma.be/kultuur/kultur.htm	Error: 404: Not Found	Culturele Gi
http://www.dma.be/bedrijf/bedryf.htm	Error: 404: Not Found	Bedrijvig Ar
http://www.dma.be/p/person.htm	Error: 404: Not Found	Stad voor&i
http://www.dma.be/leeft/omgvrg.htm	Error: 404: Not Found	Antwerpen
http://www.dma.be/top10/top10.htm	Error: 404: Not Found	Oprit Internet
http://www.dma.be/deze_www/about.htm	Error: 404: Not Found	Over de&nt

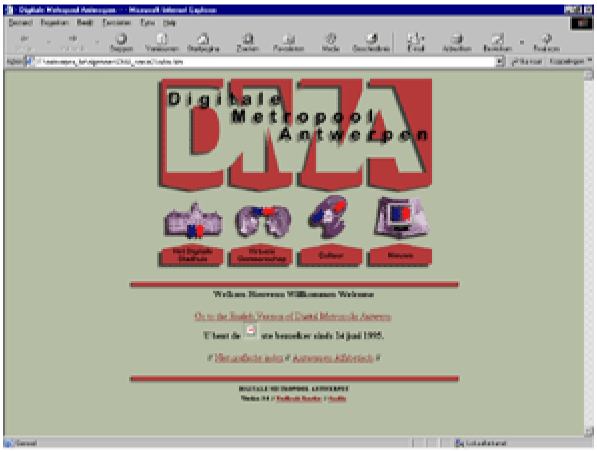
With the application Linkbot, the webpages with the broken links were localized. The rightscreen showed a list with broken links for every webpage. By clicking on the filename, the sourcecode of the webpage was opened. Subsequently the broken links were searched and corrected one by one. All the broken links refer to the webserver in this example. The absolute links had to be converted into document-relative links.

The Linkbot programme also brought other flaws into the open. Both DMA-versions contained broken external links, but that is inevitable and does not need further follow-up immediately. Linkbot also found webpages with broken ankers (links within the same webpage) and with missing or wrong attributes in the HTML-sourcecode. None of these kinds of mistakes pose a noteworthy problem when consulted, so that the correction is not an immediate necessity. In the future however, some time will be foreseen to correct this errors.

A number of downloads in an outdated fileformat were emigrated to a more recent fileformat or a more suitable archivingformat. A number of texts which were only available in WordPerfect 5.1, were transferred to XML-files. After the conversion, the links to these documents were adjusted.



Startpage of the DMA version 1
The website could be viewed in two ways: a textual version (without images) and a graphic version (with images).



Startpage of the DMA version 2
Frames were used since the 2.0 version. A frameless version was available for visitors whose browser did not support frames.

4. Lessons

A number of lessons can be learned from the archiving experience from the first two DMA-versions. These conclusions are useful when archiving electronic files in general, and not only for websites.

Put under no circumstances your trust in backuptapes as a kind of electronic record-keeping. Eventually, it ended well in the case of the DMA version 1.0 and 2.0, but DMA 3.0 which was also stored on backuptape was lost[6]. Backuptapes may meet the requirements of the IT concerning data-preservation in the short run, but they do not meet the needs for long-term record-keeping and this because of several reasons: the use of compression, the dependency on the backup programme when decompressing correctly, the bondage of the backup programme to a certain kind of operatingsystem, the dependence of the tapes to a certain filesystem, the need for a suitable type of tapestation and tapedriver, the tapes aren't accessible directly, the storage of same data on the backupcomputer, etc. On the basis of this, all backuptapes should be declined for archiving purposes. Make sure that all the electronic records are transferred as soon as possible in decompressed form to exchangeable and durable storagemedia which are suitable for preservation in the long run, such as the cd-rom. In most cases, measures will have to be taken to archive the data in a platform independent way, because backuptapes contain the computerfiles as they are used in the active informationsystem. This wasn't necessary in the case of the DMA, but the internal links needed to be adjusted.

The record-keeping on basis of the backuptapes meant that, on the one hand too few files, and on the other hand too many files were archived. The website control with Linkbot showed that a number of links were broken because they referred to files that were missing. These files were, for some reason, not on the backuptape. Seeing that the webserver with the original websites isn't available anymore, these files cannot be recuperated and are therefore irretrievable lost. This could have been avoided by starting the record-keepingproces when the original configuration was still available. Missing files could have been retrieved from the webserver.

The backups were made from the harddisks of the websserver. These harddisks are rarely the example of rational and efficient filemanagement. Harddisks contain outdated as wel as active files. In the case of the DMA this undoubtedly means that computerfiles were archived which weren't part of that particular webversion, but which were left on the webserverdisk in their folder. This means that too many files were archived.

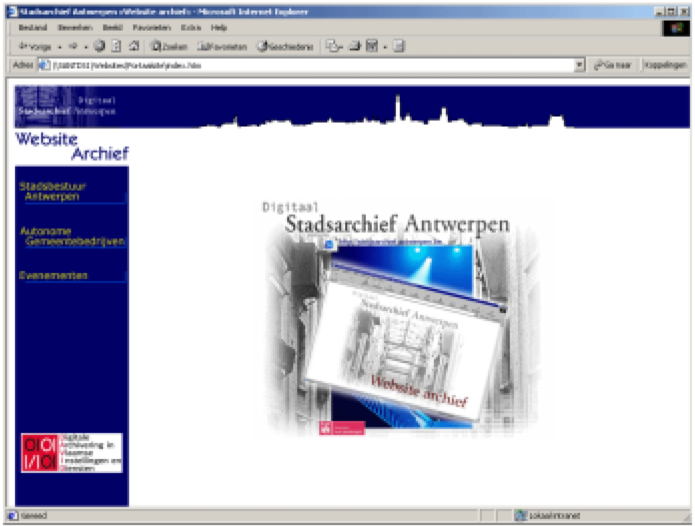
Making snapshots of the website with the aid of an off line browser could offer a solution. Only the linked files are placed off line. Files to which no other webpage refer, aren't placed off line because an off line browser follows automatically the links which it encounters. Eventually this option wasn't used, to ensure that as many files of outdated versions as possible could be saved.

Starting the archival process during the active fase of the informationsystem should not only lead to the record-keeping of the records, but also to the immediate preservation of the necessary contextual metadata. Contextual data on websites isn't kept explicitly anywhere. The gathering of the metadata had to be done completely afterwards and the archived websites were the only useful sources. It is normal for the technical metadata that the record itself serves as source, but this is usually a problematic situation for the other metadata. For many electronic records, this means that important metadata is lost. This was also the case for the archived websites. A number of information concerning the history was published on one webpage. It was possible to determine who the designers of the website were, the new features version 1.5 had to offer etcetera. But counternumbers weren't kept in a statistical way. It remains a mystery how many people visited the first two versions. We can only guess which were the new features on versions 1.1, 1.2, 1.3 and 1.4, which unlinked files that weren't linked anymore, belongs to which version, and till when version 1.5 was on line. This clearly shows the importance to act actively from the moment of creation of electronic documents and to keep the necessary documentation immediately. Ofcourse, the first step is to define which metadata needs to be preserved. The fact that the only known metadata on the webserver could be deduced from the website itself, illustrates this.

[6] On March 18, 2002 Wim Verstraeten (Telepolis Antwerp) told the Antwerp City Archives that the backuptape on which DMA 3.0. was recorded, had been overwritten. Consequently, DMA 3.0. cannot be recuperated and archived.

5. Conclusion

In January 2002 the website archive of the Antwerp City Archives was operational and the visitors could consult the archived websites and their metadata in the reading-room. The websites of the city of Antwerp take a prominent place in this website archive. Allthough the preservation of the first versions of the DMA trembled in the balance. Re-composing a computerconfiguration which could read the backuptapes was a small labour of Hercules. Besides, it was shere luck that the tapes weren't damaged and still contained nearly all files. Manually adjusting all the links was a substantual labour. The necessary conclusions were drawn from this experience, because the record-keeping of the present DMA-version (version 5, on line since February 1th, 2001) is at present in implementation.



The portalsite through which the users get access to the archived websites and their metadata.