KANSALLISARKISTO
RIKSARKIVET

2 March
2018

REQUIREMENTS SET BY THE NATIONAL ARCHIVES OF FINLAND FOR DIGITISATION PROCESS ENABLING DISPOSING THE ANALOGUE MANIFESTATION (DRAFT)

Content Requirements set by the National Archives of Finland for digitisation process enabling disposing of the analogue manifestation after it has been converted into digital format.

Limitations This document describes the digitisation process and its results. This document does not describe the actual storage package saved in the long-term storage system. Packages saved in the long-term storage system can be generated from the results of the digitisation process presented here. This document does not describe the process to dispose analogue manifestations.

Purpose The purpose of this document is to ensure the preservation of the data content of records which are part of the cultural heritage and the usability of such documents after their analogue manifestation has been destroyed.

Target group Requirements presented in this document are intended for the National Archives of Finland and other parties operating in the public sector, aiming to dispose the analogue manifestation after it has been digitised.

Regulations that authorise the issuance of the regulation

Section 14a of the Archives Act (831/1994)

Validity Until further notice

National Archives of Finland Rauhankatu 17 P.O. Box 258, 00171 Helsinki Tel. +358 29 533 7000 kirjaamo@arkisto.fi
Riksarkivet Fredsgatan 17 PB 258, 00171 Helsingfors Fax +358 9 176 302 http://www.arkisto.fi

Table of contents

1 Terms and concepts

The terms used in this document are based on the specification [RFC 2119] prepared by the Internet Engineering Task Force.[1]

Table 1. Concepts

CONCEPT	DESCRIPTION
Digital record	Digital version of an analogue record produced by means of the digitisation process.
Digital manifestation	A digital version of an analogue entity decided to be digitised. Digital record is a part of this entity.
Digitisation process	A group of functions, with which an analogue manifestation is converted into digital format.
Analogue manifestation	An analogue version of an analogue entity decided to be digitised. In this document analogue manifestation mainly contain A4/folio size paper documents but which may also include larger or smaller documents.
Archival master file[2]	An archival master file is a bitmap image produced in the digitisation process. A digital object of the highest quality in terms of its technical properties produced in the digitisation process.
Access file[3]	An access file is a bitmap image produced in the digitisation process that is offered for use, for example, via an online interface. In general, an access

[1] https://www.ietf.org/rfc/rfc2119.txt, Accessed 22 February 2018
[2] Federal Agencies Digital Guidelines Initiative -> Archival Master.
http://www.digitizationguidelines.gov/term.php?term=archivalmasterfile, Accessed 28 December 2017
[3] Federal Agencies Digital Guidelines Initiative -> Derivative file.
http://www.digitizationguidelines.gov/term.php?term=derivativefile, Accessed 28 December 2017

	file has content identical to that of an archival master file, but its information is presented in a compressed file format.
Production day	A day during which digital records are produced using equipment.
Main orientation	The main orientation allows the written content of a digitised record to be interpreted without turning the image file. If the record includes written content in several orientations, the main orientation is the orientation in which most of the written content can be read.

2 Introduction

Digitisation enabling disposing means that the analogue manifestation is disposed at the end of the digitisation process. The purpose is not to destroy the data content of analogue records , but to convert the data content into a different format (digital). When a record defined for permanent retention is converted into digital format in the digitisation process, its data content defined for permanent retention is preserved. Disposing an analogue manifestation requires that the digitisation process has been carried out using methods that do not decrease the evidential value, integrity or authenticity of the record.

The criteria presented here MUST be followed when government agencies digitise records defined for permanent retention and the analogue version is disposed after digitisation. The acceptance of digitised material in the data systems of the National Archives of Finland requires that the material meets the requirements set out in this document. Material that does not fulfil the requirements set out in this document will not be accepted in the data systems of the National Archives of Finland.

Standards used generally in the archives sector and the quality requirements set by other national archives for digitisation have been taken into account in the preparation of this document. In addition, the specifications of The National Digital Library's long-term storage specifications have been taken into account in sections 5 (Accepted formats) and 6 (Package generated in the digitisation process).[4]

This document focuses on the digitisation of various records into image files, and the processing and saving of content recognised from them by means of various technologies. This document does not describe the digitisation of audio or video. This document is obligatory under section 14a of the Archives Act (831/1994).[5]

3 General requirements for the digitisation process

The analogue material to be digitised MUST have a screening decision issued by the National Archives of Finland, specifying the preservation format for document data. If no such decision

[4] National Digital Library -> Long-term storage -> Specifications http://www.kdk.fi/en/digital-preservation/specifications, Accessed 19 February 2018

[5] Archives Act https://www.finlex.fi/fi/laki/ajantasa/1994/19940831, Accessed 19 February 2018

exists, records MUST be retained also in analogue format after digitisation, even if the process and results were as described in this document.

The conversion of analogue material into digital format is a process (digitisation process) that MUST be documented in the manner and to the extent defined here. The goal of the process is to produce authentic and intact digital manifestations.

The digitisation process MUST ensure that material to be digitised is digitised whole and with a full content. In practice, this means that all collections/series/units/records MUST be digitised so that no information is left non-digitised due to a technical or functional error.

Every single image file related to specific material MUST contain the same information as its analogue manifestation and it MUST be visually available. The image file MUST NOT include any elements that are not included in the analogue record. Exceptions are made by test targets scanned/photographed in a single image file that indicate the colours, grey tones, dimensions and resolution of the analogue record that is digitised. These MUST be positioned so that they do not cover the analogue record that is digitised.

Any blank reverse sides of documents SHOULD be removed during the digitisation process. A blank reverse side is a document page that does not contain any information. Pages that contain any information MUST NOT be removed. Image files produced during the digitisation process MUST be turned in the main orientation. Files produced during the digitisation process MAY only be turned in steps of ninety degrees after scanning.

Before scanning, the performance of the digitisation infrastructure SHOULD be optimised. After optimisation, the quality of the digital archival master files produced by the infrastructure MUST be verified by using test targets designed for this purpose. Quality MUST be verified once every production day, and quality values must fulfil their reference values. Acceptable reference values are Metamorfoze Extra Light and FAGDI 2 Star (Unbound).[6] The reference values will be specified during the further preparation of requirements in 2018.

[6] Values defined in the following documents: FADGI (Documents Unbound: General collections, 2 Star): http://www.digitizationguidelines.gov/guidelines/FADGI%20Federal%20%20Agencies%20Digital%20Guidelines%20Initiative-2016%20Final_rev1.pdf (accessed 10 January 2018) and Metamorfoze (Extra light): https://www.metamorfoze.nl/sites/metamorfoze.nl/files/publicatie_documenten/Metamorfoze_Preservation_Imaging_Guidelines_1.0.pdf (accessed 10 January 2018)

4 General recommendations and good practices for the digitisation process

In general, digitisation is regarded as a process that includes the stages presented in Figure 1.

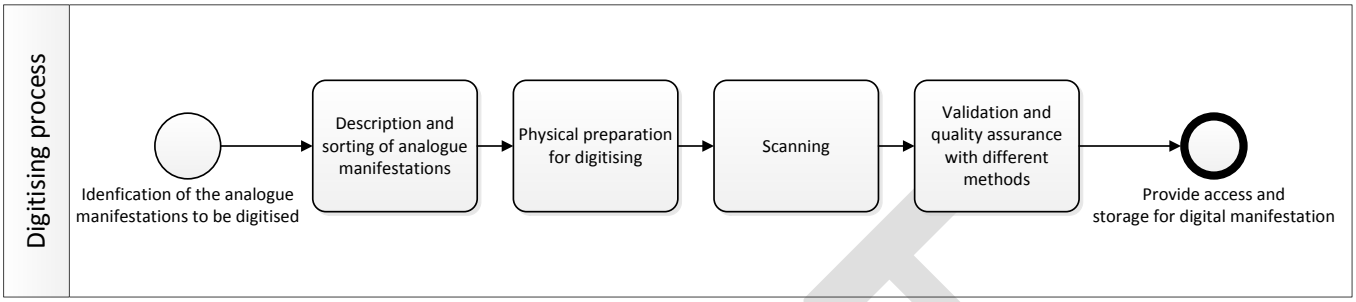

Figure 1. Conversion into digital format (general)

This section does not include any requirements. Instead, good practices related to scanning and quality assurance are presented here.

The process presented in Figure 1 consists of several stages. In general, digitisation is carried out so that records are described in the metadata system before their conversion into digital format is started. This also enables the documentation of the workflow for analogue material. After scanning, material metadata may be enriched manually or automatically. These actions include the analysis of the bitmap image produced during scanning.

The scanning infrastructure usually consists of various hardware and software that enable the digitisation of various analogue materials. In addition, a large group of systems and technologies, each of which have a different role in enabling the process, is included in the workflow.

Scanning quality assurance can be roughly divided into actions taken before scanning and quality assurance after scanning, i.e. validation.

As stated in section 3, the performance of the infrastructure should be optimised before scanning so that the digital record produced represents the best possible version that can be produced using the specific configuration. After optimisation, the performance of the infrastructure must be monitored as planned so that the quality of digital records produced during the process remains stable. For monitoring purposes, a test target, reference values for the test target and analysis software are needed. In addition to image quality, attention should be paid, with regard to the equipment infrastructure, to ensuring that all the information presented in records is converted into digital format. This means that, when acquiring equipment, special attention should be paid to the ability of devices to separate records from one another in order not to feed two overlapping records through the scanner (document scanners, open-track scanners and other scanning solutions with which documents are scanned in large groups and the separation is done by devices).

After scanning, validation can be carried out using samples. The sample size depends on the reliability of the scanning process. There are various reference values and recommendations available in general. The goal of validation is to ensure that the requirements presented in section 3 are fulfilled. If material is machine-encoded, text can be recognised using optical character recognition (OCR) methods. This stage can also be used as an indicator of the success of scanning if limits for the success of recognition can be set in the application used.

If image files are processed after scanning, a detailed image processing history should be saved, at least in the metadata of image files. If possible, it should also be saved in XML data that document the generation of the digital object.

5 Accepted formats

This format section has been divided into three subsections:

1. Image formats
2. Storage format for recognised text
3. Metadata and structure of the metadata concerning archival master files and their processing

5.1 Image formats

File formats suitable for different purposes of use MUST be produced in the digitisation process. All image files MUST be saved as 24-bit RGB images. Both the access file and the archival master file MUST be produced so that the quality of neither is lower than what is defined in sections 5.1.1 and 5.1.2 at any processing stage. Tables 3 and 4 present required data that image files MUST contain in machine-readable format. If the "Element" column is not specified, the data must be indicated but, at this stage, a metadata field required for the data has not been specified. The access file MUST contain the same information as the archival master file when inspected visually. This limits the need to use archival master files, apart from migration and any other special action.

5.1.1 Access file

Table 3 below present's metadata that MUST be written in the access file (JPEG file) generated in the digitisation process. In addition to the data presented in the table, the access file MAY include other metadata fields.

Table 2. Metadata required for access files

Element	Specifier	Required value	Metadata schema	Metadata field
Format	No unambiguous field. Example: Dublin core.	JPEG	Dublin core	dc:format
Image name				
Image file size				
Colour model	RGB		Exif.Image	PhotometricInterpretation (262)
ICC profile	Colour profile saved in image file metadata	sRGB	ICC	profileDescription

Bit depth	The number of bits per sample	8 8 8	Exif.Image	BitsPerSample (258)
	The number of samples per pixel	3	Exif.Image	SamplesPerPixel (277)
File compression	JPEG quality 60 percent			
Image width	Image width as pixels per row		Exif.Image	ImageWidth (256)
Image length	Image length as the number of pixel rows in the image		Exif.Image	ImageLength(257)
Digitisation device (scanning or photography)	The make of the equipment used to convert an analogue record into digital format (name of manufacturer)		Exif.Image	Make (271)
Digitisation device model (scanning or photography)	Specifying information about digitisation equipment by indicating the name of the model		Exif.Image	Model (272)
Serial number of digitisation device			Exif.Image	CameraSerialNumber (50735)
Program used to generate image file	The application and version with which the access file was generated		Exif.Image	Software (305)
Date and time when digital image file was generated (scanning date)	In format: YYYY:MM:DD HH:MM:SS		Exif.Image	DateTimeOriginal (36867)
Orientation	File orientation (horizontal or vertical)		Exif.Image	Orientation (274)
Resolution unit	Inches		Exif.Image	Image.ResolutionUnit (296)
XResolution	The number of pixels per resolution unit in the image width direction	300	Exif.Image	Image.XResolution (282)
YResolution	The number of pixels per resolution unit in the image length direction	300	Exif.Image	Image.YResolution (283)
Image file processing software	The name of processing software if the image file is processed after scanning		Exif.Image	Image.ProcessingSoftware (11)

5.1.2 Archival master file

Table 4 below presents metadata that MUST be written in the archival master file (TIFF file) generated in the digitisation process. In addition to the data presented in the table, the archival master file MAY include other metadata fields.

Table 3. Metadata required for archival master files

Element	Specifier	Required value if can be expressed unambiguously	Metadata schema	Metadata field

Format	No unambiguous field. Example: Dublin core.	TIFF 6.0	Dublin core	dc:format
Image name				
Image file size				
Colour model	Image file colour model	2 = RGB	TIFF tag, baseline	PhotometricInterpretation (262)
ICC profile		sRGB, eciRGB v2, ProPhoto RGB, AdobeRGB (1998)	ICC	ICC Profile (34675)
Bit depth	The number of bits per sample	8 8 8	TIFF tag, baseline	BitsPerSample (258)
	The number of samples per pixel	3	TIFF tag, baseline	SamplesPerPixel (277)
File compression		5 = LZW	TIFF tag, baseline	Compression (259)
Image width	Image width as pixels per row		TIFF tag, baseline	ImageWidth (256)
Image length	Image length as the number of pixel rows in the image		TIFF tag, baseline	ImageLength (257)
Digital image file created by	Organisation (required), person (recommended)		TIFF tag, baseline	Artist (315)
Digitisation device (scanning or photography)	The make of the equipment used to convert an analogue object into digital format (name of manufacturer)		TIFF tag, baseline	Make (271)
Digitisation device model (scanning or photography)	Specifying information about digitisation equipment by indicating the name of the model		TIFF tag, baseline	Model (272)
Serial number of digitisation device			Private TIFF tags	CameraSerialNumber (50735)
Program used to generate digital image file	The application and version with which the archival master file was generated (required). Any file processing software separated ";" (recommended).		TIFF tag, baseline	Software (305)
Date and time when digital image file was generated (scanning date)	In format: YYYY:MM:DD HH:MM:SS		TIFF tag, baseline	DateTime (306)
Orientation	File orientation (horizontal or vertical)		TIFF tag, baseline	Orientation (274)
Resolution unit	Measuring unit used to interpret the X and Y resolution	2 = inch	TIFF tag, baseline	ResolutionUnit (296)
XResolution	The number of pixels per resolution unit in the width direction	300/1	TIFF tag, baseline	XResolution (282)
YResolution	The number of pixels per	300/1	TIFF tag,	YResolution (283)

	resolution unit in the length direction		baseline	
Byte order		Big-endian or little-endian		ByteOrder

5.2 Storage format for recognised text

If text is recognised from image files using, for example, OCR (optical character recognition) or HTR (handwritten text recognition) methods, it MUST be saved in Analyzed Layout and Text Object (ALTO) format[7] (version 3.0 or greater). A separate ALTO file MUST be saved for each record converted into digital format.

5.3 Metadata and structure of the metadata concerning archival master files and their processing

The metadata described in this section represents the creation history of archival master files. This history also verifies the authenticity of the digital manifestation generated in the process. The technical metadata required for archival master files MUST be presented in accordance with the MIX metadata schema, version 2.0.[8]

Table 5 below presents, from left to right, the name of the MIX field, the purpose of the field and the obligation. The Obligation field indicates whether or not the specific field and data in accordance with its schema are obligatory as follows:

- R = Required – this data MUST be described
- O = Optional – this data SHOULD be described but it is not required

The MIX metadata schema contains two types of fields: containers and data elements. Data elements contain a specific value, while containers contain one or more data elements, and they can also contain other containers containing data elements. Table 5 only presents fields that contain a specific value, i.e. data elements.

Table 4. Metadata describing archival master files and their processing (the table only presents fields that contain data that MUST be presented in a structure in accordance with the MIX metadata schema, version 2.0)

MIX field name	Field purpose	Obligation
objectIdentifierType	Data element that designates the system or domain in which the ID of a digital record is unique	R
objectIdentifierValue	A set of characters identifying a digital image	R
fileSize	File size in bytes, e.g. 72839	R
formatName	File format, e.g. image/TIFF	R
formatVersion	File version, e.g. 6.0	R, if possible

[7] The Library of Congress » Standards » ALTO. Website of the Library of Congress. Accessed 19 December 2017. https://www.loc.gov/standards/alto/

[8] The Library of Congress » Standards » MIX. Website of the Library of Congress http://www.loc.gov/standards/mix/

byteOrder	A data element that defines the order of bytes. The value is either big-endian or little-endian.	R
compressionScheme	Compression used. Example: uncompressed or LZW	R
compressionRatio	A data element that indicates the compression ratio	R, if possible
messageDigestAlgorithm	A data element that identifies the algorithm with which the value in the messageDigest field has been created. The field value is any of the following: MD5, SHA-1, SHA256, SHA384, SHA512	R
messageDigest	A data element that specifies the output by the algorithm defined in the messageDigestAlgorithm field, e.g. e8064dc0	R
imageWidth	Image width in pixels, e.g. 1330	R
imageHeight	Image height in pixels, e.g. 1600	R
colorSpace	A data element that defines the image colour space, e.g. RGB	R
iccProfileName	A data element that defines the generally used ICC profile name, e.g. eciRGB	R
iccProfileVersion	A data element that indicates the ICC profile version used, e.g. v.2 [eciRGB v.2]	R
iccProfileURL	If the ICC profile has not been documented properly, the URL/URN of the profile must be saved in this field.	R, if possible
dateTimeCreated	A data element that indicates when the image was created. In format: YYYY-MM-DD HH:MM:SS	R
imageProducer	A data element that indicates the organisation that created the image	R
scannerManufacturer	A data element that identifies the manufacturer of the scanning device used to create the image	R
scannerModelName	A data element that identifies the model name of the scanning device used to create the image	R
scannerModelNumber	A data element that identifies the model number of the scanning device used to create the image	R
scannerModelSerialNo	A data element that identifies the serial number of the scanning device used to create the image	R
scanningSoftwareName	Name of the scanning software used	R
scanningSoftwareVersionNo	Version of the scanning software used	R
orientation	A data element that indicates the image orientation	R
samplingFrequencyUnit	A data element that indicates the measuring unit used to interpret X and Y resolutions. Required value 2 = inch	R
xSamplingFrequency	The number of pixels per resolution unit in the width direction. Required value 300/1	R
ySamplingFrequency	The number of pixels per resolution unit in the height direction. Required value 300/1	R
bitsPerSampleValue	A data element that defines the number of bits per sample, e.g. 8 or 8 8 8	R
bitsPerSampleUnit	A data element that defines the interpretation of	R

	bits. The value is either integer or floating point.	
samplesPerPixel	A data element that defines the number of samples per pixel	R
targetType	A data element that indicates whether the scanning quality table is part of the image or whether it has been scanned in a separate image	O
targetManufacturer	A data element that indicates the manufacturer of the test target	O
targetName	A data element that indicates the name of the test target	O
targetNo	A data element that indicates the serial number of the test target	O
externalTarget	A data element that indicates the location of the digital image of the test target identified in the TargetID container	O
performanceData	A data element that indicates the location of the measured data of the test target identified in the TargetID container	O

6 Packages generated in the digitisation process

The different files created in the digitisation process as presented in section 5 and its subsections MUST be saved in the directory structure presented in Figure 2 so that they can be transferred to the National Archives of Finland. The digital manifestation MUST be produced in the directory structure, regardless of when it is transferred to the National Archives of Finland. If material is not transferred to the National Archives of Finland at any stage, the directory structure is OPTIONAL. In addition to the directory structure defined here, organisations MAY, for example, save access files in their own data systems using the data structure required by each specific system. Therefore, the structure defined in this document does not exclude the use of any other storage structures.

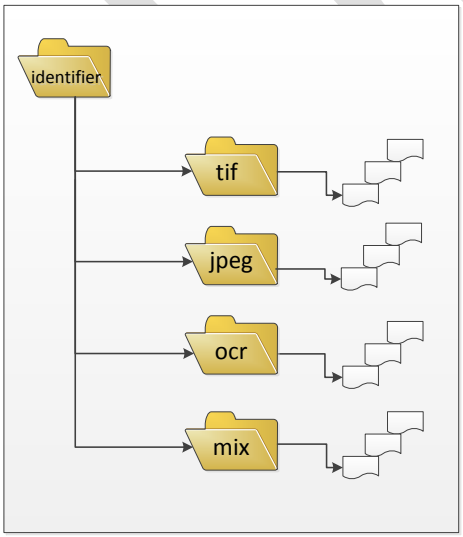

Figure 2. The transfer package structure required for the digitisation process

Table 6 presents how files MUST be named inside the directory structure presented in Figure 2. Digital formats produced in the process MUST match each other. In other words, access file 0001.jpg MUST contain the same information as archival master file 0001.tif. Furthermore, AltoXML file 0001.xml MUST contain content recognised from bitmap image 0001. Finally, 0001.xml file in accordance with the MIX metadata schema MUST represent archival master file 0001.tif.

Table 5. Content of transfer package directories

Directory	Description
identifier	An identifier of a digital manifestation, with which it MUST be possible to identify the document entity in question (e.g. archival unit).[9] The directory includes object directories.
tif	The archival master files presented in Table 3 MUST be saved in the directory as individual files. The files MUST be named containing four digits, starting from 0001.tif.
jpeg	The access files presented in Table 4 MUST be saved in the directory as individual files. The files MUST be named containing four digits, starting from 0001.jpg.
ocr	The AltoXML file presented in section 5.2 MUST be saved in the directory so that a separate XML file exists for each digitised document. The files MUST be named containing four digits, starting from 0001.xml.
mix	The required data presented in Table 5 concerning all archival master files located in the tif directory MUST be saved in the directory. Other data presented in the table SHOULD also be saved in each file. Other data in accordance with the MIX metadata schema MAY also be saved in files following a structure in accordance with the schema. The files MUST be named containing four digits, starting from 0001.xml.

If material is transferred to the National Archives of Finland, each transfer package MUST be formed into a TAR package. The content of TAR packages MUST NOT be compressed at this stage. A checksum MUST be created for TAR packages in MD5 format, and it MUST be sent in conjunction with the transfer.

7 Appendices

1. Example_package.zip[10]

[9] The analogue manifestation decided to be digitised should be described (descriptive metadata produced) before its digitisation. Using the identifier, it MUST be possible to connect digital objects created in the digitisation process to the aforementioned descriptive metadata.

[10] Example images are not references of the image quality. Files contain the metadata defined as requirements in this document. AltoXML is an example, indicating that a separate Alto file MUST be created for each file. MIX.xml is an example of the TIF file in this package, apart from elements indicating otherwise. Directories are not included in the TAR package.