
Guideline for Archiving Government Websites

Richtlijn archiveren overheidswebsites

This English translation of the [Richtlijnen archiveren overheidswebsites](#) is for information purposes only and has no legal status. The Dutch version prevails.

Colophon

The original Dutch version can be found here:

<https://www.nationaalarchief.nl/archiveren/kennisbank/Richtlijn-Archiveren-Overheidswebsites>

Translated by Kees Teszelszky with machine help.

Disclaimer: This English translation of the [Richtlijnen archiveren overheidswebsites](#) is for information purposes only and has no legal status. The Dutch version prevails.

Contact:

Telephone: +31-70-331 5460

E-mail: contact@nationaalarchief.nl

P.O. Box 90520

2509 LM The Hague

Contents

1. Introduction
2. About this document
3. Why website archiving?
4. Requirements for website archiving

Appendices:

5. Appendix: Step-by-step plan for implementing web archiving
 6. Appendix: Terminology
 7. Appendix: Motivation of the choices made
 8. Appendix: Background and assignment of this standard
-

1 Introduction

Websites are one of the most important forms of communication used by government organizations with citizens, businesses, institutions, and fellow governments. Therefore, it is also important to be able to retrieve which information was present on a website at a specific moment in time. This does not happen automatically. When a web page is changed or removed, the old version is often no longer available for consultation.

Special measures are required to keep previous versions of web pages accessible. This is referred to as website archiving.

Because websites change continuously, web archiving is an ongoing process. It must be arranged as soon as possible after a website becomes active. It is too late to take measures only at the moment when an organization no longer maintains, removes, or replaces a website. In that case, past changes are difficult or impossible to reconstruct.

Website archiving is best taken into account already at the moment when a new website is developed. At that point, there is financial and technical room to arrange it properly.

For whom is this document intended?

But how do you archive a website? Where do you start and which components do you preserve? Who must do this and how often? [This document](#) provides answers to those questions. It is intended for everyone who, as a manager, project leader, or advisor, is involved in the management and archiving of websites.

Recognized working method

[This document](#) is a standard. This means that it describes a recognized working method. This standard applies to the public parts of existing and new government websites.

Each government organization is responsible for archiving its own websites. Applying this standard is voluntary but not without obligation. If a government organization does not apply this standard, it must comply in another way with the obligation under the Archiving Act 1995 to archive its websites.

Structure of this standard

[This standard](#) is composed of several parts and several appendices:

- **Questions or suggestions?**
Applying this standard may raise questions in practice, cause problems, and lead to new insights. With these practical experiences, the National Archives aims to improve and expand the standard. Organizations that have questions or suggestions for improvements or wish to contribute practical experiences may send an e-mail to contact@nationaalarchief.nl. The National Archives will update this standard if practical experience provides reason to do so.
- **About this document**
Explanation of the nature, status, scope, objective, and target audience.
- **Why website archiving?**
Explanation of why archiving websites is important and which legal obligations apply.

- **Requirements for website archiving**

This is the core of this document. It consists of a list of requirements, with explanations, on the basis of which a government organization can set up the archiving of its websites.

For readability, the requirements do not contain practical instructions or explanations of terminology and choices made. These are included in separate appendices:

- **Appendix: Step-by-step plan for implementing web archiving**
A practical guide for implementing web archiving within a government organization.
 - **Appendix: Terminology**
Definitions with explanations of the most important terms used in this standard.
 - **Appendix: Motivation of the choices made**
Explanation of the choices made when drafting the requirements.
 - **Appendix: Background and assignment of this standard**
Explanation of how this document came into being.
-

2 About this document

2.1 Standard

The [original Dutch version of this document](#) is a standard. A standard describes a recognized working method for a particular subject and scope of application. A standard is developed and maintained through consensus by a recognized institution. A standard is intended for common use and aimed at uniform application by users.

Applying a standard results in savings because users do not need to develop their own working methods, can collaborate more easily, and can make use of products and services that are based on the standard.

2.2 Subject

This standard concerns the archiving of websites.

A website is a collection of related web pages. A web page is online information that is available via a URL and can be displayed as a unit using an application such as a web browser or media player. This includes HTML pages, as well as other information available via a URL, such as text documents, images, videos, and audio fragments.

Archiving is the total set of activities required to make information (in this case websites) sustainably accessible from the moment of creation and for as long as necessary. In this way, stakeholders can consult the information as it was accessible at some point in the past. Website archiving is therefore more than merely storing web pages.

This standard applies only to archiving websites as they are published online. It does not include requirements for the websites themselves, for archiving information received via a website, or for archiving information related to the website.

2.3 Scope

This standard applies to government websites that are online or will become online. Of these, only the public and non-personalized web pages are included. For websites that are no longer online, this standard applies insofar as this is still technically possible.

A government website is a website that is managed by or on behalf of a government organization. A government organization is defined here as a government body as referred to in the Archiving Act 1995. This includes all public-law organizations within central and decentralized government.

These government bodies also include private-law organizations and natural persons insofar as they perform a public task. This standard applies to them only if a website arises entirely from that public task.

Public means that a web page is accessible via the internet, not via a restricted intranet. The visitor does not need to log in to view it.

Non-personalized means that the content of a web page does not depend on data about the user requesting the web page, for example data stored in cookies or in a database associated with the website.

2.4 Status

Applying this standard is voluntary but not without obligation. Applying the standard is a way to comply with archival legislation requirements for websites.

Archival legislation largely leaves the method of archiving to government organizations themselves. They must therefore make their own choices regarding how to do this. For information that must be permanently preserved (including public websites), Article 16 of the Archiving Regulation requires compliance with verifiable requirements. This standard is intended as an interpretation of those verifiable requirements. If a government organization does not follow the standard, it must itself determine the verifiable requirements.

Website archives of the central government that comply with this standard are accepted by the National Archives as archival repositories for transfer. Other archival repositories may use this standard as an acceptance condition for transfer but are not required to do so.

2.5 Archiving Act

The obligation to archive websites arises from the [Archiving Act 1995](#). Article 3 of the Archiving Act 1995 obliges government organizations “to bring the records under their control into a good, orderly, and accessible condition and to preserve them, as well as to ensure the destruction of records eligible for destruction.”

The law does not explicitly mention websites. However, records are understood to mean all information that a government organization receives or creates in the performance of its tasks, regardless of the form of that information. Websites therefore also fall under this definition.

The government organization whose tasks give rise to government information is responsible for archiving that information. For websites, this is the content owner. This is the government organization that is ultimately responsible for the content of the website. Archiving may be outsourced, but the content owner remains responsible.

Some government organizations collaborate on a single website. In that case, there are multiple content owners. Each is responsible for archiving its own share of the content. It is impractical to do this separately. It is therefore recommended to jointly determine who will take responsibility for archiving. Often there is a coordinating party that is well suited to this role.

2.6 Objective and target group

The objective of this standard is that government organizations have the knowledge to archive their websites and also apply that knowledge. As a result, the content of government websites is and remains accessible to everyone who has an interest in it.

The standard is intended for government organizations that archive (or have archived) their own websites. It is not intended for archival and heritage institutions that archive websites of others as part of their collection development.

This standard is intended for the following persons within a government organization:

- Managers responsible for websites or archiving, such as CIOs, heads of records management, or heads of communications. In mutual consultation, they must issue the assignment for archiving websites. By requiring that archiving complies with this standard, they can specify that assignment.
- Project leaders who must carry out the assignment to archive websites. The step-by-step plan in this standard can serve as the basis for an action plan.
- Advisors who translate assignments for archiving websites into concrete plans, such as information managers and administrators, functional administrators, architects, (procurement) advisors, and digital archiving advisors. The requirements in this standard can serve as the basis for specifications, for example in a procurement document or in the design of an in-house solution.

The so-called care provider is administratively responsible for archiving within a government organization. Usually this is the executive or daily governing body. For a ministry, for example, this is the minister; for a municipality, it is the mayor and aldermen.

The care provider will not deal directly with the standard. In practice, one or more managers are responsible. Who this is may differ per organization.

2.7 Development

The National Archives developed this standard as part of its role as a knowledge centre in the field of archiving. On behalf of the Ministry of Education, Culture and Science (OCW), the National Archives (NA) develops standards to support government organizations in implementing archival legislation.

This standard was developed in close consultation with stakeholders and experts. A public review of a draft version was conducted, allowing all interested parties to indicate ambiguities and propose improvements.

After processing all review comments, this standard was adopted by the Standardization Council of the National Archives. The Standardization Council includes representatives from all levels of government and various archival institutions. The council ensures that the standards of the National Archives are substantively acceptable and practically usable for all government organizations.

3 Why website archiving?

Government organizations use websites to provide information to citizens, businesses, institutions, and other governments. That information can change over time or be removed from the site. For stakeholders, it can be important that the information remains accessible. The interests involved are very diverse:

- **Execution of business processes**

In carrying out tasks, it may be necessary for government organizations to consult old web pages. For example, to answer questions from citizens or to draw up reports. The content of old websites can also sometimes be reused. Many civil servants use websites as reference material in their daily work.

- **Accountability for government action**

Government organizations must be able to account for their actions. Websites are an important record of those actions. The original content of websites must therefore be accessible at the moment when actions are questioned. This may be necessary for handling requests under the Freedom of Information Act or during a debate. It may also be required in parliamentary or council investigations, audits, or audits by audit offices.

- **Citizens seeking rights and evidence**

Citizens and businesses may derive rights from information on websites. For example, when applying for services, permits, subsidies, or starting a business. Citizens and businesses base decisions and actions on government information. That information may therefore be required as evidence in appeal and objection procedures. If a government organization loses a case because it does not possess this evidence, this can lead to high costs and loss of credibility.

- **Research**

Government websites give researchers insight into how government organizations carried out their tasks and presented themselves. For example, when research is conducted into the handling of a major crisis, how certain decisions were made, or the relationship between government and citizens.

- **Heritage**

Websites are also an expression of our culture. Not only the content, but also, for example, the design and language use. When a website is still actively used, this may be difficult to imagine. But one day websites will evoke the same feeling as the Polygoon newsreels and discolored photographs from the 1970s. We would not want to miss those now, because they remind us of the past and help define how we see ourselves.

Because of this combination of interests, government organizations must archive their websites.

4 Requirements for website archiving

Government organizations that archive their public websites in accordance with this standard must meet the following requirements.

4.1 Daily harvesting

Requirement: Each website is harvested daily.

Once per year, a full harvest is carried out. Between these, daily incremental harvesting takes place of only the new and modified web pages.

In the case of a major change to a website, a full harvest is performed immediately before and immediately after this change. A major change is a change in which more than half of the web pages change.

Explanation:

Harvesting is the basis of web archiving. It captures a website so that it can be consulted later, even after it has changed.

Daily harvesting ensures that most changes are archived. It may occur that multiple changes take place between two successive harvests. In that case, only the last change is recorded.

With full harvesting, all web pages of a website are preserved. With incremental harvesting, only new and modified web pages are preserved. This keeps the impact of incremental harvesting low for websites that change little.

Repeating a full harvest annually or in the case of major changes provides additional certainty in case an earlier archive has been lost.

4.2 Complete harvesting

Requirement:

For each website, all web pages that fall within the website are fully harvested. The exception is content that is included in the list of non-harvestable content (see requirement 4.3).

For the harvesting scope, the following applies:

- All embedded content (embedded web resources) is harvested.
- No limitations on the number of hops.
- No limitations on path length.
- No “scope+1”. This means that no web page outside the website is harvested.
- Web pages for which the harvester cannot determine whether they belong to the site are included in the list of orphan pages.
- All web pages in the blacklist are documented in the list of non-harvestable content.

Explanation:

All web pages contribute to the meaning of a website. Therefore, all are harvested, except web pages included in the list of non-harvestable content.

“All web pages” means the start page and all web pages that can be reached from it via hyperlinks, insofar as the web address starts with the site URL or one of the secondary site URLs.

4.3 Non-harvestable content

Requirement:

For web content (an entire web page or part thereof) that cannot or may not be harvested, a description is available.

The descriptions are recorded per website in an OASIS Open Document Format (ODF) spreadsheet (.ods).

Each description occupies one row in this table, with the following columns:

- **URL:** The URL of the web page whose content is not harvestable.
- **Component:** The part of the web page that is not harvestable, for example “entire web page” or “streaming video”.
- **Reason:** The reason why the content is not harvestable. This may be technical or legal.
- **Description:** A short description of the non-harvestable content, so that it is known afterwards what was not harvested.
- **Alternative representations:** Description and reference to archived alternative representations of the non-harvestable content (if available).

Explanation:

With harvesting, most web pages can be archived properly. However, there are situations in which web pages or parts thereof are not harvestable. Even though these are exceptions, they must be documented. This allows a future user of the archive to see which components are missing and why.

Examples of situations in which content is not harvestable include:

- Some interactive content determines which content is retrieved from a server based on user interaction. This cannot always be determined during harvesting.
- With streaming video (such as a YouTube video), the information retrieved from the video server is only determined during playback and depends, among other things, on how the video player is operated.
- A police wanted notice on a website may contain personal data that may not be preserved.

The list of non-harvestable content may not be used to exclude content from archiving at one’s own discretion. This is only permitted when it is technically or legally impossible to harvest the content.

Often an alternative representation of the non-harvestable content is available that can be archived properly, such as an .AVI file for a streaming video or a table with the underlying data for interactive content. If an alternative representation exists, it is archived and referenced in the description.

Archiving an alternative representation can be done, for example, by adding it as a file to the website (so that it is archived along with the website), or by including it in a document management system (DMS) of the organization. Inclusion on the website is preferred, because the file then becomes an integral part of the archived website. If this is not possible, archiving elsewhere is also acceptable, as long as the description clearly states where it can be found.

4.4 Restricted-availability web pages

Requirement:

For each website, there is an up-to-date list of archived web pages whose availability is restricted on the basis of laws and regulations.

The list of restricted-availability web pages is recorded in an OASIS Open Document Format (ODF) spreadsheet (.ods). This file is included with the web archive files (see requirement 4.6).

Each restriction occupies one row in the table, with the following columns:

- **URL:** The URL of the web page to which the restriction applies.
- **Component:** The part of the web page to which the restriction applies, for example “entire web page” or “wanted notices”.
- **Reason:** The reason why a restriction applies, including a reference to the legal basis and any documentation of the decision to restrict availability.
- **Access:** Description of how and under which conditions access to the web page is possible.
- **Start date:** Start of the period during which the web page is restricted (if applicable).
- **End date:** End of the period during which the web page is restricted (if applicable).

Explanation:

This standard concerns public websites. Therefore, archived web pages may in principle be made available to everyone and in every way. Nevertheless, there may be reasons to restrict the availability of certain web pages. Rules apply to restricting access to information, such as the General Data Protection Regulation (GDPR) and copyright law. These rules must also be respected for archives.

The list of restricted-availability web pages makes this possible. For example, in the case of personal data that was accidentally published or image material subject to copyright.

Archived web pages whose availability is restricted for privacy reasons are not destroyed. Only availability is restricted for as long as the privacy considerations apply.

All website archives that are restricted or become restricted within the meaning of the Archiving Act 1995 are also restricted in availability. The reverse is not necessarily true. A restriction that applies due to copyright, for example, has nothing to do with public access. Moreover, the concept of “restricted public access” applies only to transferred archives. Archives that have not yet been transferred may also have restricted availability.

4.5 Web archive files

Requirement:

The results of a harvest are recorded in one or more files (hereinafter referred to as “web archive files”) that meet the following requirements.

4.5.1 WARC standard**Requirement:**

Each web archive file complies with the WARC standard NEN-ISO 28500, using a version that is valid no more than three years prior to the moment of harvesting.

Existing web archive files are not converted to a new version when a new version of the WARC standard appears.

Explanation:

The WARC format is the international standard for web archive files. By using this standard, it is clear what the contents of a web archive file mean, and tools are available to process the files.

Existing web archive files are not converted in order to prevent loss of information.\

4.5.2 Size**Requirement:**

Each web archive file is a maximum of 1 GB in size.

If the total web archive exceeds 1 GB, it must be split into web archive files that are each no larger than 1 GB. The file names must make clear which web archive files together form a single web archive. Only when an individual WARC record is itself already larger than 1 GB may it be stored in its own web archive file that exceeds 1 GB.

Explanation:

This follows the recommendations on file size in Annex C of the ISO 28500:2017 standard.

4.5.3 File names**Requirement:**

File names comply with the following file name pattern:

<name or code of archive creator>-<date and time start harvest>-<sequence number>-<domain name of seed>.warc

Explanation:

Uniform naming helps with recognizing and managing web archive files. The format used here follows the recommendations on file naming in the WARC standard.

An example:

The first WARC of a web archive of the GahetNA website of the National Archives that started on 31 October 2017 at 10:15 is called:

NL-HaNA-20171031101500-001-www.gahetna.nl.warc

4.5.4 Compression**Requirement:**

A web archive file is not compressed as a whole.

Explanation:

This prevents a web archive file from becoming completely unreadable in the future if only part of the file becomes corrupted.

Compression per record within a web archive file is permitted.

4.5.5 Checksum**Requirement:**

A checksum is calculated and recorded for each web archive file immediately after creation.

Periodically, and prior to transfer, all web archive files to be transferred are validated by checking the checksums. If a corrupted file is found, the file is restored (if still possible, for example from a backup).

Explanation:

A checksum makes it possible at any later time to determine that a file has not been modified since its creation. This increases confidence in the correctness of the content.

The earlier in the preservation process a checksum is added, the better the reliability of the web archive files can be verified.

At transfer, the checksum is included as part of the metadata (element 21.7).

4.5.6 All requests and responses**Requirement:**

The web archive files contain, unchanged and in full, all requests sent to the web server and all responses received from the web server during harvesting. No filtering is applied and no parts of the web pages are rewritten. URL rewriting is also not permitted. Error messages, redirects, and server errors are also all included in the web archive file.

Explanation:

To be able to determine the authenticity and reliability of the information, full insight into the communication between the web server and the harvesting service during the harvesting process is required.

This makes it possible, for example, to record that a web page was not harvested because a server error occurred or because the server was offline for maintenance.

4.5.7 Mandatory fields**Requirement:**

In addition to the fields already mandatory under the WARC standard, each web archive file contains at least the following fields (as described in ISO 28500):

- **In the WARC info block:**
 - Software
 - Format

- Robots
- **In the header of each WARC record:**
 - WARC-Block-Digest
 - Content-Type
 - WARC-WARCinfo-ID
- **In the request and response blocks:**
 - WARC-Target-URI

Explanation:

These additional fields help protect the authenticity and relationships between WARC files.

It is permitted to use additional fields beyond those required by the WARC standard.

4.6 Folders

Requirement:

The combined web archive files and associated lists are grouped in the following hierarchical folder structure:

1. **Organization:**
All websites of a single organization. Within this, one folder per website.
2. **Website:**
All files of one website of the organization. Within this, one folder per period.
3. **Period:**
All files of the website for a period beginning with a full harvest and all subsequent incremental harvests. This includes the list of restricted-availability web pages (see 4.4) as of the last day of that period. Within this, one folder per harvest.
4. **Harvest:**
All web archive files created during the harvest on a single day. This includes the list of non-harvestable content (see 4.3) for that day.

Explanation:

This folder structure is necessary for clear findability of the files. In addition, it makes it possible to assign metadata to web archive files per folder.

4.7 Metadata

Requirement:

Metadata is recorded for each web archive file and the higher-level folders in accordance with the “Application Profile Metadata Central Government” or the “Application Profile Metadata Local Governments”.

In addition to the application profiles, the metadata meets the following requirements.

4.7.1 Elements

Requirement:

For each folder or web archive file, at least the following metadata elements are recorded.

No.	Element name	Map organisation	Map website	Map period	Map harvest	Web archive file
2	Identification mark	x	x	x	x	x
3	Aggregation level	x	x	x	x	x
4	Name	x	x	x	x	x
5	Classification		x			
6	Description		x		x	
9.1	Coverage in time		x	x	x	
11	Language		x			
12	Event history		x	x	x	x
15	Relation	x	x	x		
19	Form		x			x
20	Integrity		x	x	x	
21	Format					x

Explanation:

By including this contextual information in the metadata, web archive files can be properly managed and made accessible. In addition, it helps future users to search, interpret, and assess the reliability of the website archives.

The numbering and element names are based on the “Guideline Metadata Government Information: Description of Entities and Elements”. Both application profiles are based on this guideline.

The list of non-harvestable content (requirement 4.3) and the list of restricted-availability web pages (see 4.4) are also metadata. These metadata do not fit well into the application profiles. Therefore, it was chosen to include these lists separately as tables.

4.7.2 Values

Requirement:

For the values of the elements, the following applies:

No.	Element name	At map of file	Value
3	Aggregation level	organisation	Fixed value 'serie'.
3	Aggregation level	website; period	Fixed value 'dossier'.
3	Aggregation level	harvest; web archive file	Fixed value 'archiefbestanddeel'.
5	Classification	website	The classification as it applies within the organisation for the website in question.
6	Description	website	Brief description of the website. Including: • Site URL • Homepage • Secondary site URLs
6	Description	harvest	Description of the software with parameters used to create the harvest.
9.1	Coverage in time	website	Period during which the website has been online.
9.1	Coverage in time	period	Period during which the underlying harvests took place.
9.1	Coverage in time	harvest	Date of harvest.
11	Language	website	Language of the website.
15	Relation	organisation	<u>Content owner</u> .
15	Relation	website	Organisational unit responsible for the content of this website.
15	Relation	harvest	Organisatie die de harvesting in deze periode heeft uitgevoerd.
19	Form	website	Fixed value 'websitearchief'.
19	Form	Webarchive file	Fixed value 'webarchiefbestand'.

Explanation:

These restrictions are intended for uniform interpretation of the fields that are specific to websites.

4.7.3 Delivery of metadata

Requirement:

When transferring web archive files, the metadata is delivered in the technical format determined by the relevant archival repository.

Explanation:

There is currently no national standard for the technical format in which metadata is exchanged. Each archival repository determines this itself. The National Archives uses the XML format ToPX for metadata delivery.

It is recommended to record metadata immediately during harvesting in the format prescribed for transfer. This prevents additional effort and possible errors when exchanging metadata.

4.8 Quality control

Requirement:

It is periodically checked whether website archiving complies with the requirements of this standard. Users of the web archive can report errors online.

If, through checks or reports, a deviation from the requirements is identified, the cause is investigated and measures are taken to prevent recurrence. It is investigated which web pages have not been included, have been included incorrectly, or have been included incompletely in the already created web archive files. This is documented in the metadata of those web archive files (field 20 “integrity”). The already created web archive files themselves are not modified.

Deviations, their causes, and the measures taken are reported to the manager who is responsible for website archiving at the content owner. This manager determines whether sufficient measures have been taken.

Explanation:

A website archive that does not meet the requirements may be partially or entirely unusable. It may be substantively incomplete or incorrect, technically unmanageable, or unreadable. Errors are often difficult or impossible to repair afterwards. Therefore, it is important to detect errors in time.

How periodic quality control is carried out is not specified in this standard, because no best practices are yet known for this. This means that each government organization can determine for itself how it carries out quality control, provided it complies with generally accepted quality management standards, such as those laid down in ISO 9001.

4.9 Security

Requirement:

The web archive files and associated metadata are stored in an information system that is secured to prevent loss or alteration, in accordance with the security standard NEN-ISO/IEC 27002 and the implementation of this standard as applicable to the organization.

Explanation:

This guarantees that archived websites are exactly the same at the moment of harvesting as at the moment when someone consults them in the future.

Government organizations apply different implementations of NEN-ISO/IEC 27002. For the central government, this is, for example, the Baseline Information Security Central Government (BIR). For other governments, this is expected to be the Baseline Information Security Government (BIO) as of 1 January 2019.

4.10 Publication right

Requirement:

The content owner permanently retains the right to archive its web pages and to publish the archived web pages online.

If content is procured or management activities are outsourced, this right is contractually established.

Explanation:

This guarantees that the government organization itself can determine how and to whom it grants access to archived websites.

This requirement is primarily a point of attention with regard to copyright of procured content, such as commissioned texts, images, and stock photos. Contractual use of this content must not be limited to a specific period or server. It must also be impossible, for example, for the supplier of a streaming video service to contractually restrict publication rights to the videos.

4.11 Access

Requirement:

The harvested web pages are accessible online to everyone immediately after harvesting. At a minimum, the following functions are available for all web browsers:

- It is clearly indicated that the website is an archived version.
- When clicking a link to a website that has not been archived, a clear notification is given that the web archive is being left.
- The ability to request, for each web page in the archive, the version from any date in the past, for example using a date field or a clickable calendar.
- Full-text search across all web pages in the web archive files.
- When requesting a web page that is included in the list of non-harvestable content (see requirement 4.3):
 - a clear notification is given that the web page is wholly or partially unavailable in the website archive;
 - including the information from the list for the relevant page;
 - including a link to the partial representation as available in the website archive (only if the web page is partially non-harvestable);
 - including a link to the current version of the web page, outside the website archive (if it still exists).
- The ability to view the metadata recorded for the web archive (see requirement 4.7).

For archived web pages on the list of restricted-availability web pages, the following applies (see requirement 4.4):

- These are not findable or retrievable online. Where technically possible, this restriction applies only to the part for which the restriction applies.

- A clear notification is provided that the web page, wholly or partially, is not available online, including an explanation of why not and how access may still be obtained if applicable.

Explanation:

The purpose of this requirement is that everyone can easily view archived websites, insofar as this is possible and permitted.

Prior to transfer of web archive files, the content owner is responsible for online access. After transfer, the archival repository is responsible.

The online publication of a harvested website is itself not harvested. This would result in duplication.

4.12 Permanent preservation

Requirement:

The web archive files are preserved permanently. This applies to all different components and all harvested versions of the website.

If a government organization has not included permanent preservation of its websites in its retention schedule, this must be discussed with the National Archives at the next update of the retention schedule. Until that time, the web archive files must not be destroyed.

Explanation:

Public websites are a form of official government communication. They are an important source of information about government actions, in particular about the relationship between government and citizens and businesses. Websites also show how government presents itself to society over time. They therefore also have cultural value.

For this reason, websites are preserved permanently. It may be that not all components of a website are equally important in the long term. However, because it is difficult to make a well-substantiated distinction, the entire website is preserved permanently (see also 7.5).

4.13 Transfer

Requirement:

The web archive files, together with the associated lists and metadata, are transferred to an archival repository no later than twenty years after creation, in accordance with the folder structure of requirement 4.6.

The government organization and the relevant archival repository jointly determine after which period and how often transfer takes place.

In addition to this standard, each archival repository sets general requirements for transferring an archive. Contact the relevant archival repository for this.

Explanation:

According to Article 12 of the [Archiving Act 1995](#), government organizations must transfer archives that must be permanently preserved to an archival repository. Such repositories

specialize in permanent preservation of archives, ensuring long-term accessibility of web archives.

General transfer requirements include, for example, procedures and methods for delivering files and metadata. These requirements are not specific to web archive files and therefore fall outside the scope of this standard.

5 Appendix: Step-by-step plan for implementing web archiving

Government organizations that archive their websites in accordance with this standard can do so using the following step-by-step plan. It is intended as a practical aid, for example as a basis for drawing up a project plan.

5.1 Initiative

Action:

Bring web archiving to the attention of the managers within the organization who are responsible for websites and for archiving. Ensure that they recognize the importance of web archiving and subsequently, as commissioning parties, take the initiative to implement web archiving and supervise it.

Objective:

Willingness of management to implement web archiving. This is indispensable if costs are incurred.

Stakeholders:

Commissioners of websites (typical organizational roles for this include CIO, head of communication, head of information management), careholder (or representative thereof), records manager.

5.2 Alone or together

Action:

Determine at what scale web archiving will be implemented: per website, per organizational unit, for the entire organization, or together with other government organizations. If cooperation is chosen, go through the following steps jointly.

If there are multiple commissioners, agree in advance how decisions will be made jointly. For example, by embedding this in an existing program, project, or consultation body. When web archiving is carried out jointly, each individual organization remains responsible for archiving its own websites.

Objective:

Because harvesting is arranged independently of the CMS, it is often possible to organize web archiving jointly for multiple websites. This can yield economies of scale. For example, pooling knowledge and jointly procuring services can reduce costs.

Stakeholders:

Content owner, content manager, functional CMS manager.

5.3 Project leader

Action:

The involved managers appoint a project leader for implementing web archiving.

Objective:

Someone is made available as a point of contact, for carrying out (or commissioning) the required follow-up actions and for embedding web archiving in the organization after completion of the project.

Stakeholders:

Commissioners of web archiving (see “initiative”).

5.4 Informing

Action:

Determine who plays a role in web archiving. Inform these stakeholders about the why, what, and how of web archiving. For example, through a presentation and discussion of the web archiving standard.

Address questions and make choices where necessary.

Objective:

Stakeholders are prepared for the possible consequences of archiving, for example for activities, applications, role distribution, finances, and contracts.

Stakeholders:

Project leader web archiving, content owner, content manager, functional CMS manager, harvesting service provider (if already known), records manager, archival repository.

5.5 Assign roles

Action:

Determine who assumes which role in web archiving. Depending on how a government organization organizes archiving, roles may be distributed differently.

Objective:

It is clear to everyone what is expected of them. A typical role distribution is:

- **Content owner:**
Ultimately responsible for archiving the websites. Takes the initiative for archiving and checks whether it is done properly.
- **Content manager:**
Provides substantive and procedural information about the website (see “Website inventory”).
- **Functional CMS manager:**
Provides technical information about the website. May also provide the harvesting service and access service or manage an external supplier for these.

- **Harvesting service provider:**
Carries out harvesting.
 - **Access service provider:**
Ensures that everyone has access to the harvested websites. In practice, the access service provider is often the same as the harvesting service provider.
 - **Archiving advisor:**
Advises on all general aspects of archiving and guides the transfer of web archive files to an archival repository.
 - **Archival repository manager:**
Manages transferred web archives and provides public access to them.
-

5.6 Plan and budget

Action:

Draw up an action plan. At a minimum, describe the objective, target group, expected time investment, and budget. If no budget has yet been determined, it is important first to clarify how much funding is available. Have the commissioners approve the plan.

Objective:

Make clear within the organization which activities are required and reserve the necessary resources so that website archiving can be carried out.

Stakeholders:

Careholder, content owner, content manager, functional CMS manager, harvesting service provider (if already known), records manager, archival repository.

5.7 Create website inventory

Action:

Identify which websites the government organization is content owner of and who the direct point of contact is for those websites. Do not forget websites for which (part of) the management has been outsourced. These must also be included.

Objective:

The inventory provides information to determine which websites must be archived and who plays a role in this. The harvesting service provider needs this information for archiving, and the commissioning party needs it to have the quality of the web archives checked.

Stakeholders:

Content owner, content manager, functional CMS manager, records manager, archival repository.

5.8 Manage website inventory

Action:

Make agreements about who updates the inventory when major changes occur to websites, for example when a new website is added or one is discontinued.

Ensure that it is known within the organization that new and modified websites must be reported to the manager of the inventory, so that it can be updated for archiving.

Objective:

The website inventory is always up to date, so that the harvesting service provider is aware of changes and can take appropriate measures.

Stakeholders:

Content owner, content manager, functional CMS manager, harvesting service provider.

5.9 Harvesting service

Action:

Make agreements about carrying out the harvesting of websites (the harvesting service). This guideline assumes that a contract is concluded with an external supplier for this purpose. Harvesting requires specialized expertise that is not standard within many government organizations but is available from commercial suppliers.

An organization may also set up its own harvesting service, but this often requires greater investment than outsourcing. The harvesting service provider must indicate in advance which requirements apply to websites in order to be harvested. Ensure that harvesting does not negatively affect website availability for users.

Record the agreements, especially how the created web archives are transferred to the commissioning party when the contract with the harvesting service provider ends (whether planned or unplanned). Also record the commissioning party's rights to the created archives so that it can continue to dispose of the web archives created by the external supplier. The ARBIT and CIBIT provide examples for drafting escrow arrangements that can be used for this.

This standard does not contain information on setting up a harvesting service. For this, consult the documentation and other information about harvesting software.

Objective:

The harvesting service provider creates the web archives, and clear agreements ensure a clear division of roles and requirements for the harvesting service and the websites.

Stakeholders:

Careholder, content owner, harvesting service provider.

5.10 Access service

Action:

Make agreements about making the harvested websites available to the public (the access service). In practice, the harvesting service provider often also provides the access service, but an organization may also set up its own service for this. The access service must indicate in advance which requirements apply to websites to be harvested in order to make them properly accessible.

If you publish archived websites on your own website, exclude these from harvesting. Otherwise, all previously harvested websites will also be included in a single harvest.

Objective:

Often the same supplier provides both the access service and the harvesting service, but this is not required. Clear agreements clarify the division of roles and the requirements for the access service and the websites. This standard does not include information on setting up an access service. For this, consult the documentation and other information about the relevant software.

Stakeholders:

Careholder, content owner, access service provider.

5.11 Repository

Action:

Determine where the web archive files are stored immediately after harvesting until the moment of transfer. This may be, for example, in a facility where the careholder also stores its other archives (often referred to as a DMS, RMA, or e-depot), or with the supplier of the harvesting or access service. Verify that this repository meets the same security requirements that the careholder applies to the storage of its other archives.

Objective:

Manage web archive files in accordance with the same security standards as the careholder's other digital archives.

Stakeholders:

Harvesting service provider, access service provider, records manager.

5.12 Restrictions on availability

Action:

Make agreements on managing the list of web pages for which availability is restricted:

- Who maintains the list of restrictions?
- Where and how can requests be submitted for inclusion on the list?
- Who determines whether and how the request is honored?
- How is the access service provider informed of changes to the list?

Objective:

Prevent unlawful disclosure of information from the web archive.

Stakeholders:

Content manager, security officer, privacy officer, data protection officer (DPO), access service provider.

5.13 Start harvesting

Action:

Register all websites from the inventory with the harvesting service provider so that harvesting can begin. Provide the harvested websites to the access service provider so that they are accessible.

Inform the functional managers of the websites in advance about the start of the annual harvesting of the entire website and about the daily harvesting of changes. This allows them, if necessary, to take measures to enable harvesting without negatively affecting the website.

Objective:

Preserve websites and make preserved sites accessible. Prevent websites from malfunctioning due to harvesting or harvesting from failing.

Stakeholders:

Content owner, content manager, functional CMS manager, harvesting service provider.

5.14 Reference to archived version on live website

Action:

Place references on the websites to the archived versions of the websites, for example near the search function.

Objective:

Users of the website can easily find archived web pages when searching for older versions of web pages.

Stakeholders:

Functional managers of websites.

5.15 Quality control

Action:

Have the quality of the web archive files checked regularly. The supplier must explain the cause of errors and take appropriate measures to prevent recurrence.

Objective:

Ensure that the web archive files meet quality requirements, without requiring costly repairs afterward or making recovery impossible.

Stakeholders:

Content owner, harvesting service provider, records manager.

5.16 Changes to websites

Action:

Include in website management procedures that the harvesting service provider is informed in time when a website starts, stops, or is no longer updated. This also applies to major changes that cause many web pages to change at once.

Stop harvesting when a website is discontinued or its content no longer changes. It may occur that a website is registered for harvesting even though it is no longer updated. In that case, only a one-time full harvest is required. Incremental harvesting then has no added value.

Objective:

Ensure that the harvesting service provider is prepared for major changes to websites and prevent unnecessary harvesting.

Stakeholders:

Content owner, content manager, functional CMS manager, harvesting service provider.

5.17 Adjust the retention schedule if necessary

Action:

Check whether the retention schedule clearly states that public websites are (or can be) permanently preserved and have it amended if this is not the case. See also requirement 4.12.

Objective:

Clarity regarding the retention period of websites.

Stakeholders:

Careholder, records manager, National Archives.

5.18 Transfer

Action:

Transfer the web archive files of years whose transfer period has been reached to the organization's archival repository.

The statutory transfer period is a maximum of twenty years. An archival repository may allow earlier transfer of web archive files but is not obliged to do so.

Which organization serves as the archival repository differs per government organization (see 6.1). The method of transfer may also differ per repository. Therefore, already during the setup of web archiving, make agreements with the archival repository on when, how often, and how the files will be transferred. Contact your organization's archiving advisor, who knows which repository applies and what the transfer arrangements are.

Objective:

Transfer guarantees long-term accessibility of website archives without dependence on the content owner.

Stakeholders:

Careholder, records manager, archival repository.

6 Appendix: Terminology

6.1 Archival repository

A repository designated by or pursuant to the Archiving Act 1995 for the permanent preservation of archival records.

Source: [Archiving Act 1995](#).

Explanation

An archival repository has the task of managing government information that must be preserved permanently. After a government organization has transferred its archival records (archived government information), the archival repository is responsible for the sustainable accessibility of the archival files.

There are different types of archival repositories:

- **National Archives**
Manages archives of central government organizations that operate nationally, such as ministries and independent administrative bodies (ZBOs), for example the Chambers of Commerce, Staatsbosbeheer, and the Social Insurance Bank. The National Archives is also the Regional Historical Center for the province of South Holland.
- **Regional Historical Center (RHC)**
Manages the archives of affiliated provinces, municipalities, and water authorities, as well as regionally operating central government organizations such as courts. RHCs are cooperative arrangements between the State, provinces, municipalities, and water authorities. RHCs are located in provincial capitals.
- **Municipal archival repository**
Some municipalities have their own archival repository. Other municipalities cooperate, for example by participating in an RHC or another form of cooperation.
- **Water authority archival repository**
Some water authorities have their own archival repository; others cooperate, for example by participating in an RHC or another cooperative arrangement.

6.2 Archiving

The totality of activities required to make and keep information sustainably accessible.

Source: Own definition. The Dutch [Archiving Act 1995](#) does not contain a definition of archiving. Therefore, an own definition is used that aligns with the provisions on archival records in the Archiving Act 1995.

Explanation

Accessible means findable, available, readable, interpretable, and reliable for those entitled to access, from the moment of creation and for as long as necessary. Sustainable means that the accessibility of the information is resistant to all types of change.

Government organizations are legally required to archive. Article 3 of the [Archiving Act 1995](#) states the following:

“Government organizations are obliged to bring the archival records under their control into a good, orderly, and accessible condition and to preserve them, as well as to ensure the destruction of archival records eligible for destruction.”

Archival records mean all information that a government organization has created or received in the performance of its tasks. The form of that information is irrelevant.

The duration of archiving is determined under the Archiving Act 1995 in retention schedules.

According to this definition, archiving already begins at the moment information is created and is independent of the place where the information is stored. In this respect, the definition based on the Archiving Act differs from the commonly used meaning that archiving refers to storing completed information in a specially designed archival facility.

6.3 Checksum

Data derived from a block of digital data for the purpose of detecting errors that occurred during transport or storage of the data block.

Source: [English Wikipedia.](#)

Explanation

A checksum is usually a short sequence of numbers and letters. The block of digital data is usually a file or part of a file (such as one record in a container file). The essence of a checksum is that it is unique to each data block but much smaller than the data block itself. It can be seen as a digital fingerprint of the data block.

A checksum is calculated using an algorithm at the time of creation of the data block or just before transmission over a network. There are various checksum algorithms, such as MD5, SHA1, SHA256, or SHA512.

The checksum can be recalculated at any later time using the algorithm, for example after receipt of the digital file. If the second checksum matches the first, it is virtually certain that the file has not been modified or damaged between the two calculations. If the checksums differ, this indicates that something has gone wrong with the file.

6.4 Content / web content

Information on a website that is delivered to the user by a user agent, including code or markup that defines the structure, presentation, and interactions of the content.

Source: [Web Content Accessibility Guidelines \(WCAG\) 2.0.](#)

Explanation

The content of a web page is the information that the user can perceive when consulting the web page using a user agent.

6.5 Content management (of a website)

Creating and maintaining web content.

Source: Own definition.

6.6 Content owner (of a government website)

The government organization from whose tasks the content of the website arises.

Source: Own definition.

Explanation

The content owner is responsible for the content and therefore also for archiving a government website. Both content management and website archiving may be outsourced. However, the content owner remains ultimately responsible.

6.7 Content Management System (CMS)

A software application that makes it possible for people to publish documents and data on the internet easily, without much technical knowledge (content management). Usually a web application.

Source: Dutch Wikipedia.

Explanation

A CMS is the software used to carry out content management.

6.8 Embedded content (embedded web resource)

A web resource that forms part of a web page.

Source: Derived from the definition of web page (see definition 6.18).

Explanation

Embedded content is part of a web page that is not included in the code of the web page itself, but refers to information on a server. Examples of embedded content include images on a web page and style sheets in which the fonts of the web page are defined.

See also the definition of web page.

6.9 Harvesting (of a website)

Retrieving all web pages of a website via HTTP for storage or analysis.

Source: Adapted from English Wikipedia.

Explanation

Special software (the harvester) retrieves web pages from the web server, including images, layout, and attachments. The harvesting scope determines which web pages are retrieved (see definition 6.11).

Harvesting is used, among other things, for web archiving and for creating search indexes for search engines.

6.10 Harvesting frequency

How often harvesting of a website takes place.

Source: Own definition.

Explanation

The harvesting frequency largely determines the completeness of harvesting. There are several possibilities:

6.10.1 One-time full harvesting

Harvesting takes place only once. Changes that occur afterward are not detected.

6.10.2 Periodic full harvesting

Harvesting is repeated at fixed intervals, for example daily or weekly. If the same item changes more frequently between two successive harvests, only the result of the last change is recorded.

6.10.3 Periodic incremental harvesting

In this variant, the harvester traverses the entire website but stores only those web pages that have changed since the previous (incremental) harvest.

The advantage of incremental harvesting is that storage and processing capacity are saved for web pages that have not changed.

6.10.4 Real-time harvesting

The CMS of the website sends a signal to the harvester when a web page is added, modified, or deleted. The harvester then retrieves the web page.

The advantage of real-time harvesting is that all changes to a website are detected. Moreover, storage and processing capacity are used only for web pages that change. The disadvantage of incremental harvesting is that the CMS must be configured to send such signals. With harvesting at fixed intervals, the harvester is not dependent on the CMS.

6.11 Harvesting scope

The delimitation of which web pages are retrieved during harvesting of a website and which are not.

Source: Own definition.

Explanation

A harvester must know which web pages belong to a website. In theory, this could be done by providing a complete list of all web pages of the website in advance. In practice, this is usually impossible because such a list can be very long and constantly changing. Therefore, a harvester works with a harvesting scope determined by a number of parameters.

The parameters of the harvesting scope may differ per harvester. For most harvesters, they are variants of the following parameters.

6.11.1 Start page (seed URL)

The harvester starts by retrieving the specified start page. From the start page, the harvester then retrieves all web pages that are reachable via hyperlinks.

The start page for harvesting is usually the main page of the website (also called the homepage), for example www.rijksoverheid.nl. However, it may also be a subdomain, such as the Actors Register: actorenregister.nationaalarchief.nl, or a sub-website, for example www.overheidsorganisatie.nl/over-ons/wie-is-wie.

6.11.2 Orphan pages

Orphan pages are web pages within the website that the harvester cannot discover and therefore cannot retrieve. For example, because the hyperlink to an orphan page is calculated by a script only at the moment a web page is displayed.

Orphan pages do belong to the website and must therefore be harvested. If it is known which pages are orphan pages, they are placed on a list so that the harvester knows these web pages must also be archived.

6.11.3 Site URL

The main URL with which all web pages of a website begin, such as <http://www.overheid.nl>.

6.11.4 Secondary site URLs

In some cases, part of a website may begin with a different URL, for example attachments stored on another server. In such cases, one or more secondary site URLs can be specified.

6.11.5 Blacklist

This is a list of web pages that cannot be retrieved for technical reasons. The harvester skips these web pages.

An example of a technical reason for not harvesting a web page is a calendar on a website. Such a calendar can extend infinitely into the past and future, making it impossible for the harvester to determine where to stop.

The blacklist is part of the list of non-harvestable content.

6.11.6 Scope for embedded content (embedded web resources)

The delimitation of which embedded content of web pages is retrieved. This may include all content or be limited to specific hosting domains, file types, or multimedia formats.

A web page usually consists of several components: HTML, CSS, JavaScript, images, multimedia, and links to attachments such as PDF documents. Some of these components may be hosted on a server other than the website itself.

6.11.7 Maximum number of hops

The number of hops is the number of links the harvester follows to reach a web page starting from the start page. If the number of hops is limited, the entire website may not be harvested.

This standard does not permit a maximum number of hops.

6.11.8 Maximum path depth

Path depth is the number of URL elements that appear in the URL of a web page after the domain name. For example, the homepage www.rijksoverheid.nl has a path depth of 0, and www.rijksoverheid.nl/actueel/agenda has a path depth of 2 (“actueel” and “agenda”).

If path depth is limited too much, the entire website may not be harvested.

This standard does not permit a maximum path depth.

6.11.9 Scope+1

Web pages often contain links to web pages outside the website. These may be other websites of the same government organization or websites of other organizations. With scope+1, the harvester also retrieves these external web pages, but does not follow hyperlinks on those external pages further.

Without scope+1, the harvester does not retrieve external pages.

This standard does not permit the use of scope+1.

6.12 Transfer

The mandatory procedure under the Archiving Act for transferring government information that must be permanently preserved to a designated archival repository no later than twenty years after creation, for management and access.

Source: Own definition.

Explanation

See the [DUTO wiki](#) and the website of the [National Archives](#).

6.13 Retention schedule

A systematic list of categories of information objects of a government body, established in accordance with the Archiving Act, indicating for each category whether the information objects are preserved permanently or after what period they are eligible for destruction, including accompanying explanation and justification.

Source: Derived from Article 5 of the Archiving Act 1995 and Article 5 of the Archiving Decree 1995.

Explanation

With retention schedules, a government organization makes public how it selects information

for destruction or permanent preservation. This has two objectives. First, it makes clear to stakeholders how long information is preserved. Retention schedules are therefore official publications. Second, a retention schedule is a practical instrument to determine whether and when an information object must be destroyed.

Information that must be preserved temporarily according to the retention schedule must be destroyed after that period under the Archiving Act 1995. Information that must be preserved permanently must be transferred by the government organization to an archival repository, no later than twenty years after creation.

The handbook *Valuation and Selection (Interests in Balance)* explains what a retention schedule is and how it is drawn up. Archival legislation does not set destruction periods for web archiving. The law only describes the procedure for drawing up and adopting a retention schedule. Each government organization must determine destruction periods itself by weighing interests. This is called valuation. Archival institutions to which permanently preserved information is transferred assist in this process.

6.14 URL (Uniform Resource Locator)

The unique name by which a web resource can be retrieved from a web server.

Source: Adapted from Wikipedia.

Explanation

For example: <https://en.wikipedia.org/wiki/URL>. See further the definition on Wikipedia.

6.15 User agent

All software that retrieves and presents web content to users.

Source: Web Content Accessibility Guidelines (WCAG) 2.0.

Explanation

Examples include web browsers, media players, plug-ins, and other programs and assistive technologies that help retrieve, render, and interact with web content.

A web browser is a computer program for viewing web pages. Popular browsers include Internet Explorer, Microsoft Edge, Mozilla Firefox, Google Chrome, Safari, and Opera.

6.16 WARC (Web ARChive format)

ISO standard (ISO 28500) for combining multiple web resources into a composite archive file, together with information related to the creation of the archive file.

Source: Adapted from Wikipedia.

Explanation

Typical of a WARC file is that it contains web resources retrieved by harvesting (although WARC can also be used for other archiving methods).

A WARC file contains:

- Requests sent by the harvester to the web server
- Responses returned by the web server to the harvester
- Metadata about the harvesting process

For further explanation, see ISO standard [ISO 28500](#).

6.17 Web resource

An information object or service available via a network and identified by a URL.

Source: Translation of the [W3C Glossary of Terms for Device Independence](#).

Explanation

A web page is a web resource, but images on a web page are also web resources.

6.18 Web page

A collection of information consisting of one or more web resources, identified by a single URL and intended to be displayed together. More specifically, a web page consists of a web resource with zero, one, or more forms of embedded content intended to be displayed as a unit and referenced by the URL of the non-embedded resource.

Source: Translation of the [W3C Glossary of Terms for Device Independence](#).

Note:

The W3C definition refers to a URI instead of a URL. The distinction is not relevant here. URL is used because it is customary to refer to a web page by a URL.

Explanation

A web page is an online document with a web address (URL). Based on the URL, a user agent can request the web page from a web server and then display it.

A web page can be in any format. Many web pages consist of HTML code, but this is not mandatory. A text file or an image is also a web page if a user agent can retrieve it from a web server via a URL.

A web page usually consists of multiple web resources for different components of the page, such as main text, images, videos, script code, and style sheets (instructions for layout). The main resource is identified by the URL; the URLs of the other resources are included in the main resource. This is called embedded content. The embedded content is retrieved separately from the web server but forms an integral part of the web page. When archiving a web page, all embedded content is therefore also included.

A characteristic of a web page is that it usually contains hyperlinks to other web pages. In this way, all web pages together form a web of information through which one can easily navigate.

A government web page is often hosted on a server of a government organization, but this is not mandatory. It may also be hosted on a web server of a company or another non-government

organization. This does not change the legal responsibility for archiving the web page. If a government organization is responsible for the content, it is also responsible for archiving it.

6.19 Website

A collection of interlinked web pages, including a start page, with the same base URL (the site URL).

Source:

Translation of *Web Characterization Terminology & Definitions Sheet*. In this translation, “residing at the same network location” has been rendered as “with the same base URL”, as this aligns better with the definition of a web page.

Explanation

A website is a coherent collection of web pages that belong together, for example because they belong to the same organization or concern the same subject.

The start page and the site URL determine which web pages belong to a website. Often the URL of the start page is the same as the site URL, for example <https://nos.nl/> for the NOS website.

“Interlinked” means that every web page of the website can be reached by following a sequence of hyperlinks, starting from the start page via zero, one, or more web pages on the same site.

The same base URL means that the URL of every web page of the website begins with the site URL, for example <https://nos.nl/nieuws/>.

It may occur that part of the web pages of a website begins with a different URL than the site URL, for example attachments stored on another server. In that case, a website also has one or more secondary site URLs. Web pages that are reachable from the start page but begin with a secondary site URL are then also considered part of the website.

The embedded content of a web page on a site always belongs to that website, regardless of the URL of the embedded content. This is because embedded content is an inseparable part of a web page and therefore also of the website.

6.20 Website archiving

The archiving of websites.

Source: Own definition.

Explanation

Website archiving is the sustainable accessibility of the web pages of a website.

Sometimes the definition “retrieving and storing a website” is used. However, that is only one of the methods of website archiving.

Because web pages often change and disappear, website archiving does not happen automatically. Additional measures are required. There are various ways to approach this. Client-side archiving is the most commonly used.

6.20.1 Client-side archiving (harvesting)

With this method, a harvester approaches the website in the same way as a user and therefore “sees” the website only from the front end: the client side (see also definition 6.9).

An advantage is that it does not interfere with the technology of the website, as contact is made only in the same way as any visitor to the website. However, this method of harvesting also has a number of limitations:

- Harvested web pages are retrieved by one (anonymous) user. Web pages whose content depends on user characteristics (such as a completed form or the content of cookies) cannot be archived in this way. In particular, web pages that require user login cannot be harvested.
- During consultation of a harvested web page, the underlying web server is no longer available in the state it was in during harvesting. If interaction with the web server is required to display a web page, the result is therefore not guaranteed to be the same as at the moment of harvesting. In many cases, this may even result in an error message, for example with streaming video and interactive maps. This limitation does not apply as long as rendering is handled entirely within a web browser.
- Websites that are no longer online cannot be harvested. This can sometimes be resolved by temporarily bringing the website back online.

6.20.2 Server-side archiving

With this method, the content of a website is retrieved from the CMS, or the CMS itself performs the archiving. This is only possible if the CMS administrator cooperates. Agreements must be made separately for each CMS, and the CMS may need to be modified.

The advantage of this approach is that all information in the CMS can be archived, including any history information, text changes, and other information that is present in the CMS but not visible on the front end (client side). A disadvantage is that this approach is highly dependent on the technology of the CMS.

Section 7.3 explains why this form of archiving was not chosen in this standard.

6.20.3 Transaction archiving

With transaction archiving, the exchange of data between the user agent and the web server is archived during use of a website. This archives all interaction between users and the website.

Transaction archiving is (as yet) rarely applied.

The advantage of transaction archiving is that it archives exactly all information that users receive and send, including user-dependent content and information sent back by users (for example forms). It also captures interaction with the server that determines web page behavior. Transaction archiving makes it possible to reconstruct an entire website visit by a user.

However, transaction archiving also has major disadvantages. The privacy of website users is at stake. In addition, it generates a very large amount of data, requiring extensive storage capacity.

An alternative to harvesting may be web recording, where a session of a single (anonymous) user is recorded. This is impractical because the user must manually click all links, and this must be repeated for every change to the website.

6.21 Careholder

The person who, by or pursuant to the Archiving Act 1995, is charged with the care of archival records.

Source: Archiving Act 1995.

The Archiving Act 1995 determines per type of organization who the careholder is. Usually this is the executive or daily governing body. For a ministry, this is for example the minister; for a municipality, it is the mayor and aldermen.

The term careholder is also often used to refer to the organization under which the archive falls. Strictly speaking, this is incorrect.

7 Appendix: Motivation of the choices made

When drafting this standard, the National Archives made a number of choices in consultation with experts. Some of these choices led to questions. These choices are motivated below.

7.1 Scope delimitation

On several points, the scope has been limited because no generally accepted solutions are currently known. Future versions of this standard may expand these points.

7.1.1 Only public web pages

For non-public parts of a website, there are restrictions on accessibility and possibly technical limitations on harvesting (because the harvester must gain access to non-public sections). In addition, it is unclear whether non-public parts of a website must be preserved permanently. This requires additional effort to determine how to handle non-public website sections.

7.1.2 Only non-personalized web pages

If the content of a web page depends on personal data, that data must be known for archiving. This raises several issues that require further investigation, such as how to determine which personalized web pages were requested, how long they must be archived, and how privacy is respected.

7.1.3 No social media

Both the role and technology of social media differ fundamentally from websites, and they differ per type of social media. This means that requirements for archiving websites are not automatically sufficient for archiving social media.

At present, the National Archives cannot recommend a standard working method for archiving social media expressions. Additional work is required to determine this.

7.1.4 No related information

This standard concerns only information in the form of a website. It does not concern information related to the website, such as underlying communication policy, information about the CMS used to create the website, the contents of CMS databases, information sent by users via a website, or the processing thereof.

All such information must also be archived if there is sufficient interest, but these are different forms of information and therefore fall outside the scope of this standard.

7.1.5 Limitation to client-side harvesting

This first version of the standard limits itself to harvesting as an archiving technique, because this is currently the only commonly accepted, simple, and generic technique. In addition, well-configured harvesting has little to no impact on websites and the underlying processes and applications.

Open-source software is available for harvesting, and there are multiple commercial service providers. This also makes it an affordable technique that is widely supported.

7.2 Use of the WARC standard

The [WARC standard](#) is the commonly accepted standard for storing harvested web pages. The standard is not under discussion and is widely supported.

7.3 Why not only archive the underlying data?

This standard requires harvesting (retrieving and storing) web pages as they are received by users of a website. These web pages are generated by a CMS based on data recorded in the CMS or elsewhere. An alternative strategy is to store only the underlying data.

Archiving only the data is insufficient. The appearance of the web page must also be known afterward. The underlying data alone does not provide sufficient insight into what information users of the website actually received. The Archiving Regulation states the following:

“The careholder ensures that for each of the archival records it can always be established: [...] the content, structure, and appearance at the time of receipt or creation by the government body.”

It is possible to preserve the functionality of the CMS so that web pages can be recreated at any later time based on the old underlying data. This usually requires modifications to the CMS and additional effort when modifying or replacing the CMS. This must be done for all different types

of CMS. In addition, it must always be verified whether the CMS still produces the old web pages correctly.

Harvesting is therefore a simpler, cheaper, and more controllable way of archiving.

7.4 Archiving embedded content of others

Some web pages contain embedded content created by an organization other than the content owner of the website. An example is an embedded Twitter timeline. Archiving this embedded content is necessary to be able to fully reproduce the web pages later as they once were.

It might seem logical for the owner of the embedded content to archive it. However, the government cannot rely on that owner actually doing so. Therefore, it is better to also archive embedded content created by others.

7.5 Permanent preservation of websites

How long government information must be archived is laid down as a destruction period (also referred to as a retention period) in so-called retention schedules. These are formal decisions established per government organization or group of government organizations (for example all municipalities). The destruction periods for websites are not always unambiguously recorded in retention schedules and may also differ between schedules.

This standard assumes that retention schedules determine that public government websites must be preserved permanently. This assumption is based on the Generic Valuation Model Central Government (GWR) and the Retention Schedule for municipalities and inter-municipal bodies. Together, these cover the largest part of government.

GWR 2.0 states that “final products of public information provision” (which include websites) must be preserved permanently (category 5.1). The municipal retention schedule 2017 states that “online content for external communication” must be preserved permanently (no. 20.1.2).

If a government organization has not included permanent preservation of its websites in its retention schedule, this must be discussed with the National Archives at the next update of the retention schedule.

7.6 Preserving all web pages for the same duration

Within a single website, all web pages and all their changes are preserved for the same duration, even if some web pages are of lesser importance.

Differences in destruction periods would lead to incompletely archived websites in which parts are missing. This is comparable to selectively tearing pages out of a book. This makes future use of the archived website more difficult. Missing components may be important for interpreting components that are still available.

In addition, a single destruction period is a clear rule. This prevents misunderstandings that could lead to archived web pages being destroyed incorrectly.

The choice not to distinguish between destruction periods is also consistent with the Generic Valuation Model Central Government 2.0 (GWR) and the Retention Schedule for municipalities and inter-municipal bodies 2017. These do not distinguish between destruction periods for parts of websites.

There is a possible argument against preserving all web pages for the same duration: the costs of harvesting and managing may not outweigh the added value of permanent preservation. However, at the time this standard was established, there were no indications of this. Moreover, different destruction periods within a website would also lead to additional costs, because selection would then be required.

Preserving all changes to web pages sometimes presents a practical problem. Web pages can be changed or removed at any time, even multiple times per day. Periodic harvesting of websites takes place at fixed times. Therefore, the standard requires harvesting once per day. Real-time harvesting would be even better, in which the CMS immediately reports every change to the harvesting service so that it can be recorded immediately.

At the time this standard was established, real-time harvesting was not yet a standard option in CMSs and harvesting software. Therefore, it was not included as a requirement.

7.7 Why does the National Archives not perform web archiving itself?

It is not a statutory task of the National Archives to create archives. Under the Archiving Act 1995, each government organization is responsible for this itself. Permanently preserved (web) archives of the central government are transferred to the National Archives at a later stage. After transfer, the National Archives assumes management.

In addition, the National Archives is a knowledge center for archiving and assists government organizations in setting up their (web) archiving. This standard arises from that role.

7.8 The Royal Library already archives government websites, doesn't it?

The Royal Library (KB) does indeed archive government websites. In addition, there are other institutions that archive websites, such as the Netherlands Institute for Sound and Vision (NIBG) and the Internet Archive. The KB web archive archives a selection of Dutch websites from 2007 onward. These sites are harvested once or twice per year. Anyone with a KB annual pass can consult the KB web archive in the KB reading room. NIBG archives websites of broadcasters.

Why must government organizations archive their own websites if other institutions already do so? The reason is that these institutions do not archive for the purposes of archival legislation, but to fulfill their own objectives. They make their own assessments of which websites they find interesting or important to archive, with self-chosen frequencies, selections, destruction periods, and quality characteristics. These choices will rarely coincide with those arising from archival legislation.

For these reasons, web archiving by cultural and scientific institutions cannot replace statutory web archiving. However, these institutions often have years of knowledge and experience with web archiving that they are willing to share with government organizations.

7.9 Archiving web browsers outside the scope

Web pages are consulted using many different web browsers and devices, which also change continuously. The display of web pages is not standardized and therefore differs for each combination of browser and device. Often the differences are small, but relevant differences cannot be ruled out, especially between display now and in the distant future.

This raises the fundamental question of what is actually archived during harvesting. Not the actual web page as perceived by the user, but only the code used to create the web page. Should the functionality used to render the web page from the code also be archived? It may seem like a solution to archive the code of web browsers themselves, but then the functionality of the devices and operating systems on which those browsers run would also have to be preserved. There is currently no reliable and usable solution for this.

8 Appendix: Background and assignment

8.1 Background

Government websites fall under the scope of the Archiving Act 1995. This means that the government must archive its websites. At the end of 2016, the Heritage Inspectorate (EGI) investigated to what extent this actually occurs within the central government. The conclusion was that governments archive their websites and social media insufficiently. This leads to loss of information and constitutes a (legal) risk.

The EGI drew the following conclusions:

- Organizations of the central government archive their websites insufficiently or not at all, with a few exceptions.
- The same applies to an even greater extent to official expressions on social media.
- This leads to loss of information and poses a risk in the event of a legal dispute.
- In the longer term, digital information with cultural-historical, social, and/or scientific value threatens to disappear.
- There is a lack of a policy framework and specifications for archiving government websites and expressions on social media.
- As a result, government organizations have insufficient guidance for archiving their websites and official expressions on social networking sites.
- There is insufficient knowledge about web archiving within archiving organizations.
- The division of responsibilities between the National Archives, the Royal Library, and the Public Communication Service is insufficiently clear to archiving organizations with regard to web archiving.

The report led to parliamentary questions and subsequently to a parliamentary letter in which the Minister of Education, Culture and Science made the following commitments:

- I will ask the General State Archivist to draw up and manage a guideline aimed at archiving websites and expressions on social media.

- In cooperation with CIO-Rijk, he will also make a proposal for implementing an archiving practice for these media within organizations of the central government.
-

8.2 Elaboration

In this standard, the National Archives elaborates on the first commitment of the minister for the entire government. Implementation of the standard for the central government is a separate project. This is carried out on behalf of CIO-Rijk in cooperation with the National Archives and the Public Communication Service of the Ministry of General Affairs. This standard does not address that further.

The National Archives chose to shape the guideline as a standard with concrete requirements for website archiving, with appendices consisting of a step-by-step plan, an explanation of the terminology used, and motivation of the choices made. In this way, the standard meets the need for both uniformity (the requirements), practical guidance (the step-by-step plan), and explanation (the terminology and motivation).

The definition “website archiving is the archiving of websites” in this standard differs from the definition in the EGI report (“web archiving is a container concept for retrieving and storing websites and expressions on social media”). “Retrieving and storing websites” is only one technique of archiving, usually referred to as harvesting. Moreover, it does not cover all aspects of archiving, such as findability and accessibility. Harvesting is used in this standard as the only technique for website archiving. As a result, the difference in definition has no practical consequences here. Archiving of social media falls outside the scope of this standard (see section 7.1.3).

8.3 Assignment

The National Archives develops and manages the standard on behalf of the Minister of Education, Culture and Science. The standard is a knowledge product and part of the National Archives’ professional services.

8.4 Standardization Council

This standard is submitted for adoption to the Standardization Council of the National Archives. This consultation body consists of representatives from all levels of government and archival institutions and serves as the substantive steering group for the knowledge products of the National Archives. The Standardization Council ensures that knowledge products are usable for government organizations.

8.5 Further development

The National Archives develops the standard in phases. The starting point is to make existing knowledge available as quickly as possible. Subsequently, various archival creators can test it directly in practice. The National Archives incorporates these practical experiences into the standard.

Publication details

A publication of the National Archives
P.O. Box 90520
2509 LM The Hague

Contact

Telephone: +31-70-331 5460
E-mail: contact@nationaalarchief.nl

Website: <https://www.nationaalarchief.nl/archiveren>