Specification

PROS 99/007 S4

Long term preservation formats

Version number: 3.0
Issue Date: 15 January 2016
Expiry Date: 15 January 2020

Public Record Office Victoria

VICTORIA State Government

This page is blank

Table of Contents

Copyright Statement

Disclaimer

The State of Victoria gives no warranty that the information in this version is correct or complete, error free or contains no omissions. The State of Victoria shall not be liable for any loss howsoever caused whether due to negligence or otherwise arising from the use of this Specification.

Acknowledgements

Public Record Office Victoria would like to acknowledge and thank the valuable contributions of people who took the time to comment during the development of this Specification.

Executive Summary

Public Record Office Victoria (PROV) has recently issued an updated version of the standard supporting the Victorian Electronic Records Strategy (VERS). The new standard includes a revised list of approved long term preservation formats (PROS 15/03S3). The revision:

- covers more types of digital object (e.g. web pages, email);
- reduces the cost of data migration by eliminating the need to migrate from common formats (e.g. Word) that are unlikely to have long term preservation issues;
- includes all the formats that were previously accepted by PROV.

PROV has updated this specification to make the revised list of approved long-term preservation formats available for use by agencies using the older VERS standard (PROS 99/007).

Where an agency has a type of document that is not suitable for conversion to one of these standard formats, the agency must negotiate with PROV to select an appropriate format.

This page is blank

1 Introduction

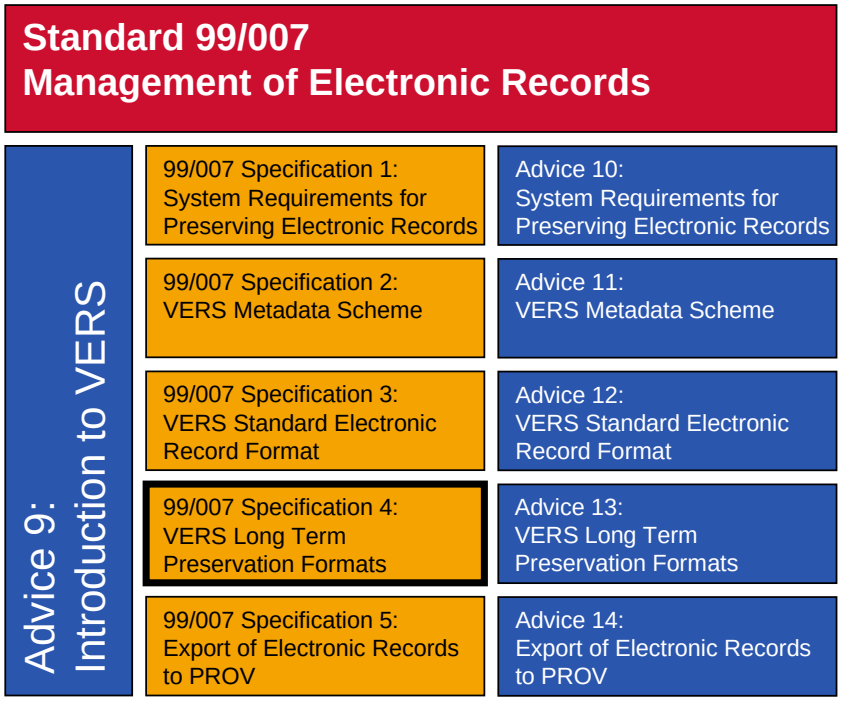
The purpose of this Specification is to specify long-term preservation formats that will be accepted by PROV.

A key digital preservation risk is that a file format will become obsolete, and files in that format will not be able to be opened by future computer systems. If this happens the information in that file is effectively lost. PROV manages this risk by requiring records to be in long term preservation formats that are unlikely to become obsolete for a very long period of time, and when they do become obsolete are expected to be easily migrated to newer formats.

Where an agency has a type of document that is not suitable for conversion to one of these standard formats, the agency must negotiate with PROV to select an appropriate format.

1.1 Relationship with other Specifications

This document is a Specification that supports the Victorian Electronic Records Strategy (VERS) Standard (PROS 99/007). The relationship between the VERS Standard, the Specifications that support this Standard, and the Introduction and Advices that explain VERS is shown below.



These documents have the following purposes:

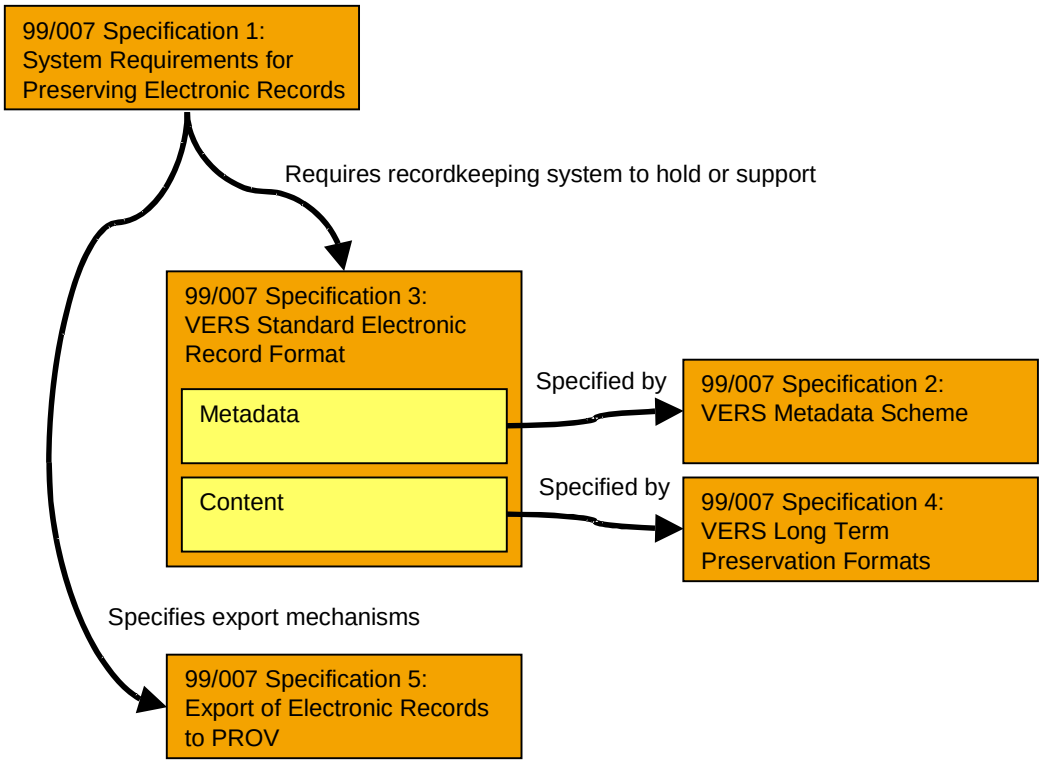
- *Management of Electronic Records*. This document is the Standard itself and is primarily concerned with conformance. The technical requirements of the Standard are contained in five Specifications.
- *Introduction to VERS.* This document provides background information on the goals and the VERS approach to preservation. Nothing in this document imposes any requirements on agencies.

- *Specifications.* These five documents provide the technical requirements that support the Standard. Agencies *must* conform to the mandatory requirements of the specifications, *must* conform to the conditional requirements of the specifications if the appropriate conditions are satisfied, and *may* conform to the optional requirements. Some optional requirements are strongly recommended and these are noted as such.

  The five Specifications are:

  - *Specification 1: System Requirements for Preserving Electronic Records.* This document specifies the overall functions that a recordkeeping system must perform to preserve electronic records for a substantial period.
  - *Specification 2: VERS Metadata Scheme.* This document specifies the metadata that a recordkeeping system must hold to conform to VERS.
  - *Specification 3: VERS Standard Electronic Record Format*. This document contains the technical definition of the VERS Encapsulated Object (VEO) format; the mandatory long-term format for records.
  - *Specification 4: VERS Long Term Preservation Formats.* This document lists the data formats that PROV accepts as suitable for representing documents for a significant period.
  - *Specification 5: Export of Electronic Records to PROV.* This document lists the approved media and mechanisms by which PROV will accept an export of electronic records.
- *Advices.* These six documents provide background information, explanatory material, and examples in support of the Standard and associated Specifications. None of the information in the Advices imposes any requirement on agencies.

Relationship between Specifications. A second view of the relationship between the five Specifications is shown in the following diagram:



Specification 1 (System Requirements for Preserving Electronic Records) details the overall requirements on a recordkeeping system for preserving electronic records over a significant period. Amongst other requirements, the recordkeeping system must be capable of exporting the records in a standardised format.

The overall features of this standardised format are defined in *Specification 3 (VERS Standard Electronic Record Format)*, but some details are defined in two other Specifications. *Specification 2 (VERS Metadata Scheme)* defines the

meaning and allowed values of the metadata that appears in a record. *Specification 4 (VERS Long Term Preservation Formats)* defines the formats in which the record content must be expressed.

Specification 5 (Export of Electronic Records to PROV) defines the mechanisms by which records are exported to PROV.

1.2 Changes from Version 1 of PROS 99/007

This specification is new in Version 2 of PROS 99/007; previously these requirements were contained in Specifications 1 and 3. In addition, the following changes have been made to the specifications:

- The PDF requirements have been tightened up.
- TIFF has been added as an approved format.

1.3 Changes between Version 2 and Version 2.1

Version 2.1 added the following file formats:

- PDF/A
- JPEG
- JPEG 2000
- MPEG-4

1.4 Changes between Version 2.1 and Version 3

This specification was revised to allow agencies using the original VERS Standard (PROS 99/007) to take advantage of the new long term preservation formats allowed under the new VERS Standard (PROS 15/03).

2 Use of Encryption/ Passwords/ Copy Protection

The encodings within a VERS Encapsulated Object must not be encrypted, protected by passwords, or subject to any form of software copy protection (e.g. digital rights management).

3 Standard Long Term Preservation Formats

Formats available for use

Agencies may use any of the formats specified in PROS 15/03S3 (Long term preservation formats) as a long term preservation format for PROS 99/007.

Encoding of formats

Binary data

Formats containing binary data must be encoded using Base64[1].

Textual data

Formats containing only textual data may be encoded using Base64 or represented as plain text.

If the data is represented as plain text, it must conform to the latest version of the Unicode standard[2] and be encoded as UTF-8. Characters that are the XML predefined entities '&', '<' and '>' must be encoded as '&', '<' and '>' respectively[3].

Suggested value of File Encoding (M128)

Appendix 1 contains suggested values to be used in the M128 metadata element (File Encoding).

It is not mandatory to use these values, but any value of M128 should contain the following information:

- The citation of the format specification (a URL may be included).
- The encoding (or representation) of the format used.

[1] IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'

[2] The Unicode Consortium. The Unicode Standard, Version 8.0.0, Mountain View, CA: The Unicode Consortium, 2015. ISBN 978-1-936213-10-8 http://www.unicode.org/versions/Unicode8.0.0/ visited 2 Sept 2015).

[3] Note that it is not illegal for XML character date to contain single and double quotation marks. See Extensible Markup Language (XML) 1.0 (Fifth Edition) Section 2.4, http://www.w3.org/TR/REC-xml/#charsets visited 2 Sept 2015.

4 References

Other Resources

You can obtain relevant publications, supplies of relevant forms, and answers to any enquiries you may have by first contacting your agency's records manager or Public Record Office Victoria:

For more information on digital preservation, please contact:

Government Services
Public Record Office Victoria
Ph: (03) 9348 5600
Fax: (03) 9348 5656
Email: standards@prov.vic.gov.au
Web: prov.vic.gov.au

Appendix 1: Recommended value for File Encoding (M128)

This appendix contains recommended text for inclusion in the File Encoding (M128) element for different formats.

Plain Text encoded as plain text

The content of the DocumentData element is a text file. The characters in the text file conform to UTF-8 encoded Unicode. Unicode is defined in 'The Unicode Standard', Version 8.0.0, The Unicode Consortium, or the equivalent ISO 10646:2014 and Amendment 1. The ampersand (and), single quote, double quote, greater than, and less than characters have been escaped using the standard XML escape conventions '&', '<', and '>' respectively.

Plain Text encoded as Base64

The content of the DocumentData element is a text file. The characters in the text file conform to UTF-8 encoded Unicode. Unicode is defined in 'The Unicode Standard', Version 8.0.0, The Unicode Consortium, or the equivalent ISO 10646:2014 and Amendment 1. The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

HTML (Hypertext Markup Language)

The content of the DocumentData element is an HTML file. The file conforms to 'HTML 4.01 Specification'. The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

CSS (Cascading Style Sheets)

The content of the DocumentData element is a CSS stylesheet. The file conforms to 'Cascading Style Sheets Level 2 Revision 1'. The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

XML (Extensible Markup Language)

The content of the DocumentData element is an XML document. The file conforms to 'Extensible Markup Language (XML) 1.0 (Fifth Edition)'. The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

PDF (Portable Document Format) or PDF-A (Portable Document Format - Archive)

The content of the DocumentData element is a PDF file. For details of PDF see 'ISO 32000-1:2008 – Document management – Portable document format – Part 1: PDF 1.7'. The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8, 'Base64 Content-Transfer-Encoding'.

DOC/DOCX (Microsoft Word Formats)

The content of the DocumentData element is a document represented in the format used by the Microsoft Word program. The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

XLS, XLSX (Microsoft Excel Format)

The content of the DocumentData element is a spreadsheet represented in the format used by the Microsoft Excel program. The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

PPT, PPTX (Microsoft Powerpoint Formats)

The content of the DocumentData element is a document represented in the format used by the Microsoft Powerpoint program. The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

TIFF (Tagged Image File Format)

The content of the DocumentData element is a TIFF image. The file conforms to 'TIFF Revision 6.0, Final – June 3, 1992'. The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

JPEG (encoded using the Appendix B format)

The content of the DocumentData element is a JPEG image. The file conforms to ISO 10918-1:1994, 'Information Technology – Digital compression and coding of continuous-tone still images: Requirements and guidelines' (also known as ITU-T T.81). The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

JPEG (represented using a JPEG/JFIF image)

The content of the DocumentData element is a JPEG image. The file conforms to ISO 10918-1:1994, 'Information Technology – Digital compression and coding of continuous-tone still images: Requirements and guidelines' (also known as ITU-T T.81), and the 'JPEG File Interchange Format', version 1.02. The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

JPEG (unknown format)

The content of the DocumentData element is a JPEG image. The file conforms to ISO 10918-1:1994, 'Information Technology – Digital compression and coding of continuous-tone still images: Requirements and guidelines' (also known as ITU-T T.81), and may also conform to 'JPEG File Interchange Format', version 1.02. The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

JPEG-2000

The content of the DocumentData element is a JPEG 2000 image. The file conforms to ISO 15444-1:2004, 'Information Technology – JPEG 2000 image coding system: Core coding system' (also known as ITU-T T.800). The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

MPEG-4 video (MP file format)

The content of the DocumentData element is an MPEG-4 video stream encoded in the MP file format. The file conforms to the following parts of ISO/IEC 14496 'Information Technology – Coding of audio-visual objects', Part 1:2004 (including Amd 1: 2005 and Amd 8:2004), Part 2:2004 (including Cor 1:2004, Amd 1:2004, and Amd 2:2005), Part 3:2004, and Part 14:2003 (including Cor 1:2006). The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

MPEG-4 video (AVC file format)

The content of the DocumentData element is an MPEG-4 video stream encoded in the AVC file format. The file conforms to the following parts of ISO/IEC 14496 'Information Technology – Coding of audio-visual objects', Part 1:2004 (including Amd 1: 2005 and Amd 8:2004), Part 2:2004 (including Cor 1:2004, Amd 1:2004, and Amd 2:2005), Part 3:2004, Part 10:2005, and Part 15:2004 (including Cor 1:2006, and Amd 1:2006). The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

MPEG-4 video (unknown – either MP or AVC file format)

The content of the DocumentData element is an MPEG-4 video stream encoded in the AVC or MP file format. The file conforms to the following parts of ISO/IEC 14496 'Information Technology – Coding of audio-visual objects', Part 1:2004 (including Amd 1: 2005 and Amd 8:2004), Part 2:2004 (including Cor 1:2004, Amd 1:2004, and Amd 2:2005), Part 3:2004, and either Part 14:2003 (including Cor 1:2006) OR Part 10:2005, and Part 15:2004 (including Cor 1:2006, and Amd 1:2006). The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

WARC (Web Archive)

The content of the DocumentData element is a WARC archive. The file conforms to ISO 28500:2009 'WARC file format'. The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

CSV (Comma-separated variables) encoded as Base64

The content of the DocumentData element is a text file containing a CSV. The file conforms to RFC4180 'Common format and MIME type for Comma-Separated Values (CSV) files'. The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

CSV (Comma-separated variables) encoded as text

The content of the DocumentData element is a text file containing a CSV. The file conforms to RFC4180 'Common format and MIME type for Comma-Separated Values (CSV) files'. The characters in the text file

conform to UTF-8 encoded Unicode. Unicode is defined in 'The Unicode Standard, Version 8.0.0, The Unicode Consortium, or the equivalent ISO 10646:2014 and Amendment 1. The ampersand (and), single quote, double quote, greater than, and less than characters have been escaped using the standard XML escape conventions '&', '<', and '>' respectively.

MP3 (MPEG1 and MPEG2 Audio Layer III)

The content of the DocumentData element is an MP3 file. The file conforms to Audio Layer III of ISO 11172-3:1993 (MPEG-1 Part 3) or ISO 13818-3:1998 (MPEG-2 Part 3). The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

MP4 audio (MPEG4 Audio)

The content of the DocumentData element is an MPEG-4 audio stream. The file conforms to Part 3:2004 of ISO/IEC 14496 'Information Technology – Coding of audio-visual objects'. The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

WAV/LPCM (Waveform Audio File Format)

The content of the DocumentData element is a WAV file with the contained audio stream encoded using LPCM (Linear Pulse Code Modulation). The file conforms to 'Multimedia Programming Interface and Data Specifications 1.0'. The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.

MIME (Email) encoded as text

The content of the DocumentData element is a MIME file conformant with RFC2049 'Multipurpose Internet Mail Extensions (MIME) Part Five: Conformance Criteria and Examples,' Freed & Borenstein, Nov 1996. The characters in the text file conform to UTF-8 encoded Unicode. Unicode is defined in 'The Unicode Standard, Version 8.0.0, The Unicode Consortium, or the equivalent ISO 10646:2014 and Amendment 1. The ampersand (and), single quote, double quote, greater than, and less than characters have been escaped using the standard XML escape conventions '&', '<', and '>' respectively.

MIME (Email) encoded in Base64

The content of the DocumentData element is a MIME file conformant with RFC2049 'Multipurpose Internet Mail Extensions (MIME) Part Five: Conformance Criteria and Examples,' Freed & Borenstein, Nov 1996. The file has been encoded using Base64, which is defined in IETF RFC 2045 'Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies', Section 6.8 'Base64 Content-Transfer-Encoding'.