RIGSARKIVET

# Memorandum

**Regarding:** Guide to Format Assessment
**To:** New Preservation Concept
**From:** Asbjørn Skødt

Date: 29. May 2020
Reference number: 19/06296

## Table of contents

## 1. Introduction

The purpose of the memorandum is to create a methodical framework and formalisation of the Danish National Archives' ongoing work concerning and discussing the approval of new formats for long-term preservation in the Danish National Archives' preservation system.

It is a condition that the use of data by authorities and private archive creators has a considerable range in the selection of formats, and the use is often not based on a specific strategy for selecting formats. The preservation of both the amount in formats as well as the formats' significant properties entails significant complexities for an archive's ability to safeguard digital preservation, and consequently there are two opposing forces in choosing a migration strategy. It results in that the ability to provide access to archived data in the future as well as maintain cost-effective preservation is significantly more secure if data is homogeneous and thus based on a few selected formats. However, homogeneity often occurs at the expense of authenticity, and it is in this area of tension that it makes sense to create the framework to allow new formats to be included within the collections of the Danish National Archives[1] on an informed and documented basis, so that the Danish National Archives may make authentic data available to researchers and other archive users as far in the future as possible.

For this purpose, a number of criteria for assessing formats have been selected and framed in a matrix. The memorandum serves as a guide to the methodological application of *Matrix_Format Assessment v2.1*[2]. The result of the method is a concrete recommendation that can be included in the decision to approve a new preservation format at the management level. The matrix is therefore not authoritative or a basis for automatic decisions.

[1] In the form of standalone archive versions with system independent data.

[2] See document 19/06296-32.

## 2. Method

The section describes the method itself and how it is used. The method is linked to completing the previously mentioned *Matrix_Format Assessment v2.1.*

The criteria each have a brief description, which explains what it takes to be "checked off", i.e. obtain the value 1. The indication of value is of a Boolean type, such that 0 is stated if the statement in the description is not true and 1 is stated if it is true. Estimates may be needed to assess the criteria. It is also possible to "flag" one or more criteria with an * in the form, which indicates that the formats do not initially fit in with the guide of the criteria, or there are special conditions that we need to be aware of that require elaboration. The special conditions must be elaborated in the supplementary note, which aims to explain the assessment of the formats in the form.

The weighting of the significance of the criteria ranges between values 1, 2 and 3. Value 1 means that the criterion is not considered to be particularly significant. Value 2 means that the criterion is considered to be important. Value 3 means that the criterion is decisive – and in the event that the criterion is not met, the format is considered disqualified – unless there are some special circumstances that apply. This is visible in the matrix in that the cells are marked red.

Thus, the guide primarily comes to its own in the following sections, and it will be a prerequisite for proper use of the form to consult the explanations and the meanings of the values. The form concludes with an overall assessment that adds the criteria values according to weighting. If the matrix is used digitally, the matrix itself adds the sum of the values of the criteria. Based on the overall assessment, a recommendation must be written in a memorandum (previously mentioned) explaining the results, as well as the formats approved as preservation format.

## 3. Criteria

The following sections describe the individual criteria for selecting preservation formats as well as the relative importance (weighting) of the criteria. The descriptions can use the concept of *content type*[3], which defines a delineated area of data for the same use, e.g. images and video are two different content types.

### 3.1. Prevalence

The format is frequently used within its content type. It is a strength for preservation formats to be frequently used across user segments for which the format is created. The more international and the more general the prevalence, the more established the format can be considered to be, which is a strong indicator in relation to preservation suitability.

[3] Content type is mentioned for the first time within the framework of the project's conceptual model v1.0, 2019.

The criterion is weighted 2, as general prevalence is an important indicator for how established a format is and opens up technical support for many years to come.

### 3.2. Prevalence II

The format is frequently used for preservation at cultural heritage institutions. It is a strength for preservation formats to be already in use at other cultural heritage preservation institutions. This occurs because like-minded actors in unison face the same challenges, thus creating opportunities for sparring and resource pooling as well as reducing the risk of technical obsolescence.

The criterion is weighted 2, as the prevalence of a format for preservation purposes is a strong basis for cooperation on common issues.

### 3.3. Lifespan

The format has been around for at least 10 years. An established format is more robust than a new format that has not been around for a long time, where it can be difficult to assess the direction in which the prevalence will go.

The criterion is weighted 1, as lifespan is to a larger extent an indication rather than something tangible.

### 3.4. Lifespan II

The format has good future prospects. The criterion is difficult to quantify, but “good future prospects” are based on an estimate on whether the prevalence and software support will increase over the coming years rather than decline. The criterion is a balance to the aforementioned criterion *3.3 Lifespan* and tells something more about the development of the format rather than just age.

The criterion is weighted 3, since whether a format is growing rather than dying is crucial to assessing the preservation suitability of the format.

### 3.5. Documentation

The format is standardised. It is a strength for preservation formats to be technically well-documented. The documentation makes it possible to analyse the functionality of a format and to develop system tools for characterisation, validation, and migration.

The criterion is weighted 2, as it is an insurmountable task in terms of resources to disassemble a format for the purpose of developing the system tools for receiving and storing data.

### 3.6. Documentation II

The format standard has not been updated within the last 10 years. The more stable a format, the easier it is to integrate it into the preservation planning since the prioritisation of resources for monitoring and analysing the format decreases.

The criterion is weighted 1, as it is only an indication of stability and the challenge can be solved by adding the necessary resources to monitor and analyse the format more thoroughly.

### 3.7. Documentation III

The documentation of the format is well-described and can be read easily. It is a strength for preservation formats to have well-described and clear documentation written in an understandable language. At the same time the documentation must not be too lengthy to read. Here we set the limit at 1,000 pages, which as to whether it is too high may be debatable.

The criterion is weighted 1, as we do not necessarily have to understand the documentation in all its details.

### 3.8. Licensing

The format is open source. It is a strength for preservation formats to be independent of companies including the strategic and financial interests of companies. In practice, the strength comes from the fact that working with the formats is cost-effective, the future holds fewer risks and the possibility that the format will become more prevalent over time is greater.

The criterion is weighted 2, as the free use of the format is a sign of health and an important factor behind the creation of an archival academic community around the format. Vise versa a solution for the approval of licensed formats may be to allocate the necessary resources that it costs to procure and operate licenses.

### 3.9. Structure

The format is self-supporting (not a container format). The criterion means that a format consists solely of its own format. Some formats allow multiple formats to be stored in them, making the format a container for other formats. Examples of container formats are video formats that store images in one format, audio in another format, and any subtitle tracks in a third format. Zip is also a container format that can have all sorts of formats stored in it.

The criterion is weighted 1, as container formats only add the implication (but also complexity) that we must approve all the formats of the container as preservation formats.

### 3.10. Structure II

The format can be read as plain text. It is a strength for preservation formats if both machines and people can interpret the data directly. Plain text means any basic text program can open the file, interpret the character set, and render the binary data as readable text.  Formats that cannot be read as plain text will not produce readable text after the interpretation of the character set.

The criterion is weighted 1, as binary formats are not problematic as long as the technical documentation is saved, and the migration strategy already entails data to be migrated if they are at risk of becoming technically obsolete.

### 3.11. Significant properties

The format supports most of the significant properties within its content type. In parallel with the method for investigating significant properties[4] the format must also be graded here. The investigation of the significant properties of the format is therefore a prerequisite for grading. In short, significant properties are the connection of a format's technical properties and structure with the use of the format by a designated community. Here it is an ideal if a preservation format has all the properties of the original format so that significant properties are not lost during migration.

The criterion is weighted 3, as significant properties in a format are associated with the actual use of the archives creator, and thus the properties are an essential indicator of whether data is preserved as it is created.

### 3.12. Dissemination

The format can be reused without conversion. If it is necessary to convert data from the preservation format in order to make data available, it will then add an extra work process and risk of loss of authenticity, which would be undesirable.

The criterion is weighted 1, as data conversion is recognised to be problematic but nonetheless, conversion is common practice in the migration strategy and cannot be weighted higher.

### 3.13. Searchability

The format has searchable information. If the format allows for searches within the content, it is easier to index the content and make it available. The example here is a TIFF of a digital document, which exclusively stores an image of the document, which must subsequently be OCR-treated to have searchable text in a separate text or database file. A PDF version of the same digital document will have the text stored as embedded searchable information.

The criterion is weighted 1, as the technical possibilities of processing offered by the format have an impact on the ability of an archive to preserve and make available data stored in the format in a cost-effective and authentic way. However, there are often tools that can mechanically remedy the lack of searchability.

[4] See memorandum *InSPECT framework til vurdering af signifikante egenskaber*, Danish National Archives, 2019

### 3.14. Interoperability

The format is suitable for data exchange. Some formats are more suitable for data exchange than others. This applies for instance to mark-up formats such as XML, or formats that are widely used for exporting and importing data between IT systems.

The criterion is weighted 2, as interoperability makes it easier to work with data in automated processes, and broad system support for data exchange is an indication of robustness.

### 3.15. Testing

The format has tools for identification and characterisation. The verification of format identity makes it possible to ensure that data are transferred to the Danish National Archives according to the current executive order and supports subsequent preservation actions such as migration. The purpose is to gain knowledge about whether a file has the format it claims to have through its file extension as well as gain knowledge about the metadata of the file.

It is optimal if the tool is already mature for use in a fully developed and open version, where the developers still offer bug fixes and updates for new operating systems. An open version means that the tool can be freely adapted to suit your own business needs.

The criterion is weighted 2, as identification and characterisation are important components of the testing process that, in a very basic way, allow for more complex preservation tasks.

### 3.16. Testing II

The format has tools for validation. The validation of the structure and content of files makes it possible to ensure that data is transferred to the Danish National Archives according to the current executive order, so that, for example, specified bit depths, standards and relationships are complied with.

Here it is optimal if the tool is already mature for use in a fully developed and open version, where the developers still offer bug fixes and updates for new operating systems. An open version means that the tool can be freely adapted to suit your own business needs.

The criterion is weighted 2, since validation is an important component of the testing process that allow for the performance of more complex preservation tasks in a very basic way.

### 3.17. Compression

The format is uncompressed or has lossless compression. Compression can cause the quality of data to be reduced so that they do not have the same accuracy as an original format may have had. However, compression may be allowed if the compression is lossless.

The criterion is weighted 3, as lossless compression is a crucial factor in ensuring authenticity.

### 3.18. Storage

The format occupies less storage space than the average for its content type. Storage is the capacity in the physical media that the preservation formats occupy. Storage is a relative factor that is affected by technological development, which over time offers greater capacity at the same price, but the criterion's significance is also affected by the budgetary framework of the preservation institution.

The criterion is weighted 1 because storage space does not have the same economic significance as in the past.

### 3.19. Migration

The format can be migrated with acceptable loss of significant properties to another format. Migration here is defined as converting data from one original format to another format within the same content type.

Next, the availability of data migration tools is a prerequisite for pursuing a migration strategy, and the better the tools the market can offer, the stronger the format's preservation suitability. In an ideal scenario, migration should result in the retention of all significant properties, the same intellectual content, and the same visual representation.

The criterion is weighted 2, as the ability to migrate data to other formats indicates robustness and in the short term means that we do not lock ourselves to the format.

### 3.20. Compatibility

The format is compatible with multiple operating systems and applications. In practice, this means that the format is both system and software independent without locking it to specific operating systems or programs, and several competing programs must exist in the market for rendering the data of the format.

The criterion is weighted 1, as system independence have an impact on the possibility of pursuing a migration strategy.

## 4. Overall assessment

The overall assessment is the sum of the values multiplied by their weighting and is also indicated as a normalised score between 0 and 1[5]. The format that achieves the highest score (without disqualifying criteria) is considered to be the most suitable preservation format, but if several formats achieve an overall rating close to each other, they can be juxtaposed. The overall assessment results in a recommendation written in note format, which can be taken further as the basis for a subjective

[5] Normalisation is done as a min-max normalisation, where the normalised values are calculated on a scale between 0 and 1, where the lowest value is set to 0 and the highest value is set to 1. In practice, normalisation is done by calculating the normalised value as (x-min(x))/(max(0)-min(x)).

management decision for the approval of a new preservation format. The recommendation is the end result of the format assessment.

## 5. Conclusion

The memorandum provides guidance in the use of the *Matrix_Format Assessment v2.1*, and through a quantified method, it enables the Danish National Archives to assess formats for approval as preservation formats for which data can be submitted in by authorities and private actors. The method is a contribution of an approval procedure and solely provides a professional recommendation based on a selection of criteria that relate to how the format behaves technically within a wider societal context.

## 6. Appendix

The assessment of the criteria that are usable in the context of the Danish National Archives' digital archiving strategy has been based on a study of the practice that the Danish National Archives has previously used as well as what practices are used at other digital preservation institutions in an international context. The following two appendices list the inspiration for the selection of criteria as well as the reason for the rejection of criteria.

### 6.1. Sources of inspiration

- Benjamin Yousefi (2017): *InterPARES*, The Swedish National Archives, Sweden
- Brown, Adrian (2008): *Selecting File Formats for Long-Term Preservation*, The National Archives, England
- Johansen, Kathrine Hougaard Edsen (2015): *Udvalgte formater og deres egenskaber*, retrieved on 20 August 2019 from: https://digitalbevaring.dk/viden/filformater-bevaring/
- *Katalog Archivischer Dateiformate* (2015), The Swiss Federal Archives, version 4.1, Switzerland
- *Kriterier for filformater*, The Danish National Archives
- *Filformatmatrice* (2009), The Danish National Archives
- *Sustainability of Digital Formats: Planning for Library of Congress Collections*, Library of Congress, USA
- *Digital Preservation Handbook* (2015), Digital Preservation Coalition, retrieved on 5 August 2019 from: https://www.dpconline.org/handbook/technical-solutions-and-tools/file-formats-and-standards

### 6.2. Rejected criteria

The list below specifies identified criteria from the sources of inspiration that were rejected for the form and a brief justification is given.

| Criterion | Explanation | Reason for rejection |
|---|---|---|

| | | |
|---|---|---|
| **Metadata** | Ample opportunities for recording embedded metadata. | It is a better practice to keep metadata in a separate table. |
| **Format class** | The format is well established within its content type. | Overlap with “prevalence”. |
| **Complexity** | The format is at risk of complexity resulting in deficiencies or misinterpretation of the format. | The criterion is only applied to video formats, and we want generally valid criteria. Also overlap with “significant properties” (possibly an antonym. Paradox) |
| **Scaling** | The format can be easily scaled upon migration. | Overlap with “storage” and “migration” |
| **Destination format** | The format has several specific formats that can be migrated to. | Overlap with “migration”. |
| **Authenticity** | The format can be migrated without loss of authenticity. | Overlap with “Significant properties” and “migration”. |
| **Recreation** | The format helps us to recreate data more authentically (higher scores on significant properties) than other formats. | Overlap with “Dissemination” and “significant properties”. |