The National Archives

Digitisation at The National Archives

Last updated: August 2016

Contents

1 Introduction 4
1.1 Who is this document for? 4
1.2 References 4
2 Document handling during the scanning process 5
2.1 Preparing documents for scanning 5
2.2 Document handling training 5
2.3 Support of documents 5
2.4 Page turning, unfolding corners 5
2.5 Staples, pins, paperclips 6
2.6 Handling seals 6
2.7 Keeping documents in order 6
2.8 Annotation and labelling 6
2.9 The scanning area 7
2.10 Use of gloves, tools, cleaning liquids and related items 7
3 Scanning equipment 7
4 Image capture and quality 8
5 File format 8
5.1 Colour space 8
5.2 Compression 8
5.3 Resolution 9
5.3.1 Embedded Capture Resolution Information 9
5.3.2 Result of calculations in both examples 10
5.4 Physical dimensions 11
6 JPEG2000 profile for a digitised record 11
7 Converting master images (TIFF and so on) to JP2 for digitised records 12
8 JPEG2000 profile for a digital surrogate 13
9 Converting master images (TIFF and so on) to JP2 for digital surrogates 13
10 Metadata 14

10.1 Embedded metadata 14
10.1.1 Example of embedded metadata in xml 14
10.1.2 Explanation of the URI 14
10.1.3 Creating the UUID 15
10.1.4 Validation of embedded metadata 16
10.2 External metadata 17
10.2.1 Technical metadata 18
10.2.1.1 Explanation of Checksums 18
10.2.1.2 Technical acquisition and technical environment metadata requirements for digitised records 18
10.2.1.3 Technical acquisition metadata requirements for digital surrogates 21
10.2.2 Transcription metadata 22
10.2.2.1 Ordering images 23
10.2.2.2 Dates 23
11 Validation of scanned images and external metadata 24
11.1 Metadata validation 24
11.2 Image validation 24
11.2.1 Tools for validation 27
11.2.1.1 Tools for JPEG2000 format validation 27
11.2.1.2 Tools for XML metadata validation 27
12 Folder structure 28
13 Overview of the process 30
Appendix A: Technical acquisition metadata for digitised records 31
Appendix B: Technical environment metadata for digitised records 45
Appendix C: Technical acquisition metadata requirements for digital surrogates 47
Appendix D: An example of the types of fields required in a transcription metadata CSV file 55

1 Introduction

This document sets out The National Archives' standards and requirements for the digitisation of analogue records in our collection. It covers the whole digitisation process from initial scanning through to delivery of the images for preservation, including The National Archives' scanned image specification (see sections 6 and 8).

This document covers the scanning of records where the resultant images will become the legal public record for permanent preservation. For purposes of clarity we refer to these images as **digitised records.** This document also covers the scanning of records where the resultant images will become digital surrogates with the original paper records being retained and remaining the legal public record. For purposes of clarity we refer to these images as **digital surrogates**.

1.1 Who is this document for?

We recommend that government departments who wish to digitise any of their paper records follow the processes set out in this document. Please contact your Information Management Consultant (IMC) for further information if it is likely that these records will be transferred to The National Archives at a future date. Other organisations are welcome to use this document for reference when developing their own standards for digitisation.

For further information regarding digitisation projects at The National Archives, please contact the Digitisation and Data Conversion Manager: digitisation.dataconversion@nationalarchives.gsi.gov.uk

For any queries about the technical aspects of this document please contact: digitalpreservation@nationalarchives.gsi.gov.uk

1.2 References

In preparing the technical imaging and metadata standards The National Archives have had regard to the following (and other standards referred to therein):

- BS 10008:2008 *Evidential weight and legal admissibility of electronic information. Specification*
- The Archives New Zealand/Te Rua Mahara o te Kāwanatanga *Digitisation Toolkit: http://archives.govt.nz/advice/guidance-and-standards/guidance-subject/digitisation-toolkit*
- The *Minimum Digitization Capture Recommendations* from ALA http://www.ala.org/alcts/resources/preserv/minimum-digitization-capture-recommendations
- US FADGI Guidelines: Technical Guidelines for Digitizing Cultural Heritage Materials
- Netherlands: http://www.metamorfoze.nl/english/digitization
- In drawing up our specifications for *surrogates* we have also reviewed published information by

- Desrochers and Thurgood of Library and Archives Canada, http://www.museumsandtheweb.com/mw2007/papers/desrochers/desrochers.html
- The Wellcome Trust, http://library.wellcome.ac.uk/assets/wtx056572.pdf

2 Document handling during the scanning process

The guidance in this section is drawn from widely accepted standards for handling archival records. The restrictions recommended for 'the scanning area' will be familiar from standard document reading room restrictions.

2.1 Preparing documents for scanning

Ideally a professional conservator will carry out this preparation and will:

- assess the condition of the records to ensure documents are not too fragile for scanning. As well as general fragility you should look for mould, pages stuck together and inserts obscuring records
- assess the condition of records, looking for any damage which has affected the legibility of the text
- remove any staples

2.2 Document handling training

New scanning operators should undergo document handling training by the conservator(s) prior to handling any documents and receive annual refresher training thereafter

2.3 Support of documents

- Use both hands at all times when moving boxes and documents.
- Ensure scanning beds are large enough to support the whole document.
- Never leave documents exposed on the scanner when unattended.
- Support books and other bound documents with a book cradle or book wedges.

2.4 Page turning, unfolding corners

- Turn pages from the fore edge (right edge) of the document not from the tail (bottom) edge.
- It is not acceptable to use moisture (including licked fingers) for page turning.
- Do not pinch document corners together to turn the page.
- The scanning operator should unfold folded corners but should not then fold them back on themselves.

- Where documents are attached to each other and cannot be separated, scan the document in a way which prevents the introduction of new creases.

2.5 Staples, pins, paperclips

- Ideally a conservator should have removed all varieties of staple as part of the process of preparing the documents for scanning. If any have been missed, inform the conservator(s).
- Scanning operators should remove pins, split pins and paperclips carefully but removal should not be forced if this will cause damage.
- Cut all treasury tags immediately prior to scanning and replace them with appropriate length nylon ended tags as soon as the file is scanned.
- The tag should be at least three times as long as the depth of the pile of papers.

2.6 Handling seals

- Take care with applied and pendant seals as they are fragile. They must not be knocked or have weight or pressure applied to them. Neither should they be left to hang off the edge of a work station.
- Do not use glass without adjustments approved by the conservator(s) (for example, lowering the document bed, putting blocks under the glass so there is no weight on the document). The same applies to documents with pigments.

2.7 Keeping documents in order

- The contents of boxes should stay together and stay in the sequence in which they came from the box.
- Work on only one document at a time so that boxes and documents do not get mixed up.
- Replace documents in closed boxes at the end of the day and return them to storage.

2.8 Annotation and labelling

- Annotation or labelling of any part of a document, including the box, is not permitted. Do not use sticky (Post-it®) notes or similar to mark documents. You can use paper markers, provided that you remove them from the document after scanning.

2.9 The scanning area

- Scanning operators' workstations should provide adequate surface area to ensure the full support of documents and allow for an organised workspace. Too little space can have a negative impact on document handling.
- Keep the scanning area clean and tidy - keep bags and coats in lockers and do not take them into the scanning area.
- No food or drink (including chewing gum) should be permitted in the scanning area.
- You may use pencils only - without erasers. No pens or correction fluid are permitted.
- Do not use hand and face moisturisers, moisturising wipes, lip balms or anything similar that is applied by hand.
- Hands should be clean and dry at all times whilst handling documents.

2.10 Use of gloves, tools, cleaning liquids and related items

- Do not wear cotton gloves or powdered gloves
- You may wear unpowdered nitrile/latex (or similar) gloves if instructed specifically by the conservator(s), for example, for photographic material.
- Do not use handling aids such as rubber thimbles and other tools unless approved by the conservator(s).
- Do not use cleaning liquids unless approved by the conservator(s).

Note: if any damage to documents is found during scanning bring this to the attention of the conservator(s) for repair before scanning takes place.

3 Scanning equipment

The National Archives approves scanning equipment for each project.

In general, The National Archives considers overhead cameras and scanners with a flat scanning bed suitable for scanning. You may use supported glass except in cases where the material may be at risk. You may only use flatbed scanners and automatic feed scanners with the approval of the conservator(s).

Similarly, the conservator(s) must approve the use of weights prior to scanning. Lights should not generate too much heat; ideally use cold light sources. Brightness levels must not have a negative impact on the health and safety of operators.

4 Image capture and quality

- Images should be de-skewed as necessary to achieve nominal skew of not greater than one degree.
- All digital images should be legible and at least as readable as the original image from which they are derived.
- Final images should be single page, unless information crosses both pages.
- All images should be viewed immediately after scanning as a check on satisfactory capture (for example images complete or not inverted) and rescanned if required.

5 File format

Sections 5 to 8 set out the technical specification The National Archives uses for producing scanned images of analogue records. Please note that this specification reflects the requirements of The National Archives and may not be suitable for implementation in other organisations.

From March 2013 all records digitised at, or for, The National Archives will be delivered for preservation as JPEG 2000 part 1 files conformant with the latest version of ISO/IEC 15444-1 JPEG 2000 part 1 and saved with the extension .jp2. If scanning software does not produce .jp2 files natively, images must be converted from a suitable intermediate file format to expected resolution and quality standards. See sections 7 and 9.

Access to the original images (for example, TIFFs) should be maintained until the master JP2 images are signed off.

5.1 Colour space

Scan images in 24 bit colour using the Enumerated sRGB colourspace profile, or for microform material in 8 bit grayscale using the Enumerated greyscale colourspace profile.

5.2 Compression

Use lossless compression for digitised records (where sole access is to be provided via the scanned image).

Lossy compression is acceptable for digital surrogates (where the original paper records are to be retained as the primary record). See section 6 below.

5.3 Resolution

Requirements as to Pixel per inch (PPI) vary according to the format of the material to be scanned:

- use a default of 300 PPI for ordinary documents

PPI should be considerably higher for any photographic media:

- photographs should be at 600 PPI
- photographic transparencies should be at 4000 PPI

For microform the requirement should be for a resolution equivalent to 300 PPI at the size of the original document. If this is not physically possible we would agree on the maximum feasible resolution.

5.3.1 Embedded Capture Resolution Information

Image capture resolution information should be written to the JP2's 'Capture Resolution Box'. This is held within the parent 'Resolution Box', which is located within the 'JP2 Header Box'. The Capture Resolution Box specifies the resolution at which the source was digitised, by flatbed scanner or other device, to create code-stream image samples. Resolution is detailed by way of a set of values written to the following parameters:

- vRcN = vertical grid resolution numerator
- vRcD = vertical grid resolution denominator
- vRcE = vertical grid resolution exponent

- hRcN = horizontal grid resolution numerator
- hRcD = horizontal grid resolution denominator
- hRcE = horizontal grid resolution exponent

The parameter values are used by the following calculations to state Vertical Resolution capture and Horizontal Resolution capture values ('VRc' and 'HRc'):

$VRc = VRcN / VRcD \times 10^{VRcE}$

$HRc = HRcN / HRcD \times 10^{HRcE}$

The parameter values written may vary by numerator value and the relative adjustment of denominator and exponent, but the resulting values of ‘VRc’ and ‘HRc’ by calculation must return the correct image

resolution values – measurements stated in 'Pixels Per Meter' from which Pixels Per Inch values can be derived (1 pixel per meter = 0.0254 pixels per inch.)

Two examples of different, but correct values for a 300 PPI (Pixels Per Inch) image are shown below:

Example 1:

- vRcN: 30000
- vRcD: 254
- hRcN: 30000
- hRcD: 254
- vRcE: 2
- hRcE: 2

Example 2:

- vRcN: 300
- vRcD: 254
- hRcN: 300
- hRcD: 254
- vRcE: 4
- hRcE: 4

5.3.2 Result of calculations in both examples

Example 1:

VRc = VRcN / VRcD x 10^{VRcE}

30,000/254 x 10^{2} = 11811.02362204724 PPM

11811.02362204724 x 0.0254 = 300 PPI

Example 2:

VRc = VRcN / VRcD x 10^{VRcE}

300/254 x 10^{4} = 11811.02362204724 PPM

11811.02362204724 x 0.0254 = 300 PPI

Confirmation of correct values can be made by running a Jpylyzer validation report, which will return written values and confirm image resolution detail by calculation. An example tag-set with values from such a report is given below:

<resolutionBox>
<captureResolutionBox>
<vRcN>300</vRcN>
<vRcD>254</vRcD>
<hRcN>300</hRcN>
<hRcD>254</hRcD>
<vRcE>4</vRcE>
<hRcE>4</hRcE>

The Per Meter and Per Inch values below are calculated by Jyplyzer from the above parameter values:

<vRescInPixelsPerMeter>11811.02</vRescInPixelsPerMeter>
<hRescInPixelsPerMeter>11811.02</hRescInPixelsPerMeter>
<vRescInPixelsPerInch>300.0</vRescInPixelsPerInch>
<hRescInPixelsPerInch>300.0</hRescInPixelsPerInch>
</captureResolutionBox>
</resolutionBox>

There are JP2 encoder software tools available (such as Kakadu, v7 onwards) that can automatically populate the JP2's Capture Resolution Box with the horizontal and vertical (x and y) Pixels Per Meter image resolution values, which can then be read from the file by software viewers and image editing tools. There are also imaging SDKs (such as ImageGear for .Net) that can be used to build configurable toolset implementations to achieve the same results.

5.4 Physical dimensions

- All scans should be size-for-size (for microfilm this refers to the size of the original), with a sufficient clear border/margin to demonstrate to users that the entire page has been captured.
- If a single scan cannot capture the page in its entirety, there should be sufficient overlap to allow users to determine clearly which of the separate digital images form the whole of the original paper page.

6 JPEG2000 profile for a digitised record

The most important aspect of this profile is the use of lossless compression (the 5-3 reversible transform). We have chosen not to lose any data because the images will become the legal Public Record. As individual tiles in a JPEG2000 image may use different compression methods we stipulate a single tile to make verification of the compression method more straightforward.

JPEG2000 option	Value
Standard:	JP2 Part 1
Transform:	5-3 reversible (lossless)
Compression ratio:	N/A
Levels:	7
Layers:	1
Progression:	RPCL
Tiles:	Not defined (single tile)
Bypass:	Selective
Colour-space:	Enumerated sRGB profile
Embedded Capture Resolution Information:	Vertical Resolution and Horizontal Resolution values as appropriate to the document
Code block size	N/A
Precinct size	N/A
Regions of interest	N/A
Tile length markers	N/A

7 Converting master images (TIFF and so on) to JP2 for digitised records

We have anticipated that the application of the conversion from a suitable intermediate file format of expected resolution and quality standards will be automated and that JP2 encoder parameters would result in a profile matching The National Archives standard for **digitised records**, and could be incorporated in a script.

8 JPEG2000 profile for a digital surrogate

For digital surrogates use lossy 6:1 compression. However, compression ratio values may vary, depending on the characteristics and visual complexity of the documents to be scanned.

JPEG2000 option	Value
Standard:	JP2 Part 1
Transform:	9-7 irreversible (lossy)
Compression ratio:	Default 6:1 It is expected that the minimum compression will be 4:1 and the maximum 10:1 depending on the nature of the original material
Levels:	7
Layers:	1
Progression:	RPCL
Tiles:	1024x1024 pixels
Colour-space:	Enumerated sRGB profile
Embedded Capture Resolution Information:	Vertical Resolution and Horizontal Resolution values as appropriate to the document
Bypass:	Selective
Code block size	N/A
Precinct size	N/A
Regions of interest	N/A
Tile length markers	N/A

9 Converting master images (TIFF and so on) to JP2 for digital surrogates

In the above table we have anticipated that the application of the conversion from a suitable intermediate file format of expected resolution and quality standards will be automated and that JP2 encoder parameters would result in a profile matching The National Archives standard for **digital surrogates** and could be incorporated in a script.

10 Metadata

10.1 Embedded metadata

The National Archives requires that a small amount of metadata should be embedded within the image file itself. This metadata is designed to assist in the long-term management of the images by making them easier to identify.

This metadata comprises:

- a copyright statement; this is usually a statement of Crown Copyright as used in the example below, or a statement indication of third party copyright
- a universally unique identifier (UUID) created for each image, see below
- a uniform resource identifier (URI) which allows us to reference the image uniquely

This metadata should constitute a well-formed XML document, which can be validated against an XML Schema provided by us - see example below. The UUID should be a uniquely generated Version 4 UUID, see below.

The image must remain a valid JPEG2000 image after the metadata has been embedded.

10.1.1 Example of embedded metadata in xml

```
<?xml version="1.0" encoding="utf-8"?>
<DigitalFile
xmlns="http://nationalarchives.gov.uk/2012/dri/artifact/embedded/metadata"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <UUID>c87fc84a-ee47-47ee-842c-29e969ac5131</UUID>
    <URI>http://datagov.nationalarchives.gov.uk/66/WO/409/27@1/c87fc84a-ee47-47ee-
842c-29e969ac5131</URI>
    <Copyright>&#169; Crown copyright: The National Archives of the UK</Copyright>
</DigitalFile>
```

10.1.2 Explanation of the URI

The URI (for example, see www.ietf.org/rfc/rfc3986.txt) is formed from The National Archives' reference data domain (currently datagov.nationalarchives.gov.uk) followed by a code that represents The National Archives in its Archon directory of archival repositories (or another repository if public records are held elsewhere under The National Archives' Place of deposit rules), followed by the catalogue reference to

piece level (please note that the ITEM level should not be included in file embedded metadata), and the UUID of the image:

- URI Description:
  {TNA DOMAIN}/{TNA ARCHON No.}/{DEPT}/{SERIES}/{PIECE}/{UUID}
- example Base String:
  http://datagov.nationalarchives.gov.uk/66/AIR/79/18727/
- example UUID:
  c87fc84a-ee47-47ee-842c-29e969ac5131

So the full Reference URI would be:

http://datagov.nationalarchives.gov.uk/66/AIR/79/18727/c87fc84a-ee47-47ee-842c-29e969ac5131

10.1.3 Creating the UUID

UUIDs must be compliant with the specification for Version 4 UUIDs as outlined in RFC4122: www.ietf.org/rfc/rfc4122.txt. Utilities are available to create such UUIDs. You should express UUIDs in lower-case hexadecimal format. These should be generated and associated with the images (as image metadata) and then embedded in the image files produced.

Programming implementations which output this standard exist in (at least):

- Java
- C
- JavaScript
- PHP
- Python
- Ruby

An example implemented in Java 6 might look like:

```
import java.util.UUID;

public class UuidExample {

    public static void main(String[] args) {
        final UUID uuid = UUID.randomUUID();
        System.out.println(uuid);
    }
}
```

This will generate the output:

uuid('c87fc84a-ee47-47ee-842c-29e969ac5131')

Note: only the string between the single quotes is required.

10.1.4 Validation of embedded metadata

Example of an XML schema used to validate embedded metadata:

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema
    xmlns:xs="http://www.w3.org/2001/XMLSchema"
    xmlns="http://nationalarchives.gov.uk/2012/dri/artifact/embedded/metadata"

targetNamespace="http://nationalarchives.gov.uk/2012/dri/artifact/embedded/met
adata"
    elementFormDefault="qualified"
    attributeFormDefault="unqualified"
    version="1.0">

    <xs:annotation>
        <xs:documentation xml:lang="en">
            XML Schema document for embedding metadata
            in digitised image records held by The National Archives.
        </xs:documentation>
    </xs:annotation>

    <xs:element name="DigitalFile">
        <xs:complexType>
            <xs:sequence>
                <xs:element name="UUID" type="UUIDv4StringType"/>
                <xs:element name="URI" type="uriType"/>
                <xs:element name="Copyright" type="copyrightType"
default="&#169; Crown copyright: The National Archives of the UK"/>
            </xs:sequence>
        </xs:complexType>
    </xs:element>

    <xs:simpleType name="uriType">
        <xs:restriction base="xs:string">
            <!-- Concatenation of URI eg.
http://datagov.nationalarchives.gov.uk/66/WO/40/27/1/ with UUID -->
            <!-- Ref string:
http://datagov.nationalarchives.gov.uk/66/WO/40/27/1/c87fc84a-ee47-47ee-842c-
29e969ac5131 -->
            <!-- URI: http://datagov.nationalarchives.gov.uk/{TNA ARCHON
No.}/{DEPT}/{SERIES}/{PIECE}/{UUID} -->
```

```
            <!-- {DEPT}: Always A-Z at least 2 characters -->
            <!-- {SERIES}: Numeric. Maybe in two parts with a single /
separator translated to a @ -->
            <!-- {PIECE}: Numbers and/or Characters and/or symbols - ; + $ in
any number of parts separated by a single / translated to a @ e.g. (27) (27@1)
(27@1@1) etc. -->

            <xs:pattern value="http://datagov.nationalarchives.gov.uk/66/[A-
Z]{2,}/([0-9]+)(@[0-9]+)?/([0-9A-Za-z\-;\+$]+)(@[0-9A-Za-z\-;\+$]+)*/[a-f0-
9]{8}-[a-f0-9]{4}-4[a-f0-9]{3}-[89ab][a-f0-9]{3}-[a-f0-9]{12}">
                <xs:annotation>
                    <xs:documentation>Reuses the UUIDv4StringType as the
postfix of the URI</xs:documentation>
                </xs:annotation>
            </xs:pattern>
        </xs:restriction>
    </xs:simpleType>

    <xs:simpleType name="UUIDv4StringType">
        <xs:annotation>
            <xs:documentation>Represents a Universally Unique Identifier
version 4</xs:documentation>
        </xs:annotation>
        <xs:restriction base="xs:token">
            <!-- 32 digits for example c87fc84a-ee47-47ee-842c-29e969ac5131
(Version 4 UUID as specified in RFC4122) -->
            <xs:pattern value="[a-f0-9]{8}-[a-f0-9]{4}-4[a-f0-9]{3}-[89ab][a-
f0-9]{3}-[a-f0-9]{12}"/>
        </xs:restriction>
    </xs:simpleType>

    <xs:simpleType name="copyrightType">
        <xs:restriction base="xs:string">
            <xs:minLength value="3"/>
        </xs:restriction>
    </xs:simpleType>
</xs:schema>
```

10.2 External metadata

The National Archives requires the collection of a variety of technical metadata relating to the creation of digitised images and the creation of electronic text (transcription) from such images. The aim of collecting this metadata is to allow for the long term management and preservation of the images and to enable us to describe the digitised content online.

Where the images represent digitised records rather than digital surrogates this technical metadata also captures the provenance of the digitised record allowing us to demonstrate its authenticity and integrity when the original paper document is destroyed.

The National Archives performs a variety of checks on this metadata to examine its internal consistency and compliance with our requirements.

10.2.1 Technical metadata

The technical metadata about the creation of digitised images broadly describes the hardware, software, and processes used to create the images. This allows the identification of any systemic issues uncovered during quality assurance process (if one particular scanner is producing poor quality images, or a software package appears to have a bug). Where the images represent digitised records rather than digital surrogates we also need to know when the actions were carried out.

The National Archives requires technical metadata to be delivered as UTF-8 (tools.ietf.org/html/rfc3629) encoded, CSV (Comma Separated Value) text files formatted according to the RFC 4180 specification (tools.ietf.org/html/rfc4180).

10.2.1.1 Explanation of Checksums

The National Archives requires that a checksum based on the SHA256 Cryptographic Hash Function (CHF) be calculated for each final image after metadata has been added, and supplied as part of the image metadata. See: csrc.nist.gov/publications/fips/fips180-3/fips180-3_final.pdf.

This is to ensure that data has not been corrupted in transit. A checksum must also be generated for the metadata files themselves. Upon receipt of the batch, The National Archives recalculates the SHA256 hash value for each of the images and metadata files concerned and confirms that the newly calculated values match the values supplied in the metadata and hash value text files. In the event that the values do not match, the batch or piece(s) within it may be rejected and returned to the supplier.

10.2.1.2 Technical acquisition and technical environment metadata requirements for digitised records

Two metadata files are required for each batch of digitised records; one file describing the technical acquisition metadata and the other file describing the technical environment metadata.

The metadata files are named: tech__<*type*>*_metadata_* v<*versionnumber*>*_*<*batchcode*>.csv

Where <type> is either 'acq' for acquisition metadata or 'env' for environment metadata, <versionnumber> is the version of the particular metadata file standard provided by The National

Archives and <batchcode> is a unique code for the batch being delivered and that matches the value of the batchcode field in the metadata itself and the volume label of the media the batch is delivered on.

For example: using version 1 of the acquisition metadata file standard and a batch with a batchcode "testbatchY16B001" the file name would be:

tech_acq_metadata_v1_testbatchY16B001.csv

The checksum for this file would be saved with file name:

tech_acq _metadata_v1_testbatchY16B001.csv.sha256

The contents of that file would be a string of the form "tech_acq_metadata_v1_testbatchY16B001.csv e3b0c44298fc1c149afbf4c8996fb92427ae41e4649b934ca495991b7852b855" where the long string beginning "e3b0c…" is the SHA256 hash value for the file. This should be a simple UTF-8 encoded text file.

There are up to 42 technical acquisition metadata fields for paper digitisation projects (not all will be relevant for every project):

- Batch code
- Department
- Division
- Series
- Sub series
- Sub sub series
- Piece
- Item
- Description
- Ordinal
- File UUID
- File path
- File checksum
- Resource URI
- Scan Operator
- Scan ID
- Scan location
- Scan native format
- Scan timestamp
- Image Resolution
- Image width
- Image height
- Image tonal resolution
- Image format
- Image colour space
- Image split
- Image split ordinal
- Image split other UUID
- Image split operator
- Image split timestamp
- Image crop
- Image crop operator
- Image crop timestamp
- Image de-skew
- Image de-skew operator
- Image de-skew timestamp
- Process location
- Jp2 creation timestamp
- UUID time stamp
- Embed timestamp
- QA Code
- Comments

For microform or photographic negatives the following three fields will also be added (to give a total of 45 fields):

- image_inversion
- image_inversion_operator
- image_inversion_timestamp

For microform the following field may also be added (to give a total of 46 fields):

- fiche reference

For further information see Appendix A Technical acquisition metadata for digitised records.

There are eight technical environment metadata fields for digitised records:

- Batch code
- Company name
- Image de-skew software
- Image split software
- Image crop software
- Jp2 creation software
- UUID software
- Embed software

For microform or photographic negative projects, a further field is added (giving a total of nine):

- Image inversion software

For further information see Appendix B Technical environment metadata for digitised records.

10.2.1.3 Technical acquisition metadata requirements for digital surrogates

One metadata file should be delivered with each batch of digital surrogates; this file describes technical acquisition metadata. The naming of this file and checksum file is as above.

There is no technical environment metadata required for digital surrogates.

There are up to 30 technical acquisition metadata fields for digital surrogates. They are a sub-set of those required for digitised records:

- Batch code
- Department
- Division
- Series
- Sub series

- Sub sub series
- Piece
- Item
- Ordinal
- Description
- File UUID
- File path
- File checksum
- Resource URI
- Scan operator
- Scan ID
- Scan location
- Image resolution
- Image width
- Image height
- Image tonal resolution
- Image format
- Image compression
- Image colour space
- Image split
- Image split ordinal
- Image split other UUID
- Image crop
- Image de-skew
- Comments

For further information see Appendix C Technical acquisition metadata requirements for digital surrogates.

10.2.2 Transcription metadata

Transcription metadata is more variable in content, dependent on the records to be transcribed. However, this section lays out some general principles relating to The National Archives' desired approach to transcription, particularly in relation to common pieces of data to be transcribed such as names, dates and addresses.

Previous experience has shown that two elements of transcription can be particularly difficult to deal with:

- ordering images into the correct sequence to form 'documents' of correctly ordered pages
- dates - in terms of onward processing due to the large numbers of possible combinations of format (day, month or year) being missing or partially unreadable (or occasionally nonsensical - 30 February)

The National Archives believes that the complexity of dealing with these issues can be reduced in the following ways:

10.2.2.1 Ordering images

An ordinal number in the metadata records the position of a single image within its parent Piece or Item. This allows images to be reordered at any time within the boundaries of a Piece or Item without renaming files.

However, should an image need to move from one Piece or Item to another this would be reflected in the ordinals, but would additionally require a rename and move of the image.

Ordinals are context sensitive, which is to say they are only unique within their Parent container, and as such should start from 1 within each Item or Piece and be incremented sequentially. If, as part of the transcription process, it is also required that material previously arranged only at Piece level is split into Items, there is no need to rename files; just record each new item in the CSV with the relevant ordinals.

So if piece 1 originally contained 12 images which transcription shows should be split into 3 items:

- item 1 might consist of the first 3 images, 0001.jp2, 0002.jp2, 0003.jp2 with ordinals 1, 2 and 3 respectively (within item 1)
- item 2 the next 5 images, 0004.jp2, 0005.jp2, 0006.jp2, 0007.jp2 0008.jp2 with ordinals 1, 2, 3, 4 and 5 respectively (within item 2)
- item 3 the final 4 images 0009.jp2, 0010.jp2, 0011.jp2, 0012.jp2 with ordinals 1, 2, 3 and 4 respectively (within item 3)

10.2.2.2 Dates

Transcribing the date parts separately will make it easier to check the status (missing, incomplete and so on) of each part without complicated parsing. Also the ability to disambiguate between omitted transcription dates and dates that were truly missing from the original allows for some automated quality assurance (QA) checking.

See Appendix D for an example of the types of fields required in a transcription metadata CSV file.

11 Validation of scanned images and external metadata

11.1 Metadata validation

The National Archives undertakes a variety of tests on the metadata to ensure it is internally consistent. Where values such as UUIDs can be repeated in different fields for the same image (for example, in both the UUID field and as part of the image URI) we will check that the same value is given in each case. We also check that rows are not duplicated. Where we have specified particular data types, character sets or character patterns, these will also be validated.

The primary tool used for this is our CSV Validator, working with our CSV Schema language. Full details of these can be found at http://digital-preservation.github.io/csv-validator/. The relevant schema for a project will be supplied in advance of imaging (and usually with the project ITT). For ease of reference the schema name will match that of the metadata file to which it applies, but with the numeric part of the reference set to zeroes, and with the format extension .csvs. So, for the example metadata filename given above, *tech_acq_metadata_v1_testbatchY16B001.csv*, the related schema would be *tech_acq_metadata_v1_testbatchY16B000.csvs*.

11.2 Image validation

The National Archives expects its suppliers to carry out general quality assurance on images through a defined process which tests all aspects of the specification laid out above. Suppliers give details of the process to The National Archives at the start of a project and provide regular reports on the application of the process and issues detected.

During our image QA process, if we find any missing images we will flag these to the scanning supplier and they will need to scan the missing images, insert them into the correct location within the piece, renumber all subsequent image numbers, update the .CSV files and redeliver either the whole batch or resubmit individual piece(s) as part of later batches back to us.

We use programmatic techniques to validate all images ensuring general compliance with the JPEG2000 part 1 specification and with the specific profiles laid out in this document (including the embedding of image metadata) see 11.2.1 Tools for validation, below. Non-compliant images are rejected and are regenerated (if necessary by rescanning).

The tools and scripts used by us are freely available and are listed in sections 11.2.1 and 11.2.2 of this document. Some individual scripts will also be developed for particular projects and made available to suppliers on request. We suggest suppliers to The National Archives incorporate these, or similar tools, into internal QA process in order to reduce the likelihood of images being rejected. Up to 10% of images per batch on every project may be inspected.

Evaluation may cover:

- correctness of mode
- correctness of resolution
- correctness of image size
- lack of sharpness
- loss of detail or image corruption
- correctness of orientation
- correctness of cropping
- skew
- overall too dark or light
- overall too low or high contrast
- correctness of file name, and
- correctness and completeness of metadata

If the random sampling suggests that more than 1% of the total batch fails to meet the required standards then the entire batch is returned to the suppliers for further quality control examination and rescanning as necessary. Where smaller proportions of images do not meet our standards only the piece(s) containing those images will be rejected from the batch and the rest of the batch will continue through the ingest process. Those piece(s) should then be resubmitted within later batches.

We can then inspect any re-scanned images if necessary. We notify suppliers of any errors found during the technical and visual QA processes by sending them a CSV file or alternative reporting formats as agreed, for example:

Field	Data Format	Description	Options or Example
batch_code	Up to 16 alpha-numeric characters	An identifier for each batch of records. The same batch number will be included in the first row of every metadata file related to that batch of records	testbatchY16B001
file_uuid	Universally unique identifier (UUID) - must adhere to UUID Version 4 format see www.ietf.org/rfc/rfc4122.txt	Universally unique identifier embedded in every image	daf49885-e182-4211-80f7-29bb0bb35112

Field	Data Format	Description	Options or Example
file_path	Must be a valid URI see www.ietf.org/rfc/rfc3986.txt	Location of file on storage as specified in the competition. For example: DeptCode/SeriesNo/Piece_Number_ItemNo_ImageNumber.jp2	file:///WO/409/27_1/1/27_1_0001.jp2
file_checksum	Must adhere to the SHA-256 standard and be expressed in lower-case hexadecimal characters, see csrc.nist.gov/publications/fips/fips180-3/fips180-3_final.pdf	A checksum of the image file conformant with the SHA256 standard	e3b0c44298fc1c149afbf4c8996fb92427ae41e4649b934ca495991b7852b855
error_description	From list below		

Error description
Incorrect mode
Incorrect resolution
Incorrect image size
Lack of sharpness
Loss of detail or image corruption
Incorrect orientation
Incorrect cropping
Skew
Overall too dark
Overall too light
Incorrect file name
Incorrect header information
Incomplete header information

11.2.1 Tools for validation

11.2.1.1 Tools for JPEG2000 format validation

As the JPEG2000 format is relatively new, experience has shown that not all tools/ encoders implement the standard correctly in all scenarios. Therefore, the recommended approach is to use a combination of tools to increase confidence in the validity of the image.

Tools employed by The National Archives have included:

Jasper Imginfo 1.900.1
Imginfo is part of the Jasper toolkit which is a reference implementation for the JPEG2000 standard. Imginfo parses the entire codestream to output information about the JPEG2000 codestream. The parser may fail if the image is not valid. This tool extracts minimal technical metadata such as height and width of the image in pixels. It has been found to be useful in reporting corruption in the image code-stream that can result in visual distortion or artefacts.
See www.ece.uvic.ca/~frodo/jasper/

OPF jpylyzer
Jpylyzer was produced by Johan van der Knijff for the Open Planets Foundation; it validates the JPEG2000 file structure by performing tests against the published standard and also extracts file properties. The result of this tool should indicate that the file is valid JP2 and the values extracted by the tool for levels, layers and so on meet the requirements set out in this document.
See www.openplanetsfoundation.org/software/jpylyzer

It is possible for these tools to be wrapped or incorporated within automated validation workflows. We would strongly promote the use of such tools to check conformance of generated JP2 files with the latest published standard and to ensure that they also match the relevant The National Archives profile.

11.2.1.2 Tools for XML metadata validation

The XML document generated to embed into the JPEG2000 images must be valid according to an XML Schema provided by The National Archives, as above. To ensure the validity of the XML document, various tools exist for validating it against the Schema. Some of the more popular include:

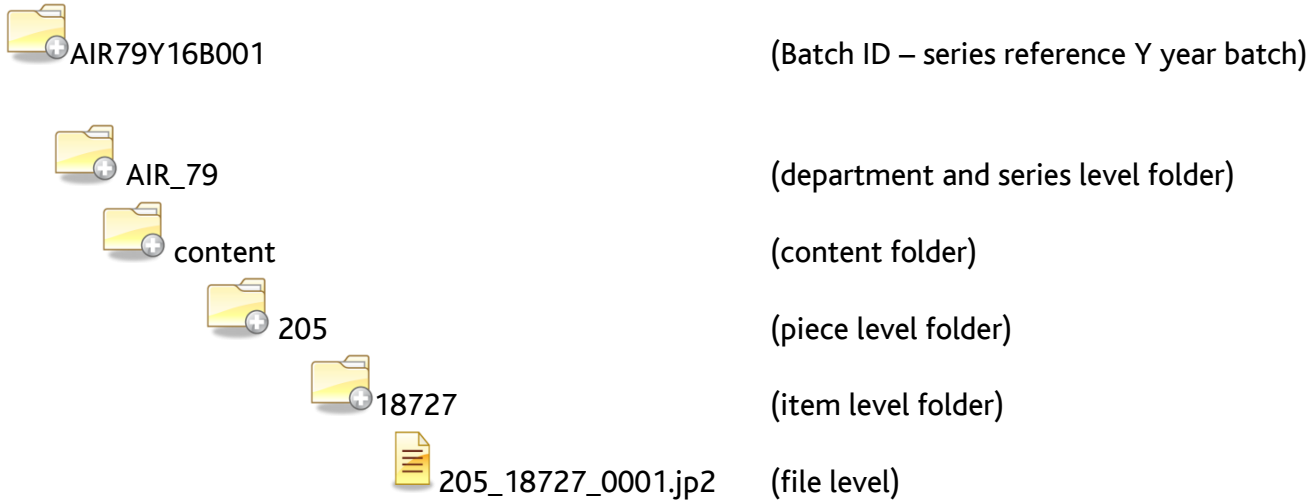
Apache Xerces: xerces.apache.org
Saxonica Saxon EE: www.saxonica.com
LibXml xmllint: xmlsoft.org/xmllint.html

12 Folder structure

The National Archives requires all images to be delivered in a folder structure which reflects the original archival hierarchy of the records.

It may sometimes be necessary to extend this hierarchy by adding more detailed cataloguing information to identify (for example) the images which relate to a single individual or those which represent a particular month's reports.

Example of such a folder structure based on a recent project:



The catalogue reference for this image would be AIR 79/205/18727.

The filenames quoted above are deliberately neutral to stress that files will be identified and managed through the folder structure and the unique identifier embedded in each file.

At The National Archives an item is defined as all the images relating to a single individual or other appropriate grouping.

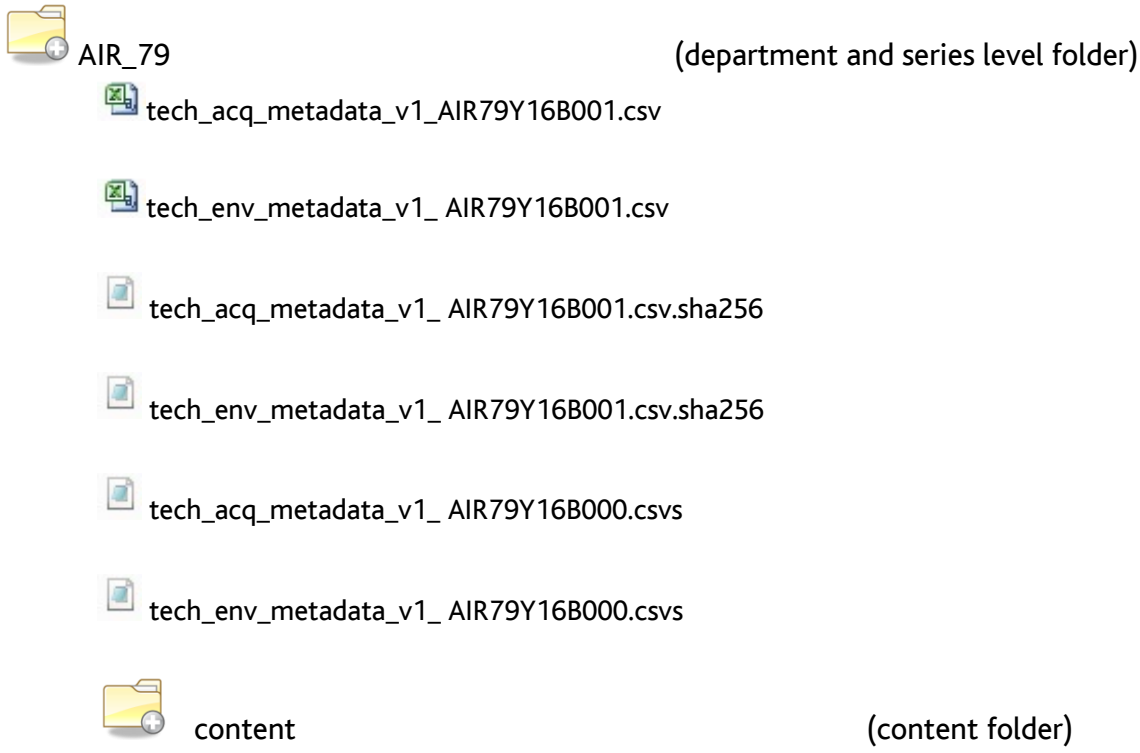
In some instances, there may be images which do not obviously relate to the previous or subsequent item. Such images are generically referred to as orphans.

As part of The National Archives' own in-house QA process to quality assure scanned images we check anything identified as an orphan. If necessary we will also advise on how material which has been missed during the scanning process should be integrated into the folder structure.

Images are delivered in one or more batches. All the metadata files for a batch should be at the content folder level of this folder structure with their respective checksum files, alongside the content folder rather than inside it. Each batch would contain a single file system of the layout specified. A batch will

normally comprise several pieces, but please note that an individual piece should arrive within a single batch, not split across multiple batches.

Example of the location of metadata files within the folder structure:



13 Overview of the process

Scan originals	The scanning process used must ensure that both sides of all pages are captured (even if blank) only once. Record details of the scanning machine used and a code for the operator.
Edit images	Perform any tasks such as cropping, de-skewing and image splitting as required. Record software used and operator.
Save as .jp2	If scanning software does not produce .jp2 files natively, then convert images. Record details of all conversion software used. Retain original images until all quality assurance (QA) is completed and The National Archives confirms they can be destroyed.
Validate jp2	
Ensure files are correctly allocated in archival hierarchy and assign final filenames	Since the final filename is required to contain elements of the catalogue reference, and to contain a number to indicate its position within an individual record, it may not be possible to construct this filename at this point of the scanning process. There should be a repeatable (and auditable) process for this allocation.
Compile external metadata	
Compile internal metadata and embed XML into .jp2	As well as external technical acquisition (for digitised records and digital surrogates) and environment metadata (for digitised records only) The National Archives requires a variety of XML elements to be embedded into the .jp2 images to assist with long-term management of the files. Since this includes the catalogue reference, allocation into the archival hierarchy must have been completed by this stage.
Validate XML	
Re-validate jp2	
Create and record hash value for each file	
Finalise relevant row in metadata .CSV file	To allow long-term assurance that the file received is the same as when originally created and has not been corrupted or tampered with a SHA256 checksum should be calculated for each image file - this should be stored in the metadata spreadsheet. A checksum for the spreadsheet itself is also required.
Submit batch to The National Archives	
Make any corrections, obtain batch signoff and delete original images	

Appendix A: Technical acquisition metadata for digitised records

All fields listed below record details of every individual image and the processes carried out on it before delivery to The National Archives. These fields will be the column headings in the metadata CSV file.

Field	Data format	Description	Options or example	Justification	Consistency check
batch_code	Up to 16 alphanumeric characters	An identifier for each batch of records	TestbatchY16B001	For consistency and cross checking with other data delivered as part of the batch	The National Archives will cross check this against the batch_code with the naming of the file and the volume label
department	Up to 8 characters	Archival hierarchy	AIR		
division	Up to 8 characters	Archival hierarchy	6		May be empty
series	Up to 8 characters	Archival hierarchy	79		
sub_series	Up to 8 characters	Archival hierarchy			May be empty
sub_sub_series	Up to 8 characters	Archival hierarchy			May be empty
piece	Up to 8 characters	Archival hierarchy	1		
item	Up to 8 characters	Archival hierarchy	2		May be empty

Field	Data format	Description	Options or example	Justification	Consistency check
description	Unstructured text	Catalogue description provided by the Authority for each piece/item. May be left blank.	2 Infantry Brigade: 2 Battalion King's Royal Rifle Corps.	Required for the Authority's ingest process, and will also support QA as the description and date range shown can be sense checked against the captured images	Must match values supplied by The Authority

Field	Data format	Description	Options or example	Justification	Consistency check
ordinal	Integer starting from 1	Describes the order of a file within an item. Should start at 1 within each item. See the note on ordering images above	1	To keep the images in order.	Expected range will usually be checked, along with a uniqueness check on the combination of piece, item and ordinal
file_uuid	Universally Unique Identifier (UUID). Adhering to UUID Version 4 format and expressed in lower-case hexadecimal characters, see: http://www.ietf.org/rfc/rfc4122.txt	Universally Unique Identifier for the image embedded in every image	daf49885-e182-4211-80f7-29bb0bb35112	QA - and unique identification of digitised records and digital surrogates for efficient processing	Aim is to ensure all image files are delivered once and only once. Check the UUID against the UUID that forms part of the URI and also against the UUID embedded in the file at the file_path provided, to ensure they match
file_path	The file path to the image. Must be a valid URI, see: http://www.ietf.org/rfc/rfc3986.txt	Location of file relative to the root of the file system containing the batch	file:///AIR_79/content/1/2/1_2_0001.jp2	QA	All image files on the file system provided must have a row in this metadata file and all file_path must have a matching file at the location given

Field	Data format	Description	Options or example	Justification	Consistency check
file_checksum	Must adhere to the SHA-256 standard and should be expressed in lower-case hexadecimal characters	A checksum of the image file conformant with the SHA256 standard	e3b0c44298fc1c149afbf4c8996fb92427ae41e4649b934ca495991b7852b855	QA- to ensure the image file was received without corruption or tampering	The National Archives will generate a checksum upon receipt of the image and expect it to match the checksum given here
resource_uri	The URI that is embedded into the Digital Image. Must be a valid URI, see: www.ietf.org/rfc/rfc3986.txt	A unique identifier with a predictable pattern	http://datagov.nationalarchives.gov.uk/66/AIR/79/1/2/daf49885-e182-4211-80f7-29bb0bb35112	QA	The Authority will check that this URI is the same as the URI embedded in the file stored at the file_path provided
scan_operator	Up to 12 alpha-numeric characters	Code representing the specific operator using the scanner that produced the image; this should be an anonymised code that the supplier can decode	ABG001	QA - the data is anonymised in order that The National Archives does not hold any personal data	Validation by The National Archives
scan_id	Up to 12 alpha-numeric characters	An individual identifier of the scanning device used to produce the	002A	QA - specific scanner id to trace back	Validation by The National Archives

Field	Data format	Description	Options or example	Justification	Consistency check
		image		problems with an image to a specific machine	
scan_location	Text	Physical location of scanner	The National Archives, Kew, Richmond, Surrey, TW9 4DU	QA	Validation by The National Archives
scan_native_format	Text	Format and version expressed as text	Cannon Raw v1.4	Provenance and QA	Validation by The National Archives
scan_timestamp	XML Schema 1.0 dateTime format with a mandatory timezone: www.w3.org/TR/xmlschema-2/#dateTime	Date and time the paper scan ends	2010-01-02T02:17:21Z	Provenance	Validation by The National Archives
image_resolution	Integer	Number in pixels per inch of the image with respect to the original object.	300 600	QA	Validation by The National Archives

Field	Data format	Description	Options or example	Justification	Consistency check
image_width	Integer	Dimensions are always in pixels	4407	QA	Validation by The National Archives against the image file stored at the file_path provided.
image_height	Integer	Dimensions are always in pixels	3030	QA	Validation by The National Archives against the image file stored at the file_path provided.
image_tonal_resolution	Value from provided enumeration		24-bit colour	QA	Validation by The National Archives against the image file stored at the file_path provided.
image_format	A PRONOM unique identifier (PUID) see: nationalarchives.gov.uk/aboutapps/pronom/puid.htm	The code used to uniquely identify a file format	x-fmt/392	QA	Validation by The National Archives against the image file stored at the file_path provided.
image_colour_space	Value from provided enumeration		sRGB	QA	The Authority will validate this against the image file stored at the file_path provided
image_split	Lower case text strings "yes" or "no" `	Specifies if the image was the result of an image split	yes	QA	Validation by The National Archives

Field	Data format	Description	Options or example	Justification	Consistency check
image_split_ordinal	Only integers allowed	For composites (see previous field), this field is used to confirm the ordering of the images. Numbering is from top left, along the top row of separate images, then from the left of each successive row (there should be overlap between adjacent images)	<table><tr><td>1</td><td>2</td></tr><tr><td>3</td><td>4</td></tr></table> Or <table><tr><td>1</td><td>2</td><td>3</td></tr><tr><td>4</td><td>5</td><td>6</td></tr><tr><td>7</td><td>8</td><td>9</td></tr></table> Etc. It may be helpful to use the comments field to provide a more human readable version of this e.g. 1=top left, 2=top middle, 3=top right, 4=middle left, 5=middle middle, 6=middle right, 7=bottom left, 8=bottom middle, 9=bottom right or similar		Validation by The National Archives

Field	Data format	Description	Options or example	Justification	Consistency check
image_split_other_uuid	One or more (separated by a comma) Universally Unique Identifier (UUID). Adhering to UUID Version 4 format and expressed in lower-case hexadecimal characters, see: http://www.ietf.org/rfc/rfc4122.txt	If the image was split, this field must contain the UUIDs of the other images that were split from the same original image as this image. If there more than two images as part of a split, this field may contain multiple UUIDs separated by a comma	0d0b88c6-9a6e-4731-ace3-b50794c1356b,a2915f99-6efa-45d4-a0c9-8fd2555643ec	QA	Field shall be empty if Image_split = "no" and populated with valid data if Image_split = "yes" QA will ensure that the other split images exist, and that they also reciprocally point back to this image through their image_split_other_uuid fields
image_split_operator	Up to 12 alpha-numeric characters	Code representing the specific operator using the split software that produced the image; this is to be an anonymised code that the supplier can decode	ABG001	Provenance and QA. The data is anonymised in order that the Authority does not hold any personal data	Field shall be empty if Image_split = "no" and populated with valid data if Image_split = "yes"

Field	Data format	Description	Options or example	Justification	Consistency check
image_split_timestamp	XML Schema 1.0 dateTime format with a mandatory timezone: www.w3.org/TR/xmlschema-2/#dateTime	The date and time the file was split	2010-01-02T06:17:21Z	Provenance	Field shall be empty if Image_split = “no” and populated with valid data if Image_split = “yes”
image_crop	Lower case text strings "auto", "manual" or "none"	Specifies if the image was cropped and if it was what type of crop was carried out	auto	QA	is("auto") or is("manual") or is("none") else must be blank
image_crop_operator	Up to 12 alpha-numeric characters	Code representing the specific operator using the crop software that produced the image; this is to be an anonymised code that the supplier can decode	ABG001	Provenance and QA. The data is anonymised in order that The National Archives does not hold any personal data	Field shall be empty if image_crop = “none” or if image_crop = “auto” and populated with valid data if image_crop = “manual”
image_crop_timestamp	XML Schema 1.0 dateTime format with a mandatory timezone: www.w3.org/TR/xmlschema-2/#dateTime	The date and time the image was cropped	2010-01-02T06:17:21Z	Provenance	Field shall be empty if image_crop = “none” and populated with valid data if image_crop = “auto” or if image_crop = “manual”

Field	Data format	Description	Options or example	Justification	Consistency check
image_deskew	Lower case text strings "yes" or "no"	Specifies if the image was deskewed	no	QA	
image_deskew_operator	Up to 12 alpha-numeric characters	Code representing the specific operator using the deskew software that produced the image; this is to be an anonymised code that the supplier can decode	ABG001	Provenance and QA. The data is anonymised in order that The National Archives does not hold any personal data	Field shall be empty if image_deskew = “no” and populated with valid data if image_deskew = “yes”
image_deskew_timestamp	XML Schema 1.0 dateTime format with a mandatory timezone: www.w3.org/TR/xmlschema-2/#dateTime	The date and time the image was deskewed	2010-01-02T06:17:21Z	Provenance	Field shall be empty if image_deskew = “no” and populated with valid data if image_deskew = “yes”
process_location	Text	Physical location of image processing procedures	The National Archives, Kew, Richmond, Surrey, TW9 4DU	Provenance and QA	

Field	Data format	Description	Options or example	Justification	Consistency check
jp2_creation_timestamp	XML Schema 1.0 dateTime format with a mandatory timezone: www.w3.org/TR/xmlschema-2/#dateTime	The date and time the JPEG 2000 image was created	2012-08-09T09:15:37+01:00	Provenance	
uuid_timestamp	XML Schema 1.0 dateTime format with a mandatory timezone: www.w3.org/TR/xmlschema-2/#dateTime	The date and time the UUID was created for file	2010-08-02T04:17:21+01:00	Provenance	
embed_timestamp	XML Schema 1.0 dateTime format with a mandatory timezone: www.w3.org/TR/xmlschema-2/#dateTimehttp://www.w3.org/TR/xmlschema-2/#dateTime	The date and time metadata was embedded in the image file	2010-01-02T05:17:21+0:00	Provenance	
image_inversion	Only valid values allowed: lower case text strings "auto", "manual"	**Microform or photographic negative projects only** Specifies if the image	auto	Provenance and QA.	Contains only valid values

Field	Data format	Description	Options or example	Justification	Consistency check
	or "none"	was inverted from negative to positive and if it was carried out by an automated process or manually			
image_inversion_operator	Up to 12 alpha-numeric characters	**Microform or photographic negative projects only** Code representing the specific operator using the inversion software that produced the image for a manual inversion; this is to be an anonymised code that the supplier can decode.	ABG001	Provenance and QA. The data is anonymised in order that The National Archives does not hold any personal data.	Field shall be empty if image_crop = “none” or if image_crop = “auto” and populated with valid data if image_crop = “manual”
image_inversion_timestamp	XML Schema 1.0 dateTime format with a mandatory timezone: http://www.w3.org/TR/xmlschema-2/#dateTime	**Microform or photographic negative projects only** The date and time the image was inverted	2010-01-02T06:17:21Z	Provenance.	Field shall be empty if image_crop = “none” and populated with valid data if image_crop = “auto”or if image_crop = “manual”

Field	Data format	Description	Options or example	Justification	Consistency check
qa_code	**For paper:** Single alphabetic character in the range A-J, or a comma separated list of such single characters **For microform:** Digits from 40 to 44, or a comma separated list of such double digits	Codes to use to indicate where information is illegible due to damage to the document. Codes are given in the next column, together with their meaning. Only the letter portion is to be used Additional codes may be defined for other types of original material	**For paper:** A. Missing Area: Corner or edge B. Missing area: hole in page C. Tears D. Text obscured by tape or other document (can't be removed/separated) E. Discolouration or staining of paper (text is difficult to read) F. Ink stains or other spill (text is obscured) G. Faint text H. Blurred or smudged text I. Offsetting of ink to facing page, or bleed through of ink from other side of page. J. Burn damage (from fire or metal corrosion) Notes: B. Including intentional holes such as tag holes, and holes from pest damage	This records pre-existing damage to the material being digitised. With surrogates, if it appears information has been lost due to damage to the original, the original may be produced to confirm this. For digitised records this option will not be available	Must be either blank or numbers between 01 and 20 (for paper) or 40-47 (for microform)

Field	Data format	Description	Options or example	Justification	Consistency check
			E. Would include staining from mould damage, discolouration, foxing, water damage etc. **For microform:** 40: microform scratched 41: illegible: image too dark 42: illegible: image too light 43: Microform breakage 44: No foliation		
comments	Text	Operator's comments, may be empty		QA	Populated at the supplier's discretion

Appendix B: Technical environment metadata for digitised records

All fields listed below record details of the technical environment used during the scanning process; this environment should be consistent for all scanned images within a batch, and as such needs only be captured once. These fields will be the column headings in the metadata CSV file.

Field	Data Format	Description	Example	Consistency Check
batch_code	Up to 16 alphanumeric characters	An identifier for each batch of records	TestbatchY16B001	The National Archives will cross check this against the batch_code with the naming of the file and the volume label
company_name	Text	Name of the company undertaking the process	Bob's Scan Ltd	
image_deskew_software	Text	Name and version of the software used for deskewing images	GNU GIMP 2.6	
image_split_software	Text	Name and version of the software used for splitting images	GNU GIMP 2.6	

Field	Data Format	Description	Example	Consistency Check
image_crop_software	Text	Name and version of the software used for image cropping	GNU GIMP 2.6	
jp2_creation_software	Text	Name and version of the software used for creating the JPEG2000 file from the acquired image	ImageMagick 6.8.0-5	
uuid_software	Text	Name and version of the software used to generate the UUID. If programmatic, use the software library name and version	Oracle Java JDK 1.6	
embed_software	Text	Name and version of the software used to embed the metadata into the image	Luratech Lurawave 11a	

image_inversion_software	Text	**Microform or photographic negative projects only** Name and version of the software used to invert image colours of negative microform		

Appendix C: Technical acquisition metadata requirements for digital surrogates

All fields listed below record details of every individual image and the processes carried out on it. These fields will be the column headings in the metadata CSV file.

Field	Data Format	Description	Options or Example	Justification	Consistency Check
batch_code	Up to 16 alphanumeric characters	An identifier for each batch of records	testbatchY16B001	For consistency and cross checking with other data delivered as part of the batch	The National Archives will cross check this against the batch_code with the naming of the file and the volume label
department	Up to 8 characters	Archival hierarchy	AIR		
division	Up to 8 characters	Archival hierarchy	6		May be empty

Field	Data Format	Description	Options or Example	Justification	Consistency Check
series	Up to 8 characters	Archival hierarchy	79		
sub_series	Up to 8 characters	Archival hierarchy			May be empty
sub_sub_series	Up to 8 characters	Archival hierarchy			May be empty
piece	Up to 8 characters	Archival hierarchy	1		
item	Up to 8 characters	Archival hierarchy	2		May be empty
ordinal	Integer starting from 1	Describes the order of a file within an item. Should start at 1 within each item. See the note on ordering images above	1	To keep the images in order.	Expected range will usually be checked, along with a uniqueness check on the combination of piece, item and ordinal

Field	Data Format	Description	Options or Example	Justification	Consistency Check
description	Unstructured text	Catalogue description provided by the Authority for each piece/item. May be left blank.	2 Infantry Brigade: 2 Battalion King's Royal Rifle Corps.	Required for the Authority's ingest process, and will also support QA as the description and date range shown can be sense checked against the captured images	Must match values supplied by The Authority
file_uuid	Universally Unique Identifier (UUID). Adhering to UUID Version 4 format and expressed in lower-case hexadecimal characters, see: www.ietf.org/rfc/rfc4122.txt	Universally Unique Identifier for the image embedded in every image	daf49885-e182-4211-80f7-29bb0bb35112	QA and unique identification of digitised records and digital surrogates for efficient processing	Aim is to ensure all image files are delivered once and only once

Field	Data Format	Description	Options or Example	Justification	Consistency Check
file_path	The file path to the image. Must be a valid URI, see www.ietf.org/rfc/rfc3986.txt	Location of file relative to the root of the file system containing the batch	file:///AIR_79/1/2/0001.jp2	QA	All image files on the file system provided must have a row in this metadata file and all file_path must have a matching file at the location given
file_checksum	Must adhere to the SHA-256 standard and should be expressed in lower-case hexadecimal characters, see: csrc.nist.gov/publications/fips/fips180-3/fips180-3_final.pdf	A checksum of the image file conformant with the SHA256 standard	e3b0c44298fc1c149afbf4c8996fb92427ae41e4649b934ca495991b7852b855	QA ensure the image file was received without corruption or tampering	The National Archives will generate a checksum upon receipt of the image and expect it to match the checksum given here
resource_uri	The URI that is embedded into the Digital Image. Must be a valid URI, see: www.ietf.org/rfc/rfc3986.txt	A unique identifier with a predictable pattern	http://datagov.nationalarchives.gov.uk/66/AIR/79/1/2/daf49885-e182-4211-80f7-29bb0bb35112	QA	The Authority will check that this URI is the same as the URI embedded in the file stored at the file_path provided

Field	Data Format	Description	Options or Example	Justification	Consistency Check
scan_operator	Up to 12 alpha-numeric characters	Code representing the specific operator using the scanner that produced the image; this should be an anonymised code that the supplier can decode	ABG001	QA- the data is anonymised in order that The National Archives does not hold any personal data	
scan_id	Up to 12 alpha-numeric characters	An individual identifier of the scanning device used to produce the image	002A	QA - specific scanner id to trace back problems with an image to a specific machine	
scan_location	Text	Physical location of scanner	The National Archives, Kew, Richmond, Surrey, TW9 4DU	QA	
image_resolution	Integer between 1 and 10000	Number in pixels per inch of the image with respect to the original object	300	QA	Validation by The National Archives
image_width	Integer	Dimensions are always in pixels	4407	QA	Validation by The National Archives
image_height	Integer	Dimensions are always in pixels	3030	QA	Validation by The National Archives
image_tonal_resolution	Value from provided enumeration		24-bit colour	QA	Validation by The National Archives

Field	Data Format	Description	Options or Example	Justification	Consistency Check
image_format	A PRONOM unique identifier (PUID) see: www.nationalarchives.gov.uk/aboutapps/pronom/puid.htm	The code used to uniquely identify a file format	x-fmt/392	QA	Validation by The National Archives
image_compression	Integer between 1 and 99	The value of N in the lossy image compression ratio N:1 used to compress the image. Note1:1 means no-compression employed	6	QA	Validation by The National Archives
image_colour_space	Value from provided enumeration		sRGB	QA	The Authority will validate this against the image file stored at the file_path provided
image_split	Lower case text strings “yes" or “no"	Specifies if the image was the result of an image split	yes	QA	

Field	Data Format	Description	Options or Example	Justification	Consistency Check
image_split_ordinal	Only integers allowed	For composites (see previous field), this field is used to confirm the ordering of the images. Numbering is from top left, along the top row of separate images, then from the left of each successive row (there should be overlap between adjacent images)	<table><tr><td>1</td><td>2</td></tr><tr><td>3</td><td>4</td></tr></table> Or <table><tr><td>1</td><td>2</td><td>3</td></tr><tr><td>4</td><td>5</td><td>6</td></tr><tr><td>7</td><td>8</td><td>9</td></tr></table> Etc. It may be helpful to use the comments field to provide a more human readable version of this e.g. 1=top left, 2=top middle, 3=top right, 4=middle left, 5=middle middle, 6=middle right, 7=bottom left etc.		Validation by The National Archives

Field	Data Format	Description	Options or Example	Justification	Consistency Check
image_split_other_uuid	One or more (separated by a comma) Universally Unique Identifier (UUID). Adhering to UUID Version 4 format and expressed in lower-case hexadecimal characters, see: www.ietf.org/rfc/rfc4122.txt	If the image was split, this field must contain the UUIDs of the other images that were split from the same original image as this image. If there are more than two images as part of a split, this field may contain multiple UUIDs separated by a comma	0d0b88c6-9a6e-4731-ace3-b50794c1356b,a2915f99-6efa-45d4-a0c9-8fd2555643ec	QA	Field shall be empty if Image_split = "no" and populated with valid data if Image_split = "yes" TNA QA will ensure that the other split images exist, and that they also reciprocally point back to this image through their image_split_other_uuid fields
image_crop	Lower case text strings "auto", "manual" or "none"	Specifies if the image was cropped and if it was what type of crop was carried out	auto	QA	
image_deskew	Lower case text strings "yes" or "no"	Specifies if the image was deskewed	no	QA	
comments	Text	Operator's comments, may be empty		QA	Populated at the supplier's discretion

Appendix D: An example of the types of fields required in a transcription metadata CSV file

Field	Data Format	Note	Options or Example	Justification	Consistency Check
batch_code	Up to 16 alpha-numeric characters	An identifier for each batch of records. Supplied by the Authority	Testbatch Y16B001	For consistency and cross checking with other data delivered as part of the batch	The National Archives will cross check this against the batch_code with the naming of the file and the volume label
department	Up to 8 characters	Archival hierarchy	AIR		
division	Up to 8 characters	Archival hierarchy	6		May be empty
series	Up to 8 characters	Archival hierarchy	79		
sub_series	Up to 8 characters	Archival hierarchy			May be empty
sub_sub_series	Up to 8 characters	Archival hierarchy			May be empty
piece	Up to 8 characters	Archival hierarchy	1		
item	Up to 8 characters	Archival hierarchy	2		May be empty
metadata_type	Up to 12 alpha-numeric characters. Taken from an enumeration	Valid metadata types are provided by The National Archives along with a list of enumerated values	ITWW01	For enabling us to validate the content of a row based on the data expected in that row	The metadata_type is a code describing which fields must and should be completed in any particular row. Rows will be validated using this code

Field	Data Format	Note	Options or Example	Justification	Consistency Check
file_path	The file path to the image. Must be a valid URI, see: www.ietf.org/rfc/rfc3986.txt	Location of file relative to the root of the file system containing the batch	file:///AIR_79/content/1/2/1_2_0001.jp2		The file must exist
ordinal	Integer starting from 1	Describes the order of a file within an item or piece. Should start at 1 within each piece or item. See the textual explanation of 'Ordering' above	1		
uuid	Must adhere to UUID Version 4 format www.ietf.org/rfc/rfc4122.txt	The UUID read from the image	c87fc84a-ee47-47ee-842c-29e969ac5131	Uniquely identifies each image	
first_date_day	2 digits - zero padded as appropriate. ? or ?? are used to indicate where individual digits are illegible in the original. If the original is blank then use a single * character	This field will also need to accept impossible dates such as 30 February or 31 April. The first and last dates are intended to capture the date range of the document and may be supplied as a fixed range	1?		

Field	Data Format	Note	Options or Example	Justification	Consistency Check
first_date_month	In full, title-case, no leading or trailing spaces, no punctuation, ? to indicate where characters are illegible. If the original is blank then use a single * character		February		
first_date_year	4 digits, no leading or trailing spaces, no punctuation, ? to indicate where characters are illegible. ???? to indicate where completely illegible. If the original is blank then use a single * character	For 2 digit years in the original The National Archives will provide advice on a Further Competition basis to establish the century.	??14		

Field	Data Format	Note	Options or Example	Justification	Consistency Check
last_date_day	2 digits zero padded as appropriate.? or ?? are used to indicate where individual digits are illegible in the original. If the original is blank then use a single * character	This field will also need to accept impossible dates such as 30 February or 31 April. The first and last dates are intended to capture the date range of the document and may be supplied as a fixed range	03		
last_date_month	In full, title-case, no leading or trailing spaces, no punctuation, ? to indicate where characters illegible. If the original is blank then use a single * character		December		

Field	Data Format	Note	Options or Example	Justification	Consistency Check
last_date_year	4 digits, no leading or trailing spaces, no punctuation, ? to indicate where characters illegible. ???? to indicate where completely illegible. If the original is blank then use a single * character	For 2 digit years in the original The National Archives will provide advice to establish the century	1897		
description		Different for every collection - could be structured in a number of fields or a single field with a short narrative			
language	Three characters representing the ISO 639-3 standard language identification code, see: www.iso.org/iso/catalogue_detail?csnumber=39534 and www.sil.org/iso639-3/default.asp	If all material is in English, this field will not be required.	eng		

Field	Data Format	Note	Options or Example	Justification	Consistency Check
comments	Text	For transcription staff's comments, may be empty		QA	Populated at the supplier's discretion