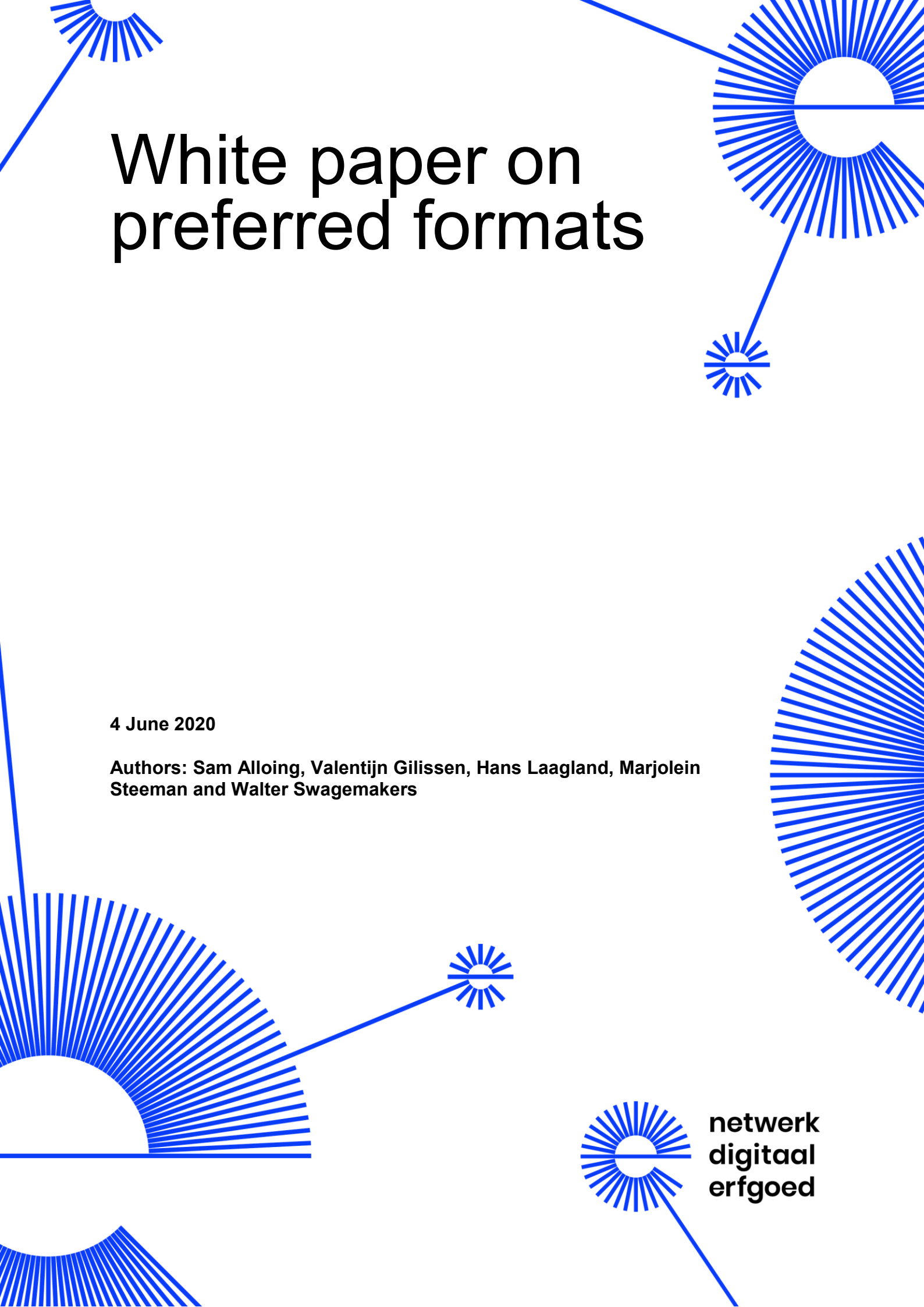
White paper on preferred formats

4 June 2020

Authors: Sam Alloing, Valentijn Gilissen, Hans Laagland, Marjolein Steeman and Walter Swagemakers



netwerk digitaal erfgoed

Table of contents

Summary

Data and their carriers work in tandem with each other. Data needs a carrier, whether it is a sheet of parchment, a glass negative plate or a PDF file.

Clearly, if you work with digital information objects, you'll need to find out as much as possible about the media that carry the information in question, i.e. file formats. The organisation curating the information needs to have a policy on file formats. That means deciding what guarantees it is willing and able to give about preservation, and how it should communicate about preservation issues with suppliers (and users) of the material.

The Dutch Digital Heritage Network (NDE) is currently preparing an on-line tool for helping archives and heritage organisations to formulate a policy on file formats. This tool will be known as the *Guide to Preferred Formats*. It will consist of two parts, i.e. explanations and examples intended to help organisations in formulating a policy on preferred formats, plus a list of file formats with detailed information on each of them.

The *Guide to Preferred Formats* will be overseen by an 'expert group on preferred formats'. The expert group's remit will be to bring together, share and refine the expertise currently available in the Netherlands.

Note for readers

All words **printed in bold type** are explained in the glossary at the end of this white paper.

1. Background information

“Help! I’ve just been sent a WordPerfect 5.1 file. What can you do with a WordPerfect 5.1 file so as to preserve the information in it?”

“Can I preserve my scans by storing them as compact JPEG2000 files, or would it be better to go for a bigger file type, such as TIFF?”

If you are planning to preserve digital material in the long term, you will need to find out about the carriers of such objects, i.e. file formats. File formats come in all sorts of shapes and sizes. During a recent stocktaking exercise, the National Library of the Netherlands found that its collection included over 1,300 different file extensions. This should give you a good idea of the extent of the problem: there are so many different options as to make it almost impossible to decide which particular file format is best suited for preserving information. On which file format should your preservation policy be based? And which file formats are definite no-no’s?

The Dutch Digital Heritage Network (NDE) is currently preparing a solution in the form of an on-line *Guide to Preferred Formats*. The idea is for this guide to help organisations reach decisions on file formats as part of their preservation policies.

This white paper analyses the problems surrounding file formats, proposes a method for describing and deciding on both preferred formats and policies on preferred formats, and outlines the contents of the on-line *Guide to Preferred Formats*, which will describe how to choose a preferred format and present up-to-date information on file formats. The aim of the white paper is to assess whether the solutions proposed meet the needs of NDE members. Hopefully, we can use the comments and feedback we receive to decide whether it is indeed worth publishing a guide for archives and heritage organisations. The publication of the white paper is also an opportunity to make improvements both to the proposed method for formulating policies on preferred formats and to the contents of the *Guide*.

The Dutch version of the white paper will be actively distributed among a range of heritage organisations, platforms and networks. It has also been translated into English in order to invite comments from organisations outside the Netherlands.

The white paper is an initiative of an NDE committee set up in 2019, the *Committee on Preservation Watch and Preferred Formats*. The committee’s aim is to encourage as many organisations as possible to incorporate preservation watch and **preferred formats** into their preservation strategies. The present white paper addresses the issue of preferred formats. A further document will be published during the course of the year to explain how to design a joint preservation watch.

2. The problem

The importance of having a preservation policy

Let's start by taking a look at the importance of a preservation policy for a heritage organisation. The NDE's (Dutch-language) *Guide to Preservation Policy* has the following to say about this:

"More and more collections curated by archives, libraries, media organisations, museums and research centres are being made available on line, in digital formats. Data loss is a constant threat. The challenges include obsolescent hardware, errors made when files are copied, file formats that are no longer readable, undocumented changes, etc. If no corrective action is taken, digital collections will be lost forever. Having a policy helps. An organisation that adopts a clear preservation policy takes a public stance in favour of the preservation and long-term accessibility of its digital collection. Policy is the connecting link between the organisation's philosophy and mission on the one hand and the activities it needs to perform in order to guarantee long-term accessibility on the other. In other words, it provides the foundations for developing a preservation strategy and justifying investments."[1]

A policy on file formats is one aspect of a preservation policy. There is a tendency for the number of different file formats used by a heritage organisation to grow unnoticed. But what does the organisation actually know about all these file formats? Which of them might be suitable vehicles for preserving the information in the long term? Which file formats could form the subject of active preservation? And what guarantees can the organisation provide for the long-term playability or usability of a particular file format? In short, on what basis should an organisation decide that a particular format is the right means of preserving a digital archive or collection, and that another format does not cut the mustard?

Digital preservation formats

If an organisation opts for **bit preservation**, the only guarantee it needs to give is that a file can be viewed 'as is'. In other words, the file displayed to the user is identical right down to individual bits to the file supplied in the ingest stage, where relevant in combination with any metadata that were supplied about the nature of the file. In its most basic form, the organisation does not even need to know anything about the file format, let alone impose certain requirements about it. The only condition is that the storage system should be able to find the file and verify its integrity by performing a checksum test.

Virus checks, and file format identification, characterisation and validation may also form part of bit preservation, however. In most cases, these activities are performed in order to collect and document knowledge. This can be done at various levels.

The more knowledge is stored in an archive, the better able it is to provide guarantees about the future usability of the files, for example, by making use of preservation strategies (such as migration

[1] *Guide to Preservation Policy* (Dutch-language publication): https://wegwijzerduurzaamheidsbeleid.nl

and emulation). Where this is the case, we see that there is a shift towards **functional preservation**. This means that extra care is taken to ensure that files remain accessible in the future, both for staff and for target users. Organisations define the practical value of their files in widely differing ways. This is because such definitions depend on the nature of the target users, the type of hardware and software they generally use, and the opportunities offered by – and the limitations of – the organisation's own infrastructure for making the file in question accessible to the target users. In other words, different organisations may make different judgements about the same formats.

Example: 'levels of knowledge' at the National Library of the Netherlands

The National Library of the Netherlands has formulated a policy on file formats based on levels of knowledge. The National Library decided on this policy because there is no way of dictating to publishers how (i.e. in which format) they should supply material to the National Library. The use of different levels of knowledge provided a solution to this problem. The idea is that the National Library should make clear how much support is provided for a given file format, in a growth path from bit preservation to functional preservation. Not every file format needs to end up at the highest level, i.e. functional preservation. The levels of knowledge are a communication vehicle enabling the National Library to clarify the degree to which it supports a particular file format.

Those file formats which are preserved (or which an organisation wishes to preserve) using functional preservation are referred to as **preservation formats**. These files, known as master files, are at the top of a digital archive's hierarchy. In order to facilitate ease of use, an organisation may decide to use them as the basis for creating other files. In some cases, these derivative files may be produced on demand in response to a user request, but they are generally produced in advance and preserved with just as much care and attention as the original master file.

Example: derivative files at the Dutch Institute for Sound and Vision

The Dutch Institute for Sound and Vision in Hilversum makes low-resolution copies (known as **proxy files**) for helping with orders for fragments. Because of the volumes and investments involved, the proxy files are stored in highly specific conditions in the Institute's archive.

Example: derivative files at the Eye Filmmuseum

Not all preservation formats are suited for immediate viewing or use. In such cases, many organisations decide to use an intermediate format known as a **mezzanine file**. This is a common practice among media archives whose preservation master files are not immediately suited for use, due to their size and nature. The Eye Filmmuseum is an example of an organisation that uses mezzanine files.

Preferred formats

Depending on the way in which the organisation organises the preservation of information objects, it may wish to designate certain formats as being preferred formats for the reception of digital material. Based on what has already been said about preservation formats, preferred formats may be defined as possessing the following characteristics:

- subjective: preferred formats are designated by the heritage organisation itself, in accordance with its own specific context and objectives. There is no such thing as an objectively defined, standard list of preferred formats.
- time-bound: the status of a format may change under the influence of technical and commercial developments. This means in practice that an organisation may decide at a later date either to designate a particular format as being a preferred format (for example, if sufficient support is available for the format in question) or to remove a format from its list of preferred formats. In the latter case, the organisation will need to think carefully about the consequences for those files it already holds in the format in question and about any guarantees that may have been given about the preservation of these files. A **preservation watch** will need to be set up to monitor the changing nature of preferred formats.
- associated with a guarantee: a preferred format always guarantees a degree of certainty about long-term preservation. In most cases, this is a guarantee of a relatively high degree of certainty about the preservation of the files in question. At the same time, an organisation is at liberty to attach any form of guarantee to one or more preferred formats. Formats associated with a lower level of certainty are known as **accepted formats**.[2]
- communicative: the aim of designating preferred formats is to inform suppliers (and users) of digital material how the archive or heritage organisation deals with certain formats. In some cases, the designation may be of a prescriptive nature (i.e. compulsory), but it can also be entirely neutral (as with the levels of knowledge applied by the National Library of the Netherlands).

The following diagram shows that, the greater the degree of certainty about long-term preservation, the more likely organisations are to refer to preferred formats. In the event of a lower degree of certainty, the organisation may decide to store files in an accepted format, for example because a particular format is in popular use. The point at which an accepted format becomes a preferred format tends to vary from one organisation to another.

[2] Guide to preferred formats published by the National Archives of the Netherlands: https://www.nationaalarchief.nl/archiveren/kennisbank/handreiking-voorkeursformaten-nationaal-archief

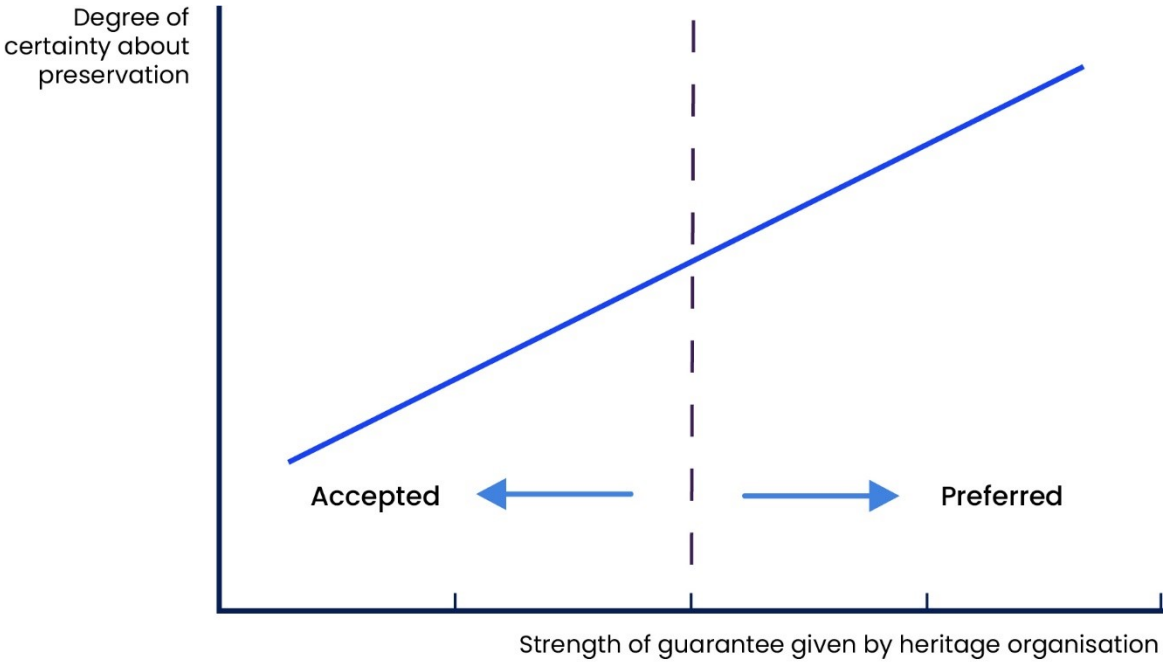


Figure 1: Relationship between the characteristics of a format and the guarantee given by the heritage organisation

The vertical axis represents the degree of certainty about long-term preservation applying to the format in question. In most cases, it is possible to specify a number of properties[3] associated with a higher degree of certainty:

- in general use: the file format may not be obscure, as this means there are unlikely to be enough tools available and also that the amount of support offered for activities such as migration is likely to be limited;
- open specifications: if the format is based on a standard, an open standard or an open specification, it should be possible, if the worst comes to the worst (i.e. if the software is no longer supported by the supplier), to tap into alternative means of support. The use of a standard, open standard or open specification also makes the file format easier to understand in technical terms;
- software-independence: this means that there is less risk of a supplier going bust, thus putting an end to support for the file format in question and making it impossible for certain functions to be performed.

The horizontal axis shows the strength of the guarantee offered by the archive or heritage organisation. It is important to bear mind the distinction already referred to above between bit preservation and functional preservation, and the various levels of preservation an organisation may decide to adopt within these two settings.[4]

In the end, the archive or heritage organisation decides for itself, on the basis of its preservation policy, which formats to accept or to prefer, and the strength of the guarantees it wishes to give in this connection. Any format may be plotted as a dot in the graph, as it were. The general expectation is that, the higher the dot is on the graph, the more likely the organisation will be to regard the format as a preferred format. Organisations may also set additional requirements about file formats. Generally speaking, the requirements that a format is expected to meet are stricter in the right-hand spectrum. The left-hand spectrum comes with less strict general specifications or requirements.

[3] DANS (Data Archiving and Networked Services) file formats: https://dans.knaw.nl/nl/over/diensten/easy/toelichting-data-deponeren/voor-het-deponeren/bestandsformaten

[4] See also the NDSA's levels of digital preservation, particularly in relation to content: https://ndsa.org//publications/levels-of-digital-preservation/

This may lead to organisations being willing to accept a broad range of file formats, irrespective of whether this is in combination with the controlled conversion of files into mezzanine or proxy files, as in the above examples involving the Dutch Institute for Sound and Vision and the Eye Filmmuseum. Also, certain archives are in a position to make advance arrangements with a creator or digitiser about the format of the digital material supplied to them. These arrangements may lead them to reject certain digital objects that do not conform with their archives' requirements or specifications.

Conclusion

The *Guide to Preservation Policy* referred to above advises the user to perform a file format check[5] as part of the process of functional preservation. This check takes the form of an inventory of the file formats present in a digital archive. It also involves studying the opportunities for ensuring that the file formats remain playable or usable in the long term. The organisation performing the check will need to assess the consequences of the findings for its own preservation policy, including for communications on how to deal with these file formats. This is a complex decision that depends on the specific context in which the organisation is operating and requires proper study.

The situation at present is that heritage organisations undertake this type of study on an individual basis. Not all of them are capable of assessing the ramifications of the findings for their own preservation policies. In other words, there is a need for collaboration in studying formats, so that organisations can benefit from each other's expertise.

[5] Wegwijzer Duurzaamheidsbeleid > Bestandsformatencheck: https://wegwijzerduurzaamheidsbeleid.nl/artikelen/bestandsformatencheck/

3. The solution: a *Guide to Preferred Formats*

Our intention in compiling the *Guide to Preferred Formats* is to help organisations to arrive at certain choices when formulating their policy on file formats and preferred formats. The *Guide* does this in two ways: first, by providing information on problematic issues surrounding the preservation of file formats. Second, it contains a comprehensive list of file formats with detailed information on each of them. The *Guide* will be produced as part of an NDE project entitled *Preservation Watch and Preferred Formats*. The *Guide* will be kept up to date by an 'expert group on preferred formats', which will be set up for this purpose in the near future.

What is the *Guide to Preferred Formats*?

The *Guide* does not claim to offer an exhaustive list of preferred formats. Rather, it is intended to be a tool that heritage organisations can use in formulating policies on file formats, based on their own experiences and knowledge coupled with those of their peers. The *Guide* should prevent organisations from having to reinvent the wheel, and instead enable them to build on existing knowledge and experience.

A further important aspect is the fact that other organisations will be able to feed their own input into the *Guide*, thus helping to build up the pool of knowledge on both preferred formats and formats in general by sharing their knowledge and experience with the rest of the community. In this sense, the *Guide* will be an open learning environment.

What does the *Guide to Preferred Formats* contain?

The *Guide*:

- explains what preferred formats, preservation formats, **access copies**, accepted formats and levels of preservation are;
- gives examples of accepted formats and preferred formats used by members of the NDE network;
- explains why certain formats are selected as preferred formats or why they are associated with a given level of preservation;
- offers archives and heritage organisations an opportunity to use its contents to formulate their own file format policies, including on their own preferred formats.

The tool is designed to make maximum use of existing information, by reusing it and bringing it together. The main sources of file information are **PRONOM**[6] and **Wikidata**.[7] The tool answers the following questions about the properties of a given file format:

[6] PRONOM: http://www.nationalarchives.gov.uk/PRONOM/Default.aspx
[7] Wikidata: https://www.wikidata.org en https://wikidp.org

- Is the format a standard, an open standard or an open specification?
- Is the format in common use?
- Is the format supported by more than one application?
- Are there any designers or suppliers whose applications support the format?

The *Guide* also lists those members of the NDE network which use a particular preferred format, stating any restrictions to which this use is subject and whether the format in question is used in combination with certain access copies. Thanks to a data link with **COPTR**, the *Guide* can also display information obtained from **COPTR**. By presenting all this information, the idea is that the *Guide* will act as a common database – as indeed is a precondition for the formation of a national preservation network.

Users consulting the *Guide* can view file formats with their file format name and Pronom Identifier (or **PUID** for short).[8] This form of identification is not sufficient for certain file formats, however, which means that the file format properties (such as the bit rate in the case of video formats) need to be used for defining preferred formats. A data link with Wikidata is used for adding these properties. Different versions of file formats may also play a role in defining preferred formats.

It also needs to be clear which particular properties are supported by the file format in question. These are referred to as their **significant properties** or **significant characteristics**. Tables and spreadsheets are both good ways of showing how this works. The question here is: does the file format allow tables to be stored with coloured cells and also enable formulae or options to be used for sorting or filtering data?

The *Guide* will contain a step-by-step plan for compiling a policy on preferred formats. This plan is based on the file format check in the *Guide to Preservation Policy* (see p. 10 above),[9] to which two steps have been added, specifically for preferred formats:

1. Make a list of the file formats present in a digital archive.
   Depending on the type of software used for the digital archive, this can be done either using the file extensions or with the aid of file identification tools.
2. Identify the preferred formats.[10]
3. Identify the properties of file formats.
4. Define a validation procedure for the ingest process.
5. Define a pre-migration validation procedure.
6. Define a post-migration validation procedure.

List with (human- and machine-readable) information on formats

As we said in the *Summary* on p. 3, the *Guide* will contain both information on file formats and a list of all the various formats currently in use.

Basically, the list will be an inventory of file formats in popular use, and will describe the durability of each one of these. This description will cover both the technical and the organisational aspects, as

[8] See for example the Data Archiving and Networked Services' (DANS) list of file formats (https://dans-labs.github.io/formats/). See also Appendix 1.

[9] *Guide to Preservation Policy*: https://wegwijzerduurzaamheidsbeleid.nl/.

[10] See the section on Preferred formats in Chapter 2: The problem.

well as information on the way in which the file format is used by the industry in question, based on a given level of knowledge about the file format.

The list of file formats should make clear to users what sort of file format they are dealing with, i.e. it should state the type and classification, include a technical description and state whether the file comes with an open specification or whether it is linked to certain software.

The purpose of the file format should be clear, as should the significant properties, i.e. the properties supported by the file format. The list should also specify which particular software is capable of opening a given format. If a file exists in a number of different versions, the list should state whether the differences have any bearing on the file's long-term accessibility.

The benefit of incorporating data links with PRONOM and Wikidata is that the list of preferred formats is readable not only by humans, but also by machines. The *Guide* can be adjusted to systematically include further (machine-readable) links with both internal and external databases and knowledge bases. One of these links will be with the NDE terminology network.[11]

One of the key requirements, if the list is to be easy to use, is that there should be a practical level of **granularity**. Let's take the TIFF and ZIP file extensions as an example. At the coarsest level of granularity, these are a file format for storing images (in the former case) and a compression format for packaging one or more files (in the latter case). At this very elementary level, there is a risk of missing certain essential information, such as properties that are dependent on a particular version. A search in the Wikidata database for files with a TIFF or ZIP extension will generate a very high level of granularity, i.e. it will produce a list of hits consisting of all versions and types of the two formats. The *Guide* will steer a middle course in these levels of granularity.

A description of all the sub-types of a particular format would not provide much more information than existing lists such as Wikidata, and would not offer users of the *Guide* a useful point of entry. It would make more sense to look for the information on TIFF at the basic level of the .tif extension, than by looking specifically for the PRONOM number of a given version such as 'fmt/353'.

For this reason, the list in the *Guide* should be geared towards a general level of granularity. Information on different versions or recommendations varying from one version to another should be given at the basic level. Detailed technical information on existing versions can be collected by creating data links to PRONOM.

Managing the *Guide to Preferred Formats*

With the exception of initiatives taken by individual heritage organisations of the 'core domain organisations',[12] there is no cross-sectoral concentration of information on file formats in the Netherlands. Given that the acquisition of a certain level of knowledge on file formats requires a big investment of both time and energy, this type of concentration would definitely be a good thing. A network making use of international knowledge of file formats would help to bridge the gap between different sectors.

[11] NDE terminology network: https://www.netwerkdigitaalerfgoed.nl/kennis-en-voorzieningen/digitaal-erfgoed-bruikbaar/termennetwerk/

[12] The 'core domain organisations' in the NDE are DANS KNAW (the Dutch institute for permanent access to digital research resources), the National Library of the Netherlands, the Dutch National Archives, the Dutch Institute for Sound and Vision, and the Cultural Heritage Agency.

In order to build this bridge, we will seek to concentrate the existing knowledge not simply by producing a *Guide to Preferred Formats*, but also by setting up an 'expert group on preferred formats'. The expert group will share and refine expertise on preferred formats within the Netherlands. It will:

- share knowledge, experiences and methods;
- promote professional development;
- coordinate activities among sectors and organisations based on a shared agenda;
- help to manage and refine the *Guide to Preferred Formats*.

By producing a *Guide to Preferred Formats* and setting up an expert group, the NDE will help to raise the level of knowledge among information professionals – from staff working for local authorities that wish to digitise planning permissions and store copies of approved documents so that they remain available for inspection on a long-term basis, to researchers wishing to ensure that research data remain available in order to guarantee the permanent reproducibility of their studies.

Glossary

List of terms relating to digital preservation, and preferred formats in particular, together with their definitions.

Accepted format

A format that a heritage organisation accepts for storage in its digital archive, subject to a certain (minimum) set of guarantees relating to long-term usability, accessibility and robustness.

Access copy

This is the same as a ‘consultation format’ and is derived from the preservation master file.

Bit preservation

Bit preservation refers to the activities that need to be performed in order to keep the bitstreams (i.e. the original sequences of zeros and ones) intact and readable.
(Source: https://wegwijzerduurzaamheidsbeleid.nl/artikelen/bit-preservering/)

COPTR

Stands for Community-Owned digital Preservation Tool Registry. COPTR is an inventory of tools used for long-term digital preservation.
(https://coptr.digipres.org/Main_Page).

Significant properties

Also known as significant characteristics, these are the properties that determine the appearance, behaviour, quality and usefulness of a digital object and which need to be preserved in order to safeguard the object’s future accessible and integrity. (https://lerenpreserveren.nl/topic/essentiele-kenmerken/)

Functional preservation

Functional preservation is intended to guarantee the long-term accessibility of digital sources by taking action to prevent data loss as a result of technological changes.
(https://wegwijzerduurzaamheidsbeleid.nl/artikelen/functionele-preservering/)

Granularity

In the context of preferred formats, granularity refers to the degree of detail in the description of a particular file format. At the lowest or most basic level of granularity, a file format is defined in terms of its extension. At the highest level of granularity, each version of the format is described in extensive detail.

Mezzanine file

See **Proxy file**.

Preferred format

A preferred format is the file format in which a digital archive that wishes to guarantee the preservation of its digital material prefers to store its digital objects.
(Source: https://lerenpreserveren.nl/woordenlijst/voorkeursformaat/)

Preservation format

A file format to which the organisation in question applies, or wishes to apply, functional preservation. The relevant files (known as master files) are at the top of a digital archive's hierarchy.

Preservation watch

Preservation watch involves monitoring new developments in relation to storage media, file formats, methods of presentation and other technological innovations in order to safeguard the authenticity, usefulness and long-term accessibility of digital collections. In addition to technological changes, other developments in a digital archive's operating environment may also pose a threat to its long-term accessibility. These could include changes in budgets, user requirements, etc.
(Source: https://wegwijzerduurzaamheidsbeleid.nl/artikelen/preservation_watch/)

PRONOM

PRONOM (http://www.nationalarchives.gov.uk/PRONOM/Default.aspx) is a database operated by the UK National Archives containing information on all file formats and their supporting software products of which the UK National Archives are aware. PRONOM holds information about software products, and the file formats which each product can read and write. You can search the database for file formats, software, vendors and life cycles, i.e. enter a date and then click 'search' to find all software products supported on that date.
(Source: https://lerenpreserveren.nl/topic/duurzame-bestandsformaten/)

Proxy file

A file derived from a file stored for the purpose of long-term preservation, i.e. a preservation master file or an archive master file. Proxy files are used for various purposes, such as viewing or generating the contents of the master file. They have a relatively low resolution and are not suited for reuse.

Other files used in addition to proxy files are **mezzanine files**. Mezzanine files (or just 'mezzanines') are intermediate format files based on a preservation master file. This intermediate format is in fact identical to the preservation master file and operates in practice as a sort of second preservation master file.
(Source: https://publications.beeldengeluid.nl/pub/387)

PUID

PUID stands for PRONOM Unique IDentifier. A PUID is the persistent, unique identifier of a record in the PRONOM database.

Wikidata

Wikidata (https://www.wikidata.org) is a free knowledge base that anyone can edit and which was designed to support Wikipedia (and other Wikimedia sister projects). It is the collaborative source of certain types of data, such as dates of birth, which are frequently used in articles in Wikipedia and other websites. Wikidata for Digital Preservation is a portal designed specifically for viewing structured data and adding structured data on software and file formats to Wikidata.
(Source: https://www.wikimedia.nl/pagina/wikipedia-en-meer)
(Source: https://wikidp.org/about)

Appendices

Related initiatives

DANS Lab op GitHub: Preferred Formats

DANS (Data Archiving and Networked Services) is planning to set up a GitHub[13] development platform to bring together expertise, information and guidelines from all over the world on preferred formats. The objective is to build a collaborative, multi-organisational infrastructure for inserting and editing information on preferred formats and exporting the same information in a systematic, standardised manner to other systems. The idea is for the platform to link up with existing initiatives and offer experts from a range of organisations an opportunity to contribute their expertise in certain formats. The GitHub will also act as a repository for general information from PRONOM and Wikidata. This will enable it to serve as an information model listing the objective and characteristic components of each format, i.e. name, extensions, relationship with other files, type of format, significant properties, openness, support, reference to documentation and level of knowledge, identifier, etc. Any organisation that has something to say about the format can be linked to it or contribute input from its own specific perspective. The GitHub will have a machine-readable data link with external sources such as PRONOM and the COPTR registry. For users, the contents of the GitHub can be rendered into a human-readable HTML or exported with the aid of SKOS.

Some examples of Dutch and international initiatives

A great deal is already known about formats and their specifications. These are a few examples of Dutch and international initiatives in this connection:

- Wikidata, especially Wikidata for Digital Preservation[14]
- PRONOM[15]
- Library of Congress > format descriptions[16]
- U.S. National Archives > Tables of File Formats[17]
- Preforma (PREservation FORMAts for culture information/e-archives) > Project Overview[18]
- Archaeology Data Service / Digital Antiquity > Guides to Good Practice[19]
- Rotterdam Municipal Archives > File formats in the e-archive[20]
- Digital Preservation Coalition > Knowledge Base: File Formats[21]

[13] DANS Lab on GitHub: https://dans-labs.github.io/formats/
[14] https://www.wikidata.org en https://wikidp.org/
[15] https://www.nationalarchives.gov.uk/PRONOM/Default.aspx
[16] https://www.loc.gov/preservation/digital/formats/fdd/descriptions.shtml
[17] https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html
[18] http://www.preforma-project.eu/project.html
[19] https://guides.archaeologydataservice.ac.uk/g2gpwiki/
[20] https://stadsarchief.rotterdam.nl/diensten/e-depot/bestandsformaten-e-depot/
[21] https://www.dpconline.org/knowledge-base/tags/109-file-formats

Preferred formats at the five core domain organisations in the NDE

So how do other people 'do it'? What choices do other organisations make in terms of their policies on formats in general, and preferred formats in particular? As a sort of 'taster' for the future full-length *Guide to Preferred Formats*, the following section describes the policies on formats in general, and preferred formats in particular, adopted by four of the five 'core domain organisations' in the NDE, i.e. DANS KNAW (the Dutch institute for permanent access to digital research resources), the National Library of the Netherlands, the Dutch National Archives, and the Dutch Institute for Sound and Vision.

DANS KNAW (Dutch institute for permanent access to digital research resources)

Preferred formats are those file formats which DANS KNAW believes provide the best long-term guarantees in terms of usability, accessibility and preservability. DANS KNAW automatically accepts all research data presented in preferred formats. […] As a general guideline, DANS KNAW states that those file formats that are best suited to long-term preservation and access:

- are frequently used;
- have open specifications;
- are not dependent on any specific software, designers or vendors.

In practice, however, it is not always possible to use formats that comply with all these requirements.[22]

National Library of the Netherlands

Although the digital repository contains a huge range of file formats, these are subjected only to limited checks and identification procedures. […] For this reason, the Library wishes to improve the procedures for checking and identifying the files held in the digital repository.

In other words, the Library needs to improve the procedures for identifying, checking and validating file formats and extracting technical metadata from them. [....] In order to do this, however, the Library needs to know more about file formats. Due to the large number of file formats in circulation, it cannot immediately upgrade its knowledge of all file formats and has decided for this reason to work with 'levels of knowledge'. This allows the Library to communicate clearly with the designated community about its knowledge of file formats and to make clear which formats will be stored and preserved.

A record is kept of the status of each file format and of the type of action that needs to be performed in order to move up to the next level of knowledge:

1. First level of knowledge: 'stored file format'
   In the case of stored file formats, the only checks performed are checksum tests to ascertain whether there has been any bit rot. Not much is known about the format as it cannot be identified.
2. Second level of knowledge: 'identified file format'
   An identified file format has a PRONOM ID. There is a tool for recognising files and adding a PRONOM ID to them. This is the first, basic step in the process of long-term preservation.

[22] Source: https://dans.knaw.nl/nl/over/diensten/easy/toelichting-data-deponeren/voor-het-deponeren/bestandsformaten

Although it is the first step on the road to functional preservation, it does not in itself guarantee functional preservation. However, it entails more than simply bit preservation.

3. Third level of knowledge: 'familiar file format'
   A familiar file format can be effectively stored as the Library is able to interpret the results of identification, validation and technical metadata extraction and has formulated guidelines for each format. […] The presence of this information means that the file may be regarded as being subject to long-term functional preservation.[23]

National Archives

The Dutch National Archives prefer to receive digital material for their e-repository in the form of 'open' formats. The National Archives also work with accepted formats in addition to preferred formats.

- Preferred formats: these are the 'open' formats. There may also be 'industry standards' for certain file types that are in common use and are sufficiently documented.
- Accepted formats: although not fully 'open', these file formats are nonetheless acceptable, for example because either the National Archives or another Dutch or foreign organisation has gained sufficient experience with them or has developed an effective preservation strategy for them.

According to the Standardisation Forum, an 'open standard' means that documentation must be easily available, there may not be any obstacles stemming from intellectual property rights (such as patent royalties), stakeholders must be able to have a say in their development, and the relevant standardisation organisation must be both independent and of a permanent nature.[24]

Dutch Institute for Sound and Vision

The Dutch Institute for Sound and Vision has designated a limited number of formats as qualifying for preservation. Its digital archive is able to issue carefully defined preservation guarantees for the formats in question. The Institute uses a number of preset criteria in order to define these preservable formats:

- The format should be a well-documented industry standard that operates on software in current use in the AV industry;
- The format must be able to be indexed within the Institute's technical and catalogue infrastructure, so that derivative files can be produced for viewing and downloading.
- In other words, the standard must be able to support media-related features such as time codes, subtitles and metadata.
- It must also be possible to transcode the format using commonly available transcoding software into other, popular formats.
- Finally, it must be possible to perform quality analyses on the format using standard analysis software.[25]

[23] Source: internal National Library of the Netherlands guidelines on file formats

[24] Source: https://www.nationaalarchief.nl/archiveren/kennisbank/handreiking-voorkeursformaten-nationaal-archief

[25] Source: https://publications.beeldengeluid.nl/pub/387

Credits

Sam Alloing
Digital Preservation Officer, National Library of the Netherlands

Valentijn Gilissen
Senior Data Manager, Preservation Officer, Data Steward, Data Archiving and Networked Services (DANS)

Hans Laagland
Director of Preservation, Tresoar, Regional Historical Centre for Friesland

Marjolein Steeman
Preservation Information Officer, Dutch Institute for Sound and Vision

Walter Swagemakers
Senior Collections Project Manager, Eye Filmmuseum

This white paper was published by the Dutch Digital Heritage Network (NDE) in June 2020.
For further information on the NDE, see: www.netwerkdigitaalerfgoed.nl/en.
Please email any comments or suggestions to: info@netwerkdigitaalerfgoed.nl.

netwerk
digitaal
erfgoed