



DAVID

Archiving websites

Filip Boudrez
Sofie Van den Eynde



FACULTEIT RECHTSGELEERDHEID
INTERDISCIPLINAIR CENTRUM VOOR RECHT EN
INFORMATICA
TIENSESTRAAT 41
B-3000 LEUVEN



Stadsarchief
Stad Antwerpen

Version 1.0

Legal depot D/2002/9.213/2

Antwerp - Leuven, July 2002

E-mail address: david@stad.antwerpen.be

Website DAVID project: <http://www.antwerpen.be/david>

TABLE OF CONTENTS

TABLE OF CONTENTS.....	3
I. INTRODUCTION	6
II. INTERNET, WWW & WEBSITES	8
III. THE IMPORTANCE OF ARCHIVING WEBSITES	13
IV. CURRENT INITIATIVES	16
V. ARCHIVING WEBSITES.....	19
A. QUALITY REQUIREMENTS FOR ARCHIVED WEBSITES	20
B. SELECTION: WHAT TO ARCHIVE?	21
B.1 The acquisition policy: what websites to archive?	21
B.2 What part of websites to archive?	23
B.2.1 Websites with static content	24
B.2.2 Websites with dynamic content.....	26
B.2.3 Conclusion.....	29
C. HOW TO ARCHIVE WEBSITES?.....	30
C.1 Websites with static content	30
C.2 Websites with dynamic content.....	31
C.2.1 Snapshots.....	31
C.2.2 Log files.....	35
C.2.3 Databases.....	35
D. FREQUENCY	35
D.1. Determination of the frequency.....	35
D.2. How to archive different versions?.....	36
D.2.1 The changes to websites with static content.....	36
D.2.2 Changes to websites with dynamic content.....	37
E. DIGITAL DURABILITY.....	37
F. MANAGEMENT OF THE ARCHIVED WEBSITES	42
F.1 Metadata	42
a) Authenticity details.....	45
b) Date on which the website was put on-line	46
F.2 Secure storage.....	47
F.3 Storage on media	47
H. AVAILABILITY	48
VI. ELABORATING THE RECORDKEEPING SYSTEM.....	49
A. ROLE OF THE ARCHIVIST.....	49
B. EFFICIENT DESIGN AND MANAGEMENT	50

C.	ARCHIVING	53
VII.	COPYRIGHT: A BARRIER WHEN ARCHIVING WEBSITES.....	55
A.	INTRODUCTION	55
B.	COPYRIGHT IN BRIEF.....	56
B.1.	Protected works	56
a)	Literary works and works of art	56
b)	And websites?.....	57
c)	Hyperlinks	58
d)	Exceptions for official government acts.....	59
B.2.	Who holds the copyrights?	59
B.3.	What do copyrights imply for their holders?.....	60
a)	Rights of property.....	60
b)	Moral rights	62
B.4.	How does one receive copyright protection?	62
C.	ARCHIVING WEBSITES: PROBLEM DEFINITION.....	63
C.1.	Acts of reproduction.....	64
C.2.	Changes	67
C.3.	Digital sustainability.....	67
C.4.	Availability.....	68
C.5.	Special regimes.....	68
a)	Computer programmes	68
b)	Databases.....	70
C.6.	Conclusion.....	72
D.	SOLUTION: EXCEPTION FOR PRESERVATION PURPOSES.....	72
VIII.	ARCHIVING PERSONAL DATA.....	75
A.	PERSONAL DATA AND THE INTERNET	75
A.1.	Organisation and management of the Internet.....	76
A.2.	IP addresses	78
A.3.	Log files.....	80
B.	PROCESSING PERSONAL DATA FOR HISTORICAL PURPOSES: COMPATIBLE WITH THE ORIGINAL GOAL?	80
IX.	THE FLEMISH GOVERNMENTS ON THE INTERNET	82
A.	THE SITUATION IN 2002	82
B.	CASE: THE ELECTRONIC PORTAL POPULATION REGISTER OF CEVI	83
X.	GOVERNMENT LIABILITY FOR ITS OWN WEBSITE	84
A.	CAN A CITIZEN OBTAIN RIGHTS FROM THE CONTENT OF A GOVERNMENT WEBSITE?	84
B.	DISCLAIMERS: USEFUL OR USELESS?.....	86
XI.	PORTAL SITES: THE FUTURE	87
A.	FROM INFORMATION TO INTERACTION AND INTEGRATION	87
B.	ONE VIRTUAL GOVERNMENT	88

C.	THE LEGAL FRAMEWORK.....	88
C.1.	Front office: the electronic identity card.....	88
C.2.	Back office: unique identification number.....	90
XII.	CONCLUSIONS & RECOMMENDATIONS.....	91
XIII.	BIBLIOGRAPHY.....	92
A.	ELECTRONIC RECORDKEEPING.....	92
B.	LEGISLATION AND RULES.....	94

I. INTRODUCTION

Websites no longer form a *terra incognita* for archivists. Archival institutions and records offices are increasingly using this Internet technology to make information and services available on-line for their customers. In some cases the interactive service is very well developed and the researcher can even consult digital records on-line. This brings the digital reading room closer into the people's homes. But, there is also a second way in which archival institutions and records offices are facing this new technology: the archiving of websites. This report focuses on the question how websites should best be archived.

It is commonly agreed that an enormous amount of information is available on the Internet. The success of the Internet is due to a variety of different factors. One of the main factors is the speed at which information is available worldwide and at which it can be adapted. This information volatility has made an investigation necessary into the ways this information should be stored. The first Swedish electronic newsletter, for example, has been lost¹. The official website of the Olympic Games of 2000 is no longer on-line. The need to archive information on the Internet became obvious at an early stage. Websites sometimes have a section "archive" where previous versions remain available on-line, or some obsolete web pages remain active but contain a banner "archived"². However these are individual initiatives of web designers or content managers who feel the need to keep older versions of web pages available for whatever reason.

This report deals with the different possibilities of archiving websites and the points that have to be taken into consideration. The first chapter introduces the Internet, the World Wide Web, the architecture and the evolution of websites. Readers who are familiar with the client-server interactions and the architecture of websites can skip this part. After a look at the importance of archiving websites, some different digital archiving strategies are dealt with. Due to the fact that, next to archival institutions and records offices, also libraries and documentation centres are active in this field, the approach taken is as wide as possible. Multiple questions arise when developing an archiving strategy for websites. The most important ones are: What to archive? How to acquire? At what frequency? How to manage websites? How to make the website archive available? Each question leads to a number of scenarios. Every institution has its own goals and will follow the option that is most convenient for them. This will result in a different archiving policy. The section of this report that deals with archiving will be closed with a practical archiving example. As always only a minimal IT infrastructure needs to be present.

When outlining an archiving strategy for websites, an organisation has to take legal copyright implications into account. A lot of reproduction acts require the permission of the author of the (content of the) website. After a short introduction into Belgian copyright legislation, we will sketch

¹ K. PERSSON, *The Kulturarw3 Project - The Swedish Royal Web Archiw*³, Lecture held in Svetlogorsk, Aug. 2000

² For example <http://europa.eu.int/ISPO/dlm/documents/guidelines.html> containing the message: "This website has been archived. Please visit the new site ...".

an overview of the relevant rules for the archivist. Also the privacy regulations will cause a range of troubles for archive institutions and records offices. We try to formulate answers to these legal issues.

Finally we will check the presence of the different Flemish governments on the Internet and then look at the plans of the Federal government regarding the development of electronic identity cards for each citizen. The development of e-government will benefit greatly from the introduction of this card sometime during 2003. The issue of archiving websites will become more predominant in the future. The aim of this report is to anticipate and speculate about the increased need to archive static and dynamic websites.

Sofie Van den Eynde wrote the legal part of this report, while Filip Boudrez wrote the text regarding digital archiving.

Antwerp - Leuven, July 2002.

II. INTERNET, WWW & WEBSITES

The Internet can most easily be described as a worldwide computer network³. The development of the Internet dates back to the sixties of last century. During the warm phase of the Cold War the American Department of Defence was looking for a means of connecting computers so that information management and rocket control would be possible from different locations. The first network was baptised “Arpanet”. The development of the father of the current Internet went hand in hand with the composition of a protocol for addressing and sending information. TCP/IP (Transmission Control Protocol / Internet Protocol) became the standard protocol and contains in fact dozens of protocols of which TCP and IP are most widely used. Each computer in the network received a unique IP-address that got linked later to a domain name (DNS: Domain Name System)⁴. Meanwhile the Internet has strongly developed itself and boasts a number of applications: gopher⁵, FTP archives, e-mail, Usenet, news groups, Telnet and the World Wide Web (WWW).

The development of the WWW is a milestone in the history of the Internet (CERN: 1989-'92). The WWW has evolved from a technology for text exchange into an interactive and dynamic client-server application that connects documents with hypertext, multimedia applications and databases to each other. The stress has shifted from pure text to graphic design and mouse control. Hypertext is text that contains hyperlinks to other documents. For the downloading of WWW documents a new protocol was developed: HTTP (Hypertext Transfer Protocol). HTTP supports the communication of hypertext files between server and client. FTP (File Transfer Protocol) is used for to exchange files between client and server.

Another standard was developed for the composition of web documents: Hypertext Mark-up Language (HTML). HTML is a fixed language that allows the definition of parts of a web page and their function via mark-up assignments. The assignments are put between < and > and are called HTML tags.

³ <http://www.isoc.org/Internet/history/brief.html>; <http://www.davesite.com/webstation/net-history.shtml>; J. HONEYCUTT (et.al.), *Het complete handboek Internet*, Schoonhoven, 1997; C.J.M. MOSCHOVITIS, *History of the Internet: a chronology, 1843 to the Present*, Santa-Barbara (California), 1999; I. ENGHOLM, *Digital design history and the registration of web development*,

⁴ When a user types in a URL (Universal Resource Locator) in the address bar of a browser, DNS transforms the domain name into an IP address.

⁵ gopher: an Internet application for exchanging information that was popular in the early 1990s. Gopher does not work with MIME types (as HTTP does) but with gopher object types. Over time, gopher has been replaced by HTTP.

Example 1: HyperText Mark-up Language

SOURCE CODE

```
<p><i>This sentence will be put in italic</i></p>


<a href= "http://www.antwerpen.be/david">
Hyperlink to DAVID website</a>
<% ASP, PHP or JSP-code %>
```

WEB BROWSER

This sentence will be put in italic



[Hyperlink to DAVID website](http://www.antwerpen.be/david)

HTML output

The client computer is the computer on which a website is viewed. The programme on the client side that generates the WWW documents is the web browser. The web browser composes the HTTP request and sends it to the server. Most recent versions are graphical browsers that support other Internet applications like FTP and gopher as well and that can automatically start other applications (for example MS Office, Acrobat Reader, MP3 player, RealPlayer, Shockwave, Flash Player, etc.) to open files in a different format than HTML. Some of these applications are plug-ins in the web browser. The computer that hosts the website and makes it available via the web is called the server. On this computer web server programmes are active (Apache, Internet Information Server, Netscape FastTrack / Enterprise Server, etc.), and also modules for scripts and executable programmes.

A website consists of a grouping of separate computer files that are stored in a specific folder structure. The link between two computer files of a website is being determined by means of hyperlinks. These links provide access to the on-line material. Web pages and connected files like pictures and download files can be linked in two ways: absolute and relative. The absolute links refer to and start from the root of the website. The root is usually indicated in the URL of the IP address: `http://www.antwerpen.be/david/eng/index.htm`. For a relative link the path indication starts from the position from which is being linked: `../eng/index.htm`. Both types of path indication refer to the folder structure, the folder names and the file names. If one of these parts is changed, also the path indication has to be changed, otherwise the link will no longer work.

The oldest WWW applications are very static. Communication between server and client is only one-way traffic. These websites are no more than a number of HTML files and images put in a certain structure on the web server, with links connecting pages and images. The HTML pages and possible connected style sheets contain the content and the layout of the web page that is being sent to the client. The interaction between server and client is limited to the sending of an HTTP request and the sending back of a web page. The instruction `http://www.antwerpen.be/index.html`

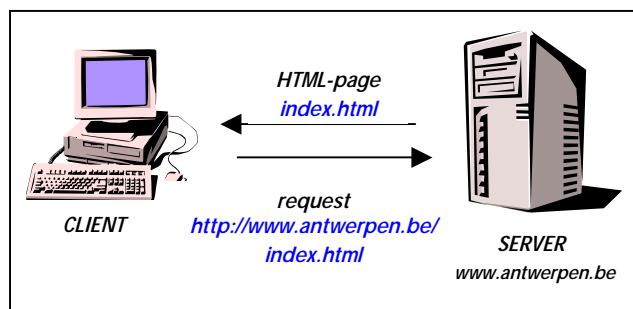


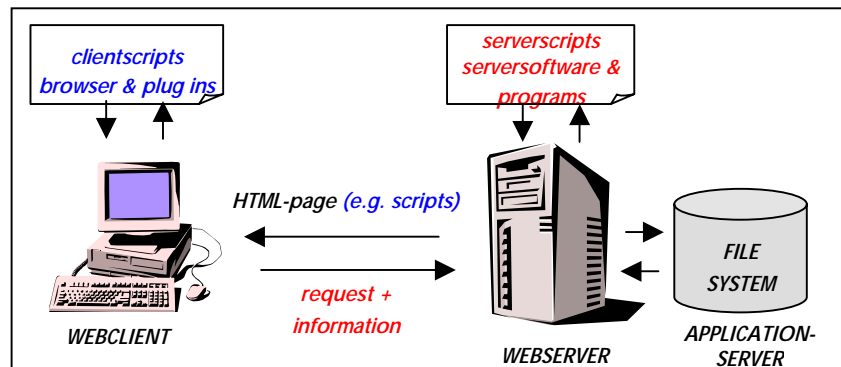
Image 1: The client-server interaction with static websites

requests the server with domain name 'www.antwerpen.be' for the HTML page `index.html`. A part of the domain name is an alias that refers to the root folder of the website on the web server. An HTTP daemon is active on the web server, waiting for requests of the web browsers, executing them and sending back the HTML pages and enclosed files to the client. The server adds an HTTP header to the files. The client browser then

generates the WWW document. The content of the web page is static and identical for every visitor. Even today, a large number of websites on the WWW consists of pages with static content. The interaction between client and server is limited primarily to exchanging requests and HTML pages. These pages are referred to in this report as ‘static web pages’.

Providing information only using static web pages has considerable limitations, and therefore new means were found to make a greater interaction possible between server and client. The earliest example of these was the use of CGI (Common Gateway Interface) server scripts. These scripts, however, are not very secure and can only be executed on the server side. A number of product linked script languages offer solutions to this and have become widespread: ASP (Active Server Pages, Microsoft), PHP (Php Hypertext Processor, Unix-Linux), JAVA Servlets, ColdFusion (Macromedia) and JSP (JavaServer Pages, Sun). Server scripts are embedded in HTML pages or are put into separate files. The execution of the server scripts requires appropriate server software. The server script often works together with a computer programme that runs on the web server (for example form.exe, query.exe, rightsite.exe). The following actions are performed when executing a server script: calling a script (programme), reading the requested files, executing the script, sending or requesting data, executing the calculations and finally sending the result as an HTML page to the web browser.

Image 2: The client-server interaction for websites with forms, that request information from databases or that are generated “on the fly”. The linked file system usually exists of a database or document management system that is active on an application server.



Common applications are the processing of sent-out form data, the querying of databases and the making available of documents via a document management system. The website is used as an interface and is no longer an autonomously functioning entity. The website is part of an information system that consists of web server software, web server configuration, script files, modules for the execution of scripts, executable programmes and databases or document management systems. The linked databases are often situated on separate application servers. These databases are referred to as the ‘deep web’ or the ‘back office system’. The linked websites are ‘database driven’. In this case the content of the web pages the client receives is depending on his or her query or on the information available in the databases at this moment. Web pages are composed “on the fly” on the web server and then sent on to the web browser. The content of these websites is dynamic and different for each visitor. Another possibility is that the ‘back office system’ publishes an HTML page at set times that can be consulted on the website.

The next step in web development, was trying to relieve the server by executing as much scripts as possible on the client side. In this type of application the server sends the client the necessary HTML pages with linked client scripts, embedded in an HTML page or in a separate file. The client subsequently executes the scripts. An absolute condition is that the client computer contains the necessary software. In most cases however, the installation of the required browser (version) will do.

Scripts that are executed on the client side can be, among others, JAVA, JAVAscript and VBscript⁶. Examples of applications with client scripts are the checking of data in a form before sending it, automatic transfer linking to another page, menu bars, roll-over images, animated buttons, automatic date and time display, adding URL to favourites, etc. The received HTML page contains client scripts in all these cases.

Applets and ActiveX applications are added to enhance the functionality of websites. Applets are little Java tools that do not function on top of an operating system, but that are being executed by the Java Virtual Machines within web browsers. Contrary to ActiveX, applets do not need access to the hard disk of the client computer. ActiveX applications enhance the web browser functionality and are installed on the hard disk, so that they can still be used after visiting a certain site.

Websites in Flash have recently become more and more popular. Flash is a Macromedia application that combines animation, text, images, interaction and sound. Websites in Flash consist of one or more ‘movies’ (*.fla files) that are published as *.swf files when spread via the web. Viewing websites in Flash requires a necessary plug-in. A website in Flash often has a static HTML version available too.

The most recent generation of WWW applications enables the spreading of personalised information. These web pages can no longer be considered as a publication or a common interface. What appears on the screen can depend on user rights, user profile (among others: who? from where? when?), time, previously consulted web pages, the applicable query and the software available to the user⁷. These websites do not have a static form and are somewhat comparable to a computer program. The output of these websites may be a static HTML page, but this page is only composed when the web server receives an HTML request. The web server learns about the user profile and his or her preferences from a cookie or the information a web client always sends along⁸. Examples of this are websites that are linked to document management systems that define who can open what documents or websites where the content depends on the user’s preferences during previous visits.

The archiving of websites encompasses both static and dynamic content. Current websites vary from static digital publications to interactive dynamic websites that are used for services or transactions. Dynamic websites may well contain a number of static pages. The DAVID website for example is mainly a static website, but a number of pages are dynamic because they are linked to the

⁶ Java is a programming language of Sun Microsystems. A compiled JAVA programme that is linked to a website is called an *applet*. JAVA runs on a virtual machine (VM). Applets are sent to the client together with the web page. An applet is not executed on the server side.

JAVAscript is another way of sending multimedia files to web browsers. JAVAscript results from a co-operation between Netscape and Sun Microsystems. The JAVAscript code of the programme is enclosed directly in an HTML page. JAVAscript codes are not being compiled, contrary to JAVA-applets.

VBscript is a combination of Visual Basic and Microsoft OLE-scripting. VBscript allows the execution of scripts both on the client and on the server side.

⁷ An example of this is the home page of the website <http://www.antwerpen.be>. A browser check is executed upon arrival. Visitors with an IE browser receive the left frame with expandible navigation as dynamic HTML. As this functionality only works in IE browsers, a more static version of the same frame is available for, say, Netscape users. This static version is composed according to the W3C standard.

⁸ Part of this information is stored in ‘cookies’. Personal data, interests, passwords, language preferences etc. are stored by the server in a cookie and sent to the browser. Upon a next visit to the same site the browser will send the cookie to the server in an HTTP header. Because of this one does not have to repeat, say, language choice. A cookie is a text file on the hard disk of the web client. Other data that the browser sends to the server are among others IP address, operating system, browser system, browser programme, screen resolution, installed plug-ins.

document management system of the Antwerp City Archives (for example: the pages Publications and Newsletters). The link between the website and the document management system is based on a Rightsite script that is executed by the server. An ASP script handles the automatic registration and unregistration on the mailing list of the newsletter.

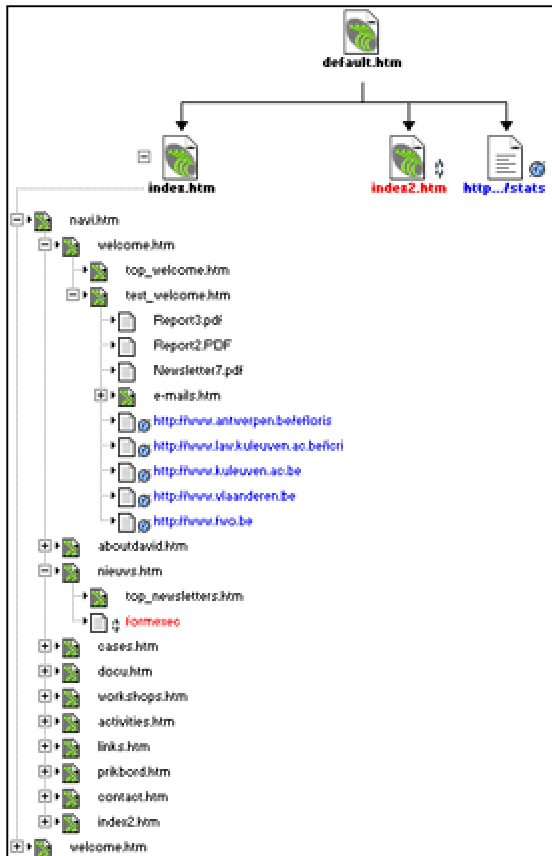


Image 3: Sitemap of the DAVID website (version 6).

A sitemap can be useful in different ways for the archivist. It shows the structure of and the links within a website, so it can be used to document the architecture of the website and the relations of the files. A sitemap is also useful tool to determine the boundaries of a website and to support the appraising process.

Many websites contain a sitemap so that information can be found quickly. There exist also computer programmes that automatically generate a sitemap.

Currently quite some work is being undertaken concerning the accessibility and management of web documents. Also in this field the common goals are user-friendliness, version management and integration in the work process. *Web content management* has grown to become a new work field within IT and has as its goal to manage the data of a website as efficiently as possible, and to keep the data up-to-date in a dynamic way. Document management systems are enhanced with modules for website content management: metadata, version management and control, management of the structure of the websites and web pages, web publishing, link management, etc. A new functionality in link management is the use of virtual links or persistent identifiers. Whereas normal website management requires a manual installation or adaptation of links, virtual links do not refer to one specific version of a document but to an object ID linked to the versions. When a document is updated the latest version is shown on the site without adjusting the link. Persistent identifiers assure that the link remains operational after the target has been moved.

One of the latest developments is the further evolution of HTML towards a real mark-up language. In the latest (X)HTML specifications the (X)HTML tags are being used less for the layout of the pages, style sheets are replacing them.

III. THE IMPORTANCE OF ARCHIVING WEBSITES

Archiving websites is important for a variety of reasons. The archival value of websites is linked to the evolution from static to interactive websites.

Firstly, archiving websites is justified by the documentary value the websites possess themselves. Archived websites are a necessary evidence material for research into the history and the evolution of this medium itself. The Internet and especially the WWW has thoroughly changed the way in which we spread and research information. Without archived websites it would be close to impossible to form an opinion of what websites looked like, what they were used for, what information they contained and what their relation with other media was. We would not know how an institution or company profiled itself via the web, what opportunities they created, what web design looked like around 1997, how HTML and scripts were used, etc.

Secondly, it is obvious that websites are being archived because of the information they contain. Websites often have a large informational value and will be useful for future research of any kind. From the first generation on websites were predominantly used for the spreading of information. Their informational value has enhanced even further in the meantime. At the beginning of the development of the WWW one could notice that information was also made available via a website in electronic form. The information on these websites (for example in HTML or PDF) was also usually available in another form (for example in a word processor file) and on another location with the records creator. Currently websites are increasingly becoming the exclusive place of publication. They are challenging the traditional channels of information transfer. A consequence of this evolution is that information destined for a website gets more immediately a format that is suited for the Internet. Another consequence is that the content of paper information carriers is changing because their previous content is now being published on a website⁹. As far as information providing is concerned, websites and paper publications are becoming more and more complementary than overlapping. Until today a large number of websites has as a primary goal the digital spreading of information. Next to their informational value, it is important to realise that websites can be frozen as on-line sources one way or another. The “freezing” of websites is a practical necessity because the volatility of the medium leads to quick adaptations. The content changes quickly and URLs change, which can make it hard to retrieve certain information created previously.

Thirdly, websites have, next to their documentary and informational value, also a cultural value. They belong to our digital heritage. They are material witnesses of our society, and without an archiving policy they will be lost for the future.

Finally, websites are being archived because they are records, because they contain records or because records are being created via a website. The current generation of websites has more functionality than the traditional information channels. Websites now play a greater role in the work process. They are not merely a means of publication, but are fully incorporated in the processes of service provision and operational management. This is illustrated by the current tendency towards e-government and e-commerce. Transactions or acts that require accountability are increasingly being

⁹ G. VOERMAN (et.al.), *Het belang van het archiveren van websites*, in *Information Professional*, 2001, p. 17

handled via the Internet or intranet sites. The (electronic) tracks of these transactions will serve as evidence or documentation and will have to be included in the recordkeeping system. Websites also lead to the creation of records via the web. These can be in the form of e-mails or databases. Websites usually contain the form via which the information was entered, the guidelines and the procedure. In this sense the websites are part of the context of these records.

This last aspect is particularly relevant for websites of government departments. Governments need to be able to justify themselves for the information they spread via the WWW, because their website can steer the actions and decisions of citizens and other departments. Government websites are in turn also a source of information about their organisation, tasks, authorities, policies, guidelines etc. They need to be archived because of later liability or accountability and also because of their historical value. What happens if a certain citizen thinks to have read something on the website and makes a claim to that? Yet there is no way of proving what they read was actually on the website, or was incorrectly published. Without archiving the website this would be impossible to prove. In Australia for example it is clearly stated that the government websites are records because the government must be liable for the information it spreads over the Internet. Any change that is made to a website is archived there¹⁰. As a shield for possible liability claims it would be good to keep archived versions of one's own websites.

Furthermore, government websites are often policy documents. They do not just form a means to make government information available, even though only this aspect of government websites is usually discussed. The Flemish decree concerning publicity of government of 18 May 1999 defines the concept "a document held by a public authority" as "the carrier, *in whatever form*, of information that a government possesses"¹¹. Even though there has never been any formal regulation in Flanders about the matter of websites being governing documents, a government website can be considered to be an electronic publication of that government. Its goal is then to inform the citizen about the policy as good as possible and to make a quick, user-friendly and transparent service possible. The reason for this silence about government information and ICT in the doctrine is obvious. This obviousness will be elaborated on below to stress the position of government websites as governing documents.

Legislation concerning access to public sector information, or access laws in short, deals with the legitimacy and credibility of the authorities as an important part of a democratic constitutional state. Because of this, every citizen has the right to access governing documents¹². The right to access governing documents aspires to put an end to the practice of secrecy and the lack of transparency within the authorities that was so typical for the pre-war period¹³. The Flemish decree concerning publicity foresees a number of measures in the framework of the so-called active and passive publicity of government, to ensure that citizens do get access to the governing documents. Active publicity encompasses the obligation to "*inform the public systematically, on time and in an understandable way about policy, decrees, decisions and other legislation, as well as about their services and about the information that is available to them*"¹⁴. This is put into practice via brochures and information that

¹⁰ *A policy for keeping records of web-based activity in the Commonwealth Government*, p. 11-12

¹¹ For more information about the concept "governing document" see: BOUDREZ, F. and VAN DEN EYNDE, S., *Archiveren van e-mail*, Stadsarchief Antwerpen – I.C.R.I., Antwerp - Leuven, October 2001, 34-35

¹² Article 32 of the Constitution

¹³ Still Belgium has only put its first steps towards publicity of government with the Act of 29 July 1991 concerning the explicit motivation of government acts (*B.S.* 12 September 1991).

¹⁴ Article 21 §1 of the decree

is being spread via a number of websites¹⁵. The task of the Flemish Information Civil Servant is to inform the population about the policy and the services of the Flemish Government and the Ministry of the Flemish Community, and to stimulate and co-ordinate the enhancement, co-ordination and realisation of the information provision of the Flemish Government¹⁶. Passive publicity implies that the citizens can consult governing documents, can ask questions about their content and ask for a transcript of them.

A government website is a government publication (a digital brochure as it were, or a collection of digital forms in the case of proactive servicing) that is by definition accessible to every citizen with an Internet connection. The question of accessibility of this governing document has therefore never been an issue. Still it is clear that also the digital publication of government information via the website makes up a governing document. Despite the fact that access laws do not contain any regulations about storing and destroying governing documents, they do oblige the administrative government in an indirect way to keep governing documents available to any citizen that may ask about them. Administrative governments therefore have to archive the different versions of their website.

When a citizen wishes to use his or her right to inspect information in person, the administrative government will, together with the requester, determine the place, date and time of the inspection¹⁷. This procedure is useless for the on-line version on the website. The user can consult the website at any moment. The government can do two things regarding the archived versions. It can add a sub-page “archive” to its website, where the older versions of the website can still be consulted. The inspection can then take place from a distance “on the spot”, via the Internet. “On the spot” then means “on the screen”¹⁸. It could also keep a local copy of the website in the recordkeeping system. In this case the decree stipulates that the citizen has to contact the department where the website is kept.

What is the scenario when the website has been archived by a careful and forward-looking government, but is not situated with the department that has received the request?¹⁹ The Commission for the Access to Governing Documents believes that an administrative government cannot claim not to possess any piece of information when the request concerns documents of which the government is the author²⁰. Furthermore we can assume that every government department keeps a transcript of all its governing documents²¹. Therefore it cannot escape its obligations by stating that it no longer possesses previous versions of its website after they have been archived.

The decree also describes the right to receive a transcript of a governing document. When viewing the (current or archived) website on a computer, the web browser automatically generates a transcript of the website. The arrival of the Internet causes the boundaries between active and passive publicity to become blurry. If the citizen requests an electronic transcript of an archived website that is no longer available on-line, for example via e-mail, the government has to comply with this request²². The

¹⁵ For example the portal site of the Flemish government <http://www.vlaanderen.be>, or the website of the Flemish information hotline <http://www2.vlaanderen.be/infolijn>

¹⁶ Article 22 §2 of the decree

¹⁷ Article 11 §4 of the decree

¹⁸ DUMORTIER, J., JANSSEN, K. a.o., *Transparante overheidsinformatie als competitief voordeel voor Vlaanderen. Literatuurstudie*, K.U.Leuven, 85

¹⁹ For example because the storage is centralised in one department.

²⁰ Commissie voor de Toegang tot Bestuursdocumenten [Commission for the Access to Governing Documents], advice 95/58

²¹ DUMORTIER, J., JANSSEN, K. a.o., *o.c.*, 84

²² DUMORTIER, J., JANSSEN, K. a.o., *o.c.*, 85

Act of 12 November 1997 concerning the publicity of governmental records in the provinces and the towns (*B.S.* 19 December 1997) determines that making a transcript of a copyright protected work is only allowed after previous permission from the author or from the person who possesses the copyright. Perhaps in this case copyright is actually creating a barrier to exercise the right to (public) information. Furthermore this limitation to publicity is not very practical any more with regard to digital governing documents on the Internet, as the web browser automatically (and thus without asking for authorisation from the author) generates a transcript.

This access law only applies to the administrative governments. This implies that neither the websites of the legislative governments²³ nor those of the judiciary²⁴ have to be archived for reasons of publicity. These websites however do have to be archived because every government needs to be able to justify the information it spreads via the Internet²⁵.

A website of an administrative government can form a governing document itself, but it can also contain or create governing documents. These have to be stored carefully too, even though they appear to be casual transactions of digital information executed via the website.

IV. CURRENT INITIATIVES

The volatility of the information on the Internet has led to the storing of websites in an early stage. Considering that the first generation of websites had the status of digital publication, it comes as no surprise that the library world started collecting (part of) the WWW.

The first real projects to archive websites in a systematic way started in 1996. During the summer of 1996 the *Internet Archive* began to collect all (textual) information on the Internet, both from news groups and from websites²⁶. Because of the scale of the project, the only way to achieve this was to have the work performed in an automated way. The *Internet Archive* co-operates with a commercial partner that collects the data and sends all data that is 6 months old to the Internet Archive²⁷. Since the

²³ About the publicity of legislation, Article 190 of the Constitution stipulates that “*no law, no decree or regulation of general, provincial or communal government is binding unless published according to the methods determined by law.*” For the acts, the royal decrees and ministerial orders this law principle has been elaborated in the Act of 31 May 1961 (*B.S.* 21 June 1961), that stipulates the Belgian Official Journal as means for the publication. Whether the government can be forced to publish government information like legislation via the Internet is a question that will not be dealt with here. This issue is studied by ICRI and the department of Communication Science of the K.U.Leuven in the PBO-project: <http://www.law.kuleuven.ac.be/icri/projects/pbo.htm>.

²⁴ The umbrella website of the judiciary in Belgium is <http://www.juridat.be>. Juridat has been created by members of the judiciary.

²⁵ See below: liability of the government for its own website

²⁶ <http://www.archive.org>; <http://www.alexacom>

²⁷ The websites in the *Internet Archive* are stored in ARC files. Alexa delivers them in this format to the *Internet Archive*. ARC files are basically no more than files with the necessary metadata added as header information to the HTML web pages (*encapsulation*). The ARC file specification has been released for Alexa (<http://www.alexacom/company/arcformat.html>). In the ARC format both metadata and HTML file are encapsulated. The metadata contains: version, URL, IP address, archiving date, MIME-type, number of

beginning of 2002 the website archive (the “Way Back Machine”) can be consulted on-line. The Swedish National Library also started a similar project in the same year. This project focuses only on the archiving of all Swedish websites (*Kulturarw³ Project*²⁸). Robots that index the sites and store them in databases are collecting the Swedish websites. The goal is not just to store the web but also the original “look and feel” and the surfing experience. In Australia the National Library has started archiving websites, news groups and mailing lists within the *Pandora project*²⁹. This project is focused towards the main Australian on-line publications and thus preselects the websites. The American *Minerva project*³⁰ had a similar approach. Minerva has collected the websites of the candidates of the 2000 presidential elections. Both projects have a selective approach and store a website by means of an off-line browser. Based on this experience there is in both countries a current shift towards the storage of their respective web space. In the Netherlands the *Occasio project*³¹ is collecting Internet news groups and the Documentatiecentrum Nederlandse Politieke Partijen [Documentation Centre for Dutch Political Parties]³² collects the websites of political parties. Soon the Bibliothèque nationale de France [National Library of France] will start a project to store the French web space³³. A pilot project has kicked off in Austria too³⁴.

Parallel to their shift of function, the storage of websites has gradually also become an archival matter. Websites have evolved from static digital publications to dynamic and interactive tools that allow provision of individually adjusted services and information. It is clear that the possibilities of websites enhance as IT evolves. In countries where e-government and digital services as the e-portal are already fully operational, one further step is taken. The records offices have compiled an archiving policy for those records that have been created via a website or for those transactions that take place via a website. In the beginning of 2000 the NARA (National Archives and Records Administration) started the archiving of the websites of all federal departments. Each department was requested to hand over a snapshot of their website to the NARA by the end of the Clinton term of office³⁵. Early 2001 the Australian National Archives Department published a policy note and guidelines for the archiving of websites within the government. The Australian government is currently researching what

characters in the HTML file, server name, date of last change. One ARC file (about 100 Mb large) contains multiple HTML files. An external database ensures the disclosure of the ARC files.

²⁸ <http://www.kb.se/eng/kbstart.htm>; <http://kulturarw3.kb.se>; K. PERSSON, *The Kulturarw3 Project - The Swedish Royal Web Archiv*³, Lecture held in Svetlogorsk, Aug. 2000; A. ARVIDSON, *Harvesting the Swedish webspace*, Lecture held in Darmstadt, 8 Sept. 2001

²⁹ <http://pandora.nla.gov.au>; W. CATHRO, C. WEBB en J. WHITING, *Archiving the web: the pandora archive at the National Library of Australia*, Lecture held during the Conference about Preserving the Present for the Future Web Archiving, Kopenhagen, 18-19 June 2001

³⁰ <http://www.cs.cornell.edu/wya/LC-web>. The website of the Minerva project is at <http://www.loc.gov/minerva> but is not accessible (yet); C. AMMEN, *MINERVA: Mapping the INternet Electronic Resources Virtual Archive -Web Preservation at the Library of Congress*, Lecture held in Darmstadt, 8 September 2001.

³¹ <http://www.iisg.nl/occasio>; J. QUAST, *Het Internetarchieff van het IISG*, in *Nederlands Archievenblad*, September 2000, p. 16-17

³² <http://www.archipol.nl>. Archipol takes a selective approach and uses HT Track and archipol.cgi to capture websites.

³³ J. MASÉNAS, *The BnF-project for web-archiving*, Lecture held in Darmstadt on 8 Sept. 2001

³⁴ A. RAUBER et.al., *Austrian on-line archive. Current status and next steps*, Lecture held in Darmstadt on 8 Sept. 2001

³⁵ <http://www.nara.gov/records/websnapshot.html>

metadata is to be stored³⁶. The Public Records office has archived the website of Downing Street 10 due to the elections of June 2001³⁷.

These first archiving experiences allow some conclusions to be drawn regarding the development of a proper recordkeeping system. All library initiatives except *Pandora* and *Minerva* want a systematic storage of (part of) the WWW. Their first assignment was the development of a suited computer programme for the set-up of a bulk collection of websites³⁸. Because of the large number of websites to be archived, all handling is automated (indexing, putting off-line, description, extraction of metadata, making available, etc.) and human interference is reduced to a minimum. These computer programmes are called “harvesters” or “web spiders” and work according to the same principle as the indexing robots of Internet search engines. They start on one specific URL and follow the hyperlinks in the web pages from there on. The harvesters need to be equipped with an extensive version and duplication control mechanism. This avoids multiple storage of the same website or mirror sites. Even with built-in checks, the systematic archiving of a national domain the size of Sweden still required about 1 000 gigabytes of storage space. The latest “harvest” of the Swedish web brought back about 30 million files and 15 million web pages. Since 1997 the Swedish project has taken 7 snapshots of the Swedish web space.

The bulk archiving option has a number of disadvantages, causing this choice not to be the most appropriate for a website recordkeeping system for archival institutions and records offices. Firstly, a harvesting operation consumes a lot of time. Downloading all websites takes several months, so that websites can only be stored at a low frequency³⁹. This low frequency contrasts with the volatility and the speed of the medium. It is too difficult to archive each version of a website. The library world is aware of this problem. In France and Austria, for example, plans exist to store a number of preselected websites at a higher frequency, next to the bulk process (for example magazines and newspapers on the web).

Secondly bulk archiving implies an absolute lack of quality control. Automatic acquisition will reduce the chance of discovering errors and there are no guarantees that websites in their active form do not require adaptations for a long-term digital legibility. This was one of the reasons why the work group of the Australian *PANDORA* project decided to focus on preselected websites only⁴⁰. The chance exists that websites in a bulk archive cannot be consulted. Checking the archived websites is

³⁶ *A policy for keeping records of web-based activity in the Commonwealth Government*, January 2001; *Guidelines for keeping records of web-based activity in the Commonwealth Government*, March 2001; S. MCKEMMISH and G. ACLAND, *Accessing essential evidence on the web: towards an Australian recordkeeping metadata standard*

³⁷ <http://www.records.pro.gov.uk/documents/prem/18/1/default.asp>; D. RYAN, *Archiving the no. 10 website - the story so far*, Lecture held in London, 25 April 2002

³⁸ Building further on the Finnish web archiving robot, a new harvester has been developed in the Nedlib project. The source code (C++) of the Nedlib harvester can be freely downloaded from <http://www.csc.fi/sovellus/nedlib>. The harvester needs to be linked to a MySQL database. The stored websites are being put in a database for two reasons: indexing (retrieval) and management of millions of files. The harvester of the university of Helsinki has Solaris as operating system. The Austrian Aola project has built its own harvester too. It built further on the Nedlib harvester.

³⁹ It is clear when a large harvesting operation starts, but never when it will end. It is hard to predict in advance how large a national domain is and how much time the storage will take. Especially large websites slow down the process.

⁴⁰ A.R. KENNEY and O.Y. RIEGER, *The National Library of Australia's Digital Preservation Agenda, an interview with C. Webb*, in *RLG-DigiNews*, 15 Febr. 2001; W. CATHRO, C. WEBB and J. WHITING, *Archiving the Web: The PANDORA Archive at the National Library of Australia* (<http://www.nla.gov.au/nla/staffpaper/2001/cathro3.html>)

necessary. Furthermore it is far from certain that one type of harvester with the same control switches is suited to store every type of website. The Way Back Machine, for example, contains a number of versions of the website of the Antwerp City Archives, but not one version is working and even the home page cannot be viewed.

Thirdly, websites that are not referred to by other websites will not be stored. Finally there is the problem of storage capacity, consulting and accessibility. The amount of stored computer files is very large so that a strong but yet quick and user-friendly solution is necessary⁴¹. Records offices and documentation centres only rarely possess this adequate infrastructure.

The *Pandora* and *Minerva* projects start from, like all other archiving initiatives, preselections of websites with archival value. As with bulk library projects, not a lot of attention is paid to the deep web. However the archiving of the deep web is an essential part of the archiving of dynamic and interactive websites. E-commerce and e-government implies dynamic and interactive websites. It is remarkable that for all these projects only one single archiving strategy is chosen, without taking the website itself into account. The architecture and nature of a website co-determines the way in which it can be archived.

V. ARCHIVING WEBSITES

A recordkeeping system for websites aims at the records that are being published on the WWW and at the records that are created through the interactions or transactions via a website. This last category of computer files is to be archived because of its contextual value, as evidence or simply to reconstruct the content of a website.

It is therefore important to start with formulating clear goals for the recordkeeping system. These goals determine WHAT will be archived of the WWW or of a specific website. The WHAT-question in turn will determine HOW websites are to be stored. Also the type of website will influence how a website is archived. Also the description (metadata), the measures for durable recordkeeping and the access to the archived website and related digital archive records need to be looked at when archiving a website.

Websites are usually archived in electronic form. The reason is obvious, because websites are primarily electronic and later consultation should take place in a way that is as close as possible to the original, on-line version. Electronic recordkeeping best guarantees the storing of as many original website characteristics as possible. Some techniques for paper archiving of websites do exist (hard

⁴¹ The existing projects mainly use tapes and hard discs as storage medium. Rarely consulted files are stored on tape, often consulted ones on hard disk. Bulk storage implies special requirements of the operating system. Conventional operating systems of personal computers have limited the number of directories and files they can contain.

copy), but too many original characteristics and too much original functionality is lost when they are applied⁴². Also, hard copy is only storing the content of the website.

The WWW has become a dynamic and interactive medium. Storing websites and their related computer files contrasts with its dynamic character so that every attempt causes some loss of functionalities and characteristics. The web is also dynamic in the sense that it is always changing and evolving. The website of a certain organisation can evolve from a static website towards a dynamic and interactive one. The recordkeeping system will therefore require a constant fine-tuning to stay up to date.

Contrary to other projects that deal with website archiving, the DAVID project will not develop a single recordkeeping system. Rather, with every step a number of different archiving solutions will be developed. The implementation of a recordkeeping system depends on the goals, the type of website and the available technological means. An example of a possible recordkeeping system will be presented in the next chapter.

A. QUALITY REQUIREMENTS FOR ARCHIVED WEBSITES

Websites have to comply with a number of quality requirements when entered into the digital repository. These quality demands are the same for static websites and for dynamic and interactive websites. An archived website has to fulfil the following quality demands:

- ☑ all files will be archived that are necessary for a detailed reconstruction of the website (text, images, style sheets, scripts, logging files, databases, user profiles, etc.).
- ☑ the start files (among others default.htm, index.htm, start.htm or welcome.htm) and the subfolders of one version are stored in one folder within the website archive.
- ☑ the file structure and the file names are copied as close to their originals as possible onto the web server. Web pages with static content can copy the same file name. Web pages with dynamic content have a name as close as possible to the original file name.
- ☑ internal links are indicated with relative path indications, external links with absolute path indications. When internal links are indicated with absolute instead of relative links, the link will call the on-line version when consulted, and therefore it is no longer the archived web page that is being viewed. Relative paths also facilitate the management (for example when moving). Links to sources outside the own website can be made with absolute path indications⁴³. Links to virtual folders should be transformed into relative links as much as possible.

⁴² The active pages in a web browser can be printed. A special capture technique for websites is offered by Adobe's Acrobat software. This programme enables to store a website as a PDF file, transforming the hyperlinks into bookmarks. It goes without saying that archiving as hard copy will cause a loss of functionality and format, and that the source code is not being stored. This technique is however remarkably quicker than making websites available off-line, but can only really be used to store the information in a quick way. This programme has the same general shortcomings as an off-line browser.

⁴³ External links could be documented so that the user can later estimate what information was lying behind the link. This can be done via HTML comments (the attributes ALT or LONGDESCR) or it is possible to stop

- ☑ active elements such as date and visitor number should be disabled. The date and the visitor number should be those of the moment of the snapshot. Furthermore, the scripting can cause error messages and infinite loops. Both elements will be added to the metadata.
- ☑ IT-dependencies (hardware, software, Internet protocols, etc.) will be limited to a minimum. The archiving is as system independent as possible. The computer files that compose a website should be standardised as much as possible. Tags and attributes of the applicable mark-up language are part of the standardised (X)HTML specification.
- ☑ all parts of a website will be archived at the same moment. The in-line images and the web pages are stored simultaneously.
- ☑ the archived website and its related records are ingested into the recordkeeping system of the organisation. The records are securely stored and described based on their metadata.

B. SELECTION: WHAT TO ARCHIVE?

The question of WHAT to archive consists of two subquestions that each co-determines the archiving strategy. Firstly, an acquisition policy needs to be determined. What websites will be archived? Then what exactly needs to be archived from these websites is determined. The answers to these questions will lead to the development of global goals for website archiving. These goals will then form the basis of the elaboration of the recordkeeping system.

B.1 The acquisition policy: what websites to archive?

As the WWW is a free medium, every individual, organisation or institution is free to publish a website on the web. At the end of 2001 an estimated 1,4 million different websites were present on the WWW⁴⁴. It is important to predetermine a clear acquisition policy because it will form the basis of the recordkeeping system. For bulk archiving the acquisition process operates in a completely automated way, harvesters are used and archiving is limited to the information the web client receives. The original web server files, the ‘deep web’, the log files and the digital transactions prints can only be archived in a selective approach.

The composition of an acquisition policy would benefit greatly from some consultation between libraries and archives. In Australia and Canada, for example, there is a division of tasks. In Australia the libraries store those websites with a status of digital publication, while the archives store those websites that are, contain or generate archive records⁴⁵. In Canada the libraries store the Internet sites

the link and redirect it to a separate HTML page with more information. These adaptations are all very labour intensive and also superfluous when the basic rules about *Web Content Accessibility* are applied. One of those rules describes that every link should be documented sufficiently on the web page itself. However, this is one of the recommendations of C. Dollar. (C. DOLLAR, *Archival preservation of smithsonian web resources: strategies, principles, and best practices*, 4.3).

⁴⁴ <http://wcp.oclc.org> ; <http://www.pandia.com/searchworld/2000-39-oclc-size.html>

⁴⁵ *A policy for keeping records of web-based activity in the Commonwealth Government*, p. 8-10.

while the archives store the intranet sites⁴⁶. The division of work is largely based on the function of the websites. In reality however, many websites will form both a digital publication and a record. Agreements between libraries and archives are needed for those websites with a mixed status. Some overlapping will be unavoidable.

Contrary to the current library projects in other countries, the Flemish documentation centres and records offices will start more from a selective acquisition profile. The selection will largely comply with their general acquisition or collection policy. In the case of the private records offices or documentation centres, it goes without saying that they will collect the websites of the institutions in their area of research. The function of the websites is not important for these records offices, because they also collect documents and publications in hardcopy.

The archive departments of public governments store the websites of their own organisation or institution. These archive departments will often also archive websites that can serve as a source or an aid for historical research, as happens with paper publications. Examples of those websites can be the election sites of local political parties, websites with relevant folklore or genealogical information, or websites belonging to persons or events that are linked to the creator in some way or another⁴⁷. It would also be probable that an archive department of a city board attempts to collect the websites that deal with the history of the city and its inhabitants. The general acquisition profile of the archive department or the documentation centre can be applied for the selection of websites to be archived.

The selective approach allows for, next to a greater quality control, some co-operation and contact with the creator and the web designers. This is important in order to store the context of the website and to stay informed about its evolution, versions and updates. The creator has a better position to keep track of this than the web designer does and he or she can also provide the archivist with the desired metadata. Belgian copyright laws require contact with the creator anyway. Without co-operation of the creator it is impossible to archive the linked computer files that are not accessible to the web client. The main disadvantage of this approach is the intensive labour required and the higher cost per archived website, especially when compared to bulk archiving.

Special cases are websites that are portal sites. Portal sites are websites that contain mainly links to other websites. These sites do not have any extra content value and face the problem that most of their links become obsolete after some time. The question needs to be posed whether these websites are worth archiving. This is not the case for links pages within a website. These pages should be archived, as they are part of a website and need to be archived in order not to disturb the running of the website.

⁴⁶ D. LÉGER, *Legal Deposit and the Internet: Reconciling Two Worlds*, Lecture held in Darmstadt, 8 Sept. 2001

⁴⁷ Examples of these with the Antwerp City Archives are the websites of the Antwerpse Vereniging voor Romeinse archeologie [the Antwerp Society of Roman Archeology], het Genootschap voor Antwerpse geschiedenis [The Fellowship for Antwerp History] (historical information), the Van Dyck Year and Mode2002 Landed | Geland (events) and KAPA (associations).

B.2 What part of websites to archive?

Archived websites need to be able to be consulted in the future. This implies that websites have to be reconstructable and interpretable. Both demands will be dealt with when answering the question of WHAT needs to be archived. Besides the typical appraisal process, there are some technological matters to deal with. After all, archiving websites isn't just a matter of preserving the parts of a website or related records with archival value.

It is important that both content and original “look and feel” of a website can be reconstructed. To reconstruct a website all files that compose a website need to be archived. A first difficulty lies with the boundaries of a website. It is not always obvious to limit the grouping of linked computer files that together form the website. Secondly, the reconstruction depends on the files that compose the website. Files with dynamic web scripting cannot be executed without webserver software and/or the deep web, which implies that it is not very useful for the reconstruction to store the website in this original format. Archiving the original ASP, PHP and JSP files can be justified for reasons of accountability or for the importance to collect original scripts. Next to the website itself, there are a number of linked computer files that together make up the content and the interaction of a website. These are in the first place the web server's log files and linked databases.

It would be beneficial to check beforehand how downloads of the website are kept in the digital repository: are they being stored together with the website or are they archived separately? Downloads are those files that are made available via the website and that usually contain text or audio-visual information (PDF, MS Word, MP3, zip, tar). The choice has an impact on the necessary storage capacities for each archived version. A possible solution could be to store those downloads that are found on the web server together with the archived website, and to store those downloads that come from a database or a content management system separate from the archived website. In this last case the metadata of the website should refer to this information, and it should be made clear what download was available on which web page. When archiving downloads the electronic recordkeeping system must be tuned towards the paper recordkeeping system. Lots of publications are still available in paper form within the organisation and will be archived accordingly. The archiving of the electronic version (for example in PDF or Word format) seems to be superfluous in these cases, unless the electronic version on the website is not identical, contains extra functionality or is considered to be the original.

Furthermore, the interpretation of a website requires knowledge of its context. Part of this information is available from the log files of the web server or in the metadata. For websites that form a part of a larger information system it might be useful to keep some documentation about this (for example technical files, description of functional demands, documentation about the technical development and system demands, installation documentation, manual for the administrator or webmaster, etc.). This also goes for documentation about linked databases with archival value.

Selection is almost only possible on website level, not on the level of the files that form a website. When archiving a website it is therefore best to ingest a complete version into the archive. It is almost impossible to limit the record keeping to that part or section of a website with archival value. Such an approach would also make it more likely that the archived website will no longer function. In a well-structured website items like scripts and images are stored in common folders so that it requires a lot of work to determine which files are to be copied and which are not. Therefore it is best to take a mirror of the complete website, in order to avoid elaborate research about what files are to be copied or to avoid the need to repeat the taking of a snapshot numerous times. This way it is insured that all

files are present and that the website will function. If a website of a local department of a political party is a part of the general website of that party, it is easier to archive the complete website. Otherwise the danger exists that files, frames or frame sets are missing and that the archived and the on-line version differ thoroughly. In the past it was suggested that archiving the textual information was most important and that images come second only (see: the cache of search robots). In the meantime however the graphical aspect of websites has gained a lot of importance, which implies that images have now become an essential part. And finally, a lot of image files do contain important textual information too.

The selection issue concerning the archiving of websites clearly shows that some archivist appraisal is necessary already at the time of creation. It is almost never possible to archive websites in retroaction.

B.2.1 Websites with static content

Archiving websites with static content causes few problems, even regarding the choice of what is to be archived. These websites only have one form and one content that are identical for every user. The website on the web server consists mainly of static HTML pages, images and style sheets. The only active element in those websites is the navigation based on fixed hyperlinks. These websites may contain some dynamic elements, but typically (for example menu bar, roll-over images, automatic date, add to favourites, etc.) they are the result of the execution of client scripts.

Websites with static content are being archived as they are available on-line on the web server. All files in a standard file format are stored in their original format. Non-standard files are preferably transformed into an appropriate archiving format before being put into the electronic repository. A copy can simply be made of the files in their original structure as they are on the web server, because the on-line and off-line versions of these websites are identical and can be consulted. Such a copy is called a mirror. The original files reconstruct the website both on-line and off-line, on any computer platform.

Copying the files from the web server will usually imply that files that are obsolete and no longer linked will be put in the archive. These files sometimes remain on the server disk after an update but no longer belong to the on-line version of the website.

The log files of a website can also have archival value because of the contextual data they contain. The server log files can determine who visited the website, what files were most often consulted, how long certain web pages were viewed, what files were downloaded, etc. The web server software keeps the log files as flat text files or directly in a database. One can determine the frequency of the log files (daily/weekly). Most web servers offer the possibility to store the log files in a common log format or in a format that is specific to the web server software⁴⁸. Another option to store this contextual data is

⁴⁸ Several formats exist for the log files: Common Logfile, Combined Logfile, NCSA Common, W3C Extended, Microsoft IIS. A log file consists of sequential lines filled with ASCII characters. Log files usually have the extension log, lf or crlf. Each line contains a directive (for example version, fields, software (browser, version, resolution, operating system), referrer, start date, end date, date, remark: a line with a directive always begins with #) or an entry. Lines with entry data contain information about the HTTP actions. The fields within one line are being separated by commas. Directives contain indications about the logging process. Fields contain a description of the information that appears in each entry. The end of an entry is indicated by a CR or a CRLF indication. An example of a log file can be consulted on the Dutch pages of the DAVID website. A distinction is usually made between the unique visits (counter +1 per

to archive the statistical information that is generated based on the log files. Archiving all log files requires a lot of storage capacity and is complex. Specific software is available for the analysis and the extraction of statistics. The common log format is being supported by most analysing tools. It's also possible to take the necessary information (abstract, statistics) in the metadata file of the archived website.

Websites with form fields are also considered to be static websites. The approach taken for static websites suffices to reconstruct the website and its form(s). If the information that is sent to a web server has any archival value, it needs to be stored as well. The sent information is usually present with the receiver in the form of an e-mail or a database. Whether the server script, the executable programme for the processing of the form data and any linked documentation are to be archived depends again on a possible liability.

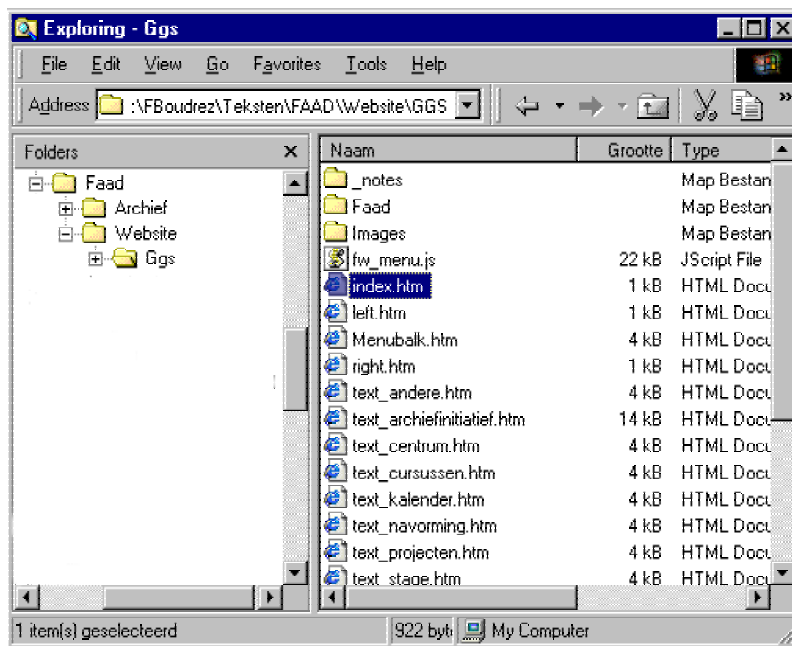


Image 4: An archived version of the website of the interuniversity archive training programme, a website with static content. The files are stored in their original format in the original folder structure. All files that compose the website are put together in one folder ('GGS'). This folder is the starting point for the relative path indications that link the files.

A log for uploading and downloading can be taken into consideration for the archiving of websites with static content. This log will keep track of what files were sent when to the web server. This file can be the basis for the deduction of the history⁴⁹, and possibly can help demonstrate what the content was at any given moment. The FTP log file for uploading and downloading can serve as a basis or the web administrator can keep this metadata in a separate file⁵⁰.

Except for the form fields, archived static websites keep their complete and original functionality.

computer that announces itself), pageviews (total number of viewed web pages) and hits (counter +1 per requested file).

⁴⁹ The Dutch webpages of the DAVID website contain an example of such a log. The following is kept in the log: date of the upload or download, time, source folder, target folder, upload or download and file name.

⁵⁰ In the file that is opened, it is possible to store all necessary data. The size of the log can be chosen. In an FTP log only a limited amount of data is registered. This file can suffice however when for example also the HTML headers contain some additional metadata.

B.2.2 Websites with dynamic content

The selection issue is more complex when archiving websites with dynamic content. The HTML pages here are only composed after the server receives a HTTP request or are being delivered via a linked application. The content of a web page can depend on the received query (for example consulting the timetable of the trains), the user profile or user preferences (for example via a cookie) or based on the information present in the linked document management system or the database. The question here is not just what part of the website is to be archived, but also whether the interaction should be stored. The content often depends on the server-client interaction and a reconstruction of the content is only possible if also this interaction is archived. Two main focal points arise: what needs to be archived for the website to be consultable later, and what parts need to be archived to store the interaction – and thus also the content.

For later consultation of most websites with dynamic content it is not enough to simply take a copy of the original files as they are on the web server. A first problem is that the website depends on the web server and the ‘deep web’. These websites are part of a complete digital information system and cannot function on their own. The website can only be displayed in its original form when the original web server configuration (among others virtual folders), the necessary software (server scripts and extra modules, server software, executable) and the linked file system (databases, document management systems, content management systems) remain active. Even though the website does not have to remain functioning, it cannot simply be disconnected from the information system that lies behind it. Dynamic and interactive websites may have a static HTML page as output in a browser, often these pages are the result of a composed and integrated information system. If this is attempted, the web browser will only display error messages.

A second difficulty that dynamic websites do not have a static content. Each visitor receives different content, depending on request or preferences. Then, what exactly is the content of the website? How can we know what information is published on a certain website? What do we have to preserve?

A variety of solutions has been suggested for both problems, but the actual solution still needs to be developed. Currently there is little or no expertise about this. Archiving websites on-line implies that the information system behind it needs to remain operational. This is no more than an application of the *computer museum strategy* and is not a solid option for long-term archiving. Once one single aspect of the whole information system is no longer available, the archived website can no longer be consulted. Emulation of these websites implies that there needs to be an emulator for the website itself and also for the web server software, the web server configuration, the scripts, the executables and the links with the databases behind. Furthermore, a different emulator would have to be made for each

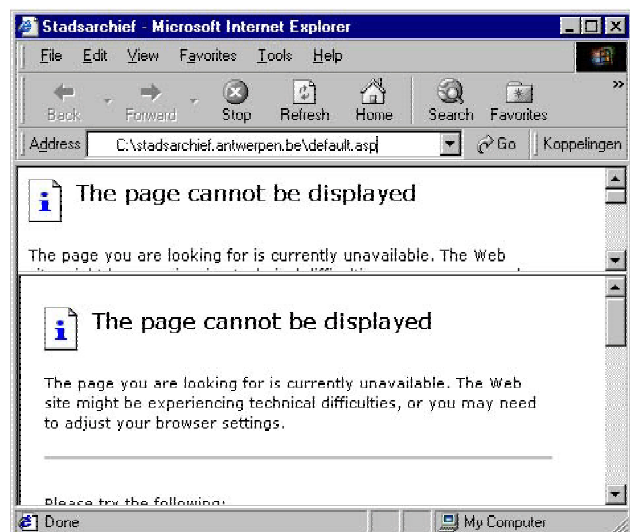


Image 5: Websites with dynamic content cannot be viewed in their original format without the server configuration and software. If one tries to view them anyway with only a web browser available, then error messages will appear where the information should have been.

website to be archived. This is not realistic. There are no known examples of application of this. Emulation can be used, however, for the web browser (see below, Digital Durability). The Universal Preservation Format⁵¹ (UPF) appeared to be another option, but due to a lack of funding a prototype has not yet been developed⁵². In Denmark and Scotland plans exist to film a computer screen while somebody is surfing the web⁵³. This is not a practical solution for large websites and would require a lot of time for consulting.

When archive these websites according to the predetermined quality demands, it is best to make a distinction between the different layers that compose a website. A starting point can be the division in layers of the whole information system described in the DAVID report *The digital recordkeeping system: inventory, information layers, and decision-making model as point of departure*⁵⁴. The three layers are the content, the structural or logical elements, and the tools. Each layer is to be archived separately, if they have archival value.

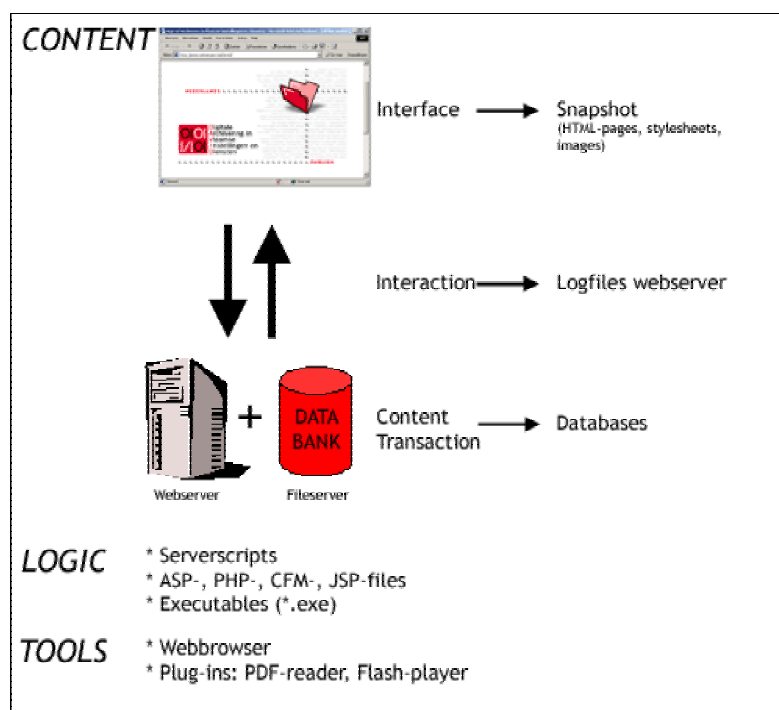


Image 6: The separate archiving of the parts that compose the information system. The interface is stored via a snapshot. Thus the way the website presented itself and the way information is shown on the WWW is archived. To know the content of the website we also archive the linked back office system and the log files of the web server. The logical elements, if required, are being copied directly from the web server. For the consultation also the matching web browser and possibly the relevant plug-ins are required.

A snapshot is an instant picture of a website as it was delivered to a web client. The dynamic web pages are transposed to static HTML pages on the web server and then sent to the web client. This static version of a website can be shown in a web browser without the web server and the ‘deep web’. This way the software dependency remains limited to the web browser and possibly some plug-ins. Similar to websites with static content, a collection of HTML pages, images and style sheets is stored.

⁵¹ The goal of the UPF is to be an archiving format for multimedia files that is independent regarding platform, application and carrier. A UPF file should also be self describing. All necessary data (metadata, technical specifications to access the content,...) are added to the file: encapsulation (<http://info.wgbh.org/upf/>).

⁵² E-mail of Tim Shepard, 23 July 2001

⁵³ B. CHRISTENSEN-DALSGAARD, *Archive Experience, not Data*, Lecture held in Darmstadt, 8 Sept. 2001; S. BORDWELL, *Objective-Archival preservation of websites*, Lecture held in London, 25 April 2002

⁵⁴ *Het digitaal archiveringssysteem: beheersinventaris, informatielagen en beslissingsmodel als uitgangspunt*, p. 15. The names of the three layers in an information system were adapted to websites (data → content, structural info → logics, tools).

The only difference between the active and the archived website is that interactive functionality like requesting data from a database is no longer available. The static HTML pages contain no more than possibly some client scripts. Archiving the original dynamic web pages and server script files is unnecessary for the off-line reconstruction of the website, but may be important because of liability or because of their proper informational value. This would imply that two different versions of the website are to be archived: a static version for consulting the website and a version with the original files as they were on the web server.

The content of active versions of dynamic websites is stored in databases, a document management system or a content management system. These linked file systems compose the ‘deep web’. A snapshot operation only rarely manages to catch the content of the ‘back office system’ too. This is possible for documents that are made available via a document management system, but not for databases that are being requested. Apart from the question whether the content of the information system behind can be stored as well during the snapshot operation, it appears to be rarely relevant to do so. For a number of reasons (such as processing time, multiple archiving of the same information when different versions of one website are being archived) it is better to keep the information within the back office system (as is the case with normal databases or content management systems) and to apply a different archiving strategy for this.

Version management will usually be vital for this type of ‘deep web’ archiving. When the content of a website comes from a linked database or web content management system, this will not be shown in the (FTP) log file about uploading and downloading of web pages. Separate log files are needed in this case within the database or the content management system.

What a client gets on the screen depends on a number of factors. Usually he or she will query the linked databases in some way or another. Archiving the query will show what information has been made available. The executed HTTP-Get commands containing the queries are stored in the log files of the web server. A web server stores among others the following information in its automated logs: visitors (IP address or domain name), date and time, pages visited, actions performed, web browser used, etc. These log files are not primarily intended to store interactions or transactions. They are usually kept for the needs of web designers and webmasters, and are not easy to decipher. Currently work is being performed on the standardisation of a log format that can be expanded according to the needs of the institution⁵⁵. It would be better to check beforehand whether the log files have a possible archival value. Firstly it is important to define what interactions have archival value so that essential data will definitely be enclosed in the log files. Secondly there is a need to compose an archiving strategy for these log files, as they may be lost otherwise. Keeping log files with archival value is a task one best does not entrust to a third party (for example an external internet provider). The log files a web administrator normally keeps will rarely suffice.

The content of the most recent generation of websites can be made dependent on the user profile, the software the user has at his or her disposal (web browser plus version), cookies etc. If this data matters, it needs to be stored together with the website. The used browser is usually stored automatically in most web server logs. The archiving of the client’s cookies can be a bit more problematic. The cookies are by default only stored on the hard disk of the web client. A format for log files exists that can incorporate the cookies as additional information, but this enlarges the log file considerably and makes it much harder to manipulate.

⁵⁵ See <http://www.w3.org/TR/WD-logfile>

The separate archiving of the different layers that compose a database-controlled website has as a disadvantage that some of the original functionality will be lost. Consulting or adding to a (document) database is no longer possible. By separating the different parts the integration and functionality are lost. That integration and functionality are based on software and logic, and keeping both operational conflicts with the desire to preserve as system independent as possible⁵⁶.

B.2.3 Conclusion

Website archiving consists of more than just copying web pages in the digital repository. For each website with archival value it needs to be examined what files are kept and in what form they are kept. This cannot be limited to the website alone: also script files, log files, user profiles and the ‘deep web’ can be taken into account for archiving. We are not just dealing with the reconstruction of the website itself but also with the storage of the necessary metadata and the living up to a liability and evidence duty.

Table 1: Summary: What part of websites need to be archived? The computer files with archival value that are not available to a web browser (*italic*) need to be copied from the web or application server.

LAYERS	STATIC CONTENT		DYNAMIC CONTENT	
	INPUT WEB SERVER	OUTPUT WEB BROWSER	INPUT WEB SERVER	OUTPUT WEB BROWSER
INTERFACE	HTML, XML GIF, JPEG, TIFF, PNG CSS, XSL	HTML, XML GIF, JPEG, TIFF, PNG CSS, XSL	HTML, ASP, PHP GIF, JPEG, TIFF, PNG, XML CSS, XSL	HTML, XML GIF, JPEG, TIFF, PNG
INTERACTION	client scripts embedded in HTML or separate files	client scripts embedded in HTML or separate files	client and server scripts <i>executables</i>	client scripts
CONTENT	HTML, txt, PDF, doc, rtf, xls, mdb, zip	HTML, txt, PDF, doc, rtf, xls, mdb, zip	HTML, txt, PDF, doc, mdb, <i>databases</i> <i>document and</i> <i>content management</i> <i>systems, log files</i> <i>web server and</i> <i>databases</i>	HTML, txt, PDF, doc, mdb
TRANSACTION	HTML forms <i>e-mails</i> <i>databases</i> <i>log files server</i>	HTML forms	HTML forms <i>e-mails and</i> <i>databases with form</i> <i>data, log files web</i> <i>server</i>	HTML forms
DOCUMENTATION	<i>description of the</i> <i>function of the</i> <i>website within the</i> <i>work process, of</i> <i>the working of the</i> <i>website, etc</i>	/	<i>description of the</i> <i>function of the</i> <i>website within the</i> <i>work process, of the</i> <i>working of the</i> <i>website, of the linked</i> <i>database, etc</i>	/

⁵⁶ The Australian Directives that archived websites have operational functionality as an important demand. The governments are required to archive dynamically generated on-line sources in a functional state. (*Archiving Web Resources: Guidelines for Keeping Records of Web-based Activity in the Commonwealth Government*, p. 12 en 26). The Directive remains vague about the actual execution, however.

C. HOW TO ARCHIVE WEBSITES?

C.1 Websites with static content

Websites with static content are archived via a mirror. A mirror is an identical copy of the files in the same file format, with the same file names and in the same structure as on the web server. Anybody who possesses the correct web browser can consult the archived website.

A first method is the direct copying of the files on the web server. Access to the hard disk of the web server is required. This can be done either by making a copy of all files on the web server itself or by using an FTP programme. Active co-operation of the creator is necessary to make a copy on the web server. The creator transfers the mirror to the archivist (“push”: tape, CD⁵⁷) or allows the archivist to access the server via FTP. An FTP programme allows placing the files on a web server from any client computer. This requires special access rights though. A common consequence is that too many computer files are being archived at once. The hard discs of websites are usually not good examples of efficient and rational file management. Obsolete and recent files are put next to each other in an unorganised way and when they are copied many files that are not linked to (any more) will be archived as well. The two main conditions for a successful application of this technique are the use of relative links for internal linking and the copying of the folder structure of the web server⁵⁸. The oldest static websites used a lot of absolute path indications for internal linking. An automatic transformation from internal absolute links into relative links is often not possible⁵⁹. In this case, the second method would be more appropriate.

This second method consists of the archivist working fully independently, copying the necessary files using an off-line browser (see below). The off-line browser can transfer absolute links into relative links, if necessary. The archivist can use an off-line browser to archive a website himself (“pull”). Contrary to FTP access, an off-line browser does not allow the archivist to modify the on-line website. However, an off-line browser can only access those files that are available to web clients. When there are no old files, linked databases or log files to be archived, an off-line browser can suffice.

Websites in Flash are considered to be websites with static content. In theory both methods can be applied, but the archiving of some Flash websites by the Antwerp City Archives has shown that the making of a mirror via an off-line browser was not always successful. For a number of sites the copying of the files on the web server turned out to be the easiest and quickest solution. There is a possibility that Flash sites can only be archived with some help from the creator. It is hard to transform absolute internal links to relative links. Current off-line browsers cannot access the binary *.swf files.

⁵⁷ The use of a CD can cause numerous difficulties. Exchangable CDs should comply with ISO-9660. This standard contains some limitations concerning the number of characters in file and folder names. Shortening the file names implies that all links need to be adapted too. Using extensions to ISO-9660 (Joliet, Rock Ridge) or restricting CDs to short term storage only can offer a solution.

⁵⁸ Because of the need to copy the identical folder structure of the web server, one could consider to transpose a static website to a TAR file and transfer it like that. TAR files are operating system dependent, however, and need to be transformed again immediately after deposit. Extraction recovers the folder structure and the separate files.

⁵⁹ Non-operational internal absolute links can be automatically localised. Transposing them to relative links however needs to be done manually. It could be an option to put the website that has been placed on a CD back on a web server and then copy it with an off-line browser.

It is usually impossible to manually adapt the links, because when *.fla files are published as *.swf files, they are usually protected against changes or import actions. The creator therefore has to adapt the internal absolute links or has to give the archivist the password or even hand over the original *.fla files.

On-line web browsers (Internet Explorer, Netscape, Opera, etc.) can be used to put one web page off-line ('save page as' command), but are not capable of archiving a complete website⁶⁰. Back-ups of the website are not sufficient archiving methods either. Back-up procedures are part of the current management and security procedures for websites, but cannot be used as recordkeeping system⁶¹.

C.2 Websites with dynamic content

C.2.1 Snapshots

In order to store interactive websites in such a way that they can be consulted without their original web server configuration and software, more than just a copy of the original files needs to be archived. Static HTML pages are being archived instead of the original ASP, PHP or JSP files. Special computer programmes are being used for the capture of interactive websites in the form of HTML files: off-line browsers.

Off-line browsers are computer programmes that can put preselected websites off-line. A separate project is being opened for the capture of each website. Within each project the settings can be defined (start URL, depth of the links, exclusions, etc.). Off-line browsers were originally designed to make local copies of commonly used websites, so that these could be consulted in the future at no extra telephone cost. The first generation of off-line browsers only served to make a copy of the website files as they were on the web server. In those days most websites were static anyway. Meanwhile the youngest generation off-line browsers has adapted itself to the evolution of websites and now contains the possibility to save dynamically composed web pages as static HTML pages. The original extensions like ASP, PHP and JSP are being transformed into HTML so that the pages can be viewed off-line from any computer with a web browser⁶². A file named 'default.asp' will be stored as 'default.asp.htm' or 'default.htm'. Some programmes will give the files a new name: the original file name is being replaced by the URL that appears in the address bar of an on-line browser. Static HTML pages on the web server are not being changed and will be put off-line as HTML files while keeping their original appearance.

⁶⁰ Only the HTML pages in the active window are stored locally. The result is that only one web page at a time is put off-line. To store a website that exists of multiple pages, each page has to be put off-line separately. Sometimes only the frame set is stored and not the web pages with the content. The predetermined quality demands are not met: internal links are absolute, common images are stored again each time, the second layer of roll-over images is not stored, the connection between the web pages of a website is lost, and the original file structure and names are not copied.

⁶¹ The experience with the first versions of the website of the City of Antwerp illustrate this clearly. The two first version were still kept on back-up tape. Archiving these versions was a complex enterprise with fortunately a happy ending. The third version could not be recuperated at all. (F. BOUDREZ, *Van backup tot gearchiveerde website. De archivering van de eerste versies van de Digitale Metropool Antwerpen*, Antwerp, 2002).

⁶² ASP, PHP or JSP files cannot be viewed off-line in a normal web browser. They require specific software like Ms Personal Web Server. This will enhance the software dependability, and this solution cannot be applied to a large number of archived websites.

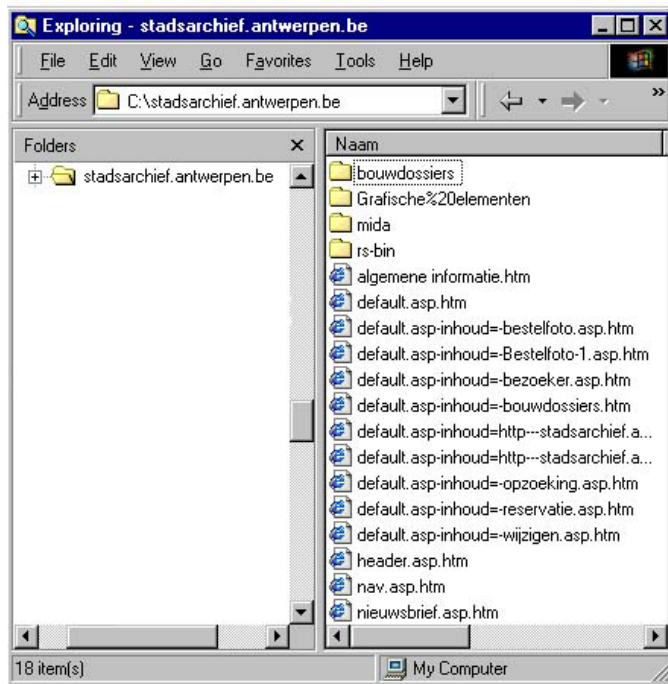


Image 7: The website of the Antwerp City Archives contains a number of web pages that are only being composed after an HTTP request has been received. The dynamic web pages are linked to a document management system and access those documents that are available to the public. These pages are archived as static HTML pages. When archiving the website all ASP files were stored as static HTML pages. Their original file name was kept but an extension .htm and the request was added.

The archived web pages show what files were available at the moment of the snapshot. The documents themselves were not included in the snapshot, but remain available within the document management system.

A harvester or an off-line browser takes a snapshot via the network. This operation may not always be without difficulties and sometimes errors occur. Choosing a good off-line browser can solve some problems, but some structural problems will still remain:

- ➔ It is hard to determine the boundaries of a website. Most programmes offer the possibility to limit a snapshot to all files that can be found within the same start URL, but folders outside it will then not be stored. A virtual folder with for example a common banner for all sites of one institution will not be included in the snapshot when the start URL is specified as that of one institution. Another problem is caused by the ‘redirects’ that occur in many websites. Most programmes ask to determine the number of levels (depth of the links) to be stored. Before an exact number is indicated here, the number of levels needs to be known as otherwise part of the website will not be archived. In most cases however more files than necessary will be archived. Also files that do not belong to the website will be put off-line. It is best to remove those superfluous files afterwards.
- ➔ Not all websites can be archived via an off-line browser or a harvester. This method is limited to (parts of) websites that are accessible to the public. Intranet sites or secured parts of a website cannot be put off-line unless the archivist has access rights.
- ➔ Only the websites will be archived. Server logs, linked databases (‘deep web’) and logs about uploading and downloading of the website cannot be captured.
- ➔ Snapshots can only be taken of active websites or files. Only those snapshot taken at certain times are available. If a website has changed multiple times between two snapshots, some versions will not be archived.
- ➔ The second layer of roll-over images, server-sided image maps, DTDs and XSL style sheets are not always stored. Current off-line browsers have great difficulty to put websites with Flash applications locally.
- ➔ Virtual folders: part of a website can be stored in a virtual folder. Most off-line browsers

will not archive the files in the virtual folder when another server name is used in the web address. Another problem is caused by the absolute path indication to reach the files in the virtual folder.

- ➡ Taking a snapshot is time consuming and takes several hours for a large website. This causes difficulties especially for very dynamic sites (for example a website of a newspaper that is constantly adapted). It is possible that already during its capture the website is being changed and that a version that never really existed is put in the archive. The first and last stored web page may belong to a different version⁶³.
- ➡ Errors occur frequently when taking a snapshot: hyperlinks no longer work, files are not available, the server gets overloaded, the server gets time-outs, scripts cause infinite loops, etc. If the site is updated while the snapshot is being taken, it is possible that an irreparable error will occur⁶⁴.

There is more than one disadvantage to the use of off-line browsers, but in a selective approach there are no alternatives to store “on the fly” generated web pages. The disadvantages can be countered by choosing a good off-line browser and repairing the errors in a snapshot.

The choice of an off-line browser is very important. A whole range of off-line browsers is available as freeware or shareware on the Internet⁶⁵. A good off-line browser should offer at least the following functions:

- Choice between an identical copy (for websites with static content), a reconstructable snapshot (for websites with dynamic content) and the putting off-line of one specific web page⁶⁶.
- Limitation of the capturing of files to a given start domain. Off-line browsers follow the links to copy the next files. If the browser does not limit itself to the given domain, external links will be followed and other websites will be copied.
- Correct follow-up of redirects within a website.
- Transformation of absolute path indications for links within the own website into relative path indications.
- Copying of the original folder structure of the web server.
- Allow selection of the type of files to be put off-line. The settings of off-line browsers allow an indication of which types of files should be copied and which should not. For example PDF files and pages with executables can be skipped when these are being archived separately.

⁶³ A known example of this is the announcement of an event in the future on the front page of a digital newspaper, while further on in the archived website a report can be found.

⁶⁴ J. HAKALA, *Collecting and Preserving the Web: Developing and Testing the NEDLIB Harvester*, in *RLG-DigiNews*, 15 April, 2001, vol.5, nr. 2

⁶⁵ An overview of off-line browsers is available on <http://www.tucows.nl>. See also <http://www.davecentral.com/browse/67/> for an overview and brief reviews.

⁶⁶ An off-line browser does not allow the production of identical copies of dynamic web pages. An off-line browser only receives HTML files and for example no ASP or PHP files. These last files can keep their original extension when put off-line, but are actually just the HTML versions of the original ASP, PHP or JSP files. The files do no longer contain any scripting.

- ✓ A number of off-line browsers add the name of the used browser programme to the HTML script as HTML comments. If comments are added, they should preferably consist of data about the snapshot operation (metadata: date, time, title website, etc.)
- ✓ Foresee version and duplication control so that active files, if necessary, can be compared to already copied files. Most programmes allow specific archiving only of those pages that have changed after a given date.
- ✓ Report the errors that occur while taking a snapshot. Errors do occur sometimes, and they should be stored in a log file.
- ✓ Possibility to announce themselves to the web server as different types of on-line browsers. A number of websites are protected and ban off-line browsers or robots. The website can have a browser check connected to it, making it desirable to announce the off-line browser as a specific browser.
- ✓ Spread the requests to the web server and allow the user to determine the number of files that are requested simultaneously. To avoid overloading and robot exclusion, the number of files requested at the same time should be limited.

Using a browser that complies with these demands is as such not yet enough a guarantee for a quality archiving. Successive snapshot operations with the same off-line browser can sometimes give different results. Factors such as server overload and time can play their role.

Snapshots have to be checked for errors and anomalies. Computer programmes can perform this quality check. A good off-line browser or harvester will report automatically on the number of captured files and the errors that occurred. Such a log file is an important indicator, but as such not sufficient. Specific programmes and on-line services exist for the check-up of websites and these can also be used for a quality check of archived or to-be archived websites⁶⁷. These programmes can perform a thorough check: validity of internal and external links, file names, HTML syntax and validity of attributes, presence of all necessary computer files, compatibility with certain web browsers, composition of forms, undefined anchors, etc⁶⁸. It is especially important that the links are still functioning. Based on these links the website is being reconstructed, the link between two computer files is determined, and access is given to matching computer files. The mirror or the snapshot cannot be functional without correct links. The correction of obsolete links is a priority. These check-up tools can localise the most common problems and maybe even solve them. E-mail addresses, forms and external links can be switched off if desired. If necessary these parts of the archived website can refer to the still active on-line version. The archivist should perform a last check.

There are problems, however, that can only be resolved manually. When archiving version 6 of the DAVID website, the snapshot had to be corrected at 4 points: 1. The second layer of the roll-over images in the navigation panel was manually copied to the matching folder, 2. The DTDs and XSL style sheets of the examples of the archived voters register and e-mails were put together with the XML files, 3. The file name ‘Grafische elementen’ [Graphical elements] that the off-line browser changed to ‘Grafische%20elementen’ had to be renamed again so that the links are working again, 4.

⁶⁷ Most HTML editors contain a link checker. Examples of on-line services are: <http://validator.w3.org/checklink> and <http://www.cast.org/bobby/>.

⁶⁸ Dead external links will be a common error when there is no Internet connection or when downloads are not stored together with the archived website.

The counter on the homepage has been disabled: the source code for the counter remained where it was but was redefined as HTML comments.

Storing the original ASP or PHP files is only possible when one has access to the hard disk of the web server (copy, FTP access).

C.2.2 Log files

Log files with archival value will be kept as flat text files or stored in a database on the web/file server and will be archived accordingly. The electronic records will be added to the recordkeeping system.

C.2.3 Databases

Archiving of linked (document) databases is a problem that we cannot deal with at length here. The versions of a database can be archived by installing a history for each object in an object-oriented database, taking snapshots of the (relational) database or keeping logs of the changes with regard to the original version (audit trails). (Document) databases with a permanent archival value need to satisfy the demands concerning system independent archiving. Normal back-ups are not taken into account for this.

D. FREQUENCY

D.1. Determination of the frequency

Websites have the advantage that they can be adapted quickly. The different versions of a website can follow up each other very quickly. This needs to be taken into account when archiving websites. The average life expectancy of a website is estimated to about 75 to 100 days⁶⁹, but this notion does not provide much help for archiving purposes. In general we can conclude that a static website undergoes less changes than a dynamic website.

When determining the frequency at which websites need to be archived, one of the biggest difficulties is that changes to websites do not occur in a fixed pattern but more at random. For each website the archiving frequency needs to be determined. Important indicators are the nature, the goal and of course the frequency of adaptation of the website. Archiving the website of a presidential candidate every day or even twice a day may be necessary for a possible historical investigation into the campaigns. The website of a political party will undergo more changes in the period before and after the elections than at other times. The website of a newspaper often changes every day or

⁶⁹ D. SHENK, *The world wide library*, in *Hotwired*, 2 Sept. 1997 (http://hotwired.lycos.com/synapse/feature/97/35/shenk1a_text.html); L. DEMBART, *Go Wayback. Internet Archive stores pages long gone from Web*, in *International Herald Tribune*, p. 13

sometimes even several times a day. Because of the documentary or historical value, archiving at a higher frequency may be necessary. This will always be the case when websites are used in a work process that one is accountable for. Archiving only the final version can be sufficient for websites of temporary organisations or initiatives.

The frequency is not only determined by the website itself but also by the acquisition policy and profile of the archive institution. The frequency may vary from the occasional archiving of a momentary snapshot to the archiving of every updated version. When archiving is required because of accountability or juridical reasons, it is absolutely essential that each version is stored or at least reconstructable.

D.2. How to archive different versions?

It is best to start with archiving the complete website. When archiving the changes there are several options: either be limited to the changes or archive the whole website again. In the first case it must be known which files have been changed and which have not. Off-line browsers can determine which files have been changed in comparison with an older version, but it may happen that older file versions become overwritten with more recent ones. It should also be considered whether taking a new mirror or snapshot is more efficient. The main parameters here are the available storage space, the reconstruction of the website and the amount of time required. Considering the fact that the file size usually is not a problem, this will be the most appropriate option in most cases. When archiving momentary snapshots at a low frequency, it might be useful to archive the complete website.

Storing all versions of a website is almost impossible without the active participation of the creator. It is often their task anyway. The creator knows when the website has been changed. Technically, archiving all versions from a distance is possible via an off-line browser, a harvester or a spider. However, this is very demanding for the network and information might be lost. It is much more efficient if the creator keeps a log of the changes or notifies the archivist when important changes occur. It will be easier to achieve this for the archive department of a public institution (archiving their own website(s)) than for records offices and documentation centres that archive websites of other institutions or societies. The Documentation Centre for Dutch Political Parties cannot count on the active co-operation of the political parties to archive their websites and therefore has to make mirrors and snapshots at their own initiative⁷⁰. In that case a webservice or a tool who notifies the archivist when a website changes, can be of interest.

D.2.1 The changes to websites with static content

For websites with static content it suffices to incorporate the changed original files in the electronic recordkeeping system. They can be added to the folder where the complete version is stored. Unique file names or placing new files in a separate folder prevent new versions from overwriting old files.

⁷⁰ Announcement Gerrit Voerman, Documentatiecentrum Nederlandse Politieke Partijen [Documentation Centre for Dutch Political Parties]

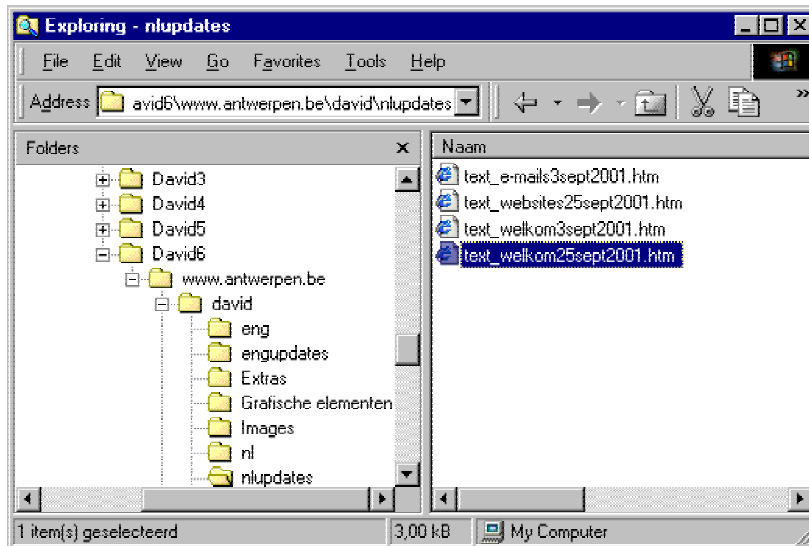


Image 8: Archiving the updates separately of version 6 of the DAVID website. The folder David6 contains a snapshot of the full website taken immediately after the website was put on-line (6 July). The files that were changed since then are kept in the folders 'nlupdates' and 'engupdates'. The date on which the update was put on-line is added to the original file name. The folders 'nlupdates' and 'engupdates' are on the same level as the folders 'nl' and 'eng' so that all links to images in the folder 'Images' remain operational and that the same image is not archived twice. As the updates have a unique name (via the date) they could also be put in the folder with the original files.

D.2.2 Changes to websites with dynamic content

A website that undergoes frequent changes is usually composed in a dynamic way. Adapting static web pages is much too time consuming if adaptations occur daily or weekly. The content is delivered via a 'back office' information system. When archiving the different versions one needs to consider a version management and an archiving of this 'back office' information system. In practice this will usually be done via some sort of database archiving.

Taking into account the speed at which dynamic websites are updated, the simplest option seems to be the incorporation of an automatic version or history management tool in the 'back office' information system. This way the labour intensity can be limited. Such a module can only rarely be incorporated in an existing system, so it is important to undertake the necessary steps during the design phase. Nowadays a lot of attention is paid to this. Keeping track of the changes to a website is one of the basic principles of *web content management*. If this option is chosen, the archivist must examine whether the records management functionalities of the web content solution fulfill the necessary requirements.

The website itself will only need to be archived when something has been changed to the website as an interface. This could be a change in layout, images or style sheets.

E. DIGITAL DURABILITY

A number of strategies can be applied to ensure the readability of archived digital documents⁷¹. The durable digital archiving of websites demonstrates that these strategies do not exclude each other but can be used next to and even in combination with each other.

Websites consist of a collection of linked computer files in a variety of file formats. To ensure digital durability in the long run, the HTML tags and attributes need to be interpreted correctly, the

⁷¹ The general archiving strategies are described in *The digital recordkeeping system: inventory, information layers, and decision-making model as point of departure*, p. 7-11 on the DAVID-website.

clientscripts need to remain executable and the linked files (images, animation, documents) have to remain readable. This will depend on their compatibility with the common web browsers and plugins⁷².

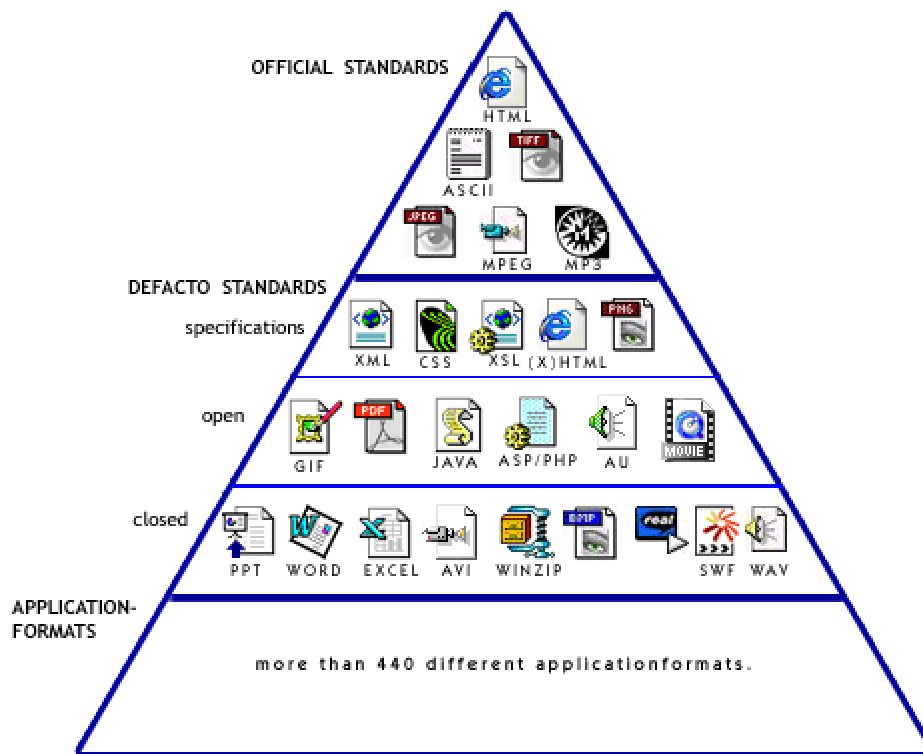
The Swedish and Finnish harvesting experiences of their own web spaces have shown that more than 440 different file formats occur in the website archive. However, this number does not give the accurate picture. There is a large standardisation going on in the field of used file formats. This is clearly because web designers are aware that as many users as possible should be able to consult the website from a wide variety of platforms. This can only be achieved when the used file formats have a status of standard. The most common file formats that compose a website (see Table 2) are all on top in the hierarchy of file formats (see Image 9). Non-standardised file formats only appear sporadically. These files are best transformed into a standardised format that copies the original characteristics as much as possible. The collection of non-standardised files contains many files that cannot be recognised because of their extension or their MIME type.

Table 2: The most common file formats that compose archived websites

TYPE	FILE FORMATS
HYPertext	htm/html, shtm/shtml, xhtml, stm/stml, xml,
IMAGES	gif, tif/tiff, jpeg/jpg, bmp, png, xbm, fif
SOUND	au, wav, mid/midi, mp2/3, riff,
VIDEO	mpg/mpeg, avi, ani, qt, mov, fli, flc
ANIMATION	animated gif, swf,
DOCUMENTS	doc, pdf, txt/text, wrl, xml, xls, xlw
DOWNLOADS	zip, rar, arj
SCRIPTS	ASP, JAVAservlets, -applets, -script (.js, .class), PHP, VB, cfm, jsp

⁷² The library world is working on the consolidation of website archives, but has no fixed solution yet for the assurance of digital sustainability. One is aware of the problem but has more than one answer to it. It boils down to the idea that all archiving strategies should be applied simultaneously: storing original hardware and software, migration and emulation.

Image 9: Hierarchy of the common file formats



Storing web pages with dynamic content as static HTML pages facilitates further. A consequence of this method is that HTML pages are being stored in the archive instead of ASP or PHP files, which avoids platform dependence. Consulting these original files will take place via a normal text editor anyway. The Swedish and Finnish harvest operations show that HTML, GIF and JPEG files make up more than 95 % of the total amount of files.

Table 3: Percentage of file formats in number and size in the Swedish website archive (2000)⁷³.

TYPE	NUMBER	SIZE
HTML	52 %	13,9 %
GIF	24 %	6,8 %
JPEG	20 %	15,9 %
TIFF		3,5 %
PNG	0,3 %	
PDF	1,3 %	10,3 %
ASCII	2 %	9,8 %
Octet-stream	0,9 %	11,5 %
Postscript	0,3 %	3,8 %
MSWord	0,3 %	
Real audio	0,2 %	
WAV	0,07 %	
Zip	0,4 %	6,5 %
MPEG		2 %

⁷³ K. PERSSON, *The Kulturarw3 Project - The Swedish Royal Web Archiv3e*, Lecture held in Svetlogorsk, Aug. 2000. <http://kulturarw3.kb.se/html/statistik.html>. The total amount of HTML, GIF and JPEG files in this table is 94%. During later harvest operations this percentage has grown to about 97 to 98%. The statistics of the archiving of the Finnish and Austrian web space confirm these figures. (A. ARVIDSON, *Harvesting the Swedish webspace*; J. HAKALA, *Harvesting the Finnish Web space - practical experiences*; A. ASCHENBRENNER, *Long-Term Preservation of Digital Material*, p. 76).

Whether archived websites can be consulted in the future in their current format depends largely on the available web browsers⁷⁴. The youngest generations of web browsers are backward compatible, support XML, show images and execute client scripts. HyperText Mark-up Language is always developing. HTML is a mark-up language that defines the presentation of a web page. HTML files contain both content and layout information. The standards for the HTML language have been defined by the W3C at the MIT Laboratory for Computer Science. HTML is the standard for publishing hypertext on the Internet. HTML exists in a number of versions. The first version was spread from March 1993 onwards. Widespread versions are 2.0 (1994), 3.2 (1996), 4.0 (1997), and 4.01 (1999, ISO 15445:2000). The most recent versions of HTML are Dynamic HTML (DHTML) and XHTML 1.0. XHTML is a rephrasing of HTML in XML⁷⁵. DHTML is the marketing term for a combination of HTML, style sheets, DOM and scripting. DHTML is not a formal standard.

The younger versions of HTML are largely compatible with previous versions. Each new version of HTML introduces new tags and attributes however, and some existing tags or attributes are considered to be obsolete and sometimes even removed⁷⁶. This is largely a consequence of the development of style sheets (CSS, XSL). HTML syntax was more and more being tuned towards the use of style sheets. HTML and style sheets have become complementary so that the principle of separating mark-up and layout is turning into reality. The web browsers are currently being adapted to the new HTML and style sheet evolution, and vice versa. Usually the new HTML versions are a fine-tuning and elaboration of previous versions, and web browsers are able to read HTML pages in an older version. Chances are however that a number of tags or attributes no longer can be executed and may cause problems. This can also occur when an HTML page contains non-standard, and therefore sometimes non-supported, HTML tags. This could be the case when an HTML editor generates code that is only supported by the browser of the same producer (for example Netscape: <CENTER>, Internet Explorer: <MARQUEE>). These non-supported or non-standardised HTML tags can become a problem for the readability of the websites in the long run. Replacing the obsolete or non-standardised tags by new tags or by tags with the same function would solve this. This boils down to a conversion of the HTML files.

Conversion implies the rewriting or reprogramming of the HTML syntax. As this process can (almost) not be fully automated, it is clear that this migration is not practical for large-scale operations. Conversion requires manual enhancement, requires a thorough knowledge of (X)HTML and is labour

⁷⁴ A website to read about the evolution of websites and the necessary plug-ins is <http://browserwatch.Internet.com>. It contains a good summary of the software that is being installed with a browser, automatically or not. This information is usually also available via the Help button of the web browser.

⁷⁵ More background information about the use of XML as language for web applications is available on <http://www.w3.org/MarkUp>.

⁷⁶ Example: in HTML 4.0 the tag <PRE> was introduced, replacing the tags <XMP>, <PLAINTEXT> and <LISTING>. Other new tags in HTML 4.01 are: <ABBR>, <ACRONYM>, <BDO>, <BUTTON>, <COL>, <COLGROUP>, , <FIELDSET>, <FRAME>, <FRAMESET>, <IFRAME>, <INS>, <LABEL>, <LEGEND>, <NOFRAMES>, <NOSCRIPT>, <OBJECT>, <OPTGROUP>, <PARAM>, , <TBODY>, <TFOOT>, <THEAD> and <Q>. Examples of deprecated HTML tags are: <APPLET>, <BASEFONT>, <CENTER>, <DIR>, , <ISINDEX>, <MENU>, <S>, <STRIKE>, <U>. Also attributes can be 'deprecated'. In HTML 4.01 the following attributes of the element BODY are deprecated: "background", "text", "link", "vlink", "alink".

intensive, even though tools are available⁷⁷. Conversion from HTML tags seems appropriate when adjusting non-standardised tags and incorrect attributes of HTML files before ingesting them into the website archive⁷⁸. However this can be postponed a bit: most current web browsers are very relaxed about applying the HTML syntax rules and can still display web pages with erroneous code.

Emulation of the necessary web browsers, on the other hand, seems to be a more appropriate answer to the HTML evolution in the long run. After emulation HTML files with older tags do not require conversions or migrations and can be viewed in their original form. There is no current need for emulators because the present-day browsers (Internet Explorer 6.x, Netscape 6.x) are also capable of displaying websites based on the older HTML specifications. With the arrival of (X)HTML and the further development of style sheets, it is not unthinkable that HTML support will diminish in the near future. An emulation of the most recent HTML browser that also supports older versions should suffice.

The current generation of web browsers consists of graphical browsers that are capable to put images in different formats (GIF, TIFF, JPEG, PNG) on the screen. The web browsers follow the evolution concerning file formats for images, and have to be able to perform the matching decompressions (for example JPEG2000: autumn 2001). There is a connection between the browser programmes on the one hand and the HTML versions and the images that web pages contain on the other hand.

A fourth requirement for the web browser includes the execution of scripts that take place on the client side. Older web browsers cannot execute JAVA or VB script. The necessary software for the execution of scripts can be present in the web browser as a standard or can be installed with the web browser as a plug-in.

Finally the web browser needs to be able to correctly display web pages that are linked to style sheets. The evolution is tending towards a separation of HTML mark-up and layout (CSS, XSL). A correct application of the most recent (X)HTML and XML standards makes the use of style sheets essential.

The digital durability of the website plug-ins is a different story. These plug-ins are usually system and producer dependent (for example Flash Player, Real Player, Live 3D, Shockwave, Real Audio, Acrobat Reader, etc.). Only in a few cases can the producer dependent plug-in be transposed to a neutral plug-in. There is no problem as long as new versions of the plug-in assure backward compatibility. An alternative will have to be found the moment that this compatibility is no longer available.

Websites that are Flash applications will probably require an individual approach. A web browser does not suffice to consult these websites; the specific Flash Player is required. The very specific Flash functionality in the field of animation and vector images and its producer dependency makes the archiving of Flash websites difficult. Emulation of the plug-ins could be one solution. Transposing the Flash files into a sequence of more standardised file formats could be another. It is hard to predict the viability of both possibilities. As most Flash websites also have a static HTML brother, it might be

⁷⁷ Within the framework of the *World Wide Web Consortium* a group of volunteers has developed a tool to find and enhance errors in the HTML syntax. For more information, the source code and available downloads, see <http://www.w3.org/People/Raggett/tidy/>.

⁷⁸ An analysis of the website collection of the Pandora project turned out that this is a real threat: 7 million non-standardised HTML tags and 14 million HTML tags with wrong attributes (W. CATHRO, C. WEBB and J. WHITING, *Archiving the web: the pandora archive at the National Library of Australia*).

better to be on the safe side and archive also the HTML version. Filming the website in Flash with a screenrecorder is another possibility⁷⁹.

A website's download files with archival value should preferably be transposed into a suited archiving format. Transposition causes the files to receive a new extension and that thus the links need to be adapted. In most cases the on-line digital information is stored in a standard format (for example *.doc, *.pdf). Transposing the audio-visual streaming files into normal audio-visual files should be considered. These last file formats are usually more standardised and this could avoid another instance of software dependency.

F. MANAGEMENT OF THE ARCHIVED WEBSITES

F.1 Metadata

Next to the classical metadata of electronic records, the metadata of archived websites contains some features that are specific to websites:

GENERAL

- ▶ Title
- ▶ Creator
- ▶ Topic/abstract
- ▶ Function/goal
- ▶ On-line versions
- ▶ Counter:
 - Counter when putting on-line
 - Counter when archiving
 - Counter when putting off-line
- ▶ Public

WEB SERVER

PLATFORM

- ▶ Hardware
- ▶ Operating system
- ▶ Web server configuration and software
- ▶ Server scripts
 - CGI
 - ASP
 - JSP
 - PHP
 - Others
- ▶ Executable programs
- ▶ Links with applications

⁷⁹ On the Dutch pages of the DAVID website is a moviefile available with some images of the DAVID-website.

LOG FILES

- ▶ Content
- ▶ Frequency

DOCUMENTATION (development and administration, processes, procedures, etc.)

MIRROR/SNAPSHOT

GENERAL

- ▶ Archived version (HTML, static/dynamix, Flash)
- ▶ Version number
- ▶ Webmaster
- ▶ Web design
- ▶ Content responsables
- ▶ Languages
- ▶ Modifications
- ▶ Missing files

TECHNICAL DETAILS

- ▶ URL / IP address
- ▶ Start page
- ▶ Sitemap
- ▶ Number of files
- ▶ Number of folders
- ▶ Total file size
- ▶ Password(s)

FILE FORMATS AND VERSIONS

- (X)HTML
- Text files:
 - ASCII
 - PDF
 - MS Word
 - WordPerfect
 - others
- Audio files:
 - AU
 - WAV
 - MP3
 - MIDI
 - others
- Video files:
 - MPEG
 - MOV/QT
 - AVI
 - others
- Image files:
 - GIF
 - JPEG

- TIFF
 - PNG
 - others
- ❑ Animation applications:
 - Animated gifs
 - Shockwave Flash Player / Movie
 - others
- ❑ Executable programs
- ❑ Client scripts
 - Java
 - ActiveX
 - Web browser(s)
 - Plug-ins

HISTORY

- ▶ Date of availability on-line
- ▶ Modifications/updates
- ▶ Date of removal
- ▶ Date of snapshot and ingestion into recordkeeping system

SOFTWARE NEEDED FOR CONSULTING

- ▶ Web browser
- ▶ Plug-ins

ERRORS AND REMARKS

- ▶ Errors / shortcomings
- ▶ Remarks

The metadata listed above can almost only be collected with the co-operation of the creator. It is very time consuming for the archivist to fill these metadata fields afterwards, and sometimes it is even impossible. Some metadata fields can only be communicated by the creator, as they cannot be deduced from the available files. One could consider asking the creator to fill in a (digital) form. When the archivist has to collect the metadata anyway, software that is available for the testing of websites is the best method. Some technical details (number of web pages, used file formats, etc.) are kept by these programmes as statistical information and can easily be copied into the metadata.

The metadata itself can be stored on a number of locations. There are several options. As it is possible to add header information to HTML pages, one way would be to store the metadata in the header of the web page as HTML meta-tags. Another option is to store the metadata in a separate computer file (for example an XML file⁸⁰). Even a combination of both options is possible. In both cases the metadata is stored on the website level. However if there is a desire to store some metadata for each file separately, the multipart MIME object (rfc 2045-2049⁸¹) offers an alternative. This allows

⁸⁰ An example of a model XML file, that can also be downloaded, is available on the DAVID website, together with a matching style sheet.

⁸¹ <http://www.rfc-editor.org/rfc/rfc2045-2049.txt>

the addition of a header with some metadata information to each type of file. A similar option is the extension of the standard HTTP header by adding extra metadata information via the web harvester or the off-line browser. The computer file will then contain information about the archiving process, information about the object and finally the data of the archived object itself. When choosing where to store the metadata, user-friendliness should be taken into account, and considering that storage has to be secure, efficient and sustainable. The archive users should be able to consult the metadata in an easy and clear way.

```

1  website archive City Archives of Antwerp
2  website from: DAVID-project
3  website captured on: Tuesday April, 9th
4  version: 7.0
5
6
7  HTTP/1.1 200 OK
8  Server: Microsoft-IIS/4.0
9  Content-Location: http://www.antwerpen.be/david/default.htm
10 Date: Tue, 09 Apr 2002 07:32:39 GMT
11 Content-Type: text/html
12 Accept-Ranges: bytes
13 Last-Modified: Wed, 04 Jul 2001 06:32:05 GMT
14 ETag: "12dfbca534c11:3526e"
15 Content-Length: 1389
16
17
18
19 <html>
20 <head>
21 <title>DAVID - Digitale Archivering in Vlaamse Instellingen en Diensten</title>
22 <meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
23 <meta name="AUTEUR" content="F. Boudrez, S. Van den Eynde">
24 <meta name="WEBADMINISTRATOR" content="F. Boudrez">
25 <meta name="POSTING" content="2 april 2002">
26 </head>
27

```

Image 10: Archived web page with some metadata in the header.

Also applications and functionality that are no longer operational are described in the metadata. This can consist of a link with a given application, with another website or with obsolete plug-ins.

Two types of metadata information specific to websites deserve special attention: authenticity details and the date on which the website was put on-line.

a) Authenticity details

Information about the authenticity of archived web pages needs to be stored. The relatively low cost of web creation and the ease with which existing web pages can be copied cause that sometimes professionally looking websites appear on the WWW that can give the visitor the impression that they have reached the website of a certain company, organisation, etc. Cyber criminals use this method to try to recover the credit card numbers of unaware cyber tourists. A *unique identifier* is for most websites a necessity rather than a gadget.

The *Public Key Infrastructure* based technique of the digital signature is a common method to determine the origin of a website. The website owner digitally signs his or her website. A *server ID*

confirms the identity of the website owner and the ownership of the relevant domain name⁸². The visitor can verify the digital signature via the *server ID* and can then be sure that they find themselves on the genuine website of the desired organisation or person. A *server ID* is a digital certificate that the website holder can put on their website and that has been given to them by a certificate allocation service. Before attributing such a server certificate this service will require evidence that an organisation does not operate under a false name and that it is authorised to operate the domain name that houses the website to be authenticated⁸³. A digital signature permits authentication of a website and also of the downloads a website offers, such as software.

Several methods exist to determine whether a website is equipped with a server certificate. There is usually an icon present on the website that indicates that this website is secure (see Image 11)⁸⁴. Clicking it puts information about the certificate on the screen⁸⁵.



Image 11: A "Secure Site Seal" icon that indicates that Verisign has allocated a server certificate to this website.

It is also possible to check via the browser whether a specific website has a certificate allocated to it. In Microsoft Internet Explorer one needs to click the right button in the page of the website to be verified and then select "Properties". In the Properties window one selects then the "certificates button". All information about the certificate appears on screen, including the validity and the certificate allocation service. The window of Netscape Navigator contains in the lower left corner an icon that represents a lock. When the site is secured, the lock is closed. Clicking the icon displays a window with security information, among others about possible web certificates. And finally one can check the website of the certificate allocation service to find out whether a given website has received a certificate from them⁸⁶.

b) Date on which the website was put on-line

Also of great importance is the date on which the website was available on-line. A number of projects⁸⁷ give each website a time stamp when archiving them. This date is an important issue because it can help answering questions such as who has published a certain scientific discovery for the first time and violations of copyright laws. Also for patent applications a website archive equipped

⁸² The registration of domain names type dot-be was assured by the vzw DNS België until the end of 2000. Since 11 December 2000 DNS no longer registers the .be domain names. A network of agents was set up that have contracts with DNS to register domain names (for example Planet Internet). After an agent has gone through the registration procedure and the fees have been paid, DNS gives an exclusive licence to the holder of a domain name to use the requested domain name. Third parties can check via the website of DNS (<http://www.dns.be>) whether a domain name has been registered and to whom it belongs.

⁸³ See for example the *Serversign Certificate™ Procedure* that Globalsign follows to hand out a Serversign™ certificate http://www.globalsign.net/digital_certificate/serversign/index.cfm

⁸⁴ For example <http://www.ignite.com/application-services/products/verisign/news/news/ns011.html>

⁸⁵ For example <https://digitalid.trustwise.com/secureServer/cgi-bin/haydn.exe>

⁸⁶ <http://secure.globalsign.net/phoenixng/services.cfm?id=1413967734&reset=yes>

⁸⁷ For example The Nordic Web Archive

with the time stamp functionality can be decisive regarding the decision whether or not to grant the patent. The Inventions Patent Act of 28 March 1984 (B.S. 9 March 1985) stipulates that a patent can only be attributed to *new* inventions. An invention is considered to be new when it does not form part of the state of the art of technique. That state of the art is formed by all knowledge that was made public via a written or oral description, via its application or via any other means before the date of the patent request⁸⁸. A publication of a discovery on the Internet (even on a Chinese website in Chinese) before the date of the patent application has as a consequence that the discovery is no longer new and thus cannot be considered for patenting. The novelty requirement is based on worldwide spread knowledge related to inventions.

Now, what is this time stamp and how is it added to the website to be archived? After selection of the website the recordkeeping system calculates a hashing code for this website. Then the recordkeeping system adds the relevant date and sometimes even the time to the hashing code and signs it with its private key. The result is the time stamp. Time stamping is based on the idea that time-stamped electronic information (for example a website) had to exist at the latest on the moment the time stamp was created. This is logical as a time stamp is calculated based on the content of the time-stamped information.

F.2 Secure storage

The management of the archived websites involves some consideration about their secure storage. The check-in of a snapshot in a common document management or recordkeeping system is not obvious. Chances are big that the hyperlinks will no longer function. The identification of the files in most document management or recordkeeping system is not based on file names but on the IDs of the files. Due to later availability it is vital that a mirror or a snapshot remains functional. Alternatives can be the storage of the snapshots on the server in secured folders (firewall, limited access rights) or on a read-only medium. It is also possible to combine the check-in of the snapshot in the document management or recordkeeping system with a copy for consultation on the server.

F.3 Storage on media

Archived mirrors and snapshots can be stored on hard discs, magnetic tapes and optical discs. When choosing magnetic tapes and optical discs, it is important due to digital durability that as many standards as possible are applied concerning formats, type of carriers, file system, etc. Important issues when using CD-ROMs as media for archived websites are the file and folder names. The ISO-9660 standard for CD-ROMs contains a few limitations concerning allowed characters and length of file and folder names. Adapting those names implies that also all links need to be adapted. When using CD-ROMs as information carriers it seems appropriate that the off-line browser transforms all file and folder names into the 8+3 compatible system. Hard discs and optical carriers allow quicker access and consulting than magnetic tapes.

⁸⁸ Article 5 of the Inventions Patent Act

H. AVAILABILITY

It is best that the archived websites be made available in a digital form. For reasons of later research or future liability, websites are best viewed in the same way that they were primarily available on the WWW.

A possible way to ensure accessibility is the use of a portal site. Each archived website will be available via the portal site, including metadata, sitemap and log files. A website archive can be made available from certain carriers, from the hard disk of a computer or via a network. A web server permits the on-line availability of all archived websites without any problem. A link between the web server and the recordkeeping system seems to be the best method for on-line availability. Some precautions need to be taken in this case to ensure the user realises they are looking at archived versions. It must be remembered that also search robots will index the archived on-line websites and that visitors can have direct access to them, without passing the archive's website first.

There are more issues when building an archive for on-line sources and making it available than just consultation of archived websites. These include related (electronic) records and metadata which need to be accessible and readable.

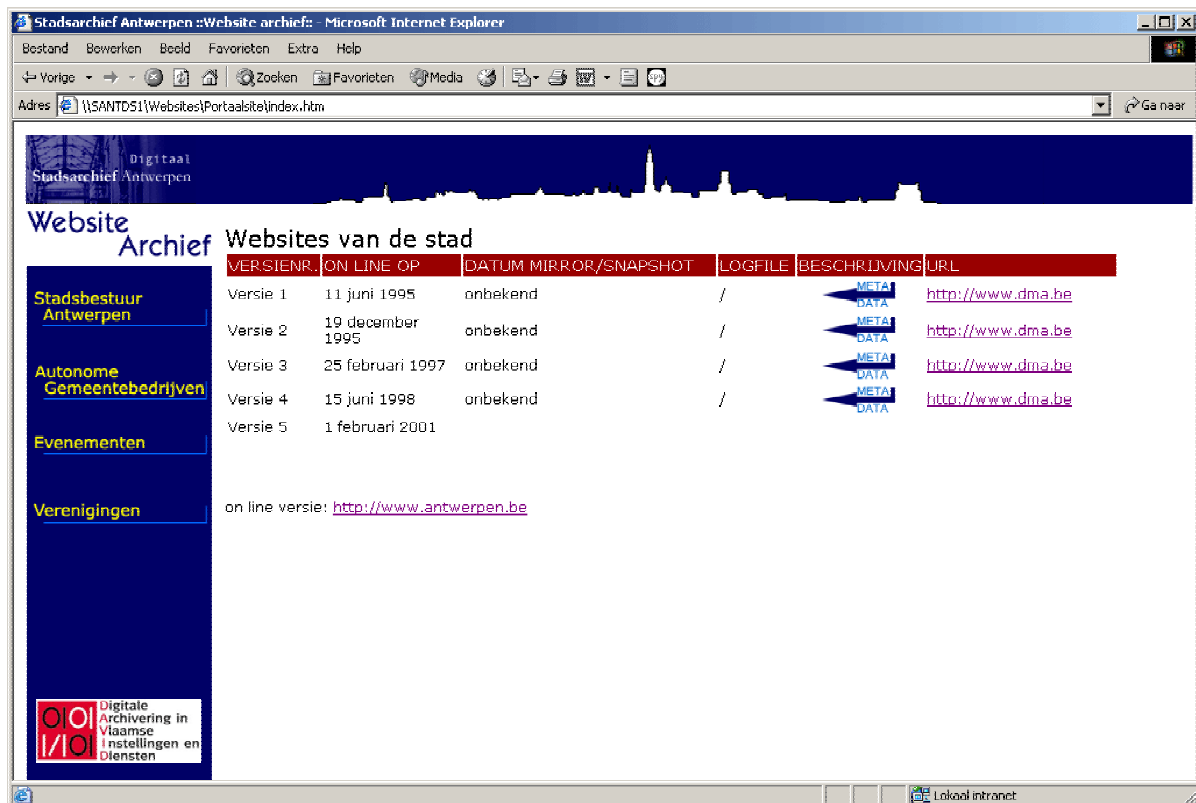


Image 12: A portal site for the website archive of the Antwerp City Archives. All collected websites and their metadata can be consulted via this portal site. The overview page of each website contains the different versions, the date on which those versions were put on-line, the snapshot date, a link to the log file of the snapshot operation, the URL with a link to the start page of each archived version, a link to the metadata of each version and finally a link to the on-line version. An on-line version of this portal site is available on the DAVID website.

VI. ELABORATING THE RECORDKEEPING SYSTEM

A. ROLE OF THE ARCHIVIST

The archivist plays an important role in the website archiving process. While most organisations and institutions have not paid much attention to website archiving, it will be the archivist's role to stress the importance of website archiving and related records, and to assure that the recordkeeping system meets the predefined quality demands.

The archivist takes care of:

- ➡ Applying the relevant legislation and rules concerning copyright and protection of personal data⁸⁹
- ➡ Determining the archival value of websites and related records (appraising). Determine as soon as possible what electronic records must be preserved.
- ➡ Defining and identifying the records in the complete information system: only these records will be put in the recordkeeping system
- ➡ Awareness of departments and IT responsables concerned of the archival value of certain websites and related records (among others prevention of unjustified deletions, registration of necessary data)
- ➡ Storing the necessary metadata
- ➡ Tuning the website archiving strategy towards the global recordkeeping system of the organisation
- ➡ Achieving the predetermined quality demands in the recordkeeping system
- ➡ Evaluation of the implementation of archiving modules in the current information systems, or when these are lacking: the start-up of local website archives
- ➡ Spreading the recordkeeping system within the organisation
- ➡ Performing a quality check of the mirrors/snapshots before they are incorporated in the digital repository
- ➡ Measures to ensure long-term consultability (accessibility and readability): migration, emulation, storage on safe and durable media, application of as many standards as possible, regular checks of the media and refreshing of the digital information to new standardised media.

The end result will be a recordkeeping system with clear procedures, routines and defining of responsibilities. Website archiving involves both the content responsables (usually the departments themselves) and the webmasters/administrators. Their exact identification depends on the

⁸⁹ See chapters VII and VIII in this report

organisational structure of the institution and the procedure for adapting and managing the website. In small enterprises both functions are usually combined by one person, whereas in larger companies both tasks will be more separated.

The archivist will be the stimulator of the recordkeeping system. When they develop website archiving within their organisation, they will usually be required to play an important role in the process. Concrete guidelines for the creation, management and archiving can serve as a starting point for the development of a recordkeeping system. Examples of those guidelines are given below. The guidelines concerning layout and management can be universally applied and are valid for all types of websites. These guidelines can be elaborated quite concretely. This is not the case for archiving guidelines however. These depend on the acquisition policy, the organisation, the type(s) of website and the available IT infrastructure. The archiving guideline in this report is limited to a listing of elements that should form part of it. Both guidelines have been more concretised in *Digitale Archivering: richtlijn & advies nr. 5 'Websitesbeheer voor archivering' [Website Management for Archiving]*⁹⁰.

B. EFFICIENT DESIGN AND MANAGEMENT

Archiving websites is easier when their design and management already takes archiving into account.

Applying a few simple rules concerning efficient website design and management results in the first place in a minimal need for adaptations at the moment of archiving and during later management. These rules deal with the usage of well-structured (X)HTML, respecting the (X)HTML syntax rules and the use of standardised (X)HTML tags and matching attributes. It is important for the website designer that these rules are applied because it avoids that part of the website may be inaccessible to possible visitors. Furthermore it is important to explicitly document the creation, modification and posting of a website when designing or managing it, so that this information can later be used as metadata. Clear guidelines and rules should prevent digital files with archival value from being deleted or overwritten.

The guidelines concerning the creation and management are partly based on the *Web Content Accessibility*⁹¹ rules of the *World Wide Web Consortium*. Assuring the exchangeability of web information has a lot of similarities with the measures for digital sustainability. However, these guidelines also contain the quality demands and the first experiences of the Antwerp City Archives concerning website archiving.

General

- ▶ Make a clear agreement on the management of the website:
 - Who is responsible for design and content?

⁹⁰ This document is available on the DAVID website.

⁹¹ Within the W3C the Web Accessibility Initiative work group deals with the accessibility problems of the web. WAI has composed among others the *Web Content Accessibility Guidelines 1.0* and a matching checklist (<http://www.w3.org/WAI/>). In Canada and the United States government administrations are obliged to apply these rules (Canada: *Government of Canada Internet Guide* http://www.cio-dpi.gc.ca/igi/index_e.asp; US: *Electronic and Information Technology Accessibility Standards Section 508*).

- ❑ Who changes the content of the website? Who makes the updates?
- ❑ How is it assured that the information is up-to-date?
- ❑ What happens with obsolete information?
- ❑ Who keeps what documentation about the design and the management of the website?
- ▶ Keep the possible archival value in mind when designing a website. Do not just take the website into account, but also consider related e-mails, log files, databases or document management systems. Predetermine the archival value of related documents and prevent that they are being deleted or overwritten.

Long term planning

When starting a website, think in the long run. On the one hand, determine a clear and controllable procedure for keeping the website up-to-date. On the other hand, make sure that old information remains available or is archived. Keep this in mind when developing a folder structure and setting up links

Folder structure

- ▶ Develop a clear folder structure for the files that compose a website
- ▶ Place all files and subfolders that compose the website in one common root folder
- ▶ Think ahead and plan the future development of the website. Remember this when developing the folder structure so that this does not need to be changed for every version. See to it that path indications are persistent (functionality of the links!).

Files

- ▶ Use as many standardised file formats as possible. An example of this could be:

Text: XML, (X)HTML, PDF, Word

Images: JPEG, GIF, PNG

Animations: GIF, Flash

Video: MPEG

Sound: MP3, WAV

Style sheets: CSS, XSL

- ▶ Give unique names to files and their versions, as much as possible.

External web pages/sources

Avoid or limit the use of web pages or images from other websites of hosted on external servers. It is not always possible to control their availability, their content and their functioning. If this is not possible, take some additional measures that these records are controlled in one way or another.

Links

- ▶ Internal links: use relative path indications
- ▶ External links: use absolute path indications
- ▶ Document the link by clearly indicating its target

Web pages

- ▶ Use mark-up and style sheets for what they were designed. Separate structure and layout. Use the (X)HTML tags for text mark-up and style sheets for text lay-out.
- ▶ Build well-organised web pages. Clearly structure them, and make them consist of three parts:
 - ❑ First line: specification of the applied (X)HTML version (and possibly specify the DTD): `<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN "http://www.w3.org/TR/html4/strict.dtd">` OR `<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "DTD/xhtml1-strict.dtd">`
 - ❑ header: `<HEAD></HEAD>`
 - ❑ body: `<BODY></BODY>`
- ▶ The root element is always: `<HTML></HTML>`
- ▶ Add a title to the header in each web page: `<TITLE></TITLE>`

Document each version: Metadata

- ▶ Describe the archived websites, their versions and their modifications. Keep the metadata of each version in a separate file or centralise all metadata.
- ▶ Produce a fixed metadata scheme. Make this scheme available in the shape of a form, an XML file or a database.
- ▶ Determine who stores the metadata

(X)HTML tags and attributes

- ▶ Use only standardised (X)HTML tags. Avoid the use of deprecated tags and attributes. The standards are: HTML 4.01 (ISO/IEC-15445:2000) and XHTML 1.0.
- ▶ Respect the order of the (X)HTML elements: correct arrangement of the (closing) tags
- ▶ Always close elements

When using:

- ▶ Non-textual information:
- ▶ Links:
- ▶ Style sheets:
- ▶ Image maps:
- ▶ Tables:

Ensure that:

- A textual alternative is available via the attributes "ALT" or "LONGDESC". Examples of non-textual elements are images, animations, image maps, applets and graphical buttons.
- The target of the link is clear
- The HTML files are also readable without the style sheets
- Mostly client image maps are used; avoid the use of server image maps
- The headers of rows and columns are clearly identified

- ▶ Frames: Each frame is identified via a meaningful name. Clarify the relation between the frames when this cannot be deduced from their names.
- ▶ Dynamic content: There is an alternative when the interaction is not functioning
- ▶ Scripts / applets: The website is still operational when the scripts/applets are switched off or are not being supported. If this is not possible, refer to a similar website with the same content.

C. ARCHIVING

The recordkeeping system depends on the organisation, the available technological infrastructure and the type of websites with record status. The most optimal situation for systematic website archiving is the integration of website creation management in a document of records management system. This results in an efficient storage of websites, records based on websites, adjustments and updates. Such a situation however is not very common within Flemish institutions. Their static or dynamic websites are not connected to the document management or recordkeeping system and most of the computer systems used do not offer the necessary functionality, so adaptations and updates escape their control and version management. Archiving websites and related records however is still in its early stages. It will be the archivist's challenge to link or incorporate archive modules to the existing information systems.

While a structured archiving procedure is being worked on, the organisation does not escape its archiving duty and therefore an intermediate solution has to be found. The only alternative will be the creation of an off-line website archive that stores website versions, their changes and related digital records⁹². This solution could be a final archiving option but is very time consuming so that websites with a higher update frequency require a more automated approach.

There are two options for the keeping of such an off-line website archive. One can delegate the responsibility for the archiving to the administrations or departments involved. This will be most appropriate for large organisations. Archiving websites should be part of the routine of adapting and updating the website. The other option is a central archiving of all websites. This option limits the responsibility of the administrations or departments to just notifying the changes that take place. The same procedure can be applied in both cases when archiving websites. The moment of archiving should be as close as possible to the moment of creation and not to the time a new version goes on-line.

A procedure or guideline for the archiving of websites should contain the following elements:

General

- ▶ Clearly agree on the archiving of the website; determine the different responsibilities: who will notify changes? Who will archive the website and the changes?
- ▶ Let the archiving take place immediately after the website update. Make sure that archiving is incorporated in the routine of updating the website

⁹² See also: *Managing Internet and Intranet Information for Long-term Access and Accountability. Implementation Guide*, 1999

- ▶ Predetermine the archival value of the website and the related records. Do not rely on back-up tapes to archive a website
- ▶ Only delete files on the web server when you are completely sure they have been archived

Where is the website archive kept?

- ▶ Provide a secured storage place for archived websites:
 - ❑ Are the websites centrally stored or not?
 - ❑ What folders are the archived websites and their versions kept in? On what medium? What folder names are used?
 - ❑ Limit access rights to these folders
 - ❑ Separate these folders from the work folders where the website is being designed
 - ❑ Make regular security copies of this folder

What to archive?

- ▶ The website:
 - ❑ Is the aim a full snapshot or mirror of each new website or version?
 - ❑ Websites with dynamic content: will only snapshots be archived or will the original server files (ASP, PHP or JSP files) also be stored?
 - ❑ What documentation about the website has any archival value?
- ▶ The server scripts:
 - ❑ Will the server scripts be archived? Each version?
 - ❑ Where are the server scripts kept?
 - ❑ How are the server scripts to be archived?
- ▶ The downloads:
 - ❑ Do the downloads have any archival value?
 - ❑ Are the downloads being archived at different places or together with the mirrors/snapshots?
- ▶ The databases:
 - ❑ Do the linked databases have any archival value? What databases or tables have to be archived?
 - ❑ How is the information in linked databases or document management systems being archived?
- ▶ The e-mails:
 - ❑ Are received e-mails to be archived?
 - ❑ How are received e-mails to be archived?
- ▶ The log files:
 - ❑ Do the log files have any archival value?
 - ❑ Are the analysis reports to be stored or the log files themselves?
 - ❑ Is any information from the log files stored in the metadata?
 - ❑ How are the log files to be archived?
 - ❑ How long are log files to be preserved?

How to archive?

- ▶ Does the website have static or dynamic content? Are mirrors or snapshots archived?
- ▶ What application will produce the snapshot of a website with dynamic content?
- ▶ What will be the quality control of mirrors or snapshots?

Versions / Frequency

- ▶ Is every change or version to be archived?
- ▶ Is the whole website to be archived after a change, or only that change?
- ▶ What is the archiving frequency of mirrors or snapshots?
- ▶ How and in what folders are the versions to be stored? How are the file names to be adapted?
- ▶ How will the different versions be documented in the metadata?

Metadata

- ▶ What additional metadata is to be stored about the archived website?

Repository

- ▶ When will the archived websites be ingested into the repository?
- ▶ How does the ingestion take place?

VII. COPYRIGHT: A BARRIER WHEN ARCHIVING WEBSITES

A. INTRODUCTION

A logical consequence of the digitisation of document traffic is that these documents will be stored in a digital form. In a previous section of this report (part IV) it has been extensively shown what technical problems appear when archiving a relatively new type of digital document: a website. Different archiving strategies have been presented, all with different implications: acquisition policy, periodicity, how to store a website for archiving, etc. The choices an organisation has to make to compose an archiving strategy will largely be dictated by the special needs of that organisation⁹³.

⁹³ See among others the section on acquisition policy

An organisation has to consider the legal copyright implications when drawing up an archiving strategy for their websites. Libraries have been sounding the alarm for a while now⁹⁴. The current copyright laws do not allow enough possibilities to make reproductions of publications without the author's permission. Other organisations as well as libraries that deal with the reproduction of information see copyright as a barrier. A lot of information is being shielded by copyright, making its reproduction subject to a number of clearly defined conditions. The copyright problem is at its most substantial with regard to websites. Immaterial as the Internet may appear, Internet transactions almost always cause some sort of copying of digital data. Surfing the Internet, FTP, 'uploading' and 'downloading' all imply an endless amount of copying⁹⁵.

Initially Cyberspace was considered to be a legal vacuum, especially by the Internet community. It was generally assumed that the existing governmental legal system and laws did not apply to the Internet. This idea became popular especially because the Internet does not have country frontiers. The new 'Internet space' was often compared to the open sea, where no state has any authority⁹⁶. This is very much incorrect. The inhabitants of Cyberspace are citizens of many countries, each with their own rights and duties, including those concerning copyright⁹⁷. Those components of the Internet with content, that is: websites and their content, do not escape from the application of copyright simply because they are present on and spread via the Internet.

B. COPYRIGHT IN BRIEF

Copyright in Belgium is controlled by two regulations. These are the Bern Convention of 9 September 1886 concerning the protection of literary works and works of art (approved by the Act of 26 June 1951, *B.S.* 13 October 1951) and the Copyright Act of 30 June 1994 concerning copyright and related rights (*B.S.* 27 July 1994, further on abbreviated as 'CA'). The Bern Convention has created a number of minimum standards of copyright protection, and Belgium has incorporated these in its Copyright Act.

B.1. Protected works

a) Literary works and works of art

The CA protects the rights of the author with regard to 'literary works and works of art'. The law does not define this term. However, the Bern Convention gives a non-exhaustive list with examples of

⁹⁴ 'Auteurswet vormt hinderpaal voor bibliotheekarchief', *NRC Handelsblad*, 3 March 1998; PRINS, J.E.J., 'Digitale duurzaamheid: een verloren geschiedenis?', 3, <http://infolab.kub.nl/till/data/topic/digiduur.html>

⁹⁵ SPOOR, J., 'The copyright approach to copying on Internet: (over)stretching the reproduction right?', in *The future of copyright in a digital environment*, HUGENHOLTZ, P. (ed.), Antwerp, Kluwer, 1996, 67 and 69

⁹⁶ BAINBRIDGE, D., *Introduction to Computer Law*, Gosport, Ashford Colour Press, 2000, 77; CLAPPAERT, W., 'Auteursrecht en Internet', in *Telecom en Internet. Recht in beweging*, DE POORTER, B. (ed.), Gent, Mys and Breesch, 1999, 359

⁹⁷ WESTERBRINK, B.N., *Juridische aspecten van het Internet*, Amsterdam, Otto Cramwinkel, 1996, 88-89

literary works and works of art⁹⁸. This list implies that the intended protection was covering a much broader area than what was understood by the term in everyday language.

The nature of a work is not relevant when determining whether a work is protected by copyright or not⁹⁹. The moment a work in its broadest definition fulfils some intrinsic characteristics, it will be protected by copyright. It suffices that the work is *original* and shaped in a *certain form*.

‘Originality’ does not imply that the idea behind the work needs to be new. A book being written on digital archiving can be copyright protected, even though many books have already been written about the topic. ‘Originality’ only indicates that the work must be the result of an intellectual effort of a person and needs to contain an individual, personal touch of the maker¹⁰⁰. A work that has been compiled by fully automated machine or technology driven procedures is therefore not protected by copyright. The extent of the intellectual effort is irrelevant, as is the quality of the result.

Furthermore, the work needs to be expressed in a form that can be observed by the senses. An idea or a concept as such can not be protected. An idea for a book (for example a scenario for an exciting book) will not be protected by copyright. No matter how unique they may be, ideas and concepts are not protected by copyright laws¹⁰¹.

b) And websites?

As indicated before, copyright is also valid on the Internet. The Internet is a worldwide network of computers (servers) that makes limited units of digital documents, so-called websites, available to the complete Internet community. A website is a unit of linked files, including text files (provided with hyperlinks or not), sound files, databases, scripts, etc. A website is a new art form that first appeared only a few years ago. The author of a website however is protected in exactly the same way as the author of a classical literary work¹⁰².

When applying the arguments about originality and observability mentioned above to a website, it must be concluded that a website, due to its nature, can fall under the regime of copyright:

⁹⁸ Article 2.1 Bern Convention: ‘*The expression “literary works and works of art” encompasses all products in the field of literature, science and art, whatever the means of expression may be, like books, brochures and other writings; lectures, speeches, sermons and other similar works, plays or dramatical-musical works, choreographic works and pantomimes, of which the means of performance has been put down in writing or by other means; musical compositions with or without words; cinematographic works and works that stem from a procedure similar to cinematography, art works of drawing, painting, constructing, sculpture, gravure and lithography; photographic works and works that stem from a procedure similar to photography; works of applied arts; illustration and geographical maps; drawings, scetches and plastical works related to geography, topography, construction or science.*’

⁹⁹ VOORHOOF, D., ‘Multimedia en auteursrecht. Afschermen en beschermen van informatie. Juridische problemen rond de beschikbaarheid en de reproductie van informatie op één drager’, in *Multimedia, Interactiviteit, Kennisspreiding*, minutes book, Symposium of the Book Society, 14 and 15 November 1995, LUC Diepenbeek, 100

¹⁰⁰ Cass., 25 October 1989, *Pas.*, 1990, I, 239; Cass., 2 March 1993, *Pas.*, 1993, 234

¹⁰¹ TRIAILLE, J.P., ‘La protection des idées. Les modes non contractuels de protection des idées en droit belge’, *J.T.*, 1994, (797), 799

¹⁰² CLAPPAERT, W., ‘Auteursrecht en Internet’, in *Telecom en Internet. Recht in beweging*, DE POORTER, B. (ed.), Gent, Mys and Breesch, 1999, 361; SPOOR, J., *o.c.*, 69

- Even though web pages are not physical documents, copyright still applies to them. To be copyright protected it suffices that a work is expressed in any form, even if that form is immaterial.
- The most important criteria when deciding on the originality of a website are its structure and composition. It is the way in which information is presented via a website that is copyright protected. Many websites share the same structure (for example welcome, presentation, contact, links, etc.). This has become common practice on the Internet and copying this structure is not a breach of copyright. Copying the *presentation* of this structure however can be a copyright violation. Websites are often constructed by the combination of ready-made ‘formats’ that their creators publish on the Internet (remark: sometimes they are illegally published on the Internet). Sometimes they are created with the aid of computer programmes: instead of programming the code personally, one clicks on the screen what the website should look like. In these cases no original work of the maker is created.

It is important to make a distinction between the website itself as copyright protected work and its components that can be individually copyright protected. Examples are photos, drawings, articles, logos, downloads, etc. Internet publications sometimes have multiple persons that hold the copyright: the photographer, the author of the article, the designer of the logo, the author of the report that can be downloaded via the website, the website designer (as far as the website is original), etc.

It is the archivist who needs to decide whether a website boasts enough originality when considering its archiving. A judge will have the final word on this.

c) Hyperlinks

The creator of a website commonly adds a collection of links to this website. These so-called *hyperlinks* are visual elements on a website that were initially used as a means to publish scientific texts. Clicking a hyperlink automatically loads the URL behind it into the browser window. This allows the user to quickly surf to related or useful sites. And thus the web of the Internet is woven. Putting a hyperlink to another website is not forbidden¹⁰³. It can be compared with the insertion of a footnote in a text, which indicates that the source that is referred to contains more information about the subject. Referring to places where copyright protected works are stored is not a breach of copyright. Care should be taken however when referring to websites where copyright protected works are illegally made available (for example illegally copied and offered software). A collection of hyperlinks on a website is copyright protected itself on the condition that the content is original enough, for example because of its structure or its completeness. The content of some portal sites really excels because of its originality.

¹⁰³ Some works consider the insertion of a link to copyright protected information as a means of secondary publication which requires the permission of the author: BRUNT, G., ‘Auteursrecht in DigiWorld’, in *Intellectueel eigendom in digitaal perspectief*, GIJRAATH, S. et.al. (ed.), Alphen aan den Rijn: Samson, 1996, 35-36, quoted by HUGENHOLTZ, P. , *Recht en Internet*, Handelingen Nederlandse Juristen-Vereniging, Deventer, Tjeenk Willink, 1998-I, 212, footnote 36.

d) Exceptions for official government acts

Article 8 §2 CA determines that official government acts are not covered by copyright. Law texts, reports of the work of parliamentary commissions, rulings and decrees etc. can therefore be used freely. All other government publications that have been established by government employees but cannot be classified as ‘official acts’ do fall under the protection of copyright. Examples are a government brochure announcing the abandoning of radio and television tax or reports written by civil servants. Considering the remark made above, also a government website, if sufficiently original, will be protected by copyright. The website itself may not be an official publication, it is a copyright protected work thanks to its unique and user inspired structure¹⁰⁴.

B.2. Who holds the copyrights?

It is important to determine who is considered to be the author of a copyright protected work. This person is the holder of the copyrights. The CA poses clearly that this has to be the person who effectively creates or has created the work. There is a refutable suspicion that the person whose name is mentioned on the work is the author¹⁰⁵.

Organisations and governments that want to be present on the web via their own website often leave its ‘construction’ to a specialised company. Whoever commands the development of a website needs to realise that it is the developer who receives all copyrights on the website. This even holds true when the developer is paid for their services¹⁰⁶. The client risks that the design used for their website may be sold on to others. To prevent this a contract can be made with the developer who then hands over his copyrights. This needs to be done according to all legislation¹⁰⁷.

The same principle applies when an employee or a civil servant establishes a website for his or her organisation while performing the tasks that result from his or her employment contract. If the employer or the government wants to obtain the copyrights from the website creator, this needs to be determined in writing. This can be arranged in advance via the employment contract or the statute.

Article 6 paragraph 1 CA clearly poses that only a human being, a natural person, can be an author. This is logical as only a human can deliver the intellectual effort that the application of copyright requires. A legal person (organisation, government institution, etc.) can as such never be the author of its own website. Therefore the natural person who represents the organisation or government institution holds the copyrights.

It is important for the archivist to know who holds the copyrights of a website and its content (articles, drawings, etc.). This is the person who has to give permission for acts of reproduction, changes and publication. A website that shows signs of originality, and therefore is protected by

¹⁰⁴ Most texts on a government website will not be categorised as an official act, even though they are legislation or jurisprudence. Disclaimers on government websites often indicate that the published information is not necessarily an exact copy of the officially approved text or that only the legislation published in the paper edition of the Belgian Official Journal is valid. See http://portal.vlaanderen.be/http://docs.portal.vlaanderen.be/channels/hoofdmenu/afw_van_aansprakelijkheid.doc

¹⁰⁵ Article 6 paragraph 2 CA

¹⁰⁶ VANHEES, H., *Auteursrecht in een notendop*, Leuven, Garant, 1998, 24

¹⁰⁷ The transfer needs to take place in writing among others for evidence purposes. See Article 3 §3 CA

copyright, usually carries the name of the web design responsible. This does not automatically imply however that this person is holder of the copyrights¹⁰⁸. These copyrights¹⁰⁹ could have been transferred to a natural person/employer-director-civil servant of the organisation or government that holds the domain name. The archivist should first contact the domain name holder to check which natural person holds the copyrights. On the DNS website it is possible to retrieve who has registered the domain of for example “antwerpen.be”. Contact details are available too so that the archivist can contact that person to identify the copyrights holder of the website.

Archive departments of public administrations often do not limit themselves to the archiving of their own organisation’s websites. The Antwerp City Archives, for example, also store the websites of other organisations that may be sources for historical research. It does not matter for the application of copyright whether the archivist wishes to archive a website of the proper organisation or another. Even for the websites of, in this case, the City of Antwerp, the archive –in theory– needs to ask the permission of the copyrights holders before being allowed to start the archiving.

B.3. What do copyrights imply for their holders?

An author possesses two types of rights: rights of property and moral rights. Rights of property enable an author to exploit their work and to gain an income from it. Moral rights protect the person of the author and the “intimate link” between them and their work¹¹⁰. Whether an author has signed contracts about the commercial exploitation of their work does not change this moral protection of their work.

a) Rights of property

The main right of property is the *reproduction right*¹¹¹. This implies that the author has the absolute and exclusive right to reproduce their work, or to have someone else reproduce it. Reproduction is defined as ‘multiplication’, or ‘the establishment of material copies’. Material does not imply that the reproduction needs to be tangible, but it does imply that it needs to be fixed in one way or another. The mere visualisation of a work on a computer screen cannot be qualified as an act of reproduction. Storing it in a computer memory *is* a type of reproduction¹¹². The used techniques are irrelevant. The reproduction can be graphic, mechanic, optical, electronic, etc.

Permission of the property rights holder is required for each reproduction of a copyright protected work, apart for the legal exceptions¹¹³. This is as a principle the maker of the work, unless property

¹⁰⁸ See further: how to receive copyright protection. The name of the web designer is usually mentioned for commercial reasons.

¹⁰⁹ This is: the property rights (see further)

¹¹⁰ CORBET, J., ‘Auteursrecht’, in *Algemene Praktische Rechtsverzameling*, Antwerp, Kluwer, 1997, 54, nr. 143

¹¹¹ Article 1 §1 paragraph 1 CA

¹¹² CORBET, J., l.c., nr. 115

¹¹³ The author’s exclusive rights of property do not have absolute value. There are some examples such as for the reproduction of sound and audio-visual works for use in the family circle (home taping, Article §1 5° CA) and for the copying for private use or for didactic use (the so-called reprography arrangement, Article 22 §1 4° CA). Archive institutions cannot apply these legally forced licences to reproduce websites (or copyright protected works in general) with the intention to archive them: *Auteursrechten, privacy en royalties in musea*,

rights have been transferred to another person. The goal of the reproduction is irrelevant. The author's permission is required even when the reproduction is undertaken to archive the work.

The reproduction right also contains, according to Article 1 CA, the 'exclusive right to grant permission to the modification or translation of the work'. The modification and translation right demonstrates that also major changes during reproduction remain under the supervision of the author, as long as the form of the original work remains recognisable¹¹⁴.

Another right of property that is important when archiving websites is the author's right to decide according to what procedure they will present their work to the public¹¹⁵. Therefore an author can decide freely whether to execute his or her work himself or to have it executed. The notion 'execute' or 'present' is defined here as any reproduction in a non-durable and short form, like the singing of a song or the screening of a movie¹¹⁶. A public presentation occurs when there is no intimate link between the people present at such an execution.

Does this exclusive publication right also apply to a website designer (or whomever the rights of property were transferred to)? The broad wording of the publication right allows incorporation of the new technological publication techniques into the copyright.

Publication of a work can take place in a personal way by performing it 'live' by performing artists for a live audience. Usually the term 'execution' is used. The legal term is 'publication' however, to indicate that it covers more than just executions, that it can also contain impersonal publications without a live audience¹¹⁷. Viewing a website takes place via the electronic transmission of digital information (HTML pages) that reaches the end user on his or her individual request.

Consulting a website does not require the 'audience' to be present at the same time at the same place. The CA however allows individual receivers of a message to form an audience even when they do not conceive the message simultaneously¹¹⁸. The WIPO Copyright Treaty of 1996 stipulates this literally in its Article 8: '*Authors of literary and artistic works shall enjoy the exclusive right of authorising any communication to the public of their works, by wire or wireless means, including the making available to the public of their works in such a way that members of the public may access these works from a place and at a time individually chosen by them*'.¹¹⁹

It is another question whether a website will ever actually be viewed. Legislation has determined that it is the *possible* access by multiple persons that is important¹²⁰.

Remark: the designer's permission is required for the publication of a website via the Internet¹²¹ (for example an organisation that has its website designed by a specialised service) but also for acts of

archieven en documentatiecentra, Minuts of the symposium held in Brussels on 23.11.97 and on 15.12.97, Vlaamse museumvereniging, Werkgroep Auteursrechten, CUYPERS, J. (ed.), Antwerp, 1999, 53 and 60.

¹¹⁴ GOTZEN, F., *Auteursrecht, tekeningen en modellen*, K.U.Leuven Law Faculty, 1998, 67

¹¹⁵ Article 1 §1 last section CA: '*Alleen de auteur van een werk van letterkunde of kunst heeft het recht om het werk volgens ongeacht welk procédé aan het publiek mede te delen.*' [Only the author of a literary work or a work of art is allowed to present the work according to any procedure.]

¹¹⁶ VANHEES, H., *o.c.*, nr.63

¹¹⁷ Verslag De Clerck, *Gedr. St.*, Chamber, B.Z., 1991-92, nr. 473/33, 64

¹¹⁸ GOTZEN, F., *o.c.*, 82

¹¹⁹ <http://clea.wipo.int/PDFFILES/English/WO/WO033EN.PDF>

¹²⁰ Court Brussels, 16 October 1996, *Auteurs en Media*, 1996, 426

reproduction. That is: when the end user (for example the archivist) starts creating a copy to be stored in the digital archive system when downloading the HTML pages.

→ Offering information via the Internet is a completely new means of communication that falls under the right of communication thanks to the broad wording of the CA¹²².

Next to reproduction right and public communication right, the author has a few other rights of property at their disposal¹²³. However these will not be discussed here as they are irrelevant for the discussion about archiving websites (and digital documents in general).

b) Moral rights

Contrary to rights of property, it is not possible to make money out of the author's moral rights or to abandon them. They are called *moral* rights because they protect the link between the author and his or her work. A copyright protected work expresses the personality of its author.

The CA has specified four moral rights for the author: the right to (first) publish the work, the right to claim fatherhood of the work, the right to be respected for the work (Article 1 §2 CA) and the right to access the work (Article 3 §1 third section, *in fine* CA).

The -for digital archiving purposes very important- right to be respected for the work, or the 'right to the integrity of the work' as it is often called, allows the author to protest against any change to the work without having to demonstrate any damage this may cause them¹²⁴.

With regard to the archiving issue the question arises to what extent the archivist can modify copyright protected websites (and the source code behind them) to make them static or resistant against technological ageing (see further).

B.4. How does one receive copyright protection?

The accomplishment of a work that fulfils the requirements of copyright protection automatically implies that that work receives copyright protection. In Belgium no formalities need to be fulfilled to receive copyright protection. Attributing copyright reservation signs (for example a copyright notice

¹²¹ The primary publication on an order-made website via the Internet will usually not cause any problems as a website is meant for publication via this medium. But to ensure this, it is beneficial to transfer the publication right to the client in writing.

¹²² That does not fall under the special conditions the CA has introduced for three forms of communication: exhibition (Article 9 CA), satellite broadcasting (Article 48 e.v. CA) and cable distribution (Article 51 e.v. CA).

¹²³ That is: the right of intention, right of renting and borrowing, and right of exhibition.

¹²⁴ Belgian copyright legislation is stricter here than the Bern Convention, that only forbids those changes that may damage the author's reputation or integrity (Article 6bis text Paris 1971).

‘©’ followed by the name of the beneficiary) does not influence the applicability of copyright¹²⁵. The lack of copyright information on a website does not imply that the website would not be copyright protected or that the archivist can make a copy without prior consent.

C. ARCHIVING WEBSITES: PROBLEM DEFINITION

One of the arguments at the basis of copyright is the idea that copyright protects the cultural heritage of a country. By attributing exclusive rights to authors of original works, these authors are being inspired to establish other works in the future. They will thus contribute to the common cultural heritage¹²⁶.

Digital creations are just as well protected by the existing copyright laws as are creations on paper, books, pamphlets, etc. There is a big paradox within the current copyright legislation concerning the archiving of websites and of digital copyright protected works in general. Acts of reproduction are inherent to the archiving of digital documents, but exactly these acts are restricted by copyright to the author. Without the author’s permission (the author of the website and/or its content) it is theoretically impossible to digitally archive websites. However not only these acts of reproduction without the author’s permission conflict with copyright legislation, as will be discussed further on.

Also the preservation of paper heritage requires acts of reproduction (copying on microfilm) when the paper is about to decay¹²⁷. Museums and libraries cannot just transfer the electronic publications they possess to other information carriers when they appear to be subject to technological ageing. Archives however assume a special position regarding the copyright problems of preservation. The problem appears much more frequently and much earlier for the archiving of digital documents and especially websites. This is so because the archivist needs to make a copy of the work before possessing it. The copyright problem does not limit itself to digital sustainability.

In the doctrine a number of attempts have been undertaken to defend a ‘right to information’ with regard to institutions or persons that query information and want to receive access to that information without the author’s permission (like archives or an art collection)¹²⁸. On the European level there are some regulations that seem to tend towards this right to information. There is the example of art 3bis

¹²⁵ Belgium has an arrangement concerning ‘legal depot’ but this is not connected to copyright. The Legal Depot Act obliges the deposition in the Belgian Royal Library of any publication published in Belgium and of any publication published abroad of which the author or one of the authors is Belgian and has an official residence in Belgium (Act of 8 April 1965, B.S. 18 June 1965).

¹²⁶ HUGENHOLTZ, P., *DIPPER Digital Intellectual Property Practice Economic Report – Copyright aspects of caching*, Institute for Information Law, University of Amsterdam, September 1999, 7

¹²⁷ Copyright survives the author for seventy years, to the benefit of their inheritants. Also with regard to paper documents, copyright will remain an obstacle for the archivist, as they remain valid for at least seventy years after the creation of the work.

¹²⁸ HUGENHOLTZ, P., *Recht en Internet*, Nederlands Juristenblad, Zwolle, Tjeenk Willink, 1998, 201; DOMMERING, P., ‘Internet: een juridische plaatsbepaling van een nieuw communicatieproces’, in *Volatilisering in de economie*, DE JONG, W., Wetenschappelijk Raad voor het Regeringsbeleid, Den Haag, SDU Uitgevers, 1997, 150, quoted by BUYDENS, M., *Auteursrechten en Internet – problemen en oplossingen voor het creëren van een online databank met beelden en/of tekst*, Brussels, Federale diensten voor wetenschappelijk, technische en culturele aangelegenheden [Federal department for scientific, technical and cultural matters], 1998, 18

section 1 of the Directive concerning the execution of television and broadcast activities¹²⁹, that allows the Member States to determine that broadcasts of certain events of considerable importance to society (meaning: sports events) should be freely available to the public. The permission to introduce such a right to information seems, because of the title of the Directive, to apply only to television, not including the Internet¹³⁰. Institutions of general importance like public archive institutions are therefore being confronted with a copyright that situates the rights mainly with the author and that does not enough realise that problems exist for the intermediate persons who save the information and make it available.

We will investigate below what problems copyright is creating with regard to the archiving of websites.

C.1. Acts of reproduction

One can wonder whether downloading a website from a web server by an archive institute is an act of reproduction that requires the author's permission.

When the archivist requests the web page they wish to archive, they load its content into the work memory of their computer. Loading web pages is called 'browsing'. There is no doubt that this browsing implies some acts of reproduction¹³¹. Jurists agree that the storage of a protected work on a digital medium is to be qualified as a reproduction in the copyright sense¹³². However, the term 'reproduction' that appears in copyright should not be considered to be a technical concept. It is a legal notion in the first place¹³³. This legal notion does not go as far as to consider every technical reproduction to be a legal reproduction. The *Legal Advisory Board* of the European Commission poses in its reaction to the *Green Paper on Copyright in the Information Society*:

*“The notions of ‘reproduction’ (...) are only fully understood if they are interpreted not as technical, but as normative (man-made) notions, i.e. they are not in a simple sense descriptive but purpose-oriented and used to define and delimit existing proprietary rights in a sensible and acceptable way. Thus, if the use of a protected work transmitted over a computer network causes (parts of the work) to be intermediately stored, this technical fact does not, in itself, justify the conclusion that an exclusive reproduction right is potentially infringed.”*¹³⁴

¹²⁹ Directive 97/36/EG of the European Parliament and the Council of 30 June 1997 changing Directive 89/552/EEG of the Council concerning the co-ordination of certain legal and administrative rules in the Member States regarding the execution of television and broadcast activities, Pb. EG nr. L 202, 3.7.1997, 60

¹³⁰ BUYDENS, M., *Auteursrechten en Internet – problemen en oplossingen voor het creëren van een online databank met beelden en/of tekst*, Brussels, Federale diensten voor wetenschappelijk, technische en culturele aangelegenheden [Federal department for scientific, technical and cultural matters], 1998, 18

¹³¹ HUGENHOLTZ, P., *Recht en Internet*, 211

¹³² This is implicitly confirmed in the Bern Convention: Article 9 (Text Paris 1971): ‘*Authors of literary and artistic works protected by this Convention shall have the exclusive right of authorizing the reproduction of these works in any manner or form.*’

¹³³ HUGENHOLTZ, P., *DIPPER Digital Intellectual Property Practice Economic Report – Copyright aspects of caching*, 14

¹³⁴ <http://europa.eu.int/ISPO/legal/en/ipr/reply/reply.html>

This is the so-called normative interpretation of the notion ‘reproduction’. The mere ‘consumption’ of information has traditionally been kept outside copyright¹³⁵. Browsing the Internet can be compared to reading through an encyclopaedia. The person who is reading does not perform acts of reproduction but only acts of usage. There is no reproduction at stake that requires the author’s permission when making a copy of the website is necessary to communicate it to others¹³⁶. This is confirmed in the new European Directive concerning copyright¹³⁷. Article 5 section 1 of the Directive foresees an exception to the reproduction right, that Member States have to introduce into their national legislation, for temporary acts of reproduction of a passing or incidental nature, that are an integral and essential part of a technical procedure and that are being applied only to allow its transfer in a network between third parties via an intermediate person. Consideration 23 of the Directive suggests that browsing is included in Article 5 section 1¹³⁸.

This exception does not apply to archiving. Both static and dynamic websites need to be taken off the web server by the archive institution via copies. These copies do not have a temporal nature. On the contrary, they are intended to be incorporated into the digital archive system for an infinite period of time. The copies have not only been made to view the website locally on a screen. *Caching* without the author’s permission could be justifiable. This is just a tool that allows a complete or partial, but always temporary storage of frequently consulted websites on the end user’s server with as only goal to speed up the passage of information via the Internet. The end user’s computer no longer needs to link to the original server to request those websites. Archiving or another method of permanent storage of websites however is not allowed without the author’s permission¹³⁹.

As far as downloading a website for archiving is concerned, there is no distinction between static and dynamic websites. Both need to be copied by the archivist to become in their possession. The exclusive reproduction right applies each time when a protected work that is stored on a digital medium is being put on a web server or removed from it¹⁴⁰. Acts of reproduction for archiving purposes therefore require the permission of the rights holder.

There are some exceptions to the exclusive reproduction right in the Copyright Act¹⁴¹. Article 22 §1 4° determines among others that the author cannot resist a partial or complete reproduction of articles or works of plastic arts, or of short fragments of works that have been stored on a graphical or similar carrier, when this reproduction is only purposed for personal or didactical usage and does not interfere with the publication of the original work (the so-called reproright). This could be an escape route for archive institutions not having to ask the rights holder’s permission each time they wish to archive a

¹³⁵ HUGENHOLTZ, P., ‘Convergence and Divergence in Intellectual Property Law: The Case of the Software Directive’, in *Information Law towards the 21st Century*, DOMMERING, E. (ed.), Deventer, Kluwer, 1992, 323

¹³⁶ KOHLER, J., *Das Autorrecht*, Jena 1880, 230, quoted by SPOOR, J., *Scripta Manent: de reproductie in het auteursrecht*, Groningen, Tjeenk Willink, 1976, 137-138

¹³⁷ More about this later. Directive 2001/29/EG of the European Parliament and the Council of 22 May 2001 concerning the harmonisation of certain aspects of copyright and neighbouring rights in the information society, <http://www.ivir.nl/wetten/eu/2001-29-EG.pdf>

¹³⁸ HUGENHOLTZ, P., *Recht en Internet*, 211, footnote 33, BUYDENS, M., ‘La nouvelle directive du 22 mai 2001 sur l’harmonisation de certains aspects du droit d’auteur et des droits voisins dans la société de l’information : le régime des exceptions’, *Auteurs en Media*, 2001/4, 434

¹³⁹ HUGENHOLTZ, P., *DIPPER Digital Intellectual Property Practice Economic Report – Copyright aspects of caching*, 46

¹⁴⁰ HUGENHOLTZ, P., *DIPPER Digital Intellectual Property Practice Economic Report – Copyright aspects of caching*, 9

¹⁴¹ See also footnote 113

copyright protected work. Reproductions here are only made for *personal usage*, being to archive them.

However this is not the type of usage the legislator intended with this notion. In the Minutes of the Senate the following has been said about the notion “personal usage”: “*Personal usage can also be applied regarding reprography: copying for personal usage by a natural person (even with professional goals) or for internal usage by a legal person of short fragments of works is considered to be legal*”¹⁴². This definition as such is very broad, especially when it is compared to the interpretation of French copyright laws. Only strictly personal usage by a natural person is allowed there, and each collective usage, for example in the framework of a legal person, is excluded. Yet it will not be possible for archive institutions to call on reprograph to make reproductions in order to archive them. Personal usage is not possible when a third person will use the reproduction, even when that usage is personal¹⁴³. This is the case for archiving. The reproduction is made for third parties to consult the archive record in the future, and not for internal consultation by the staff of the archive institution. The reprograph would be valid for the reproduction of a copyright protected work the archive institution has composed to enhance the internal working of the archive, for example a copy of the manual “digital archiving”.

The same goes for the notion “didactical usage”. Exclusive didactical usage consists of a reproduction that is used as an illustration for education or scientific research. When a person makes a reproduction for their personal study or research needs, this is personal usage of the copy. This is the situation libraries find themselves in when requested by third parties to make copies for study or research usage. The finality of the copy is all determining here. A copy made by a lawyer for professional usage is not didactical usage¹⁴⁴. A copy made by an archive department that may or may not be used for didactical (or other) purposes in the future does not qualify as a copy for didactical usage either.

Furthermore, reprograph in Belgium is much more restrictive than in most other European countries. Belgian reprograph foresees two possibilities concerning digital archiving: 1) reproduction for personal usage is limited to short fragments, while most archives will only be interested in full works; 2) the sales of original works could be damaged as non-digital archives require the possession of multiple originals while digital archives only need one original¹⁴⁵. The acts of reproduction that archive institutions undertake to archive digital works therefore do not qualify as making copies for personal usage.

In conclusion, the currently existing exceptions do not allow reproductions being made for archiving purposes without the author’s permission¹⁴⁶.

¹⁴² Report to the King for the Royal Decree concerning the reward for authors and publishers for the private or didactical usage of works stored in a graphical or similar way, *B.S.* 7 November 1997, 29 877

¹⁴³ DUBUISSON, F., ‘L’exception de reproduction d’œuvres fixées sur un support graphique ou analogue dans un but privé ou didactique’, *Journal des Tribunaux*, 1997, 657

¹⁴⁴ It is not clear how the librarian who produces the requested copy can check whether it will really be used for didactical purposes. The librarian will have to judge using their own responsibility.

¹⁴⁵ DECKER, U., ‘Electronic Archives and the Press: Copyright Problems of Mass media in the Digital Age’, *E.I.P.R. issue 7*, Sweet & Maxwell Limited, 1998, (256), 257

¹⁴⁶ LANGE, J., ‘De wet en elektronische publikaties’,

<http://nieuws.surfnet.nl/nieuws/snn-archief/achtergrond/jg97-98/kb-symposium2.html>

PRINS, J.E.J., ‘Digitale duurzaamheid: een verloren geschiedenis’, 2-3,
<http://infofab.kub.nl/till/data/topic/digiduur.html>

C.2. Changes

As described in the technical part of this report, it is not possible to archive the original files of dynamic websites in their original file structure. Contrary to that of static websites, the source code of dynamic websites needs to be altered in advance. Static HTML files need to be archived instead of the original ASP or PHP files. Only then can they afterwards be reconstructed in a platform independent manner¹⁴⁷. The transposition takes place via specially designed computer programmes called off-line browsers. Pages that were already composed in HTML do not need modifications.

Also these changes require the author's permission conform the exclusive right to the integrity of the work, as far as the website can be considered as an original creation. Copyright subjects even the smallest change to the work to the prior consent of the author. It needs to be remembered that the right to integrity cannot, as a moral right, be transferred to the employer or the customer¹⁴⁸. The archivist will need the actual author's permission for these modifications. One could pose that the result of transposing a complete website does not result in a modification but an adaptation by a computer programme. An adaptation takes place when a new original website is established based on the original one. This adaptation can even become a copyright creation itself, if the original form of the website remains clear and that the new source code is original enough. It is obvious that an archivist wants to keep the original format as much as possible. The newly made website will be a so-called 'computer-generated work'.

A computer-generated work is the result of the interaction between humans and computers. When the work of the computer becomes more important and human creativity becomes less, computer-generated works will cause troubles in the field of copyright. Copyright, once again, is only attributed to works that are assumed to be the expression of human creativity¹⁴⁹. The Anglo-American legal tradition experiences much less troubles with the phenomenon of computer-generated work than continental law. Despite the convergence of both traditions over the last few decades thanks to the digital evolution and the entrance of the United States to the Bern Convention, the Anglo-American legal tradition focuses on the work itself and examines its uniqueness. The United Kingdom is the only European country that has incorporated a specific regulation in its copyright laws for works that are computer-generated and where no human author can be attributed. The rights holder is then the (legal) person who has invested in the computer hardware or in the computer software that has generated the work. Usually however a minimal influence of the computer programmer will be distinguishable, thus protecting transformed websites by copyright anyway.

C.3. Digital sustainability

¹⁴⁷ See above

¹⁴⁸ See above

¹⁴⁹ DE COCK BUNING, M., *Auteursrecht en Informatietechnologie – Over de beperkte houdbaarheid van technologiespecifieke regelgeving*, Amsterdam, Otto Cramwinkel, 1998, 175

We will assume that the website has been archived in a static form¹⁵⁰. Emulation of the off-line browsers has been determined before to be the best solution. This is also the best choice from copyright perspective. The use of emulators does not require the HTML files to undergo changes, not even in the long run.

C.4. Availability

The exclusive communication right encompasses, as stated earlier, every type of publication. There are two ways to make archived websites available to the public. One possibility is viewing the websites in the reading room from a hard disk or an external carrier. Another option is making the archive available through a portal site (for example a sub page of the proper site) from where every archived website can be consulted. Fearing abuse, authors are less likely to authorise publication of electronic documents via the Internet. That is why the Dutch Electronic Library only allows consulting of electronic publications inside their building.

It could be argued that this problem does not pose itself for the archiving of websites. It is so that the permission to publish the website, including its content, on the Internet has already been asked from the author(s) of the website and its content (or this person may have decided to publish his or her site on the Internet). Still the archivist needs to ask permission again for this. It is not because an author has given permission to publish his or her work via a certain medium that anyone can decide autonomously to spread that work via another medium, just because the medium the author has chosen (the Internet) can potentially reach every person on earth¹⁵¹. The archivist can add this permission to the agreement that determines the reproduction right of the site.

The CA foresees an exception to the reproduction right for the consultation of works for scientific purposes. Article 22 §1 4^oter CA specifies that the reproduction of works can be allowed without the author's permission if that reproduction takes place for scientific research, is justified by a non-profit goal and does not interfere with the normal exploitation of the work. The legislator here envisages the situation when a researcher requests a copy of an archived work for the scientific research he is performing. This copy could also be a disk containing a copy of the work. When the copy will serve scientific research, this forced licence will allow the archive institution to make the reproductions.

C.5. Special regimes

Two types of creations have their own copyright arrangement, handled by a 'special law'. These two types are computer programmes and databases. Both can sometimes compose part of a website.

- a) Computer programmes

¹⁵⁰ See above

¹⁵¹ An exception has been foreseen for databases: see below

The technical part of this work has shown that the most recent generation of websites can best be compared to a computer programme. These types of websites contain scripts, next to the data that appears on the screen when loading the website and that is usually written in HTML. Scripts are micro programmes designed and stored on the site and sent to the computer (be it the client, be it the server computer) to be executed there. The result of this execution depends on the client's request, for example calling a text on the screen.

The question now is whether these websites can be classified as computer programmes in the sense of the software legislation. This question is of practical importance to the archivist, as this software legislation assumes that copyrights on computer programmes are attributed to the employer (this could also be an administration) if the site has been designed by an employee. It is not the author who holds the copyrights, as is the case with other copyright protected creations according to common copyright laws¹⁵². The archivist will have to direct himself to the employer to ask permission to archive. This issue is less relevant for public archive institutions as the employee statute will identify the employer as copyright holder, also for other creations than computer programmes.

Every European Member State had a different protection for computer programmes in the early eighties. Only few countries had incorporated the protection of computer programmes explicitly into their own copyright laws¹⁵³. The other countries did not have any specific protection for computer programmes. Worried by the differences in (future) computer programme legislation in the Member States, that would have a negative influence on the working of the common market for computer software, the European Commission published in 1988 a 'Green book' that announced a Directive concerning the copyright protection of computer programmes. This Software Directive, that became reality in 1991, has been transformed into Belgian legislation by the Act of 30 June 1994 concerning the transposition of computer programmes (the Software Act, *B.S.* 27 July 1994).

Neither the Software Directive nor the Software Act defines the concept "computer programme". Perhaps these terms have purposely remained vague, because a technology specific definition would age very quickly¹⁵⁴. Some other legal sources do have a definition:

WIPO: 'a collection of instructions purposed to have an information processing machine perform a certain task.'

American Copyright Act 1976: 'a computer program is a set of statements or instructions to be used directly or indirectly in a computer in order to bring about a certain result.'

Whether scripts that form part of a website are computer programmes in the sense of the Software Act has been made clear by doctrine. The website itself, however, cannot be classified as a computer programme, even though its source code may mainly exist of scripts and even though the website appears to be a computer programme that generates websites "on the fly". A legal argument can be found in the Copyright Directive of 22 May 2001. Consideration 23 of the Directive suggests implicitly that websites fall under the temporal reproduction right that is being arranged by Article 2 of the Directive. Therefore, websites fall under the general regime of the Copyright Directive and not

¹⁵² Article 3 Software Act

¹⁵³ Germany, France and Great Britain (1985) and Spain (1987)

¹⁵⁴ COM (88) 816 – SYN 183, *J.O.C.E.*, 12 April 1989, n° C 91/9

under the more specific Software Directives¹⁵⁵. The scripts themselves obviously are subject to the software regime.

b) Databases

Many dynamic websites are linked to a file system. This file system often contains a database that the surfer can query. For the government this can consist of forms, official acts and in the future even personal data from the population register after the implementation of the electronic identity card in the framework of the e-portal¹⁵⁶. The database can make a number of company documents available via the intranet of a company.

Databases are subject to a special copyright arrangement. This arrangement stems from a number of lawsuits in the nineties. In the case *Van Dale/Domme* the Dutch High Council decided that a database is only copyright protected when it is the result of a selection that expresses the personal vision of its creator¹⁵⁷. Along the same lines of traditional copyright laws, the *Supreme Court* of the United States ruled in the *Feist case* that a collection of mere factual data (in this case a telephone guide) cannot be considered for copyright protection due to a lack of creativity¹⁵⁸. Both judgements demonstrated that classical copyright did not provide sufficient protection to the composers of databases. Anyone could copy databases with factual data on their computer, re-arrange the data and then commercialise it as one's own product. The result was that nobody was tempted to invest in databases with an unprotected content. That is the reason why the European Database Directive¹⁵⁹ has installed a specific *sui generis* protection, regardless of whether the database itself is copyright protected. Since then the permission of the composer of a database is required to copy and/or spread its content, even if it is a collection of facts.

The applicability of this special regime has some practical consequences for the archivist as the employer holds the database's copyrights, just as for computer programmes, unless a different agreement has been made.

What protection does this regime offer databases and what are the consequences for the archivist who wants to archive a website linked to a database?

A distinction needs to be made between two levels of protection:

When the archivist wants to make a copy of the database for archiving purposes the question arises whether the material (the content) that composes the database is copyright protected. 'Works' that are protected by copyright¹⁶⁰ can only be taken into a database with the author's permission¹⁶¹. The

¹⁵⁵ Despite the fact that the regime of the Copyright Directive applies to computer programmes as far as there are no special rules in the Software Directive.

¹⁵⁶ See below

¹⁵⁷ High Council, 4 January 1991, *NJ* 1991, 608

¹⁵⁸ U.S. Supreme Court, 27 maart 1991, *Informatierecht/AMI* 1991, 179

¹⁵⁹ This Directive of 11 March 1996 concerning the legal protection of databases has been transformed into Belgian legislation by the Act of 31 August 1998 concerning the legal protection of databases (*B.S.* 14 November 1998), that has added a section 4bis with a number of special rules concerning databases to the Copyright Act of 1994. Section 4bis of the CA handles the protection of copyrighted databases. The *sui generis* right of database producers has survived in the Act of 31 August 1998.

¹⁶⁰ See above

archivist therefore needs to verify with the database creator whether its content is legally acquired. If not, the archivist breaches the copyright laws by copying this database.

Secondly they need to ask themselves the question whether the database itself is copyright protected. The database is protected as far as the content (for example its completeness) or the structuring of the material shows a sign of creativity. Databases linked to websites to answer individual requests for information (for example in the framework of e-government) will never be very original regarding the selection of their content or structuring. The archivist will have to perform this check. If the database turns out to be not original, he will still have to keep the producer's *sui generis* right into account when the acquisition, control or presentation of the content proves to be a substantial investment. Next to the copyright protection of the structure of a database, also a right to prevent extraction exists, a sort of protection of investment¹⁶².

The author of a database that has survived the originality check has according to Article 5 of the Database Directive the exclusive right to grant permission to every form of full or partial reproduction and to every publication. The archivist therefore needs the database author's authorisation to transfer its content temporarily or permanently onto another carrier, whatever means or forms are applied. He also needs his or her permission to make the database available to the public.

The Directive foresees some exceptions¹⁶³, but these do not allow an escape from the archivist's obligation to make an agreement with the database producer allowing them to copy the database published on the Internet for archiving purposes.

Databases linked to government sites are often not copyright protected and therefore no problems concerning their reproduction are caused. Their content usually consists uniquely of internal information so that some government members can take up the role of authors themselves. The database itself will usually not excel because of its originality or completeness, especially as the material that composes the database is government property. Protecting the investment and the hard work of the database author is not necessary here.

Databases fall under a more flexible regime as far as availability by archive institutions is concerned. During the preparations of the transposition of the Database Directive into Belgian legislation, the problem of institutions that offer information of possible importance to scientific research (like libraries and archives) has been presented. An exception was proposed to the Database Act with regard to reproduction and announcement of databases for scientific goals by institutions that make these databases freely or for valuable consideration available to others. The legislator remembered this sector remark and Article 22bis §1 2°-4° CA now determines that the author's permission is necessary neither to make a database available to the public (for example via the Internet), nor to reproduce it, as far as these acts take place to the benefit of scientific research. If the database's content as such is not copyright protected but is subject to the extraction right, permission needs to be asked anyway for the making available, except when the reproduction is partial¹⁶⁴. Here too it is foreseen that the database may be reproduced without the author's permission when a copy is requested for scientific purposes.

¹⁶¹ See Article 20 bis CA: '*... en laat de bestaande rechten op de werken, gegevens of andere elementen vervat in de databank onverlet.*' [... and does not interfere with existing rights on the works, data or other elements incorporated in the database]

¹⁶² It is said that the extraction right "protects the sweat of the database producer": HUGENHOLTZ, P., 'De databankrichtlijn eindelijk aanvaard: een zeer kritisch commentaar', *Computerrecht*, 1996/4, 132.

¹⁶³ Article 22bis CA

¹⁶⁴ Article 7 2° Database Act

C.6. Conclusion

Before archiving a website, an archivist has to undertake the following steps to follow the CA:

STEP 1: The archivist has to determine whether he considers the website to be copyright protected. This is a factual question that is sometimes hard to decide upon. If there is some doubt he best assumes the work to be original. He also has to decide whether the content of the website is copyright protected. If the answer to both questions is no, the archivist can run through the different archiving actions. If the answer to at least one of these questions is yes, the archivist has to go to step 2.

STEP 2: Next the archivist needs to determine what modifications can be made to the website (reproduction, changes, etc.) and qualify these either as property rights or as moral rights.

STEP 3: Finally the archivist needs to determine who holds the property rights and/or the moral rights. The person who has actually created the work normally holds the moral rights. The property rights can be handed over by agreement. Sometimes the law determines that property rights belong in principle to the employer (computer programmes and databases). Permission has to be asked to this person or these persons to perform the necessary modifications.

D. SOLUTION: EXCEPTION FOR PRESERVATION PURPOSES

The current copyright legislation does not suffice for preservation purposes, both of paper and of digital works. All actions an archivist commonly undertakes require the permission of the author, both the designer of the website and the author of its content.

Sometimes however it can be unclear who holds the copyrights or there is more than one author. Especially for websites on the often-anonymous web it can be difficult to find out who holds the copyrights. It is impossible for archive departments to ask each author for permission, as this is far too time-consuming. The risk even exists that permission is denied. We have softened the problem with regard to government websites on some points, but this does not prevent them from facing, at some stage, the obstacles of copyright as well.

Existing legislation needs to be applied as much as possible to the digital world. Only when the further development of this digital world is prevented, legislation should be adapted¹⁶⁵. The analysis made above allows us to state that copyright is forming a barrier to archive institutions, not just regarding digital archiving of websites but regarding the archiving of all copyright protected works. A law amendment will not only be necessary due to the growing digitisation, but it will also solve an old problem of the archives sector.

¹⁶⁵ LANGE, J., 'De wet en elektronische publikaties',
<http://nieuws.surfnet.nl/nieuws/snn-archief/achtergrond/jg97-98/kb-symposium2.html>

Article 46 7° CA determines that a performing artist's exclusive rights can be limited for storage by the Koninklijk Belgisch Filmarchief [the Royal Belgian Film Archives] of the cinematographic patrimony by means of copies, doubles, restorations and transpositions. The CA therefore does make an exception for the preservation of copyright protected material but unfortunately limits this exception to one single archive institution and to the rights of a performing artist. Legislative action should be undertaken to generalise this exception.

A harmonisation operation of copyright laws has started early 1994. The working group that delivered the so-called *Bangemann report* in May 1994 was convinced that '*in order to stimulate the development of new multimedia products and services, existing legal regimes - both national and Union - will have to be re-examined to see whether they are appropriate to the new information society*' and that '*where necessary, adjustments will have to be made*'¹⁶⁶. The Council of Ministers only accepted the new Copyright Directive on 9 April 2001¹⁶⁷. During the adaptation of copyright to the rise of new digital reproduction and spreading techniques, the possibility was introduced for the Member States to foresee exceptions in their copyright legislation. Harmonisation in this field seems unlikely however as most exceptions are optional. Furthermore the restrictive list of allowed exceptions to copyright is a summary of all existing exceptions in the Member States. In most Member States copyright will probably stay as it is, including exceptions.

- This approach is a missed opportunity for the archives sector and has some serious consequences. Article 5 section 2 c) of the Directive foresees the introduction of restrictions by the Member States to the author's exclusive reproduction right to the benefit of archives that do not strive towards gaining any economical or commercial benefit. The European Parliament has insisted on this stipulation, as have the Netherlands and some other Member States, to allow digital material to be stored in an archive. Only non-commercial institutions like public government archives would be targeted with this exception. Just to be clear: we repeat that this non-obligatory exception does not apply to the reproduction of already archived works for example on behalf of researchers (there is already an exception valid for this, see above). It does apply however to reproductions that are required for the archiving of the works (for example the downloading of websites) or to their storage when technological ageing would occur.

Member States are also allowed to add exceptions within their own national legislation to the exclusive right of public accessibility of their collections by libraries, museums and archives. These exceptions however need to abide to some strict conditions¹⁶⁸:

- The accessibility can only take place in the buildings of the institution via special terminals. Accessibility via the Internet will still require the rights holder's permission.

¹⁶⁶ Europe and the global information society, Recommendations to the European Council <http://www.bookmarks.de/lib/politics/bangemann/report.html#section14>

¹⁶⁷ Directive 2001/19/EG of the European Parliament and the Council concerning the harmonisation of certain legal aspects of copyright and neighbouring rights, Pb. EG 2001 L 167/10).

¹⁶⁸ Article 5.3 n) of the Directive: '*Member states can add limitations to the communication right regarding the use of works that are not for sale or that are bound by licence limitations that form part of the collection of the institutions described in part 2 under section c) (non-commercial museums, archives and libraries), consisting of a communication of the works or materials via special terminals in the buildings of the organisation of that they be made available to individual members of the public.*'

- These exceptions can only be valid for material that is not for sale elsewhere (for example software or documents that are put for sale on a website) or that is not bound by licence conditions. That is logical: copyright starts from the principle that authors are entitled to fees for their artistic creation, their contribution to the cultural heritage and their intellectual effort. Reproduction right allows the author to commercialise his or her work exclusively for him or herself. The regime of forced licences attributes some postponed wages to the author in those cases where his or her exclusive reproduction right has to be infringed upon. The Member States cannot allow that material that has already been commercialised can still be freely consulted by third parties via archive institutions.
- The consultation can only be allowed for research or private study purposes. Similar to a librarian with regard to reprodright for didactical purposes¹⁶⁹, the archivist will find it difficult to determine the intentions of the person in the reading room, but it remains his or her responsibility to check them anyway. If necessary he can ask the researchers to provide the necessary evidence.

As far as Belgium is concerned, we can pose that the legislator (so far) has taken the wrong steps. Currently there is an enactment being discussed in the Senate to change the Act of 30 June 1994 concerning copyright and neighbouring rights in the context of the developing information society¹⁷⁰. A pre-agreement was reached on 15 February 2002 and approved by the council of ministers. Both legislative initiatives erase the voluntary exceptions to reproduction right, including the one for preservation purposes. They do keep the existing exceptions however.

The preparatory works stipulated that if the other proposed exceptions were accepted, the copyright system would become useless¹⁷¹. The idea is that these exceptions would delete the essence of copyright and that an evolution would be started towards a situation where the government is subsidising the authors. Currently reproduction, communication, etc. without the author's permission is only allowed in a limited number of occasions. The new system would allow reproduction, communication, etc. as a principle because of the many exceptions, and the author in exchange receives a fair fee from the government.

The archiving matter however has nothing to do with author fees. It does not deal with measuring the interests of one individual or organisation versus those of the author. Archiving by public archive institutions measures the interests of authors and those of the public. One clearly senses that the exceptions that would breach the current legal system should be allowed. When the enactment remains as it is regarding the transition of the optional exceptions from the Directive, archivists will not be able to apply an existing legal limitation.

One can regret that the exception for preservation purposes has not been included in the Directive as an obligatory exception. It is hoped that the Belgian legislator will incorporate the exception anyway in the copyright legislation. The exceptions for preservation purposes should be limited to

¹⁶⁹ See footnote 144

¹⁷⁰ Enactment of 23 March 2001 concerning the modification of the Act of 30 June 1994 concerning copyright and neighbouring rights in the context of the developing information society, *Parl. St.*, Senate, 2000-2001, 2-704/1

¹⁷¹ Enactment of 23 March 2001 concerning the modification of the Act of 30 June 1994 concerning copyright and neighbouring rights in the context of the developing information society, *Parl. St.*, Senate, 2000-2001, 2-704/1, 4

those reproductions that are necessary to achieve the conservation tasks. This exception could be formulated as follows:

“Will not be considered to be a breach of copyright of literary works or works of art: the reproduction by institutions that have a conservation function and that do not strive towards direct or indirect economical or commercial benefits, and only to the extent that is necessary to exercise this conservation function.

Will not be considered to be a breach of copyright of literary works or works of art: the communication to the public by the institutions mentioned in the previous section, of material in their collection as far as:

- *this material is not for sale elsewhere of bound by licence conditions, and*
- *the communication is done via terminals in the buildings of the fore-called institutions, and*
- *the consultation will only take place for research of private study purposes”*

Until the day that such a stipulation will form part of the CA archivists are obliged to continuously ask permission from the copyrights holders. In the not unthinkable situation where this permission would not be granted, it can only be hoped that the judges will agree with the section on archiving of the theory of the Dutch Dior/Evora judgement by lack of a legal exception. This judgement was made after Dior was trying to prevent Kruitvat from printing the wrapping of their perfumes in their promotional flyers, based on their copyright of that wrapping. Dior tried to prevent the parallel import of its perfumes, now that this was no longer possible by its finished marks right. The Dutch Copyright Act however does not contain any exceptions for this situation. Because of the closed system of copyright exceptions Dior almost managed to stop the sales by Kruitvat. The High Council did however limit the boundaries of copyright by measuring the author’s interests versus the social or economical interests of others¹⁷².

Maybe copyright will play a less important role in the future, as is predicted by many authors. Copyright has always been the main legal tool, and still is, to limit the mass-spread of communication¹⁷³. The most typical characteristic of the Internet is in parallel to this as information can be spread worldwide and can be accessed in principle by any person with an Internet connection.

VIII. ARCHIVING PERSONAL DATA

A. PERSONAL DATA AND THE INTERNET

¹⁷² HR 20 October 1995, *NJ* 1996, 682; ALBERDINGK THIJM, C., ‘Fair use: het auteursrechtelijk evenwicht hersteld’, http://www.ivir.nl/publicaties/overig/alberdingk_thijm/fair-use.doc

¹⁷³ HUGENHOLTZ, B., *Intellectual Property rights on the Information Superhighway*, Report to the Commission of the European Communities (DG XV), August 1994

As well as the copyright issue some aspects of privacy protection when archiving websites have to be considered. The applicability of privacy arguments on the Internet is clearly demonstrated by the discontinuation of the web archiving project of the Swedish National Library by the Swedish *Data Inspection Board*. In November 2001 the Swedish National Library started automatic collection and storage of Swedish websites in order to save this part of the cultural heritage for the future. The *Data Inspection Board* however declared that keeping this kind of collection includes the processing of personal data and therefore has follow the *Personal Data Act*. Even though the *Board* recognised the need to store the cultural heritage for historical and research purposes, the decision was made that the collection and the availability of personal data for these purposes need to be arranged by a specific law. In the meantime, until such a law has been introduced, the Swedish National Library has to abandon its web archiving activities.

Processing of personal data in Belgium, also by archivists, is subject to the regime of the Act of 8 December 1992 concerning the safeguarding of personal privacy when processing personal data (further on referred to as PDPA, “Personal Data Processing Act”). “Processing” can be defined as “each action undertaken regarding personal data, including collection, requesting (downloading), storing, availability by means of transmission, etc.”. Personal data is every piece of information regarding an identified or identifiable natural person. Websites often contain a wide collection of personal data: names, addresses, phone numbers, etc. and the law considers all as personal data.

The HTML pages of dynamic websites are only being compiled on the server after an HTTP request has been received. The result the visitor gets on his or her screen depends among others on the request he has formulated or on his or her profile. Reconstructing the content of those websites is only possible when also this interaction is stored. Storing this interaction implies the archiving of log files, IP addresses and/or personalised information streams based on cookies, as described in the technical part of this report. To find out whether the processing of this data by the archive department falls under the reach of the PDPA we first need to discuss briefly the organisation and management of the Internet.

A.1. Organisation and management of the Internet

The Internet is a worldwide computer network that is used for information exchange and communication. Each computer on the Internet is identified by a unique numerical IP address in the form A.B.C.D with A, B, C and D being numbers between 0 and 255 (for example 67.154.85.243). Even though many consider the Internet as an international community of *persons*, an IP address is primarily linked to a *computer* and not to a person. A domain name is an alternative and easier way to indicate a location on the Internet. It is an easy to remember combination of letters like a company name, corresponding to one specific IP address. The domain name is for human usage, the IP address for computer usage.

As the information offered via the Internet is spread over millions of computers, *Uniform Resource Locators* (URLs) need to be used. A URL (also called an “Internet address”) is not the same as a domain name. It is the standard method to indicate the location of the different information sources. The first part of the URL indicates what protocol is to be used to call the information (for example `http://` for files on a web server, `ftp://` for files on an FTP server, `news://` for files of a Usenet news group, etc.). Next a URL contains the IP address of the computer on which the requested information is stored. A general URL only consists of the protocol indication and the IP address of the computer and usually delivers the homepage of this computer. Some computers however house multiple

homepages. They each receive a specific directory or folder on the computer that is then incorporated into the URL¹⁷⁴. And finally the URL often contains the name and the format of the requested document (for example .html, archiving-websites.pdf).

It proved to be impossible to co-ordinate millions of domain names worldwide and to keep them unique, so the *Domain Name System* (DNS) was installed. This system attributes domain names to computers that are identified via an IP address. It translates an IP address into a domain name (or vice versa) and contains a global, decentralised and hierarchical database of domain names. Domain names have the format <names>.top-level_domain, with top-level domain being a general domain (like .com for commercial websites) or a geographical domain (like .be). The top-level domain can be compared to the area code of a telephone number. The amount of domains a server has to run through to find the requested information is thus limited considerably. When a server requests a certain website via its browser, the ISP server will translate the indicated Internet address (the URL) into the IP address of the addressee. Servers on Internet crossroads, the so-called routers, will go and find the manager of the concerned site.

Every person or organisation wanting to access the Internet needs to assign an IP address to his or her systems. The *Internet Corporation for Assigned Names and Numbers*¹⁷⁵ is responsible for the assignment of IP addresses to Internet providers, who supply them then to their customers. To be connected this person or organisation needs to sign a contract with an Internet provider¹⁷⁶ containing his or her name, address and other personal data. The subscriber then receives a user name (user ID) and a password to prevent unauthorised usage of his or her subscription. The Internet provider is therefore always capable of identifying an individual network user based on the IP address.

This is also the case for the so-called “dynamic” IP addresses. Dynamic IP addresses result from the limited amount of available IP addresses. Internet providers have a limited number of IP addresses at their disposal that they can attribute to their subscribers. For this reason subscribers who are not permanently connected to the Internet will receive a different IP address each time they log on. The moment a subscriber finishes his or her Internet session, the IP address will be released to another user. An Internet provider can still identify an individual user via the dynamic IP address on the condition that he stores the traffic information that indicates what IP address was assigned to what login name at what time. For security reasons Internet providers systematically register the date, time and duration of their subscribers’ sessions and their assigned (dynamic) IP addresses.

The ISPs also keep Internet traffic log files on the web server side. Once the connection to the website has been accomplished, it will start collecting information about the visitor. All this information is registered in the server log files that thus collect a large amount of information about the communication process: who visited the website, on what date and time, duration of the visit, executed actions, *click streams*, pages visited, files downloaded etc. New surveillance software has

¹⁷⁴ The ICRI website is managed on the same computer as the general website of the Law Faculty of the University of Leuven: <http://www.law.kuleuven.ac.be/icri>

¹⁷⁵ <http://www.icann.org>

¹⁷⁶ An Internet provider is different from an Internet Service Provider (ISP). ISPs perform web hosting, putting web pages on their web server. An Internet provider gives subscribers access to the information on the computer of all different ISPs. An Internet provider acts as an access gate to the Internet and has knowledge of all the subscriber’s traffic while an ISP only overviews the action on its own server(s).

become available for ISPs that reports in real time about the types of content that is being viewed and downloaded by visitors¹⁷⁷.

Another means of retrieving and registering personalised information about visitors are *cookies*. Direct marketing companies often use cookies to learn about the interests and consumption behaviour of individual Internet users. The visited website will send a cookie to the web browser that stores it on the hard disk of the user's computer. When this person visits the same website again later, the web browser will use his or her preferences and user profile. The communication streams that occur here need to be archived as well.

A.2. IP addresses

It needs to be considered if the processing of IP addresses by an archive department falls under the application of privacy rules. To answer this it needs to be determined whether an IP address is a piece of personal data. This matter cannot be clearly solved from the perspective of Belgian legislation. Yet it is an important issue as it determines the rights of those involved with regard to the archive department (right to communicate and correct). Furthermore non-abiding to the law can lead to serious penalties. This uncertainty is caused by the interpretation of the term "identifiable". The European Directive 95/46 of 24 October 1995 concerning the protection of natural persons with regard to the processing of personal data and regarding free traffic of those data, defines an "identifiable" person as "a person who can be directly or indirectly identified by means of an identification number or by means of one or more specific characteristics of his or her physical, physiological, economical, cultural and social identity". Consideration 26 of the Directive¹⁷⁸ clarifies the "Memorie van Toelichting" [Explanatory memorandum] of the Belgian PDPA: information regarding a person needs to be considered as personal data as long as can be determined what person this information applies to by any reasonable means¹⁷⁹. To define data as personal data it therefore suffices that another person than the one responsible for the processing is capable of identifying the person using reasonable means. Or put differently: information about persons will only become anonymous when this operation is absolute and there is no way back from anonymity¹⁸⁰.

Applying this to IP addresses demonstrates that an IP address is personal data as long as it is reasonably possible to determine what person lies behind the IP address. An archive department wanting to archive the IP addresses of the visitors of a website is therefore usually processing personal data. This is obvious because the Internet provider is capable of identifying the individual surfers via their IP address, unless when, for dynamic IP addresses, traffic information about the attribution of IP addresses has not been stored or is no longer available after some time.

¹⁷⁷ *Privacy op Internet – Een geïntegreerde EU-aanpak van on-linegegevensbescherming*, Group Data Processing Article 29, approved on 21 November 2000, European Union, 50.

¹⁷⁸ "To determine whether a person is identifiable all factors need to be checked of which can be reasonably assumed that they can be used to identify the person mentioned before by the person responsible for the processing or by any other person."

¹⁷⁹ Explanatory memorandum to the enactment to transfer the Directive 95/46/EG of 24 October 1995 of the European Parliament and the Council concerning the protection of natural persons when processing personal data and concerning the free traffic of that data, *Parl. St.*, Chamber, 1997-1998, 1566/1, 12

¹⁸⁰ *Ibid.*

Other EU Member States do not follow this strict interpretation. The Netherlands for example has a very different opinion about this. This becomes visible among others in an advice of the Dutch Registration Chamber (the counterpart of the Belgian “Commissie voor de bescherming van de persoonlijke levenssfeer” [Commission for Data Protection]) of 19 March 2000¹⁸¹ that poses that an IP address should not always be considered as personal data. Regarding the notion “identifiable” the Registration Chamber focuses on the question whether the person’s identity can be reasonably, without disproportionate effort, determined by the body responsible for the processing (and therefore not by any other person)¹⁸². This depends to some extent on the holder’s possibilities and the presence or availability of additional information. It has to be assumed here that the holder is reasonably equipped. For concrete cases also the special expertise and the technical facilities of the responsible of the processing need to be taken into account, according to the Registration Chamber.

This point of view, that is probably a lot more pragmatic and contributes better to privacy protection¹⁸³, results in IP addresses only being considered as personal data if the body responsible for the processing has the possibility to identify individual Internet users via their IP addresses. However this point of view does allow taking into account the information third parties may possess. When a third party (for example an Internet provider) possesses information that allow identification, the body responsible for the processing (for example the archivist) needs to demonstrate that he does not or cannot possess that additional information that allows identification of the person involved.

Should this interpretation of the notion “identifiable” from now on be used in Belgian legislation? The Commission for Data Protection has supported this interpretation in its advice regarding the draft PDPA. It posed that the question whether a return from anonymity is reasonably possible should be judged by the body responsible for the processing¹⁸⁴. The legal framework however does not support this point of view. The royal decree confirms the very strict interpretation that had been given in the explanatory memorandum. The stipulations concerning historic, statistic and scientific research mention *coded personal data*. This implies that encoding the data, and therefore anonymity regarding the processor, does not change its qualification as personal data.

And yet there is a defensible alternative to the legal point of view that does not conflict with the text, and that allows the archive departments to escape the application of the PDPA. The definition of the concept “identifiable” in Article 1 §1¹⁸⁵ of the law can result in a refutable suspicion of identifiability. The starting point remains that a person is identifiable when a technical means exists somewhere to identify this person. However the responsible can prove that he does not or cannot possess that technical means to perform the identification¹⁸⁶. Despite that this is a negative fact, the archive department will find it easy to prove this when archiving IP numbers, as there is no relation whatsoever between the archive department and the Internet providers of the subscribers that have visited the site to be archived.

¹⁸¹ Advice z2000-0340, 19 March 2000, <http://www.registratiekamer.nl/bis/content-1-1-9-3-7-2.html>

¹⁸² The Dutch privacy act uses the term “holder” instead of “responsible for the processing”.

¹⁸³ The body responsible for the processing who does not possess the means to identify the person will hesitate his attempt to identify as he may enter the application domain of the law.

¹⁸⁴ Advice 30/96

¹⁸⁵ “*Considered as identifiable are...*”

¹⁸⁶ This opinion stems from: DE BOT, D., *Verwerking van persoonsgegevens*, Antwerp, Kluwer, 2001, 32-...

A.3. Log files

The data in the log files of ISPs, registering the browsing behaviour, does usually not lead directly to a natural person but to an IP address. If this IP address would allow identification of the user, the log files are to be considered as personal data. As long as these log files do not allow an analysis of certain users' visits to certain websites, they are not personal data.

Secondly the log files are also “traffic information” in the sense of the common point of view regarding the draft of the new Directive that will replace the current Directive 97/66/EG concerning the processing of personal data and the personal privacy protection. Traffic information as meant in the Directive is “data that is being processed for the transfer of communication via an electronic communication network or for its billing”. The common point of view determines that traffic information is in principle confidential and that storing or registering it in another way by others than the network users is forbidden. Only the technical storage necessary for the transfer of information is allowed¹⁸⁷. But also in that case traffic information must be erased or made anonymous when no longer needed for maintaining the communication, that is: when the Internet user has access to the website¹⁸⁸.

Yet we notice that ISPs store log files systematically and much longer than is needed for maintaining the communication. The new Cyber Crime Act obliges ISPs even to store their log files for at least 12 months to allow investigation and prosecution of crimes committed via a communication network¹⁸⁹. This exception to the general principle of confidentiality of telecommunication can only be called in however for matters of national security. The storage of log files by an archives department cannot qualify here. Therefore unless when an institution hosts its own website and serves as an ISP, it will find it very difficult to access the log files of the website to be archived.

B. PROCESSING PERSONAL DATA FOR HISTORICAL PURPOSES: COMPATIBLE WITH THE ORIGINAL GOAL?

The previous paragraph has shown that an archive department will encounter personal data during web archiving. Now the question arises about how this data should be processed. The answer to this question lies with the concept “historical purposes” of the PDPA.

Belgian privacy rules start from the principle that processing personal data as such is not permitted except in a number of well-defined exceptions. The cases where processing can be acceptable are defined in Article 5 of the PDPA. Even when one of the acceptability criteria has been fulfilled (for example permission of the person(s) involved) the processor can only collect personal data for well-defined and justifiable goals.

Storing personal data is a so-called new, *further* processing of personal data. What does this imply? Collecting and registering personal data as such is already a first form of processing. Next something will be done with this data. Such a further processing of personal data can be every act of processing

¹⁸⁷ Article 5 of the common point of view

¹⁸⁸ Article 6 of the common point of view

¹⁸⁹ First however a royal decree needs to be published that determines what data exactly needs to be stored.

that is performed on this data after its acquisition: storing it, consulting it, spreading it, etc¹⁹⁰. Any further processing legally needs to be in accordance with the goals for which the data has been acquired, taking all relevant factors into account, that is: the reasonable expectations of those involved¹⁹¹. When personal data has been acquired for direct marketing purposes, this data may also be stored for these purposes.

The Directive and the linked Belgian PDPA determine that a further processing of personal data *for historical purposes* can be unified with the original primary goal (and can therefore be allowed) as far as appropriate guarantees are offered¹⁹². The question is whether the processing of personal data by an archivist is a processing for historical purposes. The answer to this question is of great importance as Belgium has introduced a gradual three-step system as an appropriate guarantee for the processing of personal data for historical purposes (and for statistic and scientific purposes). This system can be found in Chapter II of the royal decree of the PDPA. If the answer to this question is positive, the archivist will have to apply the three-step system.

In principle the historical processing needs to take place based on anonymous data. If the historical goals of the processing cannot be achieved by anonymous processing, the processor is allowed to use coded personal data. If this still does not allow an achievement of the historical goals, the personal data can be processed in its original form. Even though digital data can be made anonymous much easier than paper data, this way of working would cause quite some extra handling for the archivist and furthermore would prevent the websites from being archived in their original form.

Neither the Directive's text nor its considerations however contain a definition of the notion "historical". The term "historical" is defined as the processing of personal data to analyse a past event or to allow that analysis. This is the essence of an archivist's work: allowing the future analysis of past events. The Report to the King however clearly determines that mere archiving by the responsible for the execution of his or her own files does not classify as a storage with historical purposes and therefore does not fall into the application domain of Chapter II¹⁹³. Only the archiving by another person than the responsible seems to be a processing for historical purposes. The three-step system from Chapter II of the royal decree only applies to the later processing for historical purposes that cannot be united with the original goals for which the data has been collected or received¹⁹⁴. An application to websites shows that the following method needs to be applied when archiving personal data in the framework of web archiving:

- **Storage of a proper website (or of other personal data in a similar way) by the body responsible for the execution himself**
 - = later processing (not for historical purposes) but compatible with the original goal
 - archiving of personal data is permitted and subject to the normal arrangement. All processing in the dynamic and semi-static phase belong to this + some processing in the static phase: when the archive department that possesses the files is a part of the responsible for the execution for example the archive department of a financial institution.

¹⁹⁰ DE BOT, D., *o.c.*, nr. 155, 119

¹⁹¹ Article 4 §1 2° PDPA

¹⁹² Article 6.1 b) of the Directive and Article 4 §1 2° PDPA

¹⁹³ Report to the King regarding the royal decree, *B.S.* 13 March 2001, 7846

¹⁹⁴ *Ibid.*

- **Storage of a website (or of other personal data in a similar way) by another person than the body responsible for the execution**

= later processing for historical purposes

In this situation it should be considered whether the storage is compatible with the original goal for which the data has been collected, taking into account the reasonable expectations of those involved. The possible public nature of the personal data can form a guide to determine whether the person involved could expect his or her personal data to be used later for historical purposes.

- If the answer is yes, archiving this personal data is allowed and subject to the normal regulations (for example the archiving by the Antwerp City Archives of the websites of local organisations)
- If the answer is no, archiving this personal data is allowed as far as the regulations of Chapter II of the royal decree (for example a genealogist's archiving of a website needed for historical research) are being lived up to.

IX. THE FLEMISH GOVERNMENTS ON THE INTERNET

A. THE SITUATION IN 2002

In December 1999 the European Commission launched the e-Europe initiative with the intention of bringing Europe on-line. As e-Europe was well-received by the Member States, the European Parliament and the main actors in the sector, the state and government leaders have decided during the European Council held in Lisbon on 23 and 24 March 2000 on a number of concrete measures to make e-Europe a success. It was stressed that Europe should start to use the opportunities of the new economy and especially the Internet.

On 24 May 2000 the European Commission accepted an action plan, that was agreed upon by the Member States at the European Council of Feira on 19 and 20 June 2000. The action plan is solution-driven and concentrates on who needs to do what and when. The action plan is grouped around three main goals:

- A cheaper, quicker and more secure Internet
- Investment in people and skills
- Stimulating Internet usage

In the framework of this last goal, all government levels are asked to use new technologies to make government information as accessible as possible and to realise a general electronic access to important government departments by 2003¹⁹⁵.

Recent research however does not produce a positive picture. Belgium and Flanders are mentioned in very few international studies as an example of countries with great e-government ambitions and far-reaching results¹⁹⁶. In fact, they need to excel to be able to catch up to other Member States. The presence of both federal and Flemish government on the Internet is still far too limited. This is perhaps because of the lack of a secure organisational and technical framework for digital signature applications. Currently some Federal departments are working on the implementation of the digital identity card¹⁹⁷. Due to a lack of a technical framework the current Internet services are limited. Most Flemish government websites do not yet offer any interactive services or democratic participation¹⁹⁸. Almost all sites consist of one-way traffic from the government to the citizen. They can be compared to flyers containing all sorts of useful information for the citizen such as a cultural agenda, a community newsletter, etc. The possibility arises to ask questions to the government via e-mail, but only in situations where it is not vital to securely determine the identity of the surfing citizen.

Over the last ten years the Flemish government has increased its usage of information and communication technologies for the acquisition, disclosure and spreading of information, and a lot of effort is being undertaken to enhance this even further. There is a general desire not to miss the e-Europe train. Even though the Flemish government considers the e-portal to be a political priority, it appears that an interactive usage of the Internet will not be installed in the near future. Despite this (temporary) limitation most Flemish governments are working towards being present on the Internet.

The first Flemish city sites on the Internet appeared in 1995 (Antwerp, Ghent, Kortrijk, Knokke-Heist). In the autumn of 1997 more than one Flemish community out of ten had its own website on the Internet and this had increased to 3 out of 10 by 1998¹⁹⁹. Currently almost 60% of Flemish local governments are on-line²⁰⁰. The Flemish government is present on the Internet via its portal site www.vlaanderen.be. Each of the 5 Flemish provinces has their own website²⁰¹.

B. CASE: THE ELECTRONIC PORTAL POPULATION REGISTER OF CEVI

¹⁹⁵ http://europa.eu.int/information_society/eeurope/action_plan/pdf/actionplan_nl.pdf

¹⁹⁶ GOESAERT, J., 'E-Government: Do New Technologies Build a Bridge between Government and Citizen?', in *A Decade of Research @ the Crossroads of Law and ICT*, DUMORTIER, J. (ed.), Gent, Larcier, 2001, 92

¹⁹⁷ See below

¹⁹⁸ In 1999 JO STEYAERT noticed that almost half of the community sites were hardly worth the name 'website'. They were 'under construction', not functioning or only containing one page. See STEYAERT, J., 'Het Internet in Vlaanderen: zijn steden, zijn gemeenten, zijn inwoners...', in *Digitale steden en gemeenten. Handleiding*, GOUBIN, E. (ed.), 1999, Politeia, Brussels, 36

¹⁹⁹ STEYAERT, J., GOUBIN, E en PLEES, Y., 'Digitale gemeentelijke communicatie in Vlaanderen: de cijfers', in *Digitale steden en gemeenten in Vlaanderen. Een stand van zaken*, STEYAERT, J. (ed.), Brussels, Politeia, 2000, 39

²⁰⁰ Situation on 13 June 2001: <http://belgium.fgov.be/links/1141.htm>

²⁰¹ <http://www2.cipal.be/cipal/flinks2.html>

Cevi, the subregional computer centre of the provinces of Oost and West-Vlaanderen has set up an e-portal for its customers. Each of the 93 local governments that are clients of Cevi for the Population Administration application can provide the service of this e-portal to their citizens.

Citizens consulting the website of their local government need to click on the button of the e-portal. Citizens of Ghent for example can request a transcript of their own marriage certificate via the Internet. They can also check the progress in a case. When a citizen wants to order a certain document he needs to log in with his or her national number (“rijksregisternummer” in Dutch, in the upper right corner of the social security SIS card). There will be an automatic check of the national register or the local population register. The requested documents are sent via postal mail to the official address of the citizen. Even though the national number is used (which also appears on the identity card and is therefore not very secure) and not a *unique identifier* this is currently the most advanced application of e-portals in Flanders.

The local government that wishes to offer this e-portal needs to be in the possession of a website. This site is linked to the URL of the e-portal that has been put on a secure Cevi server. It is currently impossible to provide population data via this website as a definite electronic identification of the citizen is not yet operational.

X. GOVERNMENT LIABILITY FOR ITS OWN WEBSITE

A. CAN A CITIZEN OBTAIN RIGHTS FROM THE CONTENT OF A GOVERNMENT WEBSITE?

Next to the lack of a technological framework for an interactive e-government on the Internet, there is a second reason why the Flemish government is a bit hesitant in its information policy. This may be related to the vagueness regarding the possibility of liability cases.

Government use of ICT could lead to many problems that may cause damage to the citizen. The following situations are not unthinkable:

- Something goes wrong with the transmission of a government message, or it reaches the citizen with a delay
- An unauthorised person views the content of an electronic message
- Unauthorised usage of a digital signature occurs

A citizen could also experience damage as a consequence of government information on the Internet that is incorrect, obsolete or incomplete. The question arises whether a citizen can obtain rights from such information. A distinction needs to be made between two types of information:

- Information conflicting with the law

The issue of wrong or incomplete information is more a matter of liability than of the obtaining of rights from this information by the citizen. One only possesses a certain right if there is a prior government decision that grants this right to the citizen. This decision can consist of a law, a royal decree, a communal police rule, etc. The citizen can learn about his or her rights via the Belgian Official Journal. Here all generally binding determinations that attribute rights to the citizens are published²⁰². This is also the only version that is accepted as authentic²⁰³.

The government cannot be bound by information it has provided that is conflicting with the law. A citizen cannot obtain rights from such information. Put differently: one cannot expect the government to break the law, not even when certain erroneous information has been spread. Imagine that a government website provides information about a due tax amount. The tax debtor cannot later reduce the amount to this announced level. He will eventually only pay what is determined by law and will not suffer any damage²⁰⁴.

- Other erroneous, obsolete or incomplete information

The common liability rules apply to other erroneous, obsolete or incomplete information, as such not conflicting with any legal stipulation²⁰⁵. Also the government is liable for any damage her actions may cause. This principle has been applied for the first time in the well-known Flandria judgement of the Belgian Supreme Court²⁰⁶. No constitutional or legal stipulation or general principle relieves the administrative government of its duties regarding Articles 1382 and 1383 of the Civil Code, including the duty to carefulness mentioned there²⁰⁷.

The tortuous liability of Article 1382 B.W. evolves around three notions. This implies that a citizen who was deluded by erroneous, incomplete or obsolete government information is only entitled to compensation if the government has committed an *error* and if the citizen has suffered damage *because of that (causal relation)*.

An *error* implies an illegal action. Illegal action can consist of breaking a legal rule or violating the “general carefulness norm”. There is no such thing as a rule for the government to provide correct and

²⁰² Article 190 Constitution

²⁰³ Van Orshoven, P., *Bronnen en beginselen van het recht*, Leuven, Acco, 1994, 47

²⁰⁴ GIJSSELS, J., ‘De overheidsaansprakelijkheid in verband met informatie’, *Rechtskundig Weekblad*, 1979-1980, (1202), 1212

²⁰⁵ *Overheidsinformatie: een essentiële hulpbron voor Europa. Groenboek over overheidsinformatie in de informatiemaatschappij*, European Commission, COM (1998) 585

²⁰⁶ Cass., 5 November 1920, *Pas.*, 1920, I, 223-239

²⁰⁷ Cass., 20 June 1997, *Arr. Cass.*, 1997, II, 677

complete information via its website²⁰⁸. A non-careful composition or updating of the content of the website will therefore be considered as a violation of this general carefulness norm. This norm is an unwritten rule according to which everyone, also the government, has to conduct him or herself in society as a normal careful and forward-looking person that finds him or herself in the same circumstances.

The error needs to be attributable to the government. An illegal act will not be attributable if there are softening or justifying circumstances such as an irresistible fallacy or overpower. The exemplary function of the government allows us to assume that there will only rarely be softening circumstances. Websites need to be composed with the same great care as other government information sources like flyers or television commercials.

Next to this government error, the citizen will also have to demonstrate that he suffered damage because of this error. The concept damage has received a broad definition. It could be a loss, but also a missed chance (for example with job interviews). Whether there is actual damage will depend among others on the content of the information that has been supplied. If this information is relatively vague or exists of individual comments or unbinding advice, the person concerned will have great difficulty asking justification from the government and appealing for compensation. The damage will be much simpler to prove when there is actual erroneous or incomplete information that has confused the person concerned regarding his or her rights or has damages his or her legal position. This is a factual matter however, on which a judge may have to make a final decision²⁰⁹.

To demonstrate the government error, the citizen must prove that erroneous, incomplete or obsolete information has been present at a given moment on the website. Right of evidence is therefore an important argument to regularly archive government websites.

B. DISCLAIMERS: USEFUL OR USELESS?

Many websites contain exoneration clauses (*disclaimers*). These are clauses that announce the holder of a website not to be responsible for its content²¹⁰. Also some government websites bear such an exoneration clause²¹¹. Is the government legally allowed to put its responsibility aside and thus leaving the citizen in the cold? This cannot be answered clearly.

The validity in principle of exoneration clauses is recognised in Belgium, both regarding contractual and non-contractual liability²¹². This implies that not only a debtor can limit the legal liability when not abiding to his or her commitments, but also that a damage originator can shield himself off from possible delictual liability with regard to those who might suffer the damage.

The question now is whether governments beforehand can exonerate themselves of liability caused by an illegal deed, for example by putting a disclaimer on their website. In general these exoneration

²⁰⁸ The government is not even obliged to make all sorts of information (like tourist information, contact details for government departments, etc.) available voluntarily via a website or via other means, not keeping the rules of active publicity of government into account. GIJSSELS qualifies these efforts as ‘deeds of voluntary helpfulness to the benefit of both citizens and administration.’

²⁰⁹ GIJSSELS, J., *l.c.*, 1214

²¹⁰ For example <http://www.envida.net/nl/about/terms>

²¹¹ For example http://www.europa.eu.int/geninfo/disclaimer_nl.htm

²¹² Cass., 21 februari 1907, *Pas.*, 1907, I, 135; Cass., 29 September 1972, *Arr. Cass.*, 1973, 121

clauses to the benefit of the government are judged with great carefulness²¹³. Some authors consider an exoneration clause to be justified if it does not conflict with the essential tasks the law hands over to the government²¹⁴. Setting up and maintaining a government site is not a part of the government's essential tasks and therefore such a clause would be authorised. Others situate unauthorised exoneration clauses in those cases when a citizen has no choice but to use a certain service²¹⁵. The use of government sites, including the e-portal, is far from obligatory these days.

We conclude with PRINS that a government that takes its reliable electronic image serious should not shield itself off from all possible damage claims via an exoneration clause on the Internet. These risks will need to be shielded by other means such as an insurance. The risks should be limited by attributing enough attention to availability, confidentiality and integrity both of communication and of infrastructure²¹⁶. Setting up an efficient PKI structure is the best means for this. On the other hand the content of the website needs to be followed up permanently and be composed with great care. It has to be assured that all information that is provided via the e-portal is correct.

XI. PORTAL SITES: THE FUTURE

A. FROM INFORMATION TO INTERACTION AND INTEGRATION

As mentioned earlier, government websites currently only have a limited impact. Even though most sites do separately publish electronic information, this information is always following the logic of the government and not of the citizen. Furthermore this service is not integrated. The citizen has to navigate his or her way in the chaos of websites with insufficient guarantee to completeness and correctness. And finally the government departments offer close to no electronic transactions with the citizen that would allow an electronic two-way interaction and that would integrate with the workflow lying behind the different government departments.

²¹³ VANDENBERGHE, H., 'Exoneratie- en vrijwaringsbeding bij onrechtmatige daad. Samenloop en coëxistentie', in *Exoneratiebedingen*, HERBOTS, J. (ed.), Brugge, Die Keure, 1993, 81

²¹⁴ MAUSSION, F., 'Reflexions sur la théorie de l'organe', in *La responsabilité des pouvoirs publics*, Brussels, 1991, 96 e.v.

²¹⁵ VAN HOOYDONK, E., 'De geldigheid van in havenreglementen opgenomen bevrijdingsbedingen', *Rechtskundig Weekblad*, 1990-1991, 1394

²¹⁶ PRINS, J., E-overheid: Evolutie of revolutie? <http://rechten.kub.nl/prins/Publicatnl/eoverheprev.pdf>

In the future the existing websites, with a little aid from the electronic identity card, should develop into user-friendly front office sites²¹⁷ that give the citizen a logically ordered and integrated offer of services, ranging from information to transactions. However, supplying electronic services via the Internet is only one aspect of e-government. It is the most visible part to the citizen, but also a well-organised information exchange within and between the different government departments is essential to make the portal sites function perfectly. Information would have to be communicated from the citizen to the government only once, which would allow limiting the contact with the government. This would prevent the citizen from having to leave his or her house to queue for hours to obtain a paper certificate that then needs to be handed over to another government department.

B. ONE VIRTUAL GOVERNMENT

The development of this ambitious e-government project takes currently place mostly on a federal level. A new federal government department called FEDICT is co-ordinating this. Currently two user-friendly portal sites are being planned that will be constructed according to the user's logic and that allow interaction with the internal information systems of the different federal government departments. The official assignment to build the portal was published in September 2001.

Despite the subdividing of authorities and tasks over several levels of government (local level, provinces, communities, regions, federal government, Europe) the citizen considers the government to be one unit. That is why an agreement has been reached between the federal government, the communities and the regions to create one virtual government with their electronic services. In practice the website refers the citizen to the government department responsible for the question or transaction the citizen demands. The portal sites of other government levels will also contain links to the information and transactions that are available on the federal level. Via this co-operation the different government levels commit themselves to offer the same tools (basic software, electronic identity card with digital signature, unique identification key) to the citizen.

C. THE LEGAL FRAMEWORK

C.1. Front office: the electronic identity card

Communication between the citizen and the government via the anonymous Internet needs to be protected: the government needs to be sure of the identity of the citizen who presents himself at the e-portal, and also the confidentiality and integrity of the messages needs to be guaranteed. To realise this the council of ministers has approved a proposition on 20 November 2000 to create a federal PKI infrastructure and the usage of a digital identity card by all citizens. On 19 July 2001 the decision was made to introduce the electronic identity card for all natural persons. The legal framework for the electronic identity card has been offered by the European Directive concerning electronic signatures, transposed into Belgian legislation on 9 July 2001 by the Act concerning certificate service providers.

²¹⁷ The term '*front office*' is used to indicate the electronic government services towards the citizen, as opposed to '*back office*' which is the term for the electronic traffic within and between the government departments.

There is a current enactment to adapt the national legislation on the identity card. An electronic identity card will be able to fulfil the following functions:

- Identification of the holder
- Authentication of the holder (the proof that the card holder is the person that is identified by the card)
- Means to put an electronic signature with legal validity
- Electronic proof of a presence, a mandate or a characteristic of the holder, on initiative of the holder
- Bearer of programmes that can be executed within the chip on the card (for example the creation of key pairs)

The electronic identity card will contain the following identification data, visually and electronically readable:

- Surname and first name
- Gender
- Place of birth
- Date of birth
- Nationality
- National number
- Photo

The card will therefore have two functions. The citizen can use his or her card to identify himself in a digital environment, and he or she can also use the card to place an electronic signature for example on official government documents. Two private keys²¹⁸ with matching identity certificates will be incorporated in the processor chip of the card, unless the holder resists to this. The use of the private keys will be secured by a PIN code. The identity certificates will be delivered by a certificate service provider chosen by the government via the normal government assignment procedure, with the intention of getting the lowest price for this.

The electronic identity card will be distributed by local authorities replacing the current plastic identity card and will have a validity of five years. The local population registers will sell the card for 9 EURO. The identification data will be stored on the card in an unchangeable way. If part of this data changes the citizen will require a new electronic identity card. Eleven pilot cities will test the new electronic identity card technology during 2002. The citizens of these cities are offered the possibility to immediately exchange their current card for a digital one. Cardholders whose current card has expired will receive a new digital card. This method should allow each card to be replaced within five years. After a distribution period of six months the council of ministers will decide whether the system is ready to be spread nationwide.

²¹⁸ For the principles of *Public Key Infrastructure* see the DAVID report *Wat en hoe archiveren? Op zoek naar de rol van PKI voor digitale archieven*.

The electronic identity card can form an important added value to the sealing of electronic documents by an archivist. The goal is not only to make e-government more applicable but also to improve applications that require a certified, electronic identity. The processor chip of the electronic identity card will allow the holder to store some extra key pairs linked to characteristic certificates and that will also allow him for example to digitally prove his or her identity as an archivist.

C.2. Back office: unique identification number

It obviously does not suffice to have a website where citizens can identify themselves using electronic identity cards. A first essential condition for an efficient e-government is a unique identification of the entities that information is exchanged about. Currently the legal basis is formed for the usage of a unique identification number for natural persons, companies and organisations. This number will be used in all government information systems allowing an efficient electronic data exchange between government departments and therefore a one-time collection of data from citizens and companies. The lack of such a unique identification method prevents an optimal reuse of already available information.

Today most government departments use more than one number for entities. These numbers are usually not time resistant (for example commercial register number, social security number, tax office number, etc.). In the future only one number should be kept that *is* time resistant. For natural persons this unique identification key will be the national number, which will from now on be called “personal number”. The social security identification number will be used for those persons who have never been registered into the population, foreigners or waiting register (and therefore do not have a national number). For companies and organisations the current VAT number will be used. This number is called the “company and organisation number”. Other numbers like the commercial register number and the social security number will no longer be used in the future.

For the attribution of this number the initiating government department must deliver a fixed set of basic identification data for the entity concerned. This will be stored in a database and managed by the department that attributes the identification number. The state register will fulfil this role for natural persons; for companies and organisations a Crossroads Bank for Companies will be installed. There is already a Draft Act that manages the installation of this Crossroads Bank for Companies, to be operational early 2003. Due to this generalised usage of the national number there is currently work undertaken to review the legislation about its usage.

The government also needs to assure that it no longer requests data from a citizen when this data is already available within another department. This will be achieved by linking the information systems of the federal government in a network allowing a quick and secure information exchange. The legal basis for this is formed by Article 102 of the Act of 30 December 2001 that determines that public federal departments can be obliged to make their information available, preferably electronically, to other government departments that might need this information to fulfil their tasks²¹⁹.

As most government information is spread over multiple platforms a *message engine* will be generated. This will enable an intelligent exchange of structured messages between the information systems of the federal government in its broadest sense, between those information systems and the websites or portals, and between those information systems and those of other government levels:

²¹⁹ B.S. 31 December 2001

communities, regions, provinces and cities. The different tasks to assure data storage in a unique form will be divided between the government departments in the network. The government department that manages certain information is responsible to keep it up-to-date. When a government department requires data it can consult the authentic source file. If the information is not available there, the government is allowed to obtain the data from the person involved²²⁰.

XII. CONCLUSIONS & RECOMMENDATIONS

The reader of this report should be convinced of the need to archive websites. It should also be clear that it is a complex matter with technical solutions that have not been finalised yet. Archiving static websites does not cause great problems. The latest websites however have a dynamic nature. The content of those websites can depend on factors such as the received request, the user profile or the user preferences, or the information that is available in the linked document management system or the database behind. An integrated and fully functional archiving of these websites is currently not possible. Instead the different components of the dynamic information system are being archived separately. A quality website archiving system is quite labour intensive and requires co-operation with the creator. The emulation strategy seems to be most appropriate to assure long-term readability of mirrors and snapshots. Website archiving clearly demonstrates the need to keep archiving in mind when creating the (future) archive documents.

Previous DAVID reports have concluded that the principle of secrecy of telecommunication and the privacy regulations in general do not take into account the needs of institutions that work on the preservation of our cultural heritage. It now also becomes clear that copyright laws do not escape these allegations. Archiving any copyright protected work causes practical difficulties, as reproduction rights are rigidly limited to the author of the work. The problem is increasingly present in 2002 as archiving digital works, and especially those spread via the Internet, requires reproduction from the very beginning. Our copyright laws were traditionally aimed at protecting our cultural heritage by limiting reproduction rights of copyright protected works to whoever has made the creative effort. These days however this limitation, caused by copyright, will lead to a decay of the (digital) cultural heritage. This report is an appeal to legislators to incorporate the needs of digital archive departments and institutions (and even museums or libraries) into the reforms of copyright laws that the Senate is currently developing in Belgium.

²²⁰ See Article 102 2° and 3° of the Programme Act

XIII. BIBLIOGRAPHY

A. ELECTRONIC RECORDKEEPING

A policy for keeping records of web-based activity in the Commonwealth Government, January 2001. (available on http://www.naa.gov.au/recordkeeping/er/web_records/intro.html)

C. AMMEN, *MINERVA: Mapping the INternet Electronic Resources Virtual Archive -Web Preservation at the Library of Congress*, Lecture held in Darmstadt on 8 Sept. 2001. (available on <http://www.bnf.fr/pages/infopro/dli%5Fecd12001.htm>)

W.Y. ARMS, *Web preservation project. Interim report*, 2001.

A. ARVIDSON, *Harvesting the Swedisch webspace*, Lecture held in Darmstadt on 8 Sept. 2001. (available on <http://www.bnf.fr/pages/infopro/dli%5Fecd12001.htm>)

A. ASCHENBRENNER, *Long-term preservation of digital material. Building an archive to preserve digital cultural heritage from the Internet*, Dissertation Technische Universiteit Wenen, 2001.

S. BORDWELL, *Objective-Archival preservation of websites*, Lecture held in London, 25 April 2002.

M. BURNER en B. KAHLE, *WWW Archive File Format Specification*, (available on <http://www.alexa.com/company/arcformat.html>)

W. CATHRO, C. WEBB en J. WHITING, *Archiving the web: the pandora archive at the National Library of Australia*, Lezing gehouden tijdens de conferentie *Preserving the Present for the Future Web Archiving*, Kopenhagen, 18-19 June 2001. (available on <http://www.nla.gov.au/nla/staffpaper/2001/cathro3.html>).

B. CHRISTENSEN-DALSGAARD, *Archive Experience, not Data*, Lecture held in Darmstadt, 8 Sept. 2001. (available on <http://www.bnf.fr/pages/infopro/dli%5Fecd12001.htm>)

C. DOLLAR, *Archival preservation of Smithsonian web resources: strategies, principles, and best practices*, (available on <http://www.si.edu/archives/archives/dollar%20report.html>)

I. ENGHOLM, *Digital design history and the registration of web development*, (available on <http://www.deflink.dk/eng/arkiv/dokumenter2.asp?id=695>).

IM FORUM, *Government of Canada Internet Guide*, Ottawa, 1999. (available on http://www.cio-dpi.gc.ca/ig-gi/index_e.asp)

Guidelines for keeping records of web-based activity in the Commonwealth Government, March 2001. (available on http://www.naa.gov.au/recordkeeping/er/web_records/intro.html)

J. HAKALA, *Collecting and Preserving the Web: Developing and Testing the NEDLIB Harvester*, in *RLG-DigiNews*, April 15, 2001, vol.5, nr. 2.

J. HAKALA, *Harvesting the Finnish Web space - practical experiences*, Lecture held in Darmstadt on 8 Sept. 2001. (available on <http://www.bnf.fr/pages/infopro/dli%5Fecd12001.htm>)

J. HONEYCUTT (et.al.), *Het complete handboek Internet*, Schoonhoven, 1997.

B. KAHLE, *Archiving the Internet*, in *Scientific American*, 11 April 1996.

A.R. KENNEY en O.Y. RIEGER, *The National Library of Australia's Digital Preservation Agenda, an interview with C. Webb*, in *RLG-DigiNews*, 15 Febr. 2001

D. LÉGER, *Legal Deposit and the Internet: Reconciling Two Worlds*, Lecture held in Darmstadt, 8 Sept. 2001. (available on <http://www.bnf.fr/pages/infopro/dli%5Fecd12001.htm>)

Managing Internet and Intranet Information for Long-term Access and Accountability. Implementation Guide, 1999.

Managing web resource. Management of electronic records on websites and intranets: an ERM toolkit. Kew, 2001.

J. MASÉNAS, *The BnF-project for web-archiving*, Lecture held in Darmstadt on 8 Sept. 2001. (available on <http://www.bnf.fr/pages/infopro/dli%5Fecd12001.htm>)

S. MCKEMMISH en G. ACLAND, *Accessing essential evidence on the web: towards an Australian recordkeeping metadata standard*, (available on <http://ausweb.scu.edu.au/aw99/papers/mckemmish/paper.html>)

C.J.M. MOSCHOVITIS, *History of the Internet: a chronology, 1843 to the Present*, Santa-Barbara (California), 1999

K. PERSSON, *The Kulturarw3 Project - The Swedish Royal Web Archiw³*, Lecture held in Svetlogorsk, Aug. 2000. (available on <http://kulturarw3.kb.se>) .

J. QUAST, *Het Internetarchieff van het IISG*, in *Nederlands Archievenblad*, September 2000, p. 16-17.

A. RAUBER et.al., *Austrian on-line archive. Current status and next steps*, Lecture held in Darmstadt on 8 Sept. 2001. (available on <http://www.bnf.fr/pages/infopro/dli%5Fecd12001.htm>)

D. SHENK, *The world wide library*, in *Hotwired*, 2 Sept. 1997 (available on http://hotwired.lycos.com/synapse/feature/97/35/shenk1a_text.html).

G. VOERMAN (et.al.), *Het belang van het archiveren van websites*, in *Information Professional*, 2001, p. 16-19

WORLD WIDE WEB CONSORTIUM, *Web Content Accessibility Guidelines 1.0*, 1999. (available on <http://www.w3.org/WAI/>)

B. LEGISLATION AND RULES

Books

BAINBRIDGE, D., *Introduction to Computer Law*, Gosport, Ashford Colour Press, 2000, 349 p.

BUYDENS, M., *Auteursrechten en Internet – problemen en oplossingen voor het creëren van een online databank met beelden en/of tekst*, Brussels, Federale diensten voor wetenschappelijk, technische en culturele aangelegenheden, 1998, 104 p.

CLAPPAERT, W., ‘Auteursrecht en Internet’, in *Telecom en Internet. Recht in beweging*, DE POORTER, B. (ed.), Gent, Mys en Breesch, 1999, 357-376.

CORBET, J., *Auteursrecht*, Algemene Praktische Rechtsverzameling, Antwerp, Kluwer, 1997, 173 p.

DE COCK BUNING, M., *Auteursrecht en Informatietechnologie – Over de beperkte houdbaarheid van technologiespecifieke regelgeving*, Amsterdam, Otto Cramwinkel, 1998, 281 p.

GOTZEN, F., *Auteursrecht, tekeningen en modellen*, KUL Faculteit Rechtsgeleerdheid, 1998, 249 p.

HUGENHOLTZ, P. (ed.), *The future of copyright in a digital environment*, Antwerp, Kluwer, 1996, 248 p.

SPOOR, J., *Scripta Manent: de reproductie in het auteursrecht*, Groningen, Tjeenk Willink, 1976, IV, 140 p.

VANHEES, H., *Auteursrecht in een notendop*, Leuven, Garant, 1998, 137 p.

VOORHOOF, D., ‘Multimedia en auteursrecht. Afschermen en beschermen van informatie. Juridische problemen rond de beschikbaarheid en de reproductie van informatie op één drager’, in *Multimedia, Interactiviteit, Kennisspreiding*, Minutes book, Symposium of the Book Foundation, 14 and 15 November 1995, LUC Diepenbeek.

WESTERBRINK, B.N., *Juridische aspecten van het Internet*, Amsterdam, Otto Cramwinkel, 1996, 195 p.

Articles

ALBERDINGK THIJM, C., 'Fair use: het auteursrechtelijk evenwicht hersteld', http://www.ivir.nl/publicaties/overig/alberdingk_thijm/fair-use.doc

DECKER, U., 'Electronic Archives and the Press: Copyright Problems of Mass media in the Digital Age', *E.I.P.R. issue 7*, Sweet & Maxwell Limited, 1998, 256-265.

GIJSSELS, J., 'De overheidsaansprakelijkheid in verband met informatie', *Rechtskundig Weekblad*, 1979-1980, 1202-1222.

GOESAERT, J., 'E-Government: Do New Technologies Build a Bridge between Government and Citizen?', in *A Decade of Research @ the Crossroads of Law and ICT*, DUMORTIER, J. (ed.), Gent, Larcier, 2001, 87-100.

HUGENHOLTZ, P., *DIPPER Digital Intellectual Property Practice Economic Report – Copyright aspects of caching*, Institute for Information Law, University of Amsterdam, September 1999, 46 p.

HUGENHOLTZ, P., 'De databankrichtlijn eindelijk aanvaard: een zeer kritisch commentaar', *Computerrecht*, 1996/4,

HUGENHOLTZ, P., *Recht en Internet*, Nederlands juristenblad, Zwolle, Tjeenk Willink, 1998.

LANGE, J., 'De wet en elektronische publikaties',
Beschikbaar op: <http://nieuws.surfnet.nl/nieuws/snn-archieff/achtergrond/jg97-98/kbsymposium2.html>

PRINS, J.E.J., 'Digitale duurzaamheid : een verloren geschiedenis' Available on :
<http://infolab.kub.nl/till/data/topic/digiduur.html>

SPOOR, J., 'The copyright approach to copying on Internet: (over)stretching the reproduction right?', in *The future of copyright in a digital environment*, HUGENHOLTZ, P. (ed.), Antwerp, Kluwer, 1996, 248 p.

STEYAERT, J., 'Het Internet in Vlaanderen: zijn steden, zijn gemeenten, zijn inwoners...', in *Digitale steden en gemeenten. Handleiding*, GOUBIN, E. (ed.), 1999, Politeia, Brussels, 23-36.