

Final Report on Market Consultation Web archiving

Disclaimer: This English translation is for information purposes only and has no legal status. The Dutch version 'Eindrapport Marktconsultatie Webarchivering' prevails.

1. The Dutch [Guideline on the Archiving of Public Sector Websites](#) is the standard for archiving public sector websites

a. Can you meet the requirements set out in this standard?

All market parties state that they meet the requirements of the standard for archiving public government websites. Several market parties cite ISO 28500:2017 as the standard for WARC files.

b. Does your solution offer additional functionalities; if so, which ones? And what added value do these functionalities offer the custodian?

Market players mention:

- Additional functionalities, such as the archiving of social media;
- Availability of a service portal where the custodian can play an active role in the management and quality of the archived websites.

In addition, the possibilities for searching the archived websites and the possibility of conducting academic research on the content are also mentioned.

c. Are there any requirements in the guideline that you consider unfeasible or that raise practical objections? Could you elaborate on this?

On this point, several market parties have raised the following objections:

1. Dynamic websites, the use of complex JavaScript and third-party embedded content (maps, links to external videos);
2. Metadata and the conversion from one (older) metadata standard to another (MDTO). It is important to allow sufficient time and to establish a validation process for this;
3. How to address information security in relation to public access to archived content. Further clarification of this requirement is needed;
4. The uncompressed storage of complete WARCs: greater attention should be paid to long-term file formats in order to ensure the accessibility of the content;
5. The requirement for a single, complete harvest: according to one market participant, this should be done more frequently to create manageable archives.

2. In your opinion, what are the advantages and disadvantages of dividing the contract into several plots?

Generally speaking, market participants cite more disadvantages than advantages. Many mention improved market dynamics and the ability to contract the best solutions for specific

functionalities as advantages. The main disadvantages cited are the risks associated with managing multiple contracts and ensuring the desired outcome (quality).

3. In your opinion, what methods are there for defining multiple plots?

Market stakeholders do not have a uniform view on how plots should be defined. Approaches vary, ranging from:

- Classification of plots based on functionality (e.g. Harvesting – Storage – Access – Migration and Transition);
- Classification of plots based on organisation/responsible body (e.g. clusters of ministries and autonomous administrative bodies);
- Classification of plots based on the complexity of the websites to be archived.

4. Do you prefer a single plot or multiple plots? Why?

Most market participants prefer to tender for the contract as a single lot for reasons of manageability and to ensure uniform quality and access. There are also market participants who express a preference for splitting the contract into plots due to the expertise required. Furthermore, there are parties who are flexible and can operate under both scenarios (splitting the contract into one or more plots).

5. Is your organisation ISO-certified? If so, which certification(s)?

All but one of the market players are ISO-certified.

Certifications mentioned and associated themes:

- NEN-ISO/IEC 27002 (cybersecurity, operational security)
- ISO/IEC 27001 (software development, infrastructure management, organisation, archiving, support)
- ISO 27001, SOC 2 Type II, and C5 certifications (data storage)
- ISO 9001 (mentioned by one company)
- ISO 28500:2017 (WARC standard)
- BIO (Government Information Security Baseline, Dutch)

6. Data sovereignty is a key issue in web archiving. In line with our requirement for data sovereignty, we wish to permit web archiving only on servers that are housed and hosted within the European Economic Area (EEA).

a. Can you comply with this request?

All market participants except one are able to comply with the request.

b. If not: why not?

A market player distinguishes between data sovereignty and data storage and adopts a data-centric approach to sovereignty, whereby the law follows the data; even if the data is processed in other regions.

7. What factors (e.g. retrieval, indexing, crawling, storage, etc.) do you take into account when determining the cost of archiving a website?

Market players provide a variety of answers, some more detailed than others. Generally speaking, factors such as the number and size of the websites, the complexity of the website to be archived, storage capacity and quality control are mentioned. Only one market player considers the amount of use of the archive by visitors to be a factor.

8. Could you explain which cost components are fixed and which are variable?

The cost structures of the various market players vary. For some market players, all costs are fixed (all-inclusive price, often with a limit on quantity/volume), whilst for others, everything is variable.

- Most commonly cited fixed costs: setting up a service portal, cybersecurity and compliance, licensing, Service Level Agreement (SLA) assurance, transfer to the Dutch National Archives (at an agreed rate), hosting of the viewer. Harvesting of static websites is usually fixed as well.
- Most commonly cited variable costs: number of websites, complexity (average, complex), required computing power/duration, storage volume, post-processing, import of existing web archives plus any processing, plus any processing (reharvest), support/manual processing/quality assurance, metadata management.

9. How is the automated quality control – covering both the quality of the archived website and the technical quality of WARC files – organised?

Responses range from detailed to brief to no further information, due to the public nature of the market consultation. Most market participants generate automated reports based on the archiving chain, with some market participants also making the results of these reports available to the client/responsible party via the service portal.

10. Can you describe how the web archive files and metadata are checked for accuracy and completeness during and after archiving?

a. What are the specific points of focus that are checked?

Many market players carry this out in accordance with the standard ([ISO 28500:2017](#)). One provider states that they also periodically check whether older archives continue to function correctly from a technical and operational perspective. In addition, there is a provider that periodically checks the integrity.

b. Explain how you apply [ISO 28500:2017](#) in your service provision

Market participants state in particular that they apply this standard. Two market participants are active members of the IPCC (International Internet Preservation Consortium). How this standard is applied is described in relation to other technical requirements that must be met.

c. Can you describe how the results of this audit are communicated to the website owners?

Market participants vary in terms of frequency, content and method of communication. Many offer communication via a service portal. There is also one market participant where a service portal is a possibility and where such a portal is currently under development.

11. How do you report on:

a. Potentially incorrect web archiving jobs (e.g. errors in URLs, empty websites or jobs resulting in empty WARC files, etc.)?

Almost all parties state that they log these issues automatically. Reporting takes place partly via a service portal (see also the answer to question 1b) and partly (sometimes additionally) by email (occasionally followed up by a phone call).

b. Duplicates (= two content providers submitting the same URL)?

Reporting on duplicates is usually standard, but sometimes needs to be activated. Many parties have a central URL register that prevents the addition of a duplicate, or informs both website owners immediately. One party states that compliance with the Directive makes duplicates impossible.

c. Non-harvestable content: issues with (the archiving of) websites, such as missing elements?

One market participant states that this content can be identified through quality checks. A second party states that this can be viewed on the service portal. A third party states that this is detected during the initial harvest and thereafter, and that coordination takes place with the custodian. And a fourth party states that information on this is also stored in an ODS spreadsheet. Reference is also made to the quality framework for metadata and archive files that is used.

d. Harvests that fall outside the scope, such as: crawler traps, perpetual calendars, shopping baskets, etc.?

Many market players only indicate how they resolve these issues (technically). A few state that reporting takes place via the service portal or that structural deviations are reported (without specifying how) and discussed.

12. The guidelines for archiving government websites specify which metadata must be included with the archive files during transfer. Most of the metadata must be provided by the service provider. There is also metadata that must be provided by the record-creating organisation (e.g. a description of the website and the information category from the selection list under which the website archive falls). This metadata must be compiled into a single export upon transfer of the archive.

a. Does your organisation have experience in generating this metadata? Can you specify which?

All market parties except one have experience in this area. The party that does not have this experience states that it assumes the custodian has a digital preservation or asset management system in place for the management of metadata. A number of market parties indicate that they also manage metadata, in addition to being able to generate it. One party explicitly mentions the consolidation of metadata supplied by the record-creating organisation with the automatically generated metadata from the archiving process, for the purpose of exporting the archive.

b. What options do you offer for checking and adding metadata?

One provider states that it offers checking options in various ways and at different stages:

- a service portal with validation support for the data owner;
- a validation engine that checks against the relevant metadata profile prior to export;
- a compliance report generated prior to export.

Other parties offer the option of a service portal to view metadata, for validation and enrichment. Management by period is explicitly mentioned by one party.

Two parties do not offer any options for checking or adding metadata.

c. This metadata must comply with MDTO standards. Could you explain how you intend to meet this requirement?

All market participants state that they can comply with this, in fairly general terms, but do not specify in detail how they will actually achieve this. Two participants mention provisions for the transition period from ToPX to MDTO. One participant indicates that during this period the system will generate both MDTO and ToPX and is prepared to demonstrate this. One participant states that it has an e-repository with the MDTO model.

d. Does your organisation have experience of transferring this data to an archive repository or to another service provider? Could you specify which ones?

All market players claim to have this experience. A few provide concrete details, including by mentioning an integrity check. One party refers to the use of a SIP, another to seamless transfer. One party can receive and transfer data, but has no experience with the latter.

13. In some cases, part of an archived website must be excluded from being made available in a viewer due to privacy, copyright or information security concerns.

a. How can archived content be made available on a restricted basis? Can a custodian do this themselves?

All market providers offer this option and most specify how it works. Many providers state that custodians can (also) arrange this themselves via the market provider's service portal. The level of detail regarding the restriction varies. Some refer to specific URLs and entire websites, whilst others mention granular options or exclusion from availability based on modifications to the document's metadata.

b. How is this limited availability (in accordance with section 4.4 of the [Guidelines for Archiving Government Websites](#)) managed?

Two parties refer to the guidelines and the ODF spreadsheet included therein for management purposes. A number of parties indicate that this is possible via the service portal (via role-based access for authorised users).

c. How is the administration referred to in @b transferred to a potential subsequent service provider or the archive repository?

All parties state that exporting with metadata as per @b is possible. The majority state that this is done via an export file such as a .csv or similar file. A few parties do not specify the file type for the export.

14. The tender may also include the receipt, subsequent management of, and access to data from previously created web archives.

a. Are you able to receive and process these previously created web archives? What requirements and restrictions (technical, metadata, volume of data) do you impose on the archive to be transferred?

All market participants state that they are able to accept previously created web archives. A common requirement is that WARC files must be ISO 28500:2017 compliant, and/or reference is made to the WARC 1.1 standard, and/or to the requirements set out in the [Guidelines for Archiving Government Websites](#). A few participants highlight the availability of metadata in accordance with these guidelines. Another condition mentioned on a few occasions is a secure transfer method, such as sFTP.

b. Do you have experience in receiving, indexing and providing access to such volumes of data (>1 petabyte)? Can you provide examples of this?

Virtually all market participants have experience with this, sometimes on a somewhat smaller scale (less than 1 petabyte). A few participants cite other market participants that create web archives as a source. With the exception of two participants, market participants do not provide concrete examples. One participant indicates that such examples can be provided if an NDA is agreed.

c. Can you transfer the web archives you have created (and, where applicable, previously created web archives) to the National Archives, in accordance with the [conditions](#) set by the National Archives for the transfer of digital archives; in particular, the [Export Conditions](#) document.

All parties state that they can do this. One party recommends the continuous transfer of archives throughout the term of the contract.

15. Would you be open to discussing this market consultation with the Contracting Authority?

All parties are open to discussion.