RIGSARKIVET

Memorandum

Title: P3 Testing v1.0
To: New Preservation Concept
From: Asbjørn Skødt

Date: 28. oktober 2020
Case no.: 19/06296

1. Introduction

The purpose of the step is to test existing software for identification, characterisation, conversion and validation of the file formats, which the step P2 Format Assessment finds most suitable as preservation formats. P2 Format Assessment typically leads to the recommendation of 1-2 file formats as new preservation formats.

The practical experience with test of formats in existing software must be collated in a report, which describes the approach and the results. The new knowledge can lead to revising assessment of criteria in P2 Format Assessment. In addition, the knowledge can be applied in the step P4 Consequence Assessment, which measures consequences in time, economy and quality, and in the step P5 Preservation Plan with preliminary specifications for validation requirements.

2. Test setup

The test setup must be described. The description includes what software and software settings were used and which data were used, their origin and whether the data possessed certain properties. The test setup must be described for every single test. It is recommended that the test setup includes the table of significant properties from the step P1 Migration Assessment for the purpose of testing, whether the individual significant properties can be characterized, validated and whether they are lost in conversion.

The most suitable software for the purpose should be applied. Software could be created by vendors tailoring their software to the archival business, or it could be standard software used in other contexts. It is necessary to make investigate what software are accessible and most suitable. Suitability is a judgment call but the reasons should be stated in the report.

3. Identification and characterisation

This test generates knowledge concerning the possible identification and characterization of candidate formats and this knowledge can be applied in the selection of validation procedures, extracting metadata and mapping significant properties.

Identification happens, as a minimum, through the file extension or optimally through a "magical signature". A magical signature is a unique sequence of hex-number, which only appears in a certain file format and this allows for an undisputed identification of file format. Before testing, the magical signature needs to be determined. Characterisation is typically part of the same process but besides identification, this allows the extraction of metadata from the files and potentially mapping of significant properties.

If suitable software does not exist, it must be noted in the report that the test could not be conducted. Known software are e.g. Exiftool, FITS, Siegfried, FIDO, DROID and Apache Tika.

Tests with identification and characterization does not prerequisite a large dataset but all variants of the file format must be tested e.g., Excel spreadsheets have multiple file format extensions: .xlsx, .xlsm, .xltm, .xltx and .xlsb.

4. Conversion

This test generates knowledge concerning the possible loss of significant properties that may happen when converting data to or between the candidate formats. This knowledge are applied in the step P4 Consequence Assessment for the qualitative measurements. Loss of significant properties through conversion must be collated in a table in the report.

If suitable software does not exist, it must be noted in the report that the test could not be conducted, and that it will be necessary to find funds for the development of a suitable conversion tool. Known software are e.g. ImagMagick (for images).

Tests with conversion prerequisite a large and diverse dataset.

5. Validation

This test generates knowledge concerning possible validation tools for the candidate formats. Validation means to confirm that files are created in accordance with the specifications of the format and the possible submission requirements of your archive. If the files are not compliant, the tool will inform and the files should be corrected and resubmitted for validation. Tests with validation should apply a preliminary assessment of relevant validation requirements. These requirements should be collated in a table in the report.

If suitable software does not exist, it must be noted in the report that the test could not be conducted, and that it will be necessary to find funds for the development of a suitable conversion tool. Known software are e.g. JHOVE and veraPDF.

Tests with validation prerequisite a dataset possessing the specifications and properties listed as validation requirements. Therefore, the dataset does not need to be big, but it should be specialized and precise.

6. Conclusion

P3 Testing delivers a report describing the test setup and results for the rest with software for identification, characterisation, conversion and validation of the candidate file formats, which the step P2 Format Assessment recommends as preservation formats. The test report assesses what software are available for the mentioned purposes, whether they should be customised or the development of new software are necessary.

The knowledge is applied in the steps P4 Consequence Asssessment, P5 Preservation Plan and potentially for the revision of existing criteria assessment in the step P2 Format Assessment.