

UCD Humanities Institute of Ireland

Irish Virtual Research Library and Archive

IVRLA Digitisation Processes 4.0



- CONTENTS -

1.0	INTRODUCTION	...	4
1.1	Foreword	...	4
1.2	Background	...	4
1.3	Principles	...	4
2.0	IVRLA REQUIREMENTS	...	5
2.1	General	...	5
2.2	Bit Depth	...	5
2.3	Capture Resolution	...	5
2.4	File Formats	...	6
2.5	Colour Management	...	6
2.6	Sizing	...	6
2.7	Basic Data Capture	...	6
3.0	CREATION OF PRESERVATION MASTERS (PM)		7
3.1	Introduction	...	7
3.2	PM Images	...	7
3.3	PM Text	...	7
4.0	IMAGE MANIPULATION	...	8
4.1	Orientation	...	8
4.2	Cropping	...	8
4.3	Skewing	...	9
4.4	Stitching	...	9
4.5	Blank Leaves	...	9
4.6	Scanning Summary	...	9
5.0	EDITING AND OPTIMISATION	...	10
5.1	Introduction	...	10
5.2	Procedure	...	10
6.0	CREATION OF SURROGATES	...	10
6.1	Introduction	...	10
6.2	Procedure	...	11
7.0	CREATION OF OFF-LINE IMAGE ARCHIVE		11
7.1	Introduction	...	11
7.2	DVD-ROM	...	12
7.3	LTO Drive	...	12
7.4	LTO Backup Procedures	...	13
8.0	DIGITISATION IN 10 EASY STEPS		14
APPENDIX I	- Pre-Ingestion Batch Processes	...	19
APPENDIX II	- Digitisation Sign off		20
APPENDIX III	- Original Reference Naming Conventions		17
APPENDIX IV	- IVRLA Partid Naming Conventions		23
APPENDIX V	- DVD Naming Conventions		23
APPENDIX VI	- Surrogate Files	...	24
APPENDIX VII	- oXYgen Scan Setup Settings		34

Legal Notices



This document is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 2.5 License. View the full licence [here](#).

© in the collective work – Irish Virtual Research Library and Archive (which in the context of these notices shall mean University College Dublin), 2006.

Catalogue Entry

Title	IVRLA Digitisation Processes
Creator	IVRLA
Subject	Digitisation, image manipulation, digital preservation, naming conventions, image archiving
Description	This document details digitisation procedures as used by the IVRLA Project.
Publisher	IVRLA, UCD Humanities Institute of Ireland.
Contributor	John McDonough, Audrey Drohan, Niall Shields
Date	2007-01-15
Type	Text
Format	Adobe Portable Document Format
Resource Identifier	IVRLA 1.2
Language	English
Rights	© IVRLA, University College Dublin

1.0 INTRODUCTION

1.1 Foreword

This document outlines the workflow and best practice required to successfully manage the digitisation of analogue materials as part of the IVRLA project.

It details the IVRLA scanning requirements and parameters for the creation of IVRLA Preservation Master files, and the subsequent compressed web images and compressed thumbnail images. The IVRLA is implementing a SOAP (Scan Once for All Purposes) methodology by creating high quality archival master files which will be used to generate lower quality (and facilitate optimised) surrogates for web viewing and thumbnails. The project will endeavor to minimise any intervention to the preservation master to ensure archival integrity.

The Project has drawn from documents created and made available by the Library of Congress in compiling these processes and wishes to acknowledge the work undertaken by the Library in this area. The Project would also like to acknowledge the technical assistance given by Noreen Barron, IT Project Officer, UCD Delargy Centre for Irish Folklore and the National Folklore Collection.

1.2 Background

This document should be read in conjunction with other IVRLA produced documents on the care and handling of materials for digitisation, IVRLA metadata, and the IVRLA database.

1.3 Principles

The IVRLA holds the following as the general principles in the digitisation process:

- Best quality possible within constraints of size and projected use.
- Fit for purpose – not everything is the Book of Kells.
- Capture image in full size and without compression.
- No use of interpolation.
- Archive all versions.
- Use standard file formats – TIFF / JPEG.

- Create surrogate images for delivery.
- All images will be stored in an PM format which is full size and uncompressed.
- Surrogate files (DjVu, JPEG, Preview and Thumbnail) will be created for delivery from PM only.
- Explore options for an Optimised images.
- Document everything!

2.0 IVRLA DIGITISATION REQUIREMENTS

2.1 General

The IVRLA mandatory requirement is that each original scanned item shall be reproduced as a set of five digital images:

- (1) An uncompressed preservation master (PM) image in TIFF format
- (2) A compressed reference image for web access as a download (CW1) in JPEG format
- (3) A compressed reference image for web access as a preview (CW2) in JPEG format
- (4) A compressed reference image for web access for user manipulation (CW3) in DjVu format
- (5) A thumbnail image (TN) in JPEG format

Version (1) will be created within the scanning workflow.

Versions (2), and (4) will be created within the image processing workflow.

Versions (3) and (5) will be created as separate batch processes.

In all cases, files will be named in a consistent and semi-human readable format in conformity with the IVRLA file-naming conventions (See **APPENDIX III**).

2.2 Bit Depth

All TIFF images should be 24 bits per pixel, scanned as colour RGB.

JPEG images for CW1. CW2 and TN will be created in post production.

DjVu image files for CW3 will be created in post production.

2.3 Capture resolution

A capture resolution of 450 DPI will be used.

2.4 File Formats for Images/Text

Preservation Master (PM): TIFF format in Intel byte order, uncompressed, 24 bit per pixel,

Compressed Web (CW1 and CW2): JPEG format, compressed, both as a constrained and an unconstrained version.

Compressed Web (CW3): DjVu file, compressed.

Thumbnail (TN): JPEG format, compressed and constrained.

Audio and video formats will be outlined in a separate section/document.

2.5 Colour Management

The IQ Smart will create the correct colour tables for calibrating the scanner. On the computer itself, the apple colour sync system can be used and the monitor and system calibrated to match scanner Adobe RGB 1998.

2.6 Sizing

All manuscript material is to be scanned at 100%.

Negatives are to be scanned at least 200%.

2.7 Basic Data Capture

It is intended that during scanning time, basic metadata will be entered into the IVRLA interim database. At a minimum each image or file requires the following to be logged and noted:

IVRLA Object ID	Part Dimensions
-----------------	-----------------

Title Type	Form
Title	Extent
Type of Resource	Location
Genre	Reformatting Quality
Identifier	Internet Media Type
Part ID	Digital Origin
Original Reference	DVD Number

3.0 CREATION OF PRESERVATION MASTERS (PM)

3.1 Introduction

The Preservation Master (PM) image will be used as the basis for all other iterations. It is not the intention of the project to create optimised masters. Users of the system will have the opportunity to create optimised versions as required (depending upon rights and permissions).

Filenames: See **APPENDIX III - IURLA Partid Naming Conventions.**

TIFF Header Requirements: Intel TIFF version 6.0 is the default standard.

3.2 PM Images

Image requirements in terms of format, resolution, and compression for the mandatory images are as specified below.

- Uncompressed.
- "Intel", with ver. 6.0 headers.
- Must work in IBM-compatible environment.

Mandatory Tonality (pixel-depth resolution):

- Colour: 24 bits-per-pixel
- Greyscale: 16 bits per pixel
- Black and white: 8 bits-per-pixel.

Where possible the goal is to have no sharpening or other enhancement, therefore there will be no sharpening at scanning stage for PM, scan original as is.

Sharpening increases the contrast levels (that should be done for image quality) and will not be 'true' to original. This can be done at a later stage, with caution,

for surrogates.

3.3 PM Text

As with images, a PM of each page will be made as a TIFF file with a corresponding CW version in JPEG created as part of a batch process.

Colour RGB images are required for all categories of handwritten material and for material that requires the preservation of the integrity of the original. Consequently, the majority of the material will be scanned in colour.

Greyscale images may be produced for originals that have significant tonal variation and for printed or typewritten matter with black and white or greyscale illustrations. Colour images shall be created for originals that have significant amounts of colour.

The following chart provides a summary of the image specifications:

Image Type	Description	Format/ Compression	Comment	Resolution(dpi)
Greyscale	16 bits per pixel	TIFF files, no compression	Produced by direct scanning of bound and unbound materials	450
Colour	24 bits per pixel	TIFF files, no compression	Produced by direct scanning of bound and unbound materials	450

4.0 IMAGE MANIPULATION

4.1 Orientation

- All images are to be saved in final repository in portrait, unless landscape is the orientation of the original.
- Mss and bound volume pages are not to be separated into individual files for each page.

In the delivered digital image, the top of the original document or page shall appear at the top of the display screen. Note that "right side up" for printed matter is defined as "the top of the book or magazine page" (portrait mode).

An illustration or table in a book or magazine may be printed "sideways" (landscape) to fit the page, thus aligning the top of the page with the side of the

illustration or table. In these cases, the top of the image shall be the top of the page and not the top of the illustration.

In the case of scanning multiple items simultaneously, crops can be adjusted to the correct orientation within oXYgen Scan.

4.2 Cropping

The project requires presentation of the entire original sheet or page. In no event shall the actual document be cropped.

Researchers often wish to be reassured that the entire document has been captured. A "border zone" approximately 1cm or less of the surface behind the scanned document shall be provided whenever possible.

For some combinations of document sizes and scanning equipment, capturing such a margin may not be possible for all four edges of the page but should be attempted.

4.3 Skewing

IVRLA requires that images created from unbound materials shall not be skewed.

For bound materials, it is required that images shall not be skewed; however, the tightness of the bindings may result in slightly skewed images.

Special care has to be taken by operator to optimise the position of original on scanning bed to avoid this as much as possible.

4.4 Stitching

Documents that are too large to be scanned in one piece should be scanned in two or even four parts, *with considerable overlap*, so that they may be stitched together with PhotoMerge at a later date. Each scanned part shall contain the same named as the full stitched item, but with further identifiers such as a, b, c etc to denote parts.

4.5 Blank Leaves

Interwoven blank leaves in the collection will not be scanned. Notes may be recorded in the notes field of the database indicating skipped blank pages, so as to avoid confusion, especially if leaves are physically attached.

4.6 Scanning Summary

1. Position image
2. Pre-scan
3. Crop a margin and check end-points
4. Scan and assign name(PartID)
5. Update database metadata fields

5.0 EDITING AND OPTIMISATION

5.1 Introduction

This section outlines the steps for image optimisation. It is included only to cover possible digitisation and optimisation on demand. Such actions will not take place as part of the normal workflow.

5.2 Procedure

1. Optimisation of image size - Cropping, Orientation, Perspective correction.
2. Optimisation of image tonality - White Balance (will be calibrated already), Levels and Curves.
3. Cleaning up - removing marks, spots, correcting other image problems.
4. Addition or editing of metadata to internal IPTC fields within image header – including copyright tag.
5. Conversion from RAW or 'high-bit' TIFF format to 24bit RGB Baseline uncompressed TIFF file format or JPEG.
6. Transformation to Adobe RGB 1998 colour space and profiled as such. The computer monitor should be calibrated to Adobe RGB 1998.
7. If required, some limited sharpening 'may' be applied here, although this should be recorded within technical metadata.

6.0 CREATION OF SURROGATES

6.1 Introduction

Under normal circumstances, surrogates in both compressed web and thumbnail versions will be created in batch mode as demanded.

All surrogates will have a similar filename to the archival master with the use of a .jpeg file extension to distinguish the versions, and will subsequently be renamed before ingestion into the repository.

Surrogates will be created using Adobe Creative Suite 2, and can be sent as a batch job using Image Processor.

6.2 Procedure

The surrogates will be created as part of the workflow, on demand for users and special requests, or as part of a batch process, pre-ingestion for the Repository. In such cases, the following steps should be taken:

1. Generate a medium sized (quality 5) JPEG with a watermark from the TIFF files using Image Processor in CS2 (workflow step)
2. Generate a medium sized (quality 5), 310px constrained JPEG from the watermarked JPEG files using Image Processor in CS2 (batch process)
3. Generate a medium sized (quality 5), 100px constrained JPEG from the watermarked JPEG files using Image Processor in CS2 (batch process)
4. Generate a DjVu file in Document Express (workflow step)
5. Generate PDF files for printed pamphlet material in OmniPage Pro 15 (workflow process)

CW Images

- JPEG compression to yield average compression of 10:1 for greyscale and 15:1 for colour
- JFIF format/headers
- Must work in IBM-compatible environment

Mandatory Tonality (pixel-depth resolution):

- Colour: 24 bits-per-pixel
- Greyscale: 16 bits per pixel
- Black and white: 8 bits-per-pixel.

These processes will not, under normal circumstances be used for the routine creation of CW and TN files.

7.0 CREATION OF OFF-LINE IMAGE ARCHIVE

7.1 Introduction

The Project has selected Gold DVD-ROM and Linear Tape Open (LTO) for its master archival backups. DVD-ROM will be used as the primary storage copy whilst LTO will be the preservation copy, with further working copies of the files being stored on the external hard drives.

"Recordable [DVDs] comprise a dye layer and a metallic reflective layer on a clear polycarbonate substrate. Different combinations of metals and dyes are available, giving rise to media of different colours, but recent research suggests that [DVD-Rs] which use a gold reflective layer and phthalocyanine-based dyes (often referred to as "gold/gold" disks) have the greatest life span and are most suitable for archival purposes¹. "

It has been estimated that the daily output from a scanning job is approx 45 - 100 files, depending on the size of the scanned file, totalling 4.3 GB. This should fit onto a DVD-ROM.

The following steps are necessary to back up files to DVD-ROM:

7.2 DVD-ROM

1. The files are burned onto DVD-ROM using Roxio Toast.
2. The name for the first DVD-ROM is assigned as per IVRLA naming conventions (e.g. IVRLA_00001b). This copy of the DVD-ROM is not watermarked and has no extra technical metadata.
3. Approximately 100 images will be written to each DVD-ROM.
4. A second copy of each DVD-ROM (e.g. IVRLA_00001a) will be created after the post-processing workflow is completed. This DVD-ROM has images that are watermarked and that have extra technical metadata, as applied by Bridge.
5. Images which have been burned are then moved to a 'completed' folder.
6. DVD-ROM B is stored in the Project room, and DVD-ROM A will be given to the Repository upon completion of the Project. Until then, it will have a designated storage solution.

7.3 LTO Drive

Deep archiving is done using a LTO (Linear Tape Open) Drive. The images are taken from the external HDD and are backed up onto the LTO Drive using Veritas Backup Executive.

The terminal beside the LTO Drive also includes an Excel file that lists the DVD-ROMs that have been backed up.

Approximately 50 DVD-ROMs will fit on one LTO tape.

¹ THE NATIONAL ARCHIVES (UK), 2003. *Digital Preservation Guidance Note 3, Care, Handling and Storage of Removable Media*, London. Available from: www.nationalarchives.gov.uk/preservation/advice/pdf/media_care.pdf [Accessed 17th February 2005].

7.4 LTO Backup Procedures

Once the TIFF files have been appended with metadata, they are then backed-up onto the LTO drive.

1. Select the 10 folders to be backed up from the External HDD.
2. Right click and select the option to **Back Up using Veritas Backup Exec**.
3. A new window called **Backup Settings** will open. In the Backup Options section of the window, rename **Job Name** by assigning a batch job name according to IURLA LTO naming conventions, example a_iomegaDayMonthYear (eg a_iomega230806).
4. Click **Submit Job**.
5. A pop-up dialog box will appear, called **Veritas Backup Exec**, informing you that the default device will be used. Click **Yes** to continue. The Backup Settings window will now inform you that the job has been sent and to check the Jobs Monitor. Click **Close Dialog** to exit.
6. Launch the **Veritas Backup Exec** software, and open the **Jobs Monitor** tab to view operation in real time.
7. Once the batch job has been successful, close Veritas.
8. Open the **ivrla_dvd_log** excel file on the Desktop and fill in the following details: LTO Tape name -- DVD name -- File range -- Backup date – Operator name.

The LTO now has a copy of each TIFF file with basic metadata, which will be kept in deep storage.

8.0 Digitisation Workflow in 10 Easy Steps

8.1 Introduction

This section outlines the various steps required in the Digitisation workflow.

8.2 Preparation of materials

See the IVRLA Care and Handling guidelines

Step 1: Scan to Mac

The Source Repository material is scanned using oXYgen Scan, is assigned a unique partid file name, and is saved on the Digitisation Workstation.

Scanning Summary

6. Position image
 7. Pre-scan
 8. Crop a margin and check end-points
 9. Scan and assign name(PartID)
- Update database metadata fields

Step 2: Burn DVD B with raw TIFF files

Once sufficient amount of TIFF files has been amassed to fill a 4GB DVD, the files are written to DVD-ROM using Roxio Toast, and the DVD is named using IVRLA DVD naming conventions and is given the suffix 'B'. The TIFF files on this DVD are raw files, with no metadata or watermarks.

Step 3: Save same raw TIFF files to HDD in folders denoting corresponding DVD number

These TIFF files are put into a folder, which is given the same name as the DVD just burned. Build up 10 folders worth of TIFF images before transferring them to the Image Processing machine.

Step 4: Connect HDD to Image Processing PC

The 10 folders of TIFF files are then transferred to a Repository-named folder on the HDD. The HDD is then disconnected from the Digitisation Workstation, and connected to the Image Processing Workstation.

Step 5: In Bridge, append all files with metadata

(ctrl + F to find all>select all>tools>append metadata)

1. Open the application Bridge.
2. Locate the relevant TIFF files on the HDD by going to Edit > Find, and use ‘_PM’ as a parameter.
3. Once the files have been located, go to Edit > Select All.
4. Then go to Tools > Append Metadata > IVRLA. This appends the relevant metadata to the TIFF files on the HDD.

Step 6: Backup all files to LTO (TIFF files with metadata)

Once the TIFF files have been appended with metadata, they are then backed-up onto the LTO drive.

9. Select the 10 folders to be backed up from the External HDD.
10. Right click and select the option to **Back Up using Veritas Backup Exec**.
11. A new window called **Backup Settings** will open. In the Backup Options section of the window, rename **Job Name** by assigning a batch job name according to IVRLA LTO naming conventions, example a_iomegaDayMonthYear (eg a_iomega230806).
12. Click **Submit Job**.
13. A pop-up dialog box will appear, called **Veritas Backup Exec**, informing you that the default device will be used. Click **Yes** to continue. The Backup Settings window will now inform you that the job has been sent and to check the Jobs Monitor. Click **Close Dialog** to exit.
14. Launch the **Veritas Backup Exec** software, and open the **Jobs Monitor** tab to view operation in real time.
15. Once the batch job has been successful, close Veritas.
16. Open the **ivrla_dvd_log** excel file on the Desktop and fill in the following details: LTO Tape name -- DVD name -- File range -- Backup date – Operator name.

The LTO now has a copy of each TIFF file with basic metadata, which will be kept in deep storage.

Step 7: OCR using ScanSoft OmniPage Pro 15

This stage is only applicable if the files are of printed text. If the files are graphical in nature, or contain handwritten material, skip to step 8.

1. Launch OmniPage Pro 15 from the Desktop.
2. Click the Automatic Button to begin [1-2-3]. The Load Image File dialog box appears.
3. Select the folder location and file type of the file you want to load. Files of that type in the selected location appear in the list box. OCR should only be done on non-watermarked TIFF files.

4. Select the files you want to load by opening the **Advanced** section to display the option to select all files and then add them to the Selected Files list, or in the Root Folder selection box, find the folder.
5. Click **OK** when you have selected all the files/folder you want to load.
6. Image files are loaded in the order selected and combined into one working document, OPD.

Once all the pages have been loaded (acquired) and scanned (recognised), a Save prompt box opens. Pages are saved as individual pdf files, taking their name from the input files, and are stored in the 'PDF' folder on the lomega C HDD.

Step 8: In CS2, run Image Processor to create JPGs and TIFFs with watermarks

The images are now ready to be watermarked.

Open CS2. Go to File > Scripts > Image Processor, and for each of the 10 DVD folders:

- Select:
- 1.the DVD-named image folder from the HDD,
 2. the destination folder (DVD named folder on the HDD)
 3. Save as JPEG medium 5,
 4. Save as TIFF,
 5. Run Action : Watermark(action set), watermark(action)
 6. Include ICC Profile

and click OK. Watermarked JPEGs and watermarked TIFFs are automatically saved in separate folders in the location specified.

Delete the original TIFF files from the HDD.

Move the JPEG folder into the JPEG store on the mobile external HDD, and place of copy of this folder into the JPEG storage folder 'JPEG' on the lomega C HDD.

Keep the TIFF folder where it is for now, as it will be needed to generate the DjVu images later.

Step 9: Watermarked TIFFs are used to generate DjVu files

The application Document Express Enterprise is launched from the Desktop. At the bottom of the screen are four tabs. 1. Workflow, 2. Input, 3. Output, 4. Log. Workflow, Input and Output are used to set up the batch job.

Step 1

In the Workflow tab, the following settings are applied.

Job Name = 'DVD number eg IVRLA_10063B'

Raster File = document to djvu

Make sure the box for '**perform OCR**' is ticked.

Step 2

In the Input tab, the following settings are applied.

Click on the option Choose Files beside the **Specific Files** option. This opens a new option window and allows for the selection of files. Find the corresponding DVD number folder, open it, and find the Watermarked TIFF folder, open it and then **click select** all and **ok**.

'When encoding is successful' = option **leave**

'When error occurs' = option **leave**

Step 3

In the Output tab, the following settings are applied.

'Separate documents by' option **each file**

Output location for DjVu files = choose corresponding DVD folder

No xml output

No text output

Step 4

To run the conversion process, tick the **enable** box beside the corresponding job on the left-hand pane of the Document Express window. The process then automatic begins to run. Several jobs can be set up and run – they will be run sequentially and not concurrently.

The Log tab is populated once the conversion process is enabled.

Step 10: Watermarked TIFFs are burned to DVD A for source repositories

The TIFF files, which are still on the HDD in each of the DVD folders, are returned to the Scanning Workstation and are written to DVD-ROM using Roxio Toast, and the DVD is named using IVRLA DVD naming conventions and is given the suffix 'A'. The TIFF files on this DVD have metadata and watermarks. These DVDs are given to the source repositories. Once DVD-A is burned, the folders with the watermarked TIFFs are deleted from the external HDD.

8.4 JPGs with watermark are kept on HDD for future use by cataloguers, web images, Fedora ingestion.

The watermarked JPEGs are kept on the mobile external HDD in a JPEG store folder. Once a sufficient quantity of JPEGs have amassed to fill a DVD,

one is burned using Roxio Toast at the Scanning Workstation and given to the cataloguers.

APPENDIX I – Pre-Ingestion Batch Processes

IN CS2, run Image Processor to create constrained preview JPEGs.

The watermarked JPEG files are then put through Image Processor to generate constrained Preview images for the repository.

Open CS2. Go to File > Scripts > Image Processor.

Select:

- 1.the watermarked TIFF folder from the lomega C HDD,
2. the destination folder (HDD) - Preview,
3. Save as JPEG medium 5,
4. Check the constrain box and put in the value 310 for both width and height,
5. Include ICC Profile

and click OK.

In CS2, run Image Processor to create constrained thumbnail JPEGs.

The watermarked JPEG files are then put through Image Processor again to generate constrained Thumbnail images for the repository.

Open CS2. Go to File > Scripts > Image Processor.

Select:

- 1.the JPEG image folder from the lomega C HDD,
2. the destination folder – Thumbnails on the lomega C HDD,
3. Save as JPEG medium 5,
4. Check the constrain box and put in the value 100 for both width and height,
5. Include ICC Profile

and click OK.

Checklist for Pre-Ingestion

- JPEG files, unconstrained, watermarked
- Preview JPEG files, constrained to 310 pixals, watermarked
- Thumbnail JPEG files, constrained to 100 pixals, watermarked
- PDF files
- DjVu files, watermarked

APPENDIX II – Digitisation Sign Off

The '**Sign-off**' states that the image/metadata was created, checked and found to be up to the required quality standard.

This requirement is met by the Workflow Excel file utilized by all operators.

1. Objects scanned and named.
2. Files are sent to Roxio Toast and burned onto one DVD for the Project
3. Files are then moved into a storage folder on the desktop.
4. Once 10 folders have been amassed, they are put on a mobile external Hard Disk Drive and moved to the Image Processing PC for post-processing.
5. IVRLA metadata appended using the file application Bridge
6. Watermarked TIFF and JPEG files are generated
7. 10 DVD loads of watermarked TIFFs are backed up on the LTO Drive
8. DjVu files are generated from the watermarked TIFFs.
9. Typecript/printed material is OCR'd and saved as PDF files.
10. A separate process generates constrained Preview surrogates from the watermarked JPEGs
11. A separate process generates constrained Thumbnail surrogates from the watermarked JPEGs

Once the files have been backed up on LTO and used to generate jpg surrogates, they may be deleted by the digitisation assistants, in keeping with storage constraints on the workstations.

Now the Iomega C HDD contains for each of the 10 DVD folders:

- watermarked unconstrained JPEGs in the JPEG folder
- watermarked constrained (310px) JPEGs in the PREVIEW folder
- watermarked constrained (100px) JPEGs in the THUMBNAIL folder
- watermarked DjVu files in the DjVu folder
- OCR PDFs in the PDF folder (only for printed material)

APPENDIX III- Original Reference Naming Conventions

Original Identifier

Each Archival Collection and numeric section becomes a unique original_identifier and corresponds to a unique database object

Examples:

LA30/104 becomes original_identifier la30_104 and refers to object OB_0000880_AR
LA30/105 becomes original_identifier la30_105 and refers to object OB_0000881_AR

Original Reference

The originalref field refers to the reference number hand-written on each item. This includes the original_identifier (eg la30_104) and a part number in brackets.

Examples:

LA30/104(1) becomes la30_104_1 for Archives
35.D.2/10 becomes UCD_SC_35_D_2_10 for Special Collections

Title

If this item has subsequent parts(i.e. a back to a photo or letter), the title field beside the originalref tag field is modified to reflect this and to show a relationship to the first part.

Example:

LA30/104(1) is a letter and has a back, and so becomes:
 la30_104_1 [recto]
 la30_104_1 [verso]

The brackets indicate that this title has been assigned by the cataloguer.

Multiple Page Documents

Manuscripts (eg theses, applications, statutes etc) and pamphlets are considered one item as they are (or were) bound and each page is a sub-part, (unless each loose-leaf has been assigned its own part number by the archivist or cataloguer.)

Example of a pamphlet:

UCD_SC_35_D_2_9 [Cover]
UCD_SC_35_D_2_9 Page 1
UCD_SC_35_D_2_9 Pages 2 + 3
UCD_SC_35_D_2_9 Pages 4 + 5
UCD_SC_35_D_2_9 Pages 6 + 7

UCD_SC_35_D_2_9 Pages 8 + 9
UCD_SC_35_D_2_9 Pages 10 + 11
UCD_SC_35_D_2_9 Pages 12 + 13
UCD_SC_35_D_2_9 Pages 14 + 15
UCD_SC_35_D_2_9 Pages 16 + 17
UCD_SC_35_D_2_9 Pages 18 + 19
UCD_SC_35_D_2_9 Pages 20 + 21
UCD_SC_35_D_2_9 Pages 22 + 23
UCD_SC_35_D_2_9 Page 24

There are no brackets around the title for the pages as the title is being taken from the source material.

Sequence

The collections are scanned in readable sequence, as far as possible. An example of this would be a newsletter – single piece of paper, folded in two. Scan 1 is for page one, scan 2 has pages two and three, and scan 3 has page four. However, if the paper has been unfolded and does not fold back easily, only 2 scans are done – scan 1 has pages four and one, scan 2 has pages two and three. There is a possibility of cropping the scans and placing the files in sequence at a later date.

Also, scan items as they are found in the collection, even if the pagination does not reflect the order in which they are archived. Flag any noticeable errors in the notes field.

Missing Parts

Flag all missing parts in the notes field but continue with the original reference numbers in the originalref field. This will mean gaps in the originalref table on the database but it is in keeping with how the items were archived.

Duplication of Part Reference Numbers

Comply with the convention of assigning A to the first item and B to the second item etc. They will then continue to have the original reference number, but can now be differentiated.

Omission of Part Reference Numbers

If the item is between two referenced items, then assign the same reference number as preceding item and add a B to differentiate. If the item is at the end of a sequence, simply assign the next consecutive reference number.

APPENDIX IV - Partid Naming Convention

The following Partid naming conventions have been adopted to distinguish files from the various source repositories.

Each repository has an identifying Prefix and a unique number range.

UCD Archives	AR_PM_0000001
Special Collections	SC_PM_1000001
Irish Folklore Archive	IF_PM_2000001
Irish Dialect Archive	DA_PM_3000001

APPENDIX V – DVD Naming Convention

UCD Archives	IVRLA_000001A
Special Collections	IVRLA_100001A
Irish Folklore Archive	IVRLA_200001A
Irish Dialect Archive	IVRLA_300001A

APPENDIX VI - Surrogate Files

- 1.0 Introduction
- 2.0 File Format Glossary
 - 2.1 TIFF
 - 2.2 JPEG
 - 2.3 DjVu
 - 2.4 PDF
 - 2.5 WAV
 - 2.6 MP3
 - 2.7 RealAudio ram
 - 2.8 MPEG 21
- 3.0 IURLA File Formats
 - 3.1 Images
 - 3.2 Text
 - 3.3 Audio
 - 3.4 Video
 - 3.5 Dataset

1.0 Introduction

A body of digitised content will be presented by the Irish Virtual Research Library and Archive team, as part of the pilot project. The project will be undertaking the digitisation of most of the material; however, a portion will already be in digital format. The digitised content will mainly consist of image, text, audio and video files.

Initially, Preservation Master (PM) files are created for deep storage purposes only. Subsequently, Compressed Web (CW) files are created from these for use as surrogate files in the repository and on the information web-site.

Preservation Master files must be uncompressed in order to retain archival integrity. The surrogate files are compressed file formats, but with little perceivable loss of quality.

Specific file formats are used for both preservation and surrogate files, depending on the type of content in the original resource:

Original	Preservation Master	Surrogates
Image	TIFF	JPEG, DjVu
Text	TIFF	JPEG, DjVu, PDF
Audio	Linear WAV	MP3/ RealAudio ram
Video	MPEG 21	
Dataset	Microsoft Excel File	

2.0 File Format Glossary

Definitions for the file formats used by the IVRLA project are given below:

2.1 TIFF

TIFF stands for Tagged Image File Format and it is a flexible file format.

It can handle multiple images and data in a single file through the inclusion of "tags" in the file header. Tags can indicate the basic geometry of the image, such as its size, or define how the image data is arranged and whether various image compression options are used.

The ability to store image data in a lossless format makes TIFF files a useful method for archiving images.

Unlike standard JPEG, TIFF files can be edited and resaved without suffering a compression loss.

[Adapted from www.wikipedia.org, accessed 20th September 2006]

2.2 JPEG

JPEG is a commonly used standard method of compression for photographic images. The name JPEG stands for Joint Photographic Experts Group, which is the committee that created the standard.

JPEG uses lossy compression algorithms on images. JPEG itself specifies only how an image is transformed into a stream of bytes, but not how those bytes are encapsulated in any particular storage medium.

A further standard, created by the Independent JPEG Group, called JFIF (JPEG File Interchange Format) specifies how to produce a file suitable for computer storage and transmission from a JPEG stream.

[Adapted from www.wikipedia.org, accessed 20th September 2006]

2.3 DjVu

DjVu is a computer file format designed primarily to store scanned images, especially those containing text.

It features advanced technologies such as image layer separation of text and background/images, progressive loading, arithmetic coding, and lossy compression for bitonal images. This allows for high quality, readable images to be stored in a minimum of space, so that they can be made available on the web.

Progressive loading makes the format ideal for images served over the internet.

DjVu can contain an OCRed text layer, making it easy to perform cut and paste operations.

[Adapted from www.wikipedia.org, accessed 20th September 2006]

2.4 PDF

Portable Document Format (PDF) is a file format proprietary to *Adobe Systems* for representing two-dimensional documents in a device independent and resolution independent fixed-layout document format.

Each PDF file encapsulates a complete description of a 2D document that includes the text, fonts, images, and 2D vector graphics that compose the document.

PDF files are most appropriately used to encode the exact look of a document in a device-independent way.

[Adapted from www.wikipedia.org, accessed 20th September 2006]

2.5 WAV

WAV (or WAVE), short for *Waveform* audio format, is a *Microsoft* and *IBM* audio file format standard for storing audio on PCs.

It is a variant of the RIFF bitstream format method for storing data in "chunks". WAVs are compatible with Windows and Macintosh operating systems.

The RIFF format acts as a "wrapper" for various audio compression codecs. It is the main format used on Windows systems for raw audio.

Though a WAV file can hold compressed audio, the most common WAV format contains uncompressed audio in the pulse-code modulation (PCM) format. PCM audio is the standard audio file format for CDs at 44,100 samples per second.

Since PCM uses an uncompressed, lossless storage method, which keeps all the samples of an audio track, professional users or audio experts may use the WAV format for maximum audio quality.

[Adapted from www.wikipedia.org, accessed 20th September 2006]

2.6 MP3

MPEG-1 Audio Layer 3, more commonly referred to as MP3, is a popular digital audio encoding and lossy compression format, designed to greatly reduce the amount of data required to represent audio, yet still sound like a faithful reproduction of the original uncompressed audio to most listeners.

MP3 is an audio-specific compression format. It provides a representation of pulse-code modulation-encoded audio in much less space than straightforward methods, by using psychoacoustic models to discard components less audible to human hearing, and recording the remaining information in an efficient manner.

MP3 audio can be compressed with different bit rates, providing a range of tradeoffs between data size and sound quality.

[Adapted from www.wikipedia.org, accessed 20th September 2006]

2.7 RealAudio ram

RealAudio is a proprietary audio format developed by *RealNetworks*. It uses a variety of audio codecs, ranging from low-bitrate formats that can be used over dialup modems, to high-fidelity formats for music.

It can also be used as a streaming audio format, played at the same time as it is downloaded.

In many cases, web pages do not link directly to a RealAudio file. Instead, they link to a .ram (Real Audio Metadata) or SMIL file. This is a small text file containing a link to the audio stream.

When a user clicks on such a link, the user's web browser downloads the .ram or .smil file and launches the user's media player. The media player reads the RTSP (Real Time Streaming Protocol) URL from the file and then plays the stream.

[Adapted from www.wikipedia.org, accessed 20th September 2006]

2.8 MPEG 21

The MPEG-21 standard, from the Moving Picture Experts Group aims at defining an open framework for multimedia applications.

Specifically, MPEG-21 defines a "Rights Expression Language" standard as means of sharing digital rights/permissions/restrictions for digital content from content creator to content consumer.

As an XML-based standard, MPEG-21 is designed to communicate machine-readable license information and do so in a "ubiquitous, unambiguous and secure" manner.

Among the aspirations for this standard that the industry hopes will put an end to illicit file sharing is that it will constitute: "A normative open framework for multimedia delivery and consumption for use by all the players in the delivery and consumption chain. This open framework will provide content creators, producers, distributors and service providers with equal opportunities in the MPEG-21 enabled open market."

MPEG-21 is based on two essential concepts: the definition of a fundamental unit of distribution and transaction, which is the Digital Item, and the concept of users interacting with them.

Digital Items can be considered the kernel of the Multimedia Framework and the users can be considered as who interacts with them inside the Multimedia Framework.

At its most basic level, MPEG-21 provides a framework in which one user interacts with another one, and the object of that interaction is a Digital Item. Due to that, we could say that the main objective of the MPEG-21 is to define the technology needed to support users to exchange, access, consume, trade or manipulate Digital Items in an efficient and transparent way.

The ability of a consumer to not have to pay multiple times for the same content in different formats is absent.

[From www.wikipedia.org, accessed 20th September 2006]

3.0 IVRLA File Formats

A detailed description of the type of files created by the project, together with their content and function, is given below.

3.1 Images

A substantial collection of images were identified for inclusion in the IVRLA project. Material includes photographs, negatives, slides, watercolours, maps, postcards, and drawings.

The IVRLA project creates four different files for digitised image resources:-

1. Preservation Master (PM) Image file - TIFF

The Preservation Master files for images are scanned as TIFF files in Intel byte order. It is important that these files are uncompressed, thus losing no details and remaining as true to the original as possible.

The images are scanned as colour RGB, 24 bit per pixel, and at 450 dpi (Dots per Inch).

TIFF files are then burned to DVD-Gold (2 copies) and once technical metadata has been appended, they are written to the LTO (Linear Tape Open - high capacity backup and storage) drive for deep archival storage.

TIFF file sizes for this project are quite large, typically between 2MB and 330MB, thus making them unsuitable for display purposes. Consequently they are for preservation purposes only.

2. Compressed Web 1 (CW1) Image file – JPEG

One surrogate derived from the Preservation Master is in the JPEG format, at medium quality 5, and unconstrained.

JPEGs are more suitable for display over the web, as the files sizes are much smaller. JPEG uses a lossy compression method, with some data loss.

The JPEGs are initially being used as resources for the information web site, and as an aid to cataloguing. Ultimately they will reside on the IVRLA server for use in the repository and will be controlled by disseminators.

JPEGs are watermarked and have technical metadata.

3. Compressed Web 2 (CW2) Image file – DjVu

Another surrogate derived from the Preservation Master is in the DjVu format.

This file format generates small file sizes, without the same degree of data loss as the JPEG format.

A viewer is required for this file format which provides the user with a great deal of functionality. The user can zoom to 1200%, and pan.

DjVu files will be kept on the server and are watermarked.

4. Thumbnail (TN) Image file – JPEG

The final surrogate for images is a thumbnail generated from the Preservation Master in the JPEG format. I

This is a compressed file format, at medium quality 5, and constrained (100 pixels tbc).

Thumbnails will be kept on the server and used by the repository, and are watermarked.

3.2 Text

A substantial amount of text was identified for inclusion in the IVRLA project. Material includes letters, memorabilia, notes, newspaper cuttings, and pamphlets.

Printed material is subjected to a further process of optical character recognition to generate searchable text

The IVRLA project creates four or five different files for digitised text resources, as applicable:-

1. Preservation Master (PM) Text File – TIFF

The Preservation Master files for text are scanned as TIFF files in Intel byte order.

It is important that these files are uncompressed, thus losing no details and remaining as true to the original as possible.

The images are scanned as colour RGB, 24 bit per pixel, and at 450 dpi (Dots per Inch).

TIFF files are then burned to DVD-Gold (2 copies) and once technical metadata has been appended, they are written to the LTO (Linear Tape Open - high capacity backup and storage) drive for deep archival storage.

TIFF file sizes for this project are quite large, typically between 2MB and 330MB, thus making them unsuitable for display purposes. Consequently they are for preservation purposes only.

2. Compressed Web 1 (CW) Text File – JPEG

One surrogate derived from the Preservation Master is in the JPEG format, at medium quality 5, and unconstrained.

JPEGs are more suitable for display over the web, as the files sizes are much smaller. JPEG uses a lossy compression method, with some data loss.

The JPEGs are initially being used as resources for the information web site, and as an aid to cataloguing. Ultimately they will reside on the IVRLA server for use in the repository and will be controlled by disseminators.

JPEGs are watermarked and have technical metadata.

3. Compressed Web 2 (CW) Text File - DjVu

Another surrogate derived from the Preservation Master is in the DjVu format.

This file format generates small file sizes, without the same degree of data loss as the JPEG format.

A viewer is required for this file format which provides the user with a great deal of functionality. The user can zoom to 1200%, pan, and perform a basic page text search.

DjVu files will be kept on server and are watermarked.

4. Thumbnail (TN) Image file – JPEG

Another surrogate for text is a thumbnail generated from the Preservation Master in the JPEG format.

It is a compressed file format, at medium quality 5, and constrained (100 pixels tbc).

Thumbnails will be kept on the server and used by the repository, and are watermarked.

5. OCR text file – PDF

The final surrogate derived from the Preservation Master file for text is in the PDF format. This is only used for printed material of a certain quality,

This is generated through the OCR (Optical Character Recognition) process, which scans the TIFF file, deciphers the characters on the page and creates a searchable text file.

This file is then saved as a PDF, giving an even greater degree of searchability.

Due to the acceptable percentage of error and loss of formatting, this file is for backend use only and will not be seen by the user. The PDF files will be kept on the server.

3.3 Dataset

A sample set made up of cards from the Irish Dialects Archive was deemed unsuitable for scanning. It contained data that needed to be searchable and yet was not suitable for OCR, due to its use of orthography.

Consequently the data was manually transferred from the cards into a customised *FileMaker Pro* database, and the data was then extracted out into an Excel file.

3.4 Audio

A certain amount of audio files have been identified by the IVRLA Project for inclusion into the repository.

These analogue formats will have to be converted into digital formats.

The digitisation of audio files is currently in the planning stage. The project envisages using the following three file formats:

1. Audio Preservation Master file – WAV

The Preservation Master files for audio will be converted from the current analogue format into the digital file format Linear WAV.

It is important that these files are uncompressed, thus losing no sound quality and remaining as true to the original as possible.

2. Audio Compressed Web 1 file – MP3

One surrogate derived from the Preservation Master WAV file will be in the MP3 format.

MP3 is a lossy compression format, designed to reduce the amount of data required to represent audio, while remaining faithful to the original uncompressed audio as possible.

Its smaller size makes it ideal as a presentation method of audio over the web.

3. Audio Compressed Web 2 file – RealAudio ram

Another surrogate derived from the Preservation Master WAV will be the Real Audio ram format.

This file format can be used as streaming audio that is played at the same time as it is downloaded.

3.5 Video

A certain amount of audio files have been identified by the IVRLA Project for inclusion into the repository.

The digitisation of video files is currently in the planning stage. The project envisages using the following file format:

Video Preservation Master file – MPEG 21

The Preservation Master files for video will be converted from the current 16mm, Beta, U-matic and VHS formats into the digital file format MPEG 21.

It is important that these files are uncompressed, thus losing no picture or sound quality and remaining as true to the original as possible.

APPENDIX VII - oXYgen Scan Setup Settings

Settings to be utilized with the scanning software, oXYgen, include:

- **Type:** Reflective for positives, transparent for negatives.
- **Media:** Positive or negative, as applicable.

Input Profile, Endpoints, Gradation and **Sharpness** are left at the software defaults.

Sharpness

Basic

Sharp: 6
Smooth: 5
Format: Reflective

Extended

Filter: Green
Radius: 5
Effect: Medium

Intensity

Highlight: 5
Shadow: 6

Grain

Threshold: 4
Value: 5

Clip

Highlight: No Effect
Shadow: No Effect

File Format Setup

TIFF Settings: Macintosh Byte Order

Densitometer Setting

Display: Output
Units: System Value
Sample Size: 3 x 3

Smooth Setup

Set: User

	R	G	B
85 Lpi	50	50	50
100 Lpi	50	50	50
133 Lpi	50	50	50
150 Lpi	50	50	50
175 Lpi	50	50	50
200 Lpi	50	50	50

ICC Flow Setup

Rendering Intent

Reflective

RGB: Automatic

CMYK: Automatic

Transparent

RGB: Automatic

CMYK: Automatic

Output Simulation Automatic

Prefix/Suffix Setting

Default Setting, with the option of **File Format Suffix** checked.

Language Setting

English