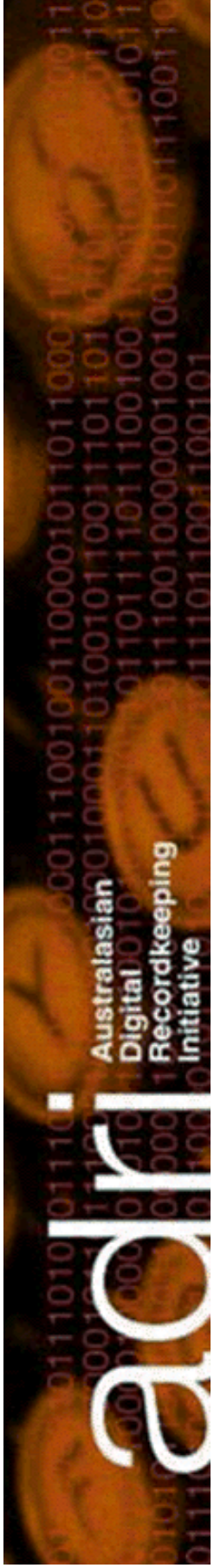

CAARA

Council of Australasian Archives and Records Authorities

Sustainable Digital File Formats

for creating and using records

Consultation Draft

November 2019

Table of Contents

1 Introduction

What this document is, who prepared it, and what we would like you to do

1.1 The problem

Will you be able to view a JPEG image in 10 years? In 100 years? In 1000 years?

Much of the digital information that is generated by the public service will have a long life. Some of the information will need to be accessible for decades. A small amount of the information will need to be kept permanently. This is just the reality of information; it is as true of information expressed in a paper form as information expressed in a digital form.

The particular challenge of digital information is that it is necessary to use software to access the information held in a digital object. It is necessary to use an image viewer to display a JPEG image, for example. Accessing digital information means having software to open the digital object for the entire lifespan of the information. For example, it may not be possible to find software that displays JPEG images in 100 years if the JPEG format has fallen out of use. If software to access a format is not available, information stored in this format will become inaccessible.

Records created by government agencies have a minimum mandated lifespan. This lifespan is set by Disposal Authorities issued by Government archival authorities. This minimum lifespan is not based on technology, instead it is based on an analysis on how long the government is likely to need the information. Typical minimum lifespans can range from 10 to 50 years. Information about people may be required to be kept for up to 100 years. Information about infrastructure usually must be kept for the life of the asset – major assets can have a life of 100 years or more. Some information is considered of permanent value and must be kept forever.

It is the agency's responsibility to maintain access to records for the minimum authorised retention period, irrespective of changing technology (unless the information has been transferred to an archive authority). If a format becomes obsolete or falls out of use, it is the agency's responsibility (and expense) to find a solution to the problem of ensuring access to the information.

This document is a draft of advice on how agencies can retain access to digital information for its expected lifespan.

1.2 The proposed solution

No-one knows what the future will hold; it is consequently impossible to eliminate the risk of a format falling out of use.

What is possible, however, is to reduce the risk by creating information in formats that are likely to remain accessible for long periods. This still leaves the challenge of selecting these 'low risk' formats.

This document presents list of formats that we believe have a low risk of becoming inaccessible over time.

We recommend that all information generated by an agency be created in a 'low risk' format, irrespective of how long it is expected that the information needs to be retained. This is for two reasons:

- It is a far simpler to implement in an agency, and to communicate to agency staff; i.e. just use these formats.
- The minimum lifespan of information can change over time.

We also recommend that agencies keep low risk file formats in mind for those records which they receive from other individuals or organisations. Where it is possible and practical, we encourage agencies to promote the use of low risk file formats for official business both internally and externally.

Please note that the goal of this document is different to that of many similar lists of formats produced by preservation institutions. Other lists are of formats accepted by the preservation institution for preservation. This list, however, is of recommended creation formats. If the staff of your agency are creating particular types of information (e.g. still

images), they should be encouraged to create the information in one of the recommended formats. We believe that the use of the recommended formats will minimise the risk of losing access to the information due to format obsolescence. As the formats are in widespread use, restricting staff to these formats is unlikely to be restrictive. Indeed, most staff would be unlikely to notice any restriction.

1.3 This document

This document is a draft of advice to agencies on preferred low risk formats for records being created or received. It contains two parts:

- A list of formats that we consider have low risk of access failure over a reasonable time, presented in summary and detailed versions. These are arranged by the type of information.
- A list of criteria we have used to select the formats. This list can be used to select a format if the type of information is not covered in the first list.

1.4 What would we like you to do?

We would like feedback on the two lists – the recommended file formats list and the criteria list.

1.5 How do you comment?

Please provide feedback to Andrew Waugh (Andrew.Waugh@prov.vic.gov.au) by 18 December 2019.

1.6 Who are we?

We are a sub-committee of the Australasian Digital Recordkeeping Initiative (ADRI). ADRI is composed of representatives of the State, Territory and National Archives of Australia and New Zealand. ADRI is a key work program of the Council of Australasian Archives and Records Authorities (CAARA).

2 Recommended formats (summary)

These formats have been identified as having low risk of access failure (e.g. through obsolescence) in the foreseeable future. They are a good choice for agency staff when creating and maintaining information.

When creating content, you should choose to save it in one of the following low risk formats. We have listed the formats by the type of information, and ordered the list so that common types of formats appear earlier in the list.

- **Word processing documents**: Microsoft Word (.docx, .doc), Open Office Document (.odt) or PDF (.pdf)
- **Spreadsheets**: Microsoft Excel (.xslx, .xls), Open Office Spreadsheet (.ods), PDF (.pdf), CSV (.csv), or TSV (.tsv)
- **Presentations**: Microsoft PowerPoint (.pptx, .ppt), Open Document Presentation (.odp), or PDF (.pdf)
- **Raster images**: TIFF (.tiff or .tif), JPEG (.jpg or .jpeg), JPEG2000 (.jp2), PNG (.png) or DNG (.dng)
- **Vector images**: SVG (.svg) or Open Document drawing (.odg)
- **Multipage scanned documents**: PDF (.pdf)
- **Audio**: WAV (.wav, .bwav, or .bwf), MP3 (.mp3), MP4 (.mp4 or .m4a), or FLAC (.flac)
- **Video (in a container)**: Ogg (.ogg or .ogv), DCP (.dcp), MPEG2 (.mpg, .mpeg), or MPEG4 (.m4v, .m4a, f4v, .or f4a)
- **Video (bare)**: Motion JPEG2000 (.mjp, .mj2 or .m2v), or DPX (.dpx)
- **Electronic publications**: PDF (.pdf) or EPUB (.epub)
- **Encapsulation**: ZIP (.zip), GZIP (.gzip), or TAR (.tar)
- **CAD**: DXF (.dxf), .DWG (.dwg), STEP (.stp, .step, or .p21), PDF/E (.pdf)
- **Geospatial**: ESRI Shapefiles, Geopackage, GEOjson (.json), DEM, GML (.gml), KML (.kml or .kmz), GEOtiff (.tiff), BigTiff (.tiff), ECW (.ecw), JPEG2000 (.jp2), SVG (.svg), LAS (.las) or IMG (.img)
- **Email**: Mime encoding (.eml), UNIX mailbox (.mbx or .mbox), or Microsoft Outlook (.msp or .pst)
- **Websites**: HTML (.htm or .html) or XML (.xml) together with supporting files (.css, .xsd, and .dtd)
- **Archived Websites**: Web archive (.warc) or Internet archive (.arc)
- **Image (graphic metafile)**: CGM (.cgm)
- **Data files**: CSV (.csv), TSV (.tsv), XML (.xml), JSON (.json or .jsn), or SIARD
- **Text files**: Plain text (.txt) files

3 Recommended formats (detail)

This section provides brief detail and our rationale for recommending each format.

3.1 Plain Text Files

Plain text files contain simple text without any markup or formatting.

Where the text files contain data (e.g. CSV files), see section 2.2 (Data). Where the text files contain website information, see Section 2.8 (Websites).

Recommended	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, standardized, 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.

3.2 Data

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

Recommended	Notes	Rationale
SIARD[1]	Software Independent Archiving of Relational Databases Use 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.

[1] https://www.loc.gov/preservation/digital/formats/fdd/fdd000426.shtml

Recommended	Notes	Rationale
CSV/TSV	Comma separated variables, tab separated variables It is not possible to represent data coding in a CSV/TSV file. This information needs to be documented separately.	CSV/TSV are very widely used textual representation of simple data structures. The mechanism has been used for a very long time – probably over 50 years.
XML[2]	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. It has been very widely adopted as the basis 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[3]	JavaScript Object Notation (JSON) is a lightweight, text-based, language-independent data interchange format. The openly documented JSON standard describes how data is encoded into text. It does not describe the meaning of the data. This information needs to be documented separately.	As a simple textual encoding of data, JSON is expected to be easily processible indefinitely. JSON is considered to be a lightweight alternative to XML. 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.

3.3 Office Documents (Word Processing)

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

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 photographs).

Recommended	Notes	Rationale
DOCX[4]	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, and is based on XML. Further, Microsoft Word is the clear market leader in office document preparation and very large numbers of documents in this format are created.
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) [5], and is consequently documented. Its main basis for inclusion, however, is simply the extraordinarily large number of documents created in this format. It is consequently unlikely that it would become inaccessible in the foreseeable future.

[2] https://www.loc.gov/preservation/digital/formats/fdd/fdd000264.shtml

[3] https://www.loc.gov/preservation/digital/formats/fdd/fdd000381.shtml

[4] https://www.loc.gov/preservation/digital/formats/fdd/fdd000397.shtml

[5] http://www.ecma-international.org/

Recommended	Notes	Rationale
ODT[6]	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	The Portable Document Format (PDF) is a file 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[7], but PDF 1.7[8] is acceptable. This format offers several forms of technical protection, including encryption, which could prevent custodians of digital content ensuring accessibility in future technological environments. These technical mechanisms should not be used.	Fully documented. Most members of the PDF family were developed by Adobe Systems Incorporated, which makes the specifications available openly and 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.

1

[6] https://www.loc.gov/preservation/digital/formats/fdd/fdd000428.shtml

[7] https://www.loc.gov/preservation/digital/formats/fdd/fdd000318.shtml

[8] https://www.loc.gov/preservation/digital/formats/fdd/fdd000277.shtml

3.4 Office Documents (Spreadsheet)

Use these formats for spreadsheets (e.g. Excel).

Recommended	Notes	Rationale
XLSX[9]	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, and is based on XML. Further, Microsoft Excel is the clear market leader in spreadsheet preparation and very large numbers of spreadsheets in this format are created.
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, and is consequently documented. Its main basis for inclusion, however, is simply the extraordinarily large number of spreadsheets created in this format. It is consequently unlikely that it would become inaccessible in the foreseeable future.
ODS[10]	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	See PDF in Section 2.3. 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.	
CSV[11]	See CSV/TSV in Section 2.3. 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.

[9] https://www.loc.gov/preservation/digital/formats/fdd/fdd000398.shtml

[10] https://www.loc.gov/preservation/digital/formats/fdd/fdd000439.shtml

[11] https://www.loc.gov/preservation/digital/formats/fdd/fdd000323.shtml

3.5 Office Documents (Presentation)

Use these formats for presentations (e.g. PowerPoints).

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).

Recommended	Notes	Rationale
PPTX[12]	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, and is based on XML. Further, Microsoft PowerPoint is the clear market leader in simple presentation preparation and very large numbers of presentations in this format are created.
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, and is consequently documented. Its main basis for inclusion, however, is simply the extraordinarily large number of presentations created in this format. It is consequently unlikely that it would become inaccessible in the foreseeable future.
ODP[13]	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	See PDF in Section 2.3. 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.	

3.6 Project Planning

There are no recommended formats to represent project planning data.

This is because we know of no project planning formats sufficiently widely used, or sufficiently well documented, to make the format likely to survive for the required periods.

[12] https://www.loc.gov/preservation/digital/formats/fdd/fdd000399.shtml

[13] https://www.loc.gov/preservation/digital/formats/fdd/fdd000435.shtml

3.7 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 that they remain accessible, it is important that each attachment be an appropriate long term preservation format.

Recommended	Notes	Rationale
EML[14]	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[15]	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[16]	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[17]	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.

3.8 Website

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 chapter.

[14] https://www.loc.gov/preservation/digital/formats/fdd/fdd000388.shtml

[15] https://www.loc.gov/preservation/digital/formats/fdd/fdd000383.shtml

[16] https://www.loc.gov/preservation/digital/formats/fdd/fdd000379.shtml

[17] https://www.loc.gov/preservation/digital/formats/fdd/fdd000378.shtml

Finally, note that 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[18]. 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.

Recommended	Notes	Rationale
HTM, HTML[19]	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[20]	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[21]	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[22]	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[23]	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.

[18] Brunelle, J. F., Kelly, M., Weigle, M. C., & Nelson, M. L. (2016). The impact of JavaScript on archivability. International Journal on Digital Libraries, 17(2), 95-117. doi:http://dx.doi.org/10.1007/s00799-015-0140-8

[19] https://www.loc.gov/preservation/digital/formats/fdd/fdd000475.shtml

[20] https://www.loc.gov/preservation/digital/formats/fdd/fdd000075.shtml

[21] https://www.loc.gov/preservation/digital/formats/fdd/fdd000482.shtml

[22] https://www.loc.gov/preservation/digital/formats/fdd/fdd000077.shtml

[23] https://www.loc.gov/preservation/digital/formats/fdd/fdd000076.shtml

3.9 Website (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.

Recommended	Notes	Rationale
WARC[24]	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[25]	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.

[24] https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml

[25] https://www.loc.gov/preservation/digital/formats/fdd/fdd000235.shtml

3.10 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.

Recommended	Notes	Rationale
WAV[26], BWAV, BWF	WAV encoding is normally used when it is desired 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.	Proprietary format, with full documentation freely available. The WAV format and its variants, and the audio encodings mentioned are extremely widely used in the sound industry for containing high quality audio information.
MP3[27]	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). 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. Beyond that, however, they extraordinarily widely used to distribute audio over the internet. It is extremely unlikely that they will become inaccessible for the foreseeable future.
MP4, M4A[28]	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[29].	These formats are part of the international standard for representing video information. As such they are unlikely to become inaccessible, even though they are not as widely used for audio.
FLAC (Free Lossless Audio Codec)[30]	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). FLAC is an open format well supported by free software.

[26] https://www.loc.gov/preservation/digital/formats/fdd/fdd000001.shtml

[27] https://www.loc.gov/preservation/digital/formats/fdd/fdd000105.shtml

[28] https://www.loc.gov/preservation/digital/formats/fdd/fdd000155.shtml

[29] https://mpeg.chiariglione.org/standards/mpeg-4/audio

[30] https://www.loc.gov/preservation/digital/formats/fdd/fdd000198.shtml

3.11 Image (Raster)

Raster images are still images typically produced by cameras or document scanners. Raster images are 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 (down sampling) 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 section 2.13 (Image - Scanned Documents).

Recommended	Notes	Rationale
TIFF[31]	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[32]	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[33]	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[34]	Portable Network Graphics. This format was also intended to replace JPEG, with similar benefits to JPEG2000. Like 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.

[31] https://www.loc.gov/preservation/digital/formats/fdd/fdd000022.shtml

[32] https://www.loc.gov/preservation/digital/formats/fdd/fdd000017.shtml

[33] https://www.loc.gov/preservation/digital/formats/fdd/fdd000138.shtml

[34] https://www.loc.gov/preservation/digital/formats/fdd/fdd000153.shtml

Recommended	Notes	Rationale
DNG[35]	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.

3.12 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 see the relevant sections (2.19 CAD; 2.20 Geospatial).

Recommended	Notes	Rationale
SVG[36]	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[37].
ODG[38]	Open Document graphics format. SVG is preferred due to its greater support and more powerful features.	This format is an international standard, and is based on XML.

[35] https://www.loc.gov/preservation/digital/formats/fdd/fdd000188.shtml

[36] https://www.loc.gov/preservation/digital/formats/fdd/fdd000020.shtml

[37] https://www.w3.org/Graphics/SVG/

[38] https://www.loc.gov/preservation/digital/formats/fdd/fdd000436.shtml

3.13 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 section 2.11 (Image - Raster). For geo-referenced images, please use section 2.20 (Geospatial).

Note that 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.

Recommended	Notes	Rationale
PDF	See PDF in Section 2.3.	

3.14 Image (Graphic Metafiles)

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

Recommended	Notes	Rationale
CGM[39]	Computer Graphics Metafile	An open, platform-independent format for the exchange of raster and vector data for technical applications.

3.15 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, agencies should avoid complete dependence on proprietary formats/compression techniques that are natively produced by capture devices; standard formats and compression techniques are preferred.

Recommended	Notes	Rationale
MJP, MJ2, M2V[40]	Motion JPEG 2000. The extension MJP is preferred	Open international standard.
DPX[41]	Digital Moving Picture Exchange (SMTPE). Used for high quality representation intended for	Reasonably widely adopted and fully documented

[39] https://www.nationalarchives.gov.uk/documents/graphic-file-formats.pdf

[40] https://www.loc.gov/preservation/digital/formats/fdd/fdd000127.shtml

[41] https://www.loc.gov/preservation/digital/formats/fdd/fdd000178.shtml

Recommended	Notes	Rationale
	theatrical distribution. The standard does not control how individual images are represented.	open standard.

3.16 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.

Recommended	Notes	Rationale
AVI[42]	Audio Video Interleave. AVI containers can use a variety of video formats (WAVE, MP3 Audio, DivX)	Widely adopted and fully documented proprietary format.
OGG, OGV[43]	Ogg container with video	Fully documented. Developed as an open source and patent-free project.
DCP[44]	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. Most significant is the format's required use in 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.

[42] https://www.loc.gov/preservation/digital/formats/fdd/fdd000059.shtml

[43] https://www.loc.gov/preservation/digital/formats/fdd/fdd000026.shtml

[44] https://www.loc.gov/preservation/digital/formats/fdd/fdd000200.shtml

3.17 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.

Recommended	Notes	Rationale
PDF	See PDF in Section 2.3. Presentations in this format will have restricted functionality – it is difficult to extract information from them and to modify them.	Open standard. Fully documented. Widely used.
EPUB[45]	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.	Popular recent standard. Fully documented. Widely used.

3.18 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.

Recommended	Notes	Rationale
ZIP[46]	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[47]	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).

[45] https://www.loc.gov/preservation/digital/formats/fdd/fdd000278.shtml

[46] https://www.loc.gov/preservation/digital/formats/fdd/fdd000354.shtml

[47] https://en.wikipedia.org/wiki/Gzip

3.19 CAD

At their basic, Computer Aided Design (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 their 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.

Recommended	Notes	Rationale
DXF[48]	(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.	The format is proprietary, but is openly published by Autodesk and made available under a CC attribution- non commercial – sharealike 3.0 license.
DWG[49]	(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[50]	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 and little experience is available as to the ability to represent all the information required in a CAD file – particularly where the CAD information is active and continues to be used.	An international standard.

3.20 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.

[48] https://www.loc.gov/preservation/digital/formats/fdd/fdd000446.shtml

[49] https://www.loc.gov/preservation/digital/formats/fdd/fdd000445.shtml

[50] https://www.loc.gov/preservation/digital/formats/fdd/fdd000448.shtml

Recommended	Notes	Rationale
ESRI Shapefile. [51]	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, and the specification is openly available. Widely used for the exchange of geospatial information. Open libraries are freely available for reading and writing to these geospatial formats.
GeoPackage[52]	GeoPackage is an open, standards-based, platform-independent, portable, self-describing, compact format for transferring geospatial information.	Fully documented, with support for many geospatial data types. GeoPackage has strong open standards support, through OGC. Has particular application to devices where internet connectivity and bandwidth are limited.
GeoJSON[53]	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.	This fully documented format is an open standard, maintained by the Open Geospatial Consortium. 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.
DEM[54] [55]	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 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.	It is an 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 large numbers of legacy files, self-containment, relatively simple field structure and broad, mature software support.
GML[56]	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[57]	Keyhole Markup Language. XML based representation of geographic data. KMZ is the Zip	Openly documented format by the Open Geospatial Consortium. Textual format based on

[51] https://www.loc.gov/preservation/digital/formats/fdd/fdd000280.shtml

[52] https://www.goepackage.org/

[53] https://www.loc.gov/preservation/digital/formats/fdd/fdd000382.shtml

[54] https://www.loc.gov/preservation/digital/formats/fdd/fdd000285.shtml

[55] https://gisgeography.com/dem-dsm-dtm-differences/

[56] https://www.loc.gov/preservation/digital/formats/fdd/fdd000296.shtml

[57] https://www.loc.gov/preservation/digital/formats/fdd/fdd000340.shtml

Recommended	Notes	Rationale
	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.	XML. Widely used (particularly in Google Earth and Google Maps).
GeoTIFF, BigTIFF[58]	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. Both are openly available specifications.	Openly available, widely supported in geospatial systems.
ECW[59]	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.	ECW is a ubiquitous and stable format. While proprietary, the specification is available and the format is vendor-supported. Backwards-compatible.
JPEG2000	Open-source raster format (refer also section 2.11 Image - Raster). Similar functionality to ECW, apart from compression types.	Openly available, widely supported in geospatial systems.
SVG[60]	Scalable Vector Graphics	This is an open standard from the W3C using XML. Reasonably widely supported, although an older format now and less relevant in the GIS industry these days. Vector caches are the current way to render vector out as image tiles. An example is the Open GIS Consortium's (OGC) GeoPackage: https://www.geopackage.org/

[58] https://www.loc.gov/preservation/digital/formats/fdd/fdd000279.shtml

[59] https://www.gdal.org/frmt_ecw.html

[60] https://www.loc.gov/preservation/digital/formats/fdd/fdd000020.shtml

Recommended	Notes	Rationale
LAS[61]	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 very clumsy to use, because of the time required to import data and convert the numbers to binary for analysis. In practice, LIDAR point cloud datasets are usually shared in LAS files or losslessly compressed derivatives of LAS (e.g. LAZ). Used to generate other geospatial products, such as digital elevation models, canopy models, building models, and contours.	Most widely supported LIDAR format. Maintained as a public specification by the American Society for Photogrammetry and Remote Sensing (ASPRS).
IMG[62]	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.

1

2

[61] https://www.loc.gov/preservation/digital/formats/fdd/fdd000418.shtml

[62] https://www.loc.gov/preservation/digital/formats/fdd/fdd000420.shtml

4 Long term preservation format criteria

1 What characteristics distinguish low risk preservation formats?

2 These criteria are aspirational and apply to born-digital records and formats for digitisation. Not all formats will meet
3 all of these criteria all of the time:

4 1. **Ubiquitous**: The format is widely used and supported around the world.
5 2. **Unrestricted**: The format is free from patents and legal encumbrance (including intellectual property rights)
6 and preferably embodies open-source principles.
7 3. **Well Documented**: The format is identifiable and is well documented (the specification is publicly available
8 and sufficiently detailed that groups or individuals *with sufficient skills and incentive can build software to*
9 *render it accurately*.
10 4. **Stable**: The file format is stable (rare releases of newer versions) and is backwards and forwards compatible,
11 or has a clear migration path.
12 5. **Platform Independent:** The format should be supported by a wide range of software or is platform-
13 independent.
14 6. **Uncompressed**: Ideally the format should be uncompressed. If compression is used, lossless compression is
15 preferred.
16 7. **Supported**: Technical support is readily available from vendors, community or third parties.
17 8. **Metadata Friendly**: File formats with metadata support are preferred.
18

19 A discussion about the meaning of these criteria can be found in the next section.

5 Commentary

Explanatory information about the characteristics

5.1 General

5.1.1 Must a format satisfy all criteria for it to be considered low risk?

No, a format could be considered low risk even though it does not satisfy all the criteria. In general, the more criteria a format satisfies, the lower the risk.

5.2 Details of specific criteria

5.2.1 Ubiquitous

A format is considered to be ubiquitous if it dominates its information category. For example, the vast majority of word processing documents are in the Microsoft Word (.doc, .docx) format. Similarly, almost all digital photographs captured are in the JPEG format.

Ubiquitous formats are considered to be low risk for longevity because the sheer amount of information in that format means that there is a significant economic incentive for applications to support that format. For example, image processing software that does not support the JPEG format is unlikely to sell many copies.

5.2.2 Unrestricted

Legal restrictions on implementing software that reads or writes a format is likely to reduce the range of software that handles the format. This, in turn, increases the risk that the format will become unsupported.

Some formats may have different restrictions on the production of format readers vs writers (e.g. PDF). Restrictions on software that writes the format are less serious than those for readers.

Open source software is preferred but not required. By open source, we mean that the source code for software that handles the format can be downloaded and is sufficiently free from restrictions that it can be used to maintain in house (or could be used to reverse engineer a specification that can be used to write a new piece of software). Open source is not required because economic demand for very widely used software often results in reverse engineered competing products that will read proprietary formats (a good example is the Autodesk DWG CAD format). The fact that a piece of software is open source does not mean that software is being actively maintained, or is of sufficient quality to provide accurate access to the information in a format.

5.2.3 Well Documented

A well-documented format allows future users to accurately implement software that reads the format. A format that has no public documentation can be extremely difficult to re-implement.

The best documentation has been formally published by independent bodies (e.g. ISO, CCITT, W3C, or the IETF). However, format descriptions produced by vendors themselves can still be valuable (e.g. the ZIP specification).

It is important that the documentation accurately documents the format. Some vendors extend the documented format and use un-documented extensions.

The best indication of the quality of a specification is that multiple independent implementations have been produced from the specification (this links to interoperability, and ubiquity). Some formats, e.g., QuickTime and MPEG-4, allow for a very wide range of implementations compared to say, MPEG-2[63].

Note that open source software code itself is an excellent form of documentation.

5.2.4 Stable

A stable or robust format means that the economic investment in writing software to deal with the format is long-lived. This encourages the production of software. It also means that the software itself has more chance of being robust and stable; that is bugs have been found and eliminated.

Where formats do change, but are backwards and forwards compatible, this means that software can read documents from newer versions of the software (possibly with some loss of functionality) and older versions. This minimises the risk that documents can become unreadable.

Formats which have a clear migration path also have lower risk, for example migrations to Office 365 from Microsoft Outlook PST files and Exchange Server are relatively straightforward whereas Office 365 migrations from Lotus Notes are possible but somewhat more complex.

5.2.5 Platform Independent

Formats that have been independently implemented by a range of vendors are lower risk than those that have only been implemented by one or two. Implementation by a range of vendors shows that 1) the format is documented, 2) the documentation can be used to accurately implement the format, and 3) that the documentation actually describes the format.

One concern is that a format may appear to be implemented by multiple vendors, when, in fact, the different products actually use a single common implementation.

5.2.6 Uncompressed

Formats that are compressed have an additional layer of complexity. Applications that need to use the data must first decompress it before it can be used. This decompression adds a risk to future use. The compression algorithm can be considered as an additional data format layered over the top of the actual data format. It can consequently be evaluated just as any other format: is it ubiquitous (e.g. JPEG); has it any restrictions (e.g. patents); and so on.

It should be noted that some types of information are routinely compressed to reduce their large size (e.g. audio and video files).

Some forms of compression are lossy; that is, they discard information to achieve a small file size and the decompressed object will have lesser quality compared with the original. Other compression algorithms are lossless, and the decompressed object will be identical to the original.

In general, lossless compression is preferred as this allows the original object to be retrieved. However, lossy compression is acceptable in the following circumstances:

- The original object is already compressed using a lossy compression. Examples of these would be most still images, sound files, and almost all video. All of these would normally be received from the capture device (e.g. camera) already compressed using lossy compression and usually in a proprietary format. Care must be taken to ensure that files taken directly from a capture device are in an interoperable format, or that they are transcoded to a suitable format as part of the capture workflow. In other words, agencies should avoid proprietary formats/compression techniques that are natively produced by capture devices; standard formats and compression techniques such as LZW, JPEG and JPEG2000 are preferred.

[63] https://www.loc.gov/preservation/digital/formats/content/video_quality.shtml

- The original object is too large to be used for business purposes and the original will not be required. Again, the examples would be still images, sound files, and video which may need to be reduced in size in order to be usable. Consideration should be given to retaining both the high-quality original and the lower-quality reduced size copy.

5.2.7 Supported

Technical support means that it is possible to get someone to assist with the format, particularly with regard to solving technical problems. Some formats are actively supported, particularly those that are commercial and public domain formats that are have an active user community.

5.2.8 Metadata Friendly

A digital object which incorporates all the metadata needed to render its data as usable information or understand its context are likely to be easier to sustain over the long term and less vulnerable to catastrophe than digital objects that are stored separately from related metadata. An example of a format with metadata embedded within the file is JPEG (where the image includes metadata that describes how the image was captured). This metadata can be a fixed set of attributes, or it can be arbitrary.