Public Record Office Victoria

SPECIFICATION

PROS 19/05 S3: LONG TERM SUSTAINABLE FORMATS

Version number: 1.1
Issue Date: 15 April 2020
Expiry Date: 29 August 2029

About this Specification

All Victorian public offices must comply with this Specification when preparing digital records for transfer to PROV. Permanent value digital records must be in one of the specified formats or able to be converted reliably and efficiently to one of these formats.

It is recommended that Victorian public offices also use these formats when temporary value records need to be kept for a long period of time.

The diagram below shows the relationship between this Specification and related documents.

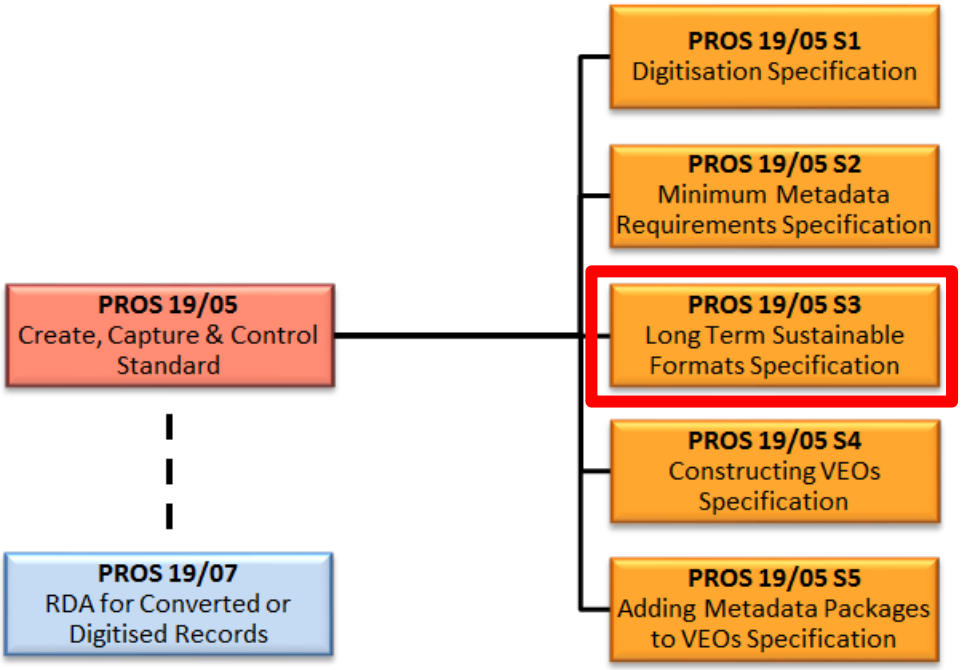

Table of Contents

1 Introduction 3

1.1 Authority of Standards and Specifications 3
1.2 Obligation 3
1.3 Applying this Specification 3

2 Format Selection 4

2.1 Criteria for Format Selection 4
2.2 Format Requirements 4
2.3 Avoiding Migration 4
2.4 Format Selection 4
2.5 Version Selection 5
2.4 File Validity 5
2.5 Inclusion of Original Format 5

3 Long Term Sustainable Formats 6

4 APPENDIX – Further Information 8

DOCUMENTS 8
TEXT 9
SPREADSHEETS 9
PRESENTATIONS 10
WEBSITES 10
WEBSITES (*ARCHIVAL*) 11
EMAIL 12
IMAGE (*RASTER*) 12
IMAGE (*VECTOR*) 13
IMAGE (*SCANNED DOCUMENTS*) 14
IMAGE (*GRAPHIC METAFILES*) 14
AUDIO 14
VIDEO (*BARE*) 15
VIDEO (*CONTAINER*) 16
DATA 17
ELECTRONIC PUBLICATIONS 17
ENCAPSULATION 18
COMPUTER AIDED DESIGN (CAD) 18
GEOSPATIAL 19

1 Introduction

1.1 Authority of Standards and Specifications

Under section 12 of the *Public Records Act 1973,* the Keeper of Public Records ('the Keeper') is responsible for the establishment of Standards for the efficient management of public records and for assisting Victorian public offices to apply those Standards to records under their control.

Heads of public offices are responsible under section 13b of the *Public Records Act 1973* for carrying out a program of efficient management of public records. The program of records management needs to cover all records created by the public office, in all formats, media and systems across the organisation.

The Standards and Specifications support the Victorian Electronic Record Strategy (VERS) Digital Forever 2018-2021[1], which is designed to ensure the creation, capture and preservation of authentic, complete and meaningful digital records.

This Specification is part of the *PROS19/05 Create, Capture and Control Standard*. This Specification, as varied or amended from time to time, shall have effect for a period of ten (10) years from the date of issue unless revoked prior to that date.

1.2 Obligation

It is mandatory for all Victorian public offices to follow the principles and comply with the requirements of the Standards and Specifications issued by the Keeper.

1.3 Applying this Specification

This Specification sets out the formats that have been assessed by the Public Record Office Victoria (PROV) as likely to remain readable and usable for a long period of time and provides guidance on format selection.

Permanent value digital records must be in one of these formats or able to be converted reliably and efficiently to one of these formats. This will help to ensure:

- records are readable and usable while held by the public office; and
- VERS encapsulated objects (VEOs) can be constructed when records are being prepared for transfer to PROV.[2]

It is recommended that these formats also be used when temporary value records need to be kept for a long period of time.[3] This will help to ensure records are readable and usable for the time they need to be retained and records can reliably and effectively be migrated during system or organisational changes (e.g. machinery of government changes, administrative mergers etc.).

---

[1] The previous *PROS15/03 Standard for the encapsulation of digital records* has been revoked and the requirements have now been included in the *PROS19/05 Create, Capture and Control Standard* and associated Specifications.

[2] If permanent value digital records planned for transfer are not in one of these formats or can't be reliably and efficiently converted to one of these formats, seek advice from PROV.

[3] For example, if the records must be retained for longer than the life of the system. As it is not necessarily known when formats are being determined, an arbitrary time period might be selected. For example, it might be decided that temporary records which must be retained for longer than 10 years will be held in one of the formats set out in this Specification.

2 Format Selection

2.1 Criteria for Format Selection

There is a risk that in future it will not be possible to obtain software to extract or present information embedded in a digital object. Over a sufficiently long period of time, all formats can be expected to become unreadable. At some point it will be necessary to undertake a preservation action for a format. This specification aims to identify formats which are likely to be available for a long period of time and for which the necessary tools are likely to be available when records need to be exported/migrated in the future.

The formats were chosen because they meet one or more of the following criteria:

- extremely widespread adoption
- dominance in a particular category
- independent use by a range of vendors for the software they have developed
- well documented
- adherence to a published formal specification
- no restrictions on usage, preferably open source
- stability over a period of time
- already accepted by the previous version of this Standard for the above reasons.

2.2 Format Requirements

All record content transferred to PROV must be represented in one of the long term sustainable formats outlined.

Public offices are strongly encouraged to also use these formats for digital records that must be retained for a long period of time. Use of the formats will aid in ensuring long term access to the content. The longer the information needs to be kept, the more important it is to use one of these long term sustainable formats.

2.3 Avoiding Migration

Public offices are strongly encouraged to adopt these long term sustainable formats for day to day business use. This will avoid the requirement to subsequently convert records to these formats.

Migration from one format to another is to be avoided. In general, migration is expensive (to obtain the necessary tools, to carry out the migration, and to conduct the necessary quality assurance to ensure the migration was carried out successfully). There is always a risk of losing information when migrating.

2.4 Format Selection

Several long term sustainable formats are provided for most categories of information.

Public offices can choose the most appropriate format for their business needs. In accordance with avoiding migration, the most appropriate format to choose will normally be the one in which the business is actually undertaken.

2.5 Version Selection

Unless otherwise indicated, PROV will accept any version or variant of the selected long term sustainable formats. This is because, in most cases, it is difficult for public offices to configure or setup software products to produce particular variants of a format. Further, most software will produce the latest version[4] of a particular format.

Where particular versions of a format are preferred over others this is indicated. Public offices are encouraged to adopt the recommended versions where possible.

2.4 File Validity

PROV reserves the right to reject records that are not valid according to the format specification.

2.5 Inclusion of Original Format

Where record content has been migrated to a long term preservation format for the purposes of transfer, a copy of the original, un-migrated format must also be included in the VEO unless otherwise agreed by PROV.

The requirement to include a copy of the original format guards against the following risks that:

- the migration did not result in an accurate representation of the original record
- the migration caused a loss of information
- a better migration approach may be available in the future.

PROV will not generally require a copy in original format where:

- the record is extremely large, and
- the migration process is a routine technology process with little chance of content loss.

A typical example of a situation where the original would not normally be required is the conversion of video or audio to a long term preservation format.

[4] At the time the creation software was produced.

3 Long Term Sustainable Formats

The following formats have been assessed by PROV as being sustainable over a long period of time. Please contact us if a commonly used format has not been included. See the Appendix for more information about each format, including reasons for its inclusion.[5]

Table 1: Sustainable formats.

TYPE	SUSTAINABLE FORMAT
Documents and text	• Plain text (.txt) • Portable Document Format (.pdf)[6] • Microsoft Word (.doc, .docx) • Open Office Document (.odt)
Spreadsheets	• Microsoft Excel (.xls, .xlsx) • Open Office Spreadsheet (.ods) • Portable Document Format (.pdf) • Comma separated values (.csv)[7] • Tab-separated values (.tsv)
Presentations	• Microsoft PowerPoint (.ppt, .pptx) • Open Document Presentation (.odp) • Portable Document Format (.pdf)
Websites	• HyperText Markup Language (.htm, .html) or Extensible Markup Language (.xml)[8] together with supporting files (.css, .xsd, .dtd)[9] • Web ARChive format (.warc)[10] • Internet archive (.arc)
Email	• Mime encoding (.eml) • UNIX mailbox (.mbx or .mbox) • Microsoft Outlook (.msg or .pst)
Image (raster)	• Joint Photographic Experts Group (.jpg, jpeg) • JPEG2000[11] (.jp2) • Tagged image file format (.tif, .tiff) • Portable Networks Graphic (.png) • Digital Negative (.dng)

[5] A good resource for more detailed information is the Library of Congress https://www.loc.gov/preservation/digital/formats/index.html

[6] It is preferred that PDF documents be conformant to PDF/A-1 (ISO 19005-1) or PDF/A-2 (ISO 19005-2). PDF/A-3 must not be used.

[7] See http://tools.ietf.org/html/rfc4180 for a non-normative definition of a CSV file.

[8] XML files will be readable for the indefinite future, but they may not be interpretable. This is because the meaning of the XML markup is defined in separate standards (e.g. SVG for vector graphics).

[9] It is recommended that HTML files conform to HTML 4.01 Specification and CSS 2.1 Specification.

[10] Note that WARC is a container format that encapsulates web objects. Each web object (e.g. web pages) in the WARC file must be in one of the long term preservation formats specified in this specification.

[11] JPEG2000 is accepted as a long term preservation format in PROS 99/007 (Version 2.0). For this reason, PROV will continue to accept files in this format.

TYPE	SUSTAINABLE FORMAT
Image (vector)	• Scalable Vector Graphics (.svg) • Open Document Graphics (.odg)
Image (graphic metafile)	• Computer Graphics Metafile (.cgm)
Image (multi-page scanned document)	• Portable Document Format (.pdf)
Audio	• Moving Picture Experts Group MPEG 1/2 Audio Layer 3 (.mp3) • Moving Picture Experts Group MPEG-4 (.mp4) • Free Lossless Audio Codec (.flac) • WAV using an LPCM codec (.wav, .bway, .bwf)
Video (in a container)	• Ogg (.ogg or .ogv) • Digital Cinema Package (.dcp) • Moving Picture Experts Group MPEG2 (.mpg, .mpeg) • Moving Picture Experts Group MPEG4 (.mp4, .m4v, .m4a, .f4v, . f4a)
Video (bare)	• Motion JPEG2000 (.mjp, .mj2, .m2v) • Digital Moving Picture Exchange DPX (.dpx)
Data files	• Comma Separated Values (.csv) • Tab-separated Values (.tsv) • Extensible Markup Language (.xml) • JavaScript Object Notation (.json or .jsn) • Software Independent Archiving of Relational Databases (.siard)
Electronic publications	• Portable Document Format (.pdf) • Electronic Publication (.epub)
Encapsulation	• ZIP (.zip) • GZIP (.gzip) • Tape Archive (.tar)
Computer Aided Design (CAD)	• Drawing eXchange Format (.dxf) • Drawing (.dwg) • Standard for the Exchnge of Product Data (.stp, .step, or .p21) • Portable Document Format/E (.pdf)
Geospatial	• ESRI Shapefiles • Geopackage • GEOjson (.json) • Digital Elevation Model, Geography Markup Language (.dem,.gml) • Keyhole Markup Language (.kml or .kmz) • GEOtiff (.tiff) • BigTiff (.tiff) • Enhanced Compression Wavelet (.ecw) • JPEG2000 (.jp2) • Scalable Vector Graphics (.svg), • Log ASCII Standard (.las) • Image Format (.img)

4 APPENDIX – Further Information

DOCUMENTS

Use these formats for documents produced by word processing programs (e.g. Word).

WARNING: Documents may have a range of other format types (e.g. images) embedded within the document. When producing documents, this embedded content should be in an appropriate long term preservation format (e.g. JPEG for photographs).

FORMAT	NOTES	RATIONALE
DOCX	Microsoft Word Document (XML). Although this is a format used by a specific proprietary software product, it is not necessary to use this product to create DOCX documents.	This format is an international standard, based on XML. Microsoft Word is the clear market leader in office document preparation. Very large numbers of documents are created in this format.
DOC	Microsoft Word Document. Although this is a format used by a specific proprietary software product, it is not necessary to use this product to create DOC documents.	This format is an Ecma standard (an industry-based standards organisation), so is documented. Microsoft Word is the clear market leader in office document preparation. Very large numbers of documents are created in this format.
ODT	OpenDocument Text Document Format. The OpenDocument format is a generic format for office documents, such as text, spreadsheets, presentations, drawings, and databases.	The format is an open standard, based on XML, and is supported by a variety of office software packages.
PDF	Portable Document Format. Developed by Adobe to deliver final 'published' documents that are primarily intended to be read by end users. It is difficult to extract data from PDF files and to modify them. These documents may be structured or simple. They can include text, images, graphics, and other multimedia content, such as video and audio, and are independent of application software, hardware, and operating systems. For recordkeeping purposes, we recommend using variants of PDF-A, but PDF 1.7 is acceptable. This format offers several forms of technical protection, including encryption, which could limit accessibility in future technological environments. These technical mechanisms should not be used.	Fully documented. Most PDF formats were developed by Adobe Systems Inc. which makes the specifications openly available at no charge. Several members of the family have been adopted as ISO international standards, e.g. PDF/X, PDF/A, and PDF version 1.7. Extremely widely adopted for disseminating page-oriented documents.

TEXT

FORMAT	NOTES	RATIONALE
TXT	A .txt file generally contains only plain text (without formatting), intended for humans to read. The Unicode character set, and its encoding as UTF-8 or UTF-16 is the standard and widespread method of representing textual characters. The UTF-8 encoding was designed to be backwards compatible so that the older ASCII encoding is also valid UTF-8. UTF-16 caters for the additional characters needed for non-English text.	Plain text is public, standardised, and universally readable. The use of plain text provides independence from programs that require their own special encoding or formatting or file format. Plain text files can be opened, read, and edited with countless text editors and utilities.

SPREADSHEETS

FORMAT	NOTES	RATIONALE
XLSX	Microsoft Excel Spreadsheet (XML). Although this is a format used by a specific proprietary software product, it is not necessary to use this product to create XLSX spreadsheets.	This format is an international standard, based on XML. Microsoft Excel is the market leader in spreadsheet preparation. Very large numbers of spreadsheets are created in this format.
XLS	Microsoft Excel Spreadsheet. Although this is a format used by a specific proprietary software product, it is not necessary to use this product to create XLS spreadsheets.	This format is an Ecma standard, so is documented. Very large numbers of spreadsheets are created in this format.
ODS	OpenDocument Spreadsheet Document Format. The OpenDocument format is a generic format for office documents, such as text, spreadsheets, presentations, drawings, and databases.	The format is an open standard, based on XML, and is supported by a variety of office software packages.
PDF	Spreadsheets saved to this format have restricted functionality – it is difficult to extract data from them. The underlying formulae that are used to calculate the information are not captured. This format is most appropriate for final 'published' spreadsheets that are exclusively intended to be read by end users.	Fully documented. Most PDF formats were developed by Adobe Systems Inc. which makes the specifications openly available at no charge. Several members of the family have been adopted as ISO international standards, e.g. PDF/X, PDF/A, and PDF version 1.7. Extremely widely adopted for disseminating page-oriented documents.
CSV	Note that spreadsheets in this format will have restricted functionality – the format only represents the data. It cannot store the underlying formulae used to calculate the data, nor can it represent the formatting of particular cells. It should only be used when only the data in the spreadsheet is of value.	As it is simple text, this format is extremely easy to access and process.

PRESENTATIONS

WARNING: Office documents may have a range of other format types (e.g. images) embedded within the document. When producing documents, this embedded content should be in an appropriate long term preservation format (e.g. JPEG for images).

FORMAT	NOTES	RATIONALE
PPTX	Microsoft Powerpoint (XML). Although this is a format used by a specific proprietary software product, it is not necessary to use this product to create PPTX presentations.	This format is an international standard, based on XML. Further, Microsoft Powerpoint is the clear market leader in simple presentation preparation. Very large numbers of presentations are created in this format.
PPT	Microsoft Powerpoint. Although this is a format used by a specific proprietary software product, it is not necessary to use this product to create PPT presentations.	This format is an Ecma standard, so is documented. Very large numbers of presentations are created in this format.
ODP	OpenDocument Presentation Document Format. The OpenDocument format is a generic format for office documents, such as text, spreadsheets, presentations, drawings, and databases.	The format is an open standard, based on XML, and is supported by a variety of office software packages.
PDF	Portable Document Format. Presentations in this format will have restricted functionality – it is difficult to extract information from them and to modify them. It is most appropriate for final 'published' presentations that are primarily intended to be read by end users.	Fully documented. Most PDF formats were developed by Adobe Systems Inc. which makes the specifications openly available at no charge. Several members of the family have been adopted as ISO international standards, e.g. PDF/X, PDF/A, and PDF version 1.7. Extremely widely adopted for disseminating page-oriented documents.

WEBSITES

The formats in this section represent a captured website. These represent the files that are transmitted to a web browser to display a web page.

Note that while several formats are presented here, many of these formats capture only an aspect of the website (e.g. CSS files capture style information that is applied to other HTML or XML files). Therefore, a website capture typically contains a range of these file types. It is not sufficient to preserve only the content (HTML or XML) files if the intent is to capture the core aspects of a website.

Further, a website typically delivers various types of content (e.g. images, audio, etc.). This section is only concerned with the formats which structure and display web information. Other delivered content may be covered by formats mentioned elsewhere in this appendix.

Finally, these web formats only relate to 'static' web pages. Many web pages also include dynamic content – a trend which is increasing. This could include code that is executed in the web browser (e.g. client side JavaScript) to achieve certain effects. Or it could include API calls to servers to retrieve content as the user performs actions (e.g. Google map embedded in the page). A preserved website which depends partly upon client-side executed scripts will likely vary in completeness and sometimes pull content from the live web ("leakage") which is likely to be chronologically inconsistent. It is essentially impossible to fully capture these types of web pages in a form that is suitable for long term preservation but the results might still be acceptable, depending on the extent and nature of the client-side scripting used.

FORMAT	NOTES	RATIONALE
HTM, HTML	Any variant acceptable, with a preference to the later versions. Different web browsers will display HMTL pages differently. Later versions of HTML are more consistent in their display. Many HTML pages have an associated CSS resource that controls how the page is displayed. It is consequently necessary to capture both the HTML page and the CSS resource to properly capture the page.	HTML is a text based standard. The various versions of HTML are standards which are widely available. HTML is ubiquitous on the internet.
XML	Any variant acceptable, with a preference to the later versions. XML pages have an associated CSS resource that controls how the page is displayed. It is consequently necessary to capture both the XML page and the CSS resource to properly capture the page.	XML is an international standard that is extremely widely used in a range of applications.
CSS	Used to represent the styles of elements in XML and HTML. There are a large range of versions of the standard, with more features being added continually.	A text based international standard, it is very widely used to style web pages.
XSD	XML Schema Definition. These files are used to control the structure of XML documents. They are not necessary to display or use XML documents, but some tools will require them.	This is a text based international standard.
DTD	XML Document Type Definition. These files are used to control the structure of XML documents. They are not necessary to display or use XML documents, but some tools will require them.	DTD documents are part of the XML international standard.

WEBSITES (*ARCHIVAL*)

The previous section described the types of files that must be captured in order to preserve a web page (or collection of webpages. This section describes formats that are used to encapsulate the files representing a webpage or website into a single archival object. They are equivalent to the encapsulation formats described below, but are targeted specifically at encapsulating web sites.

FORMAT	NOTES	RATIONALE
WARC	Web ARChive file format. This format was developed from the ARC format.	This format is commonly used internationally by archives and libraries to perform large scale harvest and capture of web sites. It is consequently well supported by tools. It is a simple encapsulation format and would be easy to migrate if necessary.
ARC	Internet Archive ARC file format. This format is acceptable but the WARC format is preferred.	Documentation and tools to use files in the format freely available.

EMAIL

These formats are used to capture sent or received emails. Some recommended formats can either represent a single email, while others capture an entire mailbox including the contained emails.

NOTE: Emails usually have attachments. The formats contained in this section will only ensure that the basic text in the email body and the email headers are accessible, and that the attachments can be extracted as they were originally sent. They will not ensure that the information in an attachment is accessible. Attachments can be of any format. In order to remain accessible, it is important that each attachment be an appropriate long term preservation format.

FORMAT	NOTES	RATIONALE
EML	MIME encoded emails. Each instance of this format represents a single email.	This is a textual representation that is identical to the standard representation of an email while it is being transferred between computers. It is consequently extraordinarily widely used and well documented.
MBX, MBOX	UNIX style mailbox format. Each instance of this format represents a collection of emails, the mailbox (or part of a mailbox) of a user.	This is a textual representation that is an extension to the standard representation of an email while it is being transferred between computers. It is widely used.
MSG	The Microsoft Outlook Item MSG file format is a syntax for storing a single Message object, such as an email, an appointment, a contact, a task, and so on, in a file. A MSG file may also include email attachments.	A proprietary file format developed by Microsoft but no patents are claimed. Full documentation is available. This format is widely used.
PST	Microsoft mailbox format (generic). Each instance of this format represents a collection of emails, the mailbox (or part of a mailbox) of a user.	This format is widely used.

IMAGE (*RASTER*)

Raster images are still images typically produced by cameras or document scanners, composed of a grid of dots.

High quality images are extremely large. For this reason, most image production processes discard information to produce an 'acceptable' image at a reasonable size. Information may be discarded by using compression (i.e. lossy compression), reducing the resolution (downsampling) or reducing the colour width (going from 48 bit colour to 24 bit, or from greyscale to bitonal).

The formats in this section should be used where the file contains a single image. For multi-page scanned images, refer to *Image - Scanned Documents*.

FORMAT	NOTES	RATIONALE
TIFF	TIFF 6.0 is a preferred format for long term preservation use. The image in a TIFF file can be either uncompressed or compressed using a variety of compression formats. LZW compression is a widely used lossless technique. Note that uncompressed images can be extremely large.	A widely supported format by many applications. Fully documented specification. Proprietary but patents not enforced for wrapper format. Patents for the common LZW compression have long expired.
JPEG, JPG	JPEG images are usually expressed as a JFIF file, but they can be a raw JPEG image. Either approach is acceptable.	JPEG is an international standard, and it is extremely widely used – it is the default capture format for digital cameras (including phones).

		There would be billions of JPEG images. It is unlikely that JPEG would be unreadable at any time in the foreseeable future.
JPEG2000, JP2	JPEG2000 was intended to replace JPEG, but JPEG is simply too well entrenched in the market. JPEG2000 achieves higher rates of compression for similar quality, and has a lossless compression option.	JPEG2000 is an international standard, though is not very widely adopted.
PNG	Portable Network Graphics. This format was also intended to replace JPEG, with similar benefits to JPEG2000. It has been unable to make significant inroads into the market, though it is becoming more widely used in the web.	PNG is an international standard, though is not very widely adopted.
DNG	The Digital Negative (DNG) Camera Raw Format can be used for storing camera raw files, which are data captured from the camera sensor and requiring further processing to generate usable images. The advantage of raw files is the increased artistic control over the resulting image, as opposed to JPEG and TIFF files. Storing the DNG file with an embedded JPEG preview image may help with digital asset management. DNG is an extension of TIFF 6.0 and is compatible with TIFF/EP (Digital Photography).	A fully documented non-proprietary format, created as a standard by Adobe Systems. The DNG format can be used by a wide variety of hardware and software applications which generate, process, manage or archive camera raw files.

IMAGE (*VECTOR*)

Vector images are still images composed of lines and uniformly coloured areas. They are easily identifiable as the quality of the image does not degrade as the image is zoomed in. Typically vector images are produced by drawing programs such as Adobe Illustrator.

It should be noted that the most widely used Image (Vector) formats are the proprietary formats AI (Adobe Illustrator) and CDR (Corel Draw).

CAD files and Geospatial files are also typically vector images, but they have additional capabilities.
For more information on these formats refer to *CAD* and *Geospatial*.

FORMAT	NOTES	RATIONALE
SVG	Scalable Vector Graphics (SVG) is a language for describing two-dimensional graphics in XML. SVG allows for three types of graphic objects: vector graphic shapes (e.g. paths consisting of straight lines and curves), images and text.	This is an open standard from the W3C using XML. (SVG) is supported by all modern browsers for desktops and mobiles. Some features, such as SMIL animation and SVG Fonts are not as widely supported. There are many SVG authoring tools, and export to SVG is supported by all major vector graphics authoring tools.
ODG	Open Document Graphics format. SVG is preferred due to its greater support and more powerful features.	This format is an international standard based on XML.

IMAGE (*SCANNED DOCUMENTS*)

The formats in this section should be used when a multipage document (e.g. a book or report) has been scanned.

For simple single images, please use the formats in *Image - Raster*. For geo-referenced images, refer to *Geospatial*.

NOTE: While JPG and TIFF are not typically recommended scanning formats for documents (neither being widely used for holding multipage documents), some large document and image repositories do use multi-page TIFF.

FORMAT	NOTES	RATIONALE
PDF	Portable Document Format.	Fully documented. Most PDF formats were developed by Adobe Systems Inc. which makes the specifications openly available at no charge. Several members of the family have been adopted as ISO international standards, e.g. PDF/X, PDF/A, and PDF version 1.7. Extremely widely adopted for disseminating page-oriented documents.

IMAGE (*GRAPHIC METAFILES*)

Graphic metafiles are file formats that can contain a mixture of different types of images (raster **and** vector images).

FORMAT	NOTES	RATIONALE
CGM	Computer Graphics Metafile.	An open, platform-independent format for the exchange of raster and vector data for technical applications.

AUDIO

These formats represent sound only, such as recorded music, radio broadcasts, telephone conversations, and podcasts.

High quality audio files are extremely large. For this reason, most audio formats discard information, either by using compression, reducing the sampling rate, or the sample quality. A great deal of skill has been used to produce formats (and compression techniques) that minimise the size of audio files while retaining sufficient audio quality for the intended purpose. In most cases, the creator of a recording can select the quality of the recording. In selecting an audio format, it is consequently important to consider the type of audio material being stored to ensure that the quality is sufficient. Telephone conversations, for example, can be stored with far lower quality than the master recording of a performance of a piece of music.

FORMAT	NOTES	RATIONALE
WAV, BWAV, BWF	WAV encoding is normally used to encode a high quality representation of audio information. WAV and its Broadcast WAVE variants are container formats, which contain an audio bitstream that is encoded using LPCM (as used on audio CDs) or an MPEG audio format. For format planning purposes, it is useful to know which encoding is to be used. The WAV format and its variants, and the audio encodings mentioned, are widely used in the sound industry for containing high quality audio information.	Proprietary format, with full documentation freely available. Widely adopted.

MP3	MPEG-1 or -2 audio layer III. MP3 is the standard audio format for consumer consumption of audio. It has been carefully constructed to give an acceptable quality of audio with a small file size (high compression). They are widely used to distribute audio over the internet. It is expected that MP3 audio files would give acceptable quality for most business purposes. Where high quality recordings are necessary, a WAV file is recommended.	MP3 audio files are an international standard. Widely adopted.
MP4, M4A	MPEG-4 Audio (AAC encoding). This audio format is intended to replace MP3, giving better quality for the same level of compression. Compared with MP3, this format is not as widely used, but AAC has seen considerable adoption by industry.	These formats are part of the international standard for representing video information.
FLAC	Open source bit stream encoding format designed for lossless compression of LPCM audio data, with many of its default parameters tuned to CD-quality music data. Generally used for final-state, end-user delivery. FLAC files are much smaller than WAV (about a third the size of the uncompressed audio).	Open format well supported by free software.

VIDEO (*BARE*)

These formats are used for moving images (which may or may not include sound). Video formats are divided into two classes: bare formats which contain only the video and sound information; and container formats that allow multiple different types of object to be included in one file. This section is about bare video formats.

High quality video streams are extraordinarily large. For this reason, actual video files almost invariably have the quality reduced to allow the video stream to be a manageable size. Normally, the video is compressed using a lossy compression algorithm, the image size (resolution) is strictly limited, and the frame rate is adjusted to the minimum acceptable. It is for this reason that decisions about the quality of the video should be based on business needs and industry standards, rather than archival decisions. In general however, avoid complete dependence on proprietary formats/compression techniques that are natively produced by capture devices; standard formats and compression techniques are preferred.

FORMAT	NOTES	RATIONALE
MJP, MJ2, M2V	Motion JPEG 2000. The extension MJP is preferred.	Open international standard.
DPX	Digital Moving Picture Exchange (SMTPE). Used for high quality representation intended for theatrical distribution. The standard does not control how individual images are represented.	Reasonably widely adopted and fully documented open standard.

VIDEO (*CONTAINER*)

These formats are used for moving images (which may or may not include sound). Video formats are divided into two classes: bare formats which contain only the video and sound information; and container formats that allow multiple different types of object to be included in one file. These could include multiple audio files (for different languages), closed captions, etc. This section is about container video formats.

It is important to note that with container formats, the actual video stream may be encoded in many different ways. This needs to be considered.

High quality video streams are extraordinarily large. For this reason, actual video files almost invariably have the quality reduced to allow the video stream to be a manageable size. Normally, the video is compressed using a lossy compression algorithm, the image size (resolution) is strictly limited, and the frame rate is adjusted to the minimum acceptable. It is for this reason that decisions about the quality of the video should be based on business needs and industry standards, rather than archival decisions. In general however, agencies should avoid complete dependence on proprietary formats/compression techniques that are natively produced by capture devices; standard formats and compression techniques are preferred.

FORMAT	NOTES	RATIONALE
AVI	Audio Video Interleave. AVI containers can use a variety of video formats (WAVE, MP3 Audio, DivX)	Widely adopted. Fully documented.
OGG, OGV	Ogg container with video.	Fully documented. Developed as an open source and patent-free project.
DCP	Digital Cinema Package. Format is widely used in final-state for use in a cinema distribution chain; may also serve as a middle-state format for archiving. Uses DCDM to actually encode the video stream. This, in turn, uses MXF_UNC to actually encode the video stream	Fully disclosed proprietary format. Documentation available at Digital Cinema Initiatives Web site. No licensing or patents identified.
MPG, MPEG	MPEG2 video (i.e. as used on DVDs). Lossy compression. Many software tools exist for encoding and decoding.	Open standard. Widely adopted for filmmaking, DVD disks, and other applications. This format must be used for digital terrestrial TV broadcasting.
M4V, M4A F4V, F4A	MPEG4 video. Huge range of options. Generally a final-state (end-user delivery) format. Widely used. MPEG-4_AVC based on H.264 is a subtype that appears to be more widely adopted.	Open standard. Fully documented as ISO standard.

DATA

Data files contain data such as database tables, scientific observations, and metadata.

FORMAT	NOTES	RATIONALE
SIARD	Software Independent Archiving of Relational Databases. Used for capturing information from relational databases. SIARD is an XML representation of the data in a database and its schema.	As a textual representation, it is expected that the information will be accessible, even if the SIARD software becomes obsolete. SIARD is widely used by Archives in Europe to archive databases.
CSV/TSV	Comma separated variables, tab separated variables. CSV/TSV are very widely used textual representation of simple data structures. It is not possible to represent data coding in a CSV/TSV file. This information needs to be documented separately.	The mechanism has been used for a very long time – probably over 50 years.
XML	The XML standard describes how data is encoded into text. It does not describe the meaning of the data. This information needs to be documented separately.	XML is an open international standard mechanism for marking up data in textual format. Widely adopted for interchange of documents and data over the Web. Many generic tools exist, including free and open source software. Major software vendors have all incorporated support for XML in some form.
JSON	JavaScript Object Notation (JSON) is a lightweight, text-based, language-independent data interchange format. It describes how data is encoded into text, but does not describe the meaning of the data. This information needs to be documented separately.	Openly documented. As a simple textual encoding of data, JSON is expected to be easily processible indefinitely. JSON is so simple that support for reading or writing it is integrated into almost every system or programming language used for applications on the Web.

ELECTRONIC PUBLICATIONS

eBooks are used to publish books in a form that they can easily be read on a digital device such as a tablet.

FORMAT	NOTES	RATIONALE
PDF	Portable Document Format. Publications in this format will have restricted functionality – it is difficult to extract information from them and to modify them.	Fully documented. Most PDF formats were developed by Adobe Systems Inc. which makes the specifications openly available at no charge. Several members of the family have been adopted as ISO international standards, e.g. PDF/X, PDF/A, and PDF version 1.7. Extremely widely adopted for disseminating page-oriented documents.
EPUB	EPUB is a format for electronic publications with reflowable text in marked up document structure with associated images for illustrations, all in a container format. Reflowable text allows the text display to dynamically resize to fit screen size.	Fully documented. Widely used.

ENCAPSULATION

Encapsulation formats group a number of files together into one 'archive' file. Probably the best known example of an encapsulation format is the 'ZIP' file format. Another name for this type of file is an 'archive file', but we prefer not to use this term as these formats are not designed for archival use.

Encapsulation formats are particularly easy to migrate, as they provide a simple representation of a file system. It is consequently easy to unpack the encapsulation and re-encapsulate in another format.

FORMAT	NOTES	RATIONALE
ZIP	GZIP and ZIP files are completely different. A program that reads one format will not necessarily be able to read the other.	ZIP is very widely used. Although not a standard, the format is openly published, and ZIP is used by a number of International standards.
GZIP	GZIP is primarily intended as a compression program. Although multiple files can be included in a GZIP encapsulation, this is not its most common use. GZIP and ZIP files are completely different. A program that reads one format will not necessarily be able to read the other.	GZIP is widely used, particularly in UNIX/LINUX environments.
TAR	The age of TAR can be judged by its meaning: 'Tape Archive'; it was originally intended to encapsulate files for writing onto magnetic tape.	TAR is widely used in UNIX/LINUX environments. It is defined in an open standard (POSIX).

COMPUTER AIDED DESIGN (CAD)

At their most basic level, CAD files represent plans of objects (technical drawings). However, modern CAD files can model complex 3D shapes and can associate properties and metadata with objects in the plans. CAD files are consequently far more complex than simple representations of paper plans.

Because of the complexity of the data, and the dominance of a few commercial products, users generally have a choice between widely used proprietary formats, and standard formats that are not widely used, or which lack features of the proprietary formats.

FORMAT	NOTES	RATIONALE
DXF	(Autodesk) Drawing Exchange Format. This is a proprietary format. DXF files come in two versions: ASCII and binary. The ASCII version of this format should be used, as the binary is not widely used. Should only be used for 2D models, not 3D models, as 3D models are not fully documented in the public releases of the specification.	This format is proprietary, but is openly published by Autodesk and made available under a CC attribution- non commercial – sharealike 3.0 license.
DWG	(Autodesk) Drawing. Extremely widely used as an exchange format, particularly for 3D models. This format is proprietary, and has not been published by Autodesk. The format has, however, been reverse engineered by a consortium and this specification is available.	While proprietary and not formally published, the format is extremely widely supported by CAD products. It is also extremely widely used. A published, reverse engineered, specification does exist.

STP, STEP, P21	Standard for the Exchange of Product Model Data (ISO 10303-28:2007)	An international standard. Does not appear to have the product or usage support that DWG/DXF have.
PDF/E	PDF/Engineering (ISO 24517-1:2008). This format is relatively new, with little evidence available as to whether it can represent all the information required in a CAD file – particularly where CAD information is active and continues to be used.	An international standard.

GEOSPATIAL

Geospatial formats capture information related to the Earth. The classic piece of geospatial data is a map, however, geospatial data is best thought of as a set of geographical labelled features (points, lines and polygons) with associated properties.

Like CAD data, there are a number of widely used proprietary geospatial formats and a set of standardised, but less powerful, geospatial formats.

FORMAT	NOTES	RATIONALE
ESRI Shapefile	An open representation of geospatial information. It is important to note that, unlike other file formats discussed in this document, ESRI geospatial information is distributed amongst a number of different files. Related files have the same name, but different file extensions (e.g. SHP, SHX, DBF, CPG, PRJ, SBN, and SBX). The ESRI format dates from the 1990s and is a defacto standard for the exchange of geospatial information. The age of the specification means that some features of modern geospatial systems are not supported.	Fully documented. Widely used for the exchange of geospatial information. Open libraries are freely available for reading and writing to these geospatial formats.
GeoPackage	GeoPackage is an open, standards-based, platform-independent, portable, self-describing, compact format for transferring geospatial information. Has particular application to devices where internet connectivity and bandwidth are limited.	Fully documented, with support for many geospatial data types. GeoPackage has strong open standards support, through the Open Geospatial Consortium.
GeoJSON	GeoJSON is an open standard data interchange format designed for representing simple geographical features, along with their non-spatial attributes. It is based on JSON, the JavaScript Object Notation. TopoJSON is an extension of GeoJSON that encodes topology. It is used by many open source spatial systems, and is being rapidly adopted for Web applications involving mapping, such as those processing high volumes of Internet of Things 'IoT' data.	Fully documented. Open standard, maintained by the Open Geospatial Consortium. Rapid adoption.
DEM	A digital elevation model (DEM) represents terrain elevations for ground positions at regularly spaced horizontal intervals (i.e. 'bare earth' with no vegetation or man-made features, referenced to a	Open standard, maintained by the US Geographical Survey (USGS) and is used throughout the world. It was superseded by the SDTS format (no longer preferred) but the format remains popular due to

	common vertical datum). Single file ASCII text format. DEM is often used as a generic term for digital terrain models (DTM) and digital surface models (DSM), which include features on the earth's surface.	large numbers of legacy files, self-containment, relatively simple field structure and broad, mature software support.
GML	Geography Markup Language. This is a full featured representation of geospatial information.	Openly documented format by the Open Geospatial Consortium, based on XML. The standard is relatively new, and adoption is still occurring.
KML	Keyhole Markup Language. XML based representation of geographic data. KMZ is the Zip version. Primarily used for the publication of data ready for visualisation, particularly for non-specialists. GML and KML are complementary. Whereas GML is a way of modelling or encoding geographic content, KML is a way of presenting that content visually.	Openly documented format by the Open Geospatial Consortium. Textual format based on XML. Widely used (particularly in Google Earth and Google Maps).
GeoTIFF, BigTIFF	Geospatial Tagged Image File Format. This is a profile of TIFF (raster image). The profile adds the ability to store georeferenced and geocoding information to a raster image (e.g. map, aerial photograph, satellite image). BigTIFF is a variant that supports very large raster images.	Openly available specifications, widely supported in geospatial systems.
ECW	ECW (Enhanced Compression Wavelet) is a proprietary wavelet compression image format optimized for aerial and satellite imagery. The lossy compression format efficiently compresses very large images with fine alternating contrast while retaining their visual quality.	Widely adopted and stable format. While proprietary, the specification is available and the format is vendor-supported. Backwards-compatible.
JPEG2000	Open-source raster format (refer also to *Image - Raster*). Similar functionality to ECW, apart from compression types.	Openly available, widely supported in geospatial systems.
SVG	Scalable Vector Graphics. This is an open standard from the W3C using XML. It is an older format so now less relevant in the GIS industry. Vector caches are the current way to render vector out as image tiles. An example is the Open GIS Consortium's (OGC) GeoPackage	Open standard. Reasonably widely supported.
LAS	LAS is a binary file format for the exchange of LIDAR (Light Detection and Ranging) 3D point cloud data. Although point cloud data can be described in ASCII text files, such files are clumsy to use, because of the time required to import data and convert the numbers to binary formats for analysis. In practice, LIDAR point cloud datasets are usually shared in LAS files or losslessly compressed derivatives of LAS (e.g. LAZ).	Most widely supported LIDAR format. Maintained as a public specification by the American Society for Photogrammetry and Remote Sensing (ASPRS).

	Used to generate other geospatial products, such as digital elevation models, canopy models, building models, and contours.	
IMG	This format is used for processing remote sensing data, since it provides a framework for integrating sensor data and imagery from many sources.	A widely used proprietary, partially documented format for multi-layer geo-referenced raster images.