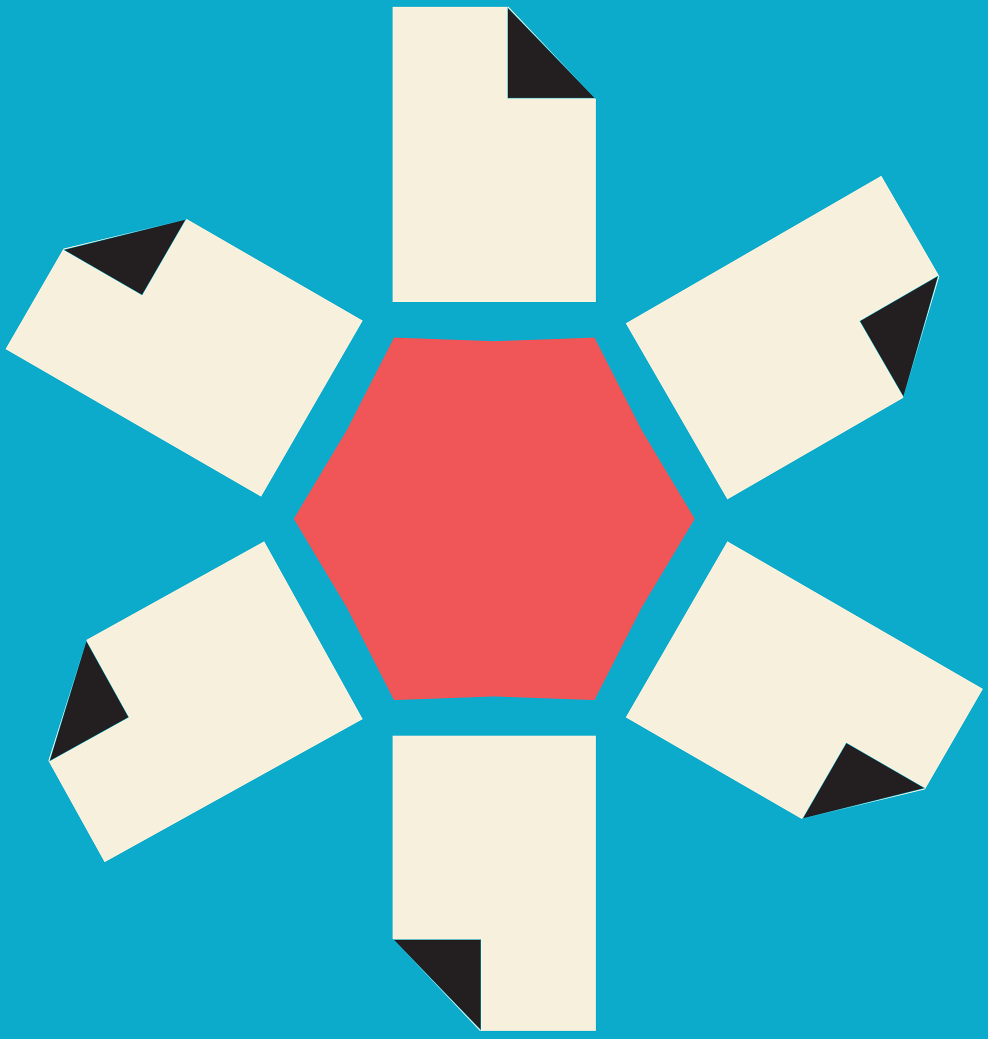

Format watch report

dri

Summary

A format review was carried out by DRI in July 2018. The results of the review are presented here. The file formats in the repository were identified and for each file type, various registries were consulted to determine its sustainability and preservability, including the UK National Archives[1], Library of Congress[2], the Digital Preservation Coalition[3] and the International Association of Sound and Audiovisual Archives[4].

The DRI Preservation Policy[5] and Factsheet 3 File Formats[6] also informed this investigation.

Required Actions

- Investigate creation of web-renderable surrogates for MS Word .doc files.
- Continue to monitor suitability of QuickTime .mov files for preservation.

[1] http://apps.nationalarchives.gov.uk/pronom/Default.aspx
[2] https://www.loc.gov/preservation/digital/formats/intro/intro.shtml
[3] https://www.dpconline.org/knowledge-base/tech-watch-reports
[4] https://www.iasa-web.org/tc06/guidelines-preservation-video-recordings
[5] Digital Repository of Ireland. DRI Preservation Policy, Digital Repository of Ireland [Distributor], Digital Repository of Ireland [Depositing Institution], https://doi.org/10.7486/DRI.2r377c523
[6] Digital Repository of Ireland. DRI Factsheet No 3: File formats, Digital Repository of Ireland [Distributor], Digital Repository of Ireland [Depositing Institution], https://doi.org/10.7486/DRI.rj43ck402

File Type Summary

Images

Type	Count
jpeg	24967
tiff	13242

While tiff is a preferred format for preservation, jpeg is also acceptable. Both are widely supported and no preservation actions are required at this time for these formats.

Documents

Type	Count
msword	341
pdf	6154
plain text	674
xml	2000

PDF, Plain text and XML files are all preferred formats for preservation by DRI. MS Word doc and docx files are acceptable formats.

Although doc is a proprietary format, it is ubiquitous and has ongoing support from Microsoft, and the specification has been published[7]. There are many different versions of Word doc files, however. The UK National Archives DROID tool[8] was used to further identify the a random sample of 15% of these files. Based on the sampling results, the 341 MS Word assets are Microsoft Word 97 - 2003 .doc files[9]. Other tools such as the Unix file tool reported that the majority were created with Word 6.1, with a small number created with Word 5.1.

Best results for conversion are usually achieved by using Word directly to perform the conversion and there are often problems in conserving the essential properties of the files when conversions performed using other tools. Because of these factors, DRI believes that no immediate preservation actions are required for these files. This is inline with how MS Word doc files are handled by other repositories[10][11].

It is worth noting that DRI does not currently produce web-renderable surrogates for Word doc files. We recommend that possible surrogate formats be investigated as this would also inform any future migration strategy.

[7] https://msdn.microsoft.com/en-us/library/cc313153.aspx
[8] http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/
[9] http://www.nationalarchives.gov.uk/pronom/fmt/40
[10] http://www.loc.gov/preservation/resources/rfs/textmus.html#digital
[11] https://wiki.archivematica.org/Word_processing_files

Moving Image

Type	Count
mp4	18
quicktime	6
x-msvideo	3

The UK National Archives DROID tool was again used to investigate the video files in the repository. They are in the formats MPEG-412, Quicktime MOV13 and AVI14. These are not preferred formats for preservation, but they are acceptable to DRI. MP4 and AVI are very well supported and we do not believe that any preservation actions are required in these cases.

Apple are no longer supporting QuickTime for Windows beyond version 2.0 but QuickTime (.MOV) files are still playable through open source media players such as VLC (available for Windows). Apple have published two versions of the QuickTime File Format specification[15] and although it is not clear whether future versions of the specification will continue to be published, the files currently in the Repository all play back correctly in VLC. DRI is monitoring the situation for any further changes in the sustainability of .MOV files.#

Audio

Type	Count
mp3	328
wav	196

WAV is a preferred format for preservation by DRI. MP3 is an acceptable preservation format. Due to the ubiquity of MP3 and its status as a de facto standard for web-based audio, DRI is confident in its ability to preserve these files in the medium to long term and no preservation actions are required at this time.

[12] http://www.nationalarchives.gov.uk/pronom/fmt/199
[13] http://www.nationalarchives.gov.uk/pronom/x-fmt/384
[14] http://www.nationalarchives.gov.uk/pronom/fmt/5
[15] https://developer.apple.com/standards/classic-quicktime/

Appendix

File Types

Mime Type	Count
jpeg (exchangeable image file format)	12831
jpeg (jpeg file interchange format)	11303
tiff (tiff exif)	8684
pdf (portable document format)	3391
tiff (tagged image file format)	2963
pdf (pdf/a)	2758
xml (extensible markup language)	2000
tiff (tiff exif, tagged image file format)	1283
jpeg (jpeg, jpeg image data)	811
plain (plain text, plain text, rtf, rich text format, rich text format, rich text format file)	673
msword (microsoft word document)	341
mpeg (mpeg 1/2 audio layer 3)	279
x-wave (waveform audio)	138
tiff (tagged image file format, exchangeable image file format (uncompressed))	132
tiff (Tagged Image File Format)	98
x-wave (wave, waveform audio)	55
mpeg (MPEG 1/2 Audio Layer 3)	49
tiff (tiff dlf benchmark for faithful digital reproductions of monographs and serials: color, tagged image file format)	36
tiff (tiff exif, exchangeable image file format (uncompressed))	23

Mime Type	Count
jpeg (JPEG File Interchange Format)	19
tiff (tiff dlf benchmark for faithful digital reproductions of monographs and serials: color)	19
mp4 (iso media, mpeg v4 system, version 2)	17
quicktime (iso media, apple quicktime movie, mov, quicktime)	6
jpeg (Exchangeable Image File Format)	3
pdf (pdf/x)	3
png (portable network graphics)	3
x-msvideo (audio/video interleaved format)	3
pdf (pdf/a, pdf exif)	2
x-wave (waveform audio, microsoft excel format)	2
zip (zip format)	2
mp4 (iso media, mpeg v4 system, itunes avc-lc, m4v)	1
plain (plain text)	1
tiff (tagged image file format, microsoft excel)	1
tiff (tiff dlf benchmark for faithful digital reproductions of monographs and serials: color, exchangeable image file format (uncompressed))	1
tiff (tiff dlf benchmark for faithful digital reproductions of monographs and serials: grayscale and white, tagged image file format)	1
tiff (tiff exif, microsoft word document)	1
vnd.ms-excel (microsoft excel, microsoft excel format)	1
vnd.openxmlformats-officedocument.spreadsheetml.sheet (microsoft excel 2007+, opendocument text, office open xml workbook)	1
x-empty (empty, tagged image file format for image technology (tiff/it))	1
x-wave (waveform audio, mpeg 1/2 audio layer 3)	1